Un abordaje matemático del aprendizaje supervisado

Alejandro Cholaquidis

Índice general

Estas notas están muy fuertemente basadas en [9] y en [15], junto con notas de Modelos Lineales de Mario Wschebor, y notas del curso de aprendizaje estadístico de Ricardo Fraiman. El objetivo es dar un panorama general del aprendizaje supervisado desde una perspectiva matemática. Se requiere tener algunos conocimientos previos de teoría de la medida (los cuales se resumen de forma muy sucinta en uno de los apéndices, pero pueden leerse en cualquier libro clásico, por ejemplo [42],[26] o [11]), y de probabilidad básica (variables aleatorias, convergencias, etc), que pueden encontrarse en cualquier libro de probabilidad, por ejemplo [29, 23].

Para hacer las notas lo más autocontenidas posibles hemos incluido además otro apéndice que incluye unas pocas desigualdades básicas de variables aleatorias que usaremos. Luego de la introducción, que da un panorama general del aprendizaje, veremos el capítulo 2, sección 7, del libro [29] que aborda el concepto de esperanza condicional, que será central para la lectura del resto de los capítulos.

Quedo infinitamente agradecido a Antonio Cuevas por las charlas esclarecedoras que tuvimos sobre diversos temas, y a Ernesto Mordecki por cederme la posibilidad de dar el curso en su versión 2024.

Finalmente, demás está decir que todas las erratas y errores que puedan contener estas notas son de mi responsabilidad. Quedo agradecido al lector que, amablemente, me las haga saber a acholaquidis@hotmail.com

La cronología que sigue es parcial, arbitraria, incompleta y caprichosa. Solo da cuenta de, por un lado, algunas publicaciones relevantes que espero estimulen al lector interesado a seguir buscando, y de mi imposibilidad (por falta de conocimientos y porque llevaría a un libro en si mismo) de escribir un listado más exhaustivo. Toda cronología sobre aprendizaje debería empezar con Turing[1], y eso haremos.

1. En 1950 Turing publica su famoso trabajo [34] con la pregunta ¿Puede una máquina pensar?, donde plantea el test de Turing: una prueba de imitación, donde un interrogador humano interactúa con dos entidades ocultas: una máquina y un ser humano. El interrogador debe determinar cuál de las dos entidades es la máquina basándose únicamente en sus respuestas a las preguntas formuladas. Si la máquina puede engañar al interrogador para que piense que es humana una proporción significativa de veces, entonces se considera que la máquina ha demostrado inteligencia comparable a la humana. En 1998 Smale propone, como último en la lista de problemas del milenio, la pregunta "¿Cuáles son los límites de la inteligencia, tanto artificial como humana?".

2. También en 1950, el padre de la teoría de la información, Shannon (ver [27]), publica [28], donde muestra que una computadora puede ser programada para jugar al ajedrez.

3. En 1951 en [24] se publica uno de los primeros algoritmos sobre el método estocástico de descenso por gradiente. Para ser justos con los desarrollos en SGD habría que hacer una cronología aparte, que contenga los trabajos, de, por ejemplo, LeCun y Hinton, entre otros.

4. En 1954 se publica el libro de Blackwell [3]. Una cronología de la teoría de juegos debería empezar, claro está, con [39]. Cito a Blackwell por pura arbitrariedad, cabe destacar que el aprendizaje por recompensas[2] está fuertemente emparentada con los MDP[3] y la teoría de juegos.

5. En 1958 Rosenblatt publica [25] donde propone el perceptrón, un clasificador lineal ("red neuronal") con una neurona. Lo veremos en estas notas. A 2024 el trabajo tiene más de 19.000 citas (no es de los primeros en proponer redes neuronales, antes de eso están, por citar algunos, los trabajos de Minsky, Rochester, etc.)

6. En 1964 Vapnik publica [37], que se centra en el análisis teórico de los perceptrones. En este año se publican dos trabajos que ponen los cimientos de lo que luego será la regresión por núcleos (o método de Nadaraya y Watson), [41] y [21].

7. En 1971 se publican dos trabajos importantes: [8], donde se introduce el concepto de NP-completitud, y Vapnik junto con Chervonenkis publican su célebre trabajo [38], donde introducen el concepto de coeficiente de fragmentación y dimensión de V.C.

8. En 1977 Stone publica en el Annals of Statistics [30], donde se da un teorema general de consistencia de los estimadores locales de la función de regresión. Esto también lo veremos en estas notas.

9. Dejamos por acá la cronología, para no aburrir, pero la lista podría seguir, y debería contener a Yoshua Bengio y a Ian Goodfellow entre otros.

[1] aquí también, para hacer justicia con la regresión lineal, habría que empezar antes con los trabajos de Galton [13] y Pearson [22], y con el discriminante lineal de Fisher, de 1936.

[2] una rama de la IA que no veremos pero que permitió, entre otras cosas, que Deep Blue le ganara a Kasparov en 1997 (para hacer justicia con Garry habría que aclarar que ganó 3-1/2,2-1/2 en 6 partidas), y Alphago a Lee Sedol en Go en 2016

[3] Markov Decision Process

1 Introducción

Hay dos referencias esenciales para entender el aprendizaje estadístico desde una perspectiva histórica. Por un lado [6], de Breiman, donde plantea que hay dos culturas en materia de aprendizaje, una basada en modelos y otra en algoritmos, y por otro, el trabajo de Vapnik: [36], que da cuenta con rigor y generalidad, de sus valiosos aportes a esta disciplina. Citando algunas partes del resumen del trabajo de Breiman:

> *Hay dos culturas en el uso del modelado estadístico para llegar a conclusiones a partir de datos. Una asume que los datos son generados por un modelo de datos estocástico dado. La otra usa modelos algorítmicos y trata el mecanismo de datos como desconocido. La comunidad estadística se ha comprometido con el uso casi exclusivo de modelos de datos [...].*

Breiman, partidario de la segunda cultura, considera en dicho trabajo a la primera como teoría irrelevante de conclusiones cuestionables. Sin entrar en esa discusión, pero citando a Breiman nuevamente, el aprendizaje es, entonces, *llegar a conclusiones a partir de datos.* Estos datos vienen dados por una muestra de una variable, que denotaremos, de forma genérica como X, que tiene asociada, para el caso de aprendizaje supervisado, una respuesta, que denotaremos Y. Cuando no se tiene la variable Y, se está ante un problema de aprendizaje no supervisado, que no abordaremos en estas notas, pero que mencionaremos brevemente más adelante. Por otra parte, y más recientemente, se incorporaron los problemas de aprendizaje por recompensas, donde hay una interacción entre un agente y un medio en el cual toma decisiones que a su vez afectan las acciones que tomará el agente.

1.1 Aprendizaje supervisado

Los dos objetivos centrales del aprendizaje (supervisado) son, por un lado la predicción, es decir, dada una nueva entrada X, asignarle un valor de salida o respuesta Y, y, por otra parte, extraer información sobre la naturaleza de la asociación entre X e Y. Los algoritmos que abordan estos problemas suelen denominarse máquinas de aprendizaje ("machine learning").

En cuanto a las dos culturas mencionadas, la **cultura basada en el modelado** de los datos plantea que la salida Y se modela como

$$Y = f\Big(\text{variables predictoras},\ \text{ruido}\ ,\ \text{parámetros}\Big).$$

y asumen hipótesis estadísticas sobre la generación de los datos. El ejemplo clásico de esto es el clasificador lineal de Fisher, donde se asume igual varianza entre los dos grupos. Los valores de los parámetros se estiman a partir de los datos, y se usa el modelo para obtener información (segundo objetivo que mencionamos), y para predecir. Otros ejemplos de esto son la regresión lineal y logística (cuando se hacen hipótesis sobre la distribución de los errores). En estos casos la validación del modelo se hace por medio de un test de bondad de ajuste (ya que asumimos una distribución para el modelo), y del estudio de los residuos.

Por otra parte, la **cultura basada en algoritmos** considera que la asociación f entre X e Y es una caja negra compleja y desconocida, sobre la cual no se hace ninguna hipótesis especial y se busca predecir Y a partir de X. Un ejemplo típico de esto es el perceptrón, que es también un clasificador lineal como el de Fisher, pero no asume hipótesis de homocedasticidad. Entre estos algoritmos se encuentran los árboles de decisión, las redes neuronales, los métodos de regresión basados en vecinos más cercanos (k-nn), y el método de regresión basado en núcleos de Nadaraya y Watson. En estos casos la validación se da por medio del error de predicción.

Los datos de que disponemos son, como dijimos, pares $(X_1, Y_1), \ldots, (X_n, Y_n)$, que usualmente se asumen que son copias independientes a idénticamente distribuidas (iid) del par (X, Y). Además, en el enfoque algorítmico, se tiene un conjunto (independiente del de entrenamiento) que se denomina de testeo, y que se va a usar para evaluar el modelo entrenado. En general lo que se hace es partir el conjunto de datos original en un porcentaje, típicamente del entorno del 70 u 80 %, para entrenamiento y el resto se deja para testeo.

El ejemplo más simple de problema que se aborda en el aprendizaje supervisado es la clasificación binaria, en la cual Y toma únicamente los valores 0 o 1, como cuando se quiere clasificar un e-mail en spam o no-spam, en base a e-mails que ya hemos clasificados. O, clasificar pacientes en sanos o enfermos, en base a resultados médicos. Cuando Y es continua se denominar problema de regresión y se plantea encontrar una función f, de modo que $f(X)$ sea una aproximación razonable, en algún sentido que hay que especificar, a Y. [1] Construir una aproximación de f basada

[1] Un criterio de proximidad usual es buscar la f medible que minimice $\mathbb{E}(d(f(X), Y)^2)$, es decir, minimizar el error cuadrático medio, o maximizar la verosimilitud como en el caso de la regresión logística que veremos.

en la muestra nos permite predecir un valor de Y desconocido a partir de un nuevo dato X. No obstante, si no ponemos hipótesis sobre la distribución del par (X, Y), como en el caso de algoritmos basados en modelos, no vamos a poder, por ejemplo, hacer intervalos de confianza o pruebas de hipótesis. Una alternativa para generar estas bandas de confianza libres de distribución, es decir, sin hipótesis sobre la distribución de (X, Y), es hacer inferencia conformal, una propuesta de Vovk et al del 98 que puede encontrarse en, por ejemplo [40] y que ha tomado gran relevancia en los últimos años en estadística y en general en machine learning.

1.1.1 Parámetros e hiperparámetros

En general los algoritmos como "support vector machine"(SVM), la regresión logística, y la regresión lineal, se los considera métodos paramétricos, mientras que k-nn, árboles y bosques de regresión y/o clasificación, o regresión por núcleos se los considera no paramétricos.

El término "no paramétrico" refiere a que el algoritmo de aprendizaje no asume una forma específica para la función de distribución de los datos o para la relación entre X e Y, ni estima un conjunto fijo de parámetros de una función determinada (como los coeficientes en una regresión lineal). En k-nn, como veremos, si bien hay un valor k, (que luego su elección se puede hacer por diferentes métodos), este se denomina parámetro de ajuste, de tuneado, o hiperparámetro. Lo mismo sucede con los métodos basados en núcleos, donde aparece el ancho de la ventana a elegir, o con las redes neuronales, donde la profundidad (cantidad de capas) es un hiperparámetro. Los hiperparámetros se utilizan para controlar el proceso de aprendizaje y se estiman usando la muestra de entrenamiento de diferentes maneras. Veamos muy brevemente dos de ellos, ℓ-fold y *holdout*.

Validación Cruzada ℓ-Fold

1. Se divide el conjunto de entrenamiento en ℓ partes (o "folds"), no necesariamente iguales.
2. Se fija un conjunto inicial de valores para el o los hiperparámetros.
3. Se entrena el modelo (es decir, se hace la estimación de sus parámetros en caso de haberlos) ℓ veces, usando los hiperparámetros antes elegidos. En cada una de estas ℓ veces se utiliza como muestra de entrenamiento $\ell - 1$ partes, mientras que la restante se usa para aproximar el error.
4. Se promedian los errores resultantes de las ℓ iteraciones.[2]
5. Se repiten los dos puntos anteriores con otro conjunto de hiperparámetros.
6. Finalmente, luego de haber probado con varios hiperparámetros, se elije aquel que minimiza el error calculado en el punto 4.

Dentro de estos métodos se encuentra el clásico "leave one-out", en el cual ℓ es del tamaño de la muestra de entrenamiento, menos uno, de allí su nombre.

Validación Cruzada Holdout

Se divide el conjunto de datos entrenamiento en dos partes: uno exclusivamente de entrenamiento (usualmente en el orden de un 60 % del total de datos) y un conjunto de validación (del orden del 20 %). El conjunto de validación se utiliza únicamente para ajustar los hiperparámetros del modelo cuyos parámetros fueron estimados con la muestra de entrenamiento.

Plug-in o no-plug-in

Otra jerarquización de algoritmos que suele hacerse, y en general de estimadores, es entre plug-in y no plug-in. En los primeros la estimación se obtiene cambiando las distribuciones teóricas desconocidas por sus distribuciones empíricas.

En el marco de clasificación binaria, se usa el término plug-in cuando la función de regresión teórica (que es, como veremos, la esperanza condicional $m^*(x) = \mathbb{P}(Y = 1|X = x)$), se cambia por una aproximación, m_n, de ella, basa en la muestra, y se toma el clasificador que vale 1 si $m_n(x) > 1/2$ y 0 si no. Es decir, se *enchufa* (plug-in) el estimador m_n en la regla de Bayes $g^*(x) = \mathbb{I}_{\{m^*(x)>1/2\}}$, que es, como veremos, la que minimiza el error de clasificación $\mathbb{P}(g(X) \neq Y)$ entre todas las posibles reglas.

En el marco de la regresión, k-nn y Nadaraya-Watson pueden considerarse también métodos plug-in. Ejemplos clásicos de métodos que no son plug-in podrían ser aquellos que se obtienen por máxima verosimilitud, o los métodos Bayesianos, donde se asume que los parámetros tienen una determinada distribución a priori.

[2] recordemos que, como conocemos las verdaderas etiquetas, podemos calcular el error medio, $d(Y_i, \hat{Y}_i)$, en cada una de estas ℓ iteraciones, siendo $\hat{Y}_i$ el valor asignado por el modelo entrenado, al dato X_i.

1.1.2 Error de Bayes

Es importante aclarar que en clasificación la probabilidad de equivocarse (asignarle a un nuevo dato una etiqueta que no es la correcta) no necesariamente se va a poder hacer arbitrariamente pequeña, aún cuando la muestra de entrenamiento se pueda hacer tan grande como se quiera. Por error nos referimos, no el calculado en la muestra de entrenamiento, que es 0 si se interpolan los pares (X_i, Y_i), sino en un nuevo par (X, Y). Es claro que un clasificador que interpole los datos tiene una capacidad de predicción (y por lo tanto utilidad) nulos, pero un error 0 en dicha muestra. En el caso de clasificación binaria la cota inferior del error de una regla de clasificación g, es decir $\mathbb{P}(g(X) \neq Y)$, está dada por la probabilidad de equivocarse de la mejor regla de clasificación (la cual es teórica y en general desconocida), que es, como dijimos y probaremos más adelante, asignar la etiqueta $Y = 1$ si $m^*(x) = \mathbb{P}(Y = 1|X = x) > 1/2$ y cero en caso contrario. Esta regla se denomina regla de Bayes. Lo mismo sucederá en regresión, el mejor predictor de $Y \in \mathbb{R}$ basado en X es $m^*(X) = \mathbb{E}[Y|X]$, donde mejor refiere a que minimiza $\mathbb{E}|Y - f(X)|^2$ al considerar todas las posibles f medibles. Por lo tanto, la cota inferior del error cuadrático medio de cualquier estimador m_n[3] es $\mathbb{E}|Y - m^*(X)|^2$, que no necesariamente es 0. Este valor mínimo de error se le conoce como error de Bayes. Los métodos de clasificación plug-in estiman m^* con m_n, basada en la muestra, y usan la regla de clasificación g_n que vale 1 si $m_n(x) > 1/2$ y 0 en caso contrario.

1.2 Aprendizaje no supervisado

En estos algoritmos los patrones a identificar son similitudes entre subgrupos ("cluster") de valores de la variable X. Es decir, no tiene por qué haber una Y subyacente, lo cual hace que definir qué quiere decir que algo es un grupo dependa del problema en cuestión. Tampoco se asume de antemano que se conoce la cantidad de grupos que existen, aunque, típicamente, a estos algoritmos se les da como dato cuantos grupos tiene que encontrar. Lo que tenemos es entonces una muestra $X_1, \ldots, X_n$ de la variable X, y queremos agrupar subgrupos de estas observaciones. Ejemplos clásicos de esto son los algoritmos jerárquicos en los cuales, si buscamos k grupos, obtenemos un conjunto de k subgrupos que va a estar contenido en el que obtendríamos si buscamos $k-1$. Es decir, tenemos una jerarquía de grupos, donde cada partición se obtiene de partir o unir la partición anterior. Si estos grupos se construyen tomando la muestra inicial como un todo que se va dividiendo en subgrupos se dice que el algoritmo es divisivo, mientras que, si se construye tomando la muestra como un conjunto de n puntos aislados que se van uniendo, se dice que es un método aglomerativo. Otro algoritmo clásico es el de k-medias, en el cual se fija de antemano el valor k de grupos, y se buscan $G_1, \ldots, G_k$, una partición de $\{X_1, \ldots, X_n\}$, que minimicen $\sum_{i=1}^{k} \sum_{X_j \in G_i} d(X_j, x_i)$, donde cada x_i son los centros en los grupos G_i. Es decir, x_i es el dato de G_i que minimiza $\sum_{X_j \in G_i} d(X_j, x)$. Si, por ejemplo, d es la norma euclidiana al cuadrado, se obtiene la partición que minimiza la varianza dentro de cada grupo ponderada por la cantidad de elementos del grupo. Problemas clásicos donde esto surge son la búsqueda de patrones de comportamiento de usuarios de redes sociales, de consumidores, etc, en base al historial su comportamiento.

1.3 Aprendizaje por refuerzos o recompensas

En el aprendizaje por refuerzos hay una función de recompensa que se quiere maximizar y el algoritmo va aprendiendo, a efectos de maximizar esta recompensa, a medida que toma diferentes acciones, las cuales, al elegirse pueden dar un cambio en las acciones posibles a tomar en el futuro. Es decir, hay una interacción con el entorno. Un ejemplo simple de esto son los "bandits", donde las acciones posibles a tomar en cada instante de tiempo $t = 1, 2, \ldots, L$, siendo L la cantidad de veces que se va a jugar, son finitas y fijas. Como cuando se juega a una máquina tragamonedas, de donde provine el nombre. Cada acción devuelve, si se la elije, una recompensa, que usualmente es aleatoria. El objetivo es encontrar la mejor acción posible (donde mejor usualmente se toma como aquella cuyo valor esperado de la recompensa es el más grande).

Si bien existe una miríada de variantes de algoritmos de aprendizaje, por ejemplo algoritmos de aprendizaje semi-supervisado, en los que se dispone de datos etiquetados y otros no etiquetados (usualmente la cantidad de datos no etiquetados es mucho mas grande que la de los etiquetados), vamos a enfocarnos en algoritmos de clasificación y regresión supervisados.

[3] es decir, $\mathbb{E}|Y - m_n(X)|^2$

2 Esperanza Condicional.

En la introducción dijimos que, dadas X e Y dos variables aleatorias, Y a valores reales y X a valores en algún espacio medible, la función medible f que minimiza el error cuadrático medio $\mathbb{E}|f(X)-Y|^2$, es la esperanza condicional

$$m^*(x) = \mathbb{E}[Y|X=x].$$

Vamos a demostrar ese hecho, pero primero, introducir de forma rigurosa su definición. La esperanza condicional tiene una motivación e interpretación geométrica muy relevante. Si consideramos

$$L^2(\Omega,\mathcal{A},\mathbb{P}) = \{Y:\Omega\to\mathbb{R} : \mathbb{E}(Y^2)<\infty, Y \text{ es medible respecto de } \mathcal{A}\},$$

es un espacio de Hilbert[1] si tomamos $\langle X,Y\rangle = \mathbb{E}(XY)$[2]. Por lo tanto, dado un subespacio cerrado $\mathcal{V}\subset L^2(\Omega,\mathcal{A},\mathbb{P})$, **existe y es única**, la proyección ortogonal, $\Pi_{\mathcal{V}}(Y)$, de Y sobre $\mathcal{V}$, es decir

$$\langle Y-\Pi_{\mathcal{V}}(Y), Z\rangle = \mathbb{E}\Big(\big(Y-\Pi_{\mathcal{V}}(Y)\big)Z\Big) = 0 \quad \text{para todo } Z\in\mathcal{V}, \tag{2.1}$$

o, lo que es lo mismo

$$\Pi_{\mathcal{V}}(Y) = \operatorname*{arg\,min}_{Z\in\mathcal{V}} \|Y-Z\|^2 = \operatorname*{arg\,min}_{Z\in\mathcal{V}} \mathbb{E}|Y-Z|^2. \tag{2.2}$$

La segunda igualdad en (2.1) es equivalente a

$$\mathbb{E}(YZ) = \mathbb{E}\big(\Pi_{\mathcal{V}}(Y)Z\big) \quad \forall Z\in\mathcal{V}. \tag{2.3}$$

Consideremos $\mathfrak{F}\subset\mathcal{A}$ una sub σ-álgebra y sea $\mathcal{V}$ el subespacio vectorial de todas las variables en $L^2(\Omega,\mathfrak{F},\mathbb{P})$. Se puede ver que $\mathcal{V}$ es un subespacio cerrado de $L^2(\Omega,\mathcal{A},\mathbb{P})$. En este caso, basta con que (2.3) se cumpla para indicatrices de conjuntos de $\mathfrak{F}$, (ya que si lo cumplen las indicatrices de conjuntos de $\mathfrak{F}$, luego cualquier función medible respecto de $\mathfrak{F}$ se aproxima por suma de indicatrices.) La ecuación (2.3) queda

$$\mathbb{E}(Y\mathbb{I}_F) = \mathbb{E}\big(\Pi_{\mathcal{V}}(Y)\mathbb{I}_F\big) \quad \forall F\in\mathfrak{F}. \tag{2.4}$$

La función $\Pi_{\mathcal{V}}(Y):\Omega\to\mathbb{R}$ es la esperanza condicional de Y respecto de $\mathfrak{F}$, y la denotaremos $\mathbb{E}(Y|\mathfrak{F})$. Como no vamos a pedir que $Y\in L^2(\Omega)$, vamos a definirla mediante el Teorema de Radón-Nykodim, pero va a verificar (2.4) y va a ser $\mathfrak{F}$-medible. Para el caso $Y\in L^2(\Omega)$ va a tener la propiedad minimizante (2.2).

2.1 Esperanza condicional respecto de una σ-álgebra

A lo largo del capítulo $(\Omega,\mathcal{A},\mathbb{P})$ es un espacio de probabilidad y las variables aleatorias que aparecen toman valores en la recta ampliada $\overline{\mathbb{R}}\equiv\mathbb{R}\cup\{\infty\}\cup\{-\infty\}$. En $\overline{\mathbb{R}}$ consideraremos la σ-álgebra de Borel $\mathcal{B}(\overline{\mathbb{R}})$, que es la menor σ-álgebra que contiene los conjuntos

$$\{(a,b)\mid a,b\in\mathbb{R}, a<b\}\bigcup\{(-\infty,a]\cup\{+\infty\} : a\in\mathbb{R}\}\bigcup\{[b,+\infty)\cup\{-\infty\} : b\in\mathbb{R}\}.$$

En general estas variables van a ser medibles respecto de $\mathcal{B}(\overline{\mathbb{R}})$, y en Ω tomaremos o bien $\mathcal{A}$ o bien una sub σ-álgebra de ella $\mathfrak{F}\subset\mathcal{A}$.
Si bien la esperanza condicional se puede definir de forma más general, vamos a suponer que Y es una variable aleatoria para la cual $\mathbb{E}(|Y|)<\infty$.

Definición 2.1. Sea $Y:\Omega\to\mathbb{R}$ una variable aleatoria (medible respecto de la σ-álgebra $\mathcal{A}$), tal que $\mathbb{E}(|Y|)<\infty$ y sea $\mathfrak{F}\subset\mathcal{A}$ una sub σ-álgebra. La **esperanza condicional** $\mathbb{E}(Y|\mathfrak{F}):\Omega\to\mathbb{R}$ es una variable aleatoria que verifica

[1] recordar que un espacio de Hilbert es un espacio vectorial dotado de un producto interno que lo haga separable y completo

[2] hay un detalle "formal": así definida $\|X\| = \sqrt{\mathbb{E}(X^2)}$ no es una norma, ya que podría ser 0 sin que la variable sea nula. Para solucionar esto se toma una relación de equivalencia, $\sim$, en el L^2 que definimos, dada por $X\sim Y$ si son iguales a menos de conjuntos de probabilidad 0. Finalmente el espacio L^2 es el de las clases de equivalencia de esta relación. En las notas asumiremos que estamos en este espacio de clases de equivalencias. Notar que no es un espacio de funciones

i. $\mathbb{E}(Y|\mathfrak{F})$ es medible respecto de $\mathfrak{F}$

ii. para todo $F \in \mathfrak{F}$

$$\mathbb{E}(Y\mathbb{I}_F) = \int_F Y d\mathbb{P} = \int_F \mathbb{E}(Y|\mathfrak{F}) d\mathbb{P} = \mathbb{E}(\mathbb{E}(Y|\mathfrak{F})\mathbb{I}_F) \tag{2.5}$$

observar que la ecuación (2.5) es (2.4).

Para demostrar que $\mathbb{E}(Y|\mathfrak{F})$ existe hay que usar el Teorema de Radon-Nikodym. Si Y es tal que $\mathbb{E}(|Y|) < \infty$, Q, definida en $\mathfrak{F}$ como

$$Q(F) \equiv \int_F Y d\mathbb{P},$$

es una medida signada finita, absolutamente continua respecto de $\mathbb{P}$ en $\mathfrak{F}$. Por lo tanto, por el Teorema de Radon-Nikodym, existe $\mathbb{E}(Y|\mathfrak{F})$, $\mathfrak{F}$-medible tal que

$$Q(F) = \int_F \mathbb{E}(Y|\mathfrak{F}) d\mathbb{P} \quad \forall F \in \mathfrak{F}.$$

Observación 2.2. $\mathbb{E}(Y|\mathfrak{F})$ *está definida a menos de conjuntos de probabilidad* 0. *Además, y esto será usado para probar propiedades de la esperanza condicional,* ***es única c.s.*** *(esto se sigue de la unicidad c.s. en el Teorema de Radon Nikodym).* [3]

Definición 2.3. Dado $A \in \mathcal{A}$, la esperanza condicional $\mathbb{E}(\mathbb{I}_A|\mathfrak{F})$, con $\mathfrak{F} \subset \mathcal{A}$, se denota $\mathbb{P}(A|\mathfrak{F})$. Se sigue de esta definición que $\mathbb{P}(A|\mathfrak{F})$ es $\mathfrak{F}$-medible y,

$$\mathbb{P}(F \cap A) = \int_F \mathbb{I}_A d\mathbb{P} = \int_F \mathbb{P}(A|\mathfrak{F}) d\mathbb{P} \qquad \forall F \in \mathfrak{F}.$$

Veamos como queda la esperanza condicional respecto de una partición de Ω, pero antes, vamos a introducir la siguiente notación

Si $\mathbb{P}(A) > 0$ se denota

$$\mathbb{E}(Y|A) \equiv \frac{\mathbb{E}(Y\mathbb{I}_A)}{\mathbb{P}(A)}. \tag{2.6}$$

Sea $\mathcal{D} = \{D_1, D_2, \dots\}$ una partición de Ω, es decir: $D_i \in \mathcal{A}$, $\mathbb{P}(D_i) > 0$ para todo i, $\cup_i D_i = \Omega$ y $\mathbb{P}(D_i \cap D_j) = 0$ para todo $i \neq j$. Vamos a usar que si Y es medible respecto de $\mathscr{G} \equiv \sigma(\mathcal{D})$, la σ-álgebra generada por $\mathcal{D}$, entonces $Y = \sum_{i=1}^{\infty} y_i \mathbb{I}_{D_i}$ *c.s* en D_i, donde la notación $Y = X$ *c.s* en D_i significa que $\mathbb{P}(\{Y \neq X\} \cap D_i) = 0$.

Teorema 2.4. *Sea $\mathscr{G} \equiv \sigma(\mathcal{D})$ la σ-álgebra generada por $\mathcal{D}$ e Y una variable aleatoria para cual $\mathbb{E}(|Y|) < \infty$, entonces*

$$\mathbb{E}(Y|\mathscr{G}) = \mathbb{E}(Y|D_i) \quad c.s \text{ en } D_i,$$

es decir, usando (2.6),

$$\mathbb{E}(Y|\mathscr{G}) = \frac{\mathbb{E}(Y\mathbb{I}_{D_i})}{\mathbb{P}(D_i)} \quad c.s \text{ en } D_i,$$

Demostración. Por la observación que hicimos antes $\mathbb{E}(Y|\mathscr{G})(\omega) = y_i$ para todo $\omega \in D_i$ (a excepción de conjuntos de probabilidad 0). Por lo tanto,

$$\int_{D_i} Y d\mathbb{P} = \int_{D_i} \mathbb{E}(Y|\mathscr{G}) d\mathbb{P} = y_i \mathbb{P}(D_i),$$

de donde se sigue que

$$y_i = \frac{1}{\mathbb{P}(D_i)} \int_{D_i} Y d\mathbb{P} = \frac{\mathbb{E}(Y\mathbb{I}_{D_i})}{\mathbb{P}(D_i)} = \mathbb{E}(Y|D_i),$$

donde en la última igualdad hemos usado (2.6). □

Observación 2.5. *Sea $A \in \mathcal{A}$, no es lo mismo $\mathbb{E}(Y|A)$ como fue definida en* (2.6) *que $\mathbb{E}(Y|\sigma(A))$, que por el Teorema 2.4 es*

$$\mathbb{E}(Y|\sigma(A)) = \mathbb{E}(Y|A)\mathbb{I}_A + \mathbb{E}(Y|A^c)\mathbb{I}_{A^c}$$

Recordemos que si $A, B \in \mathcal{A}$, $\mathbb{P}(B|A) = \mathbb{P}(B \cap A)/\mathbb{P}(A)$, tenemos el siguiente corolario

Corolario 2.6. *Si $\mathcal{D} = \{D_1, D_2, \dots\}$ es una partición finita o numerable de Ω, la esperanza condicional de $B \in \mathcal{A}$ dado $\sigma(\mathcal{D})$ es la variable aleatoria $\mathbb{P}(B|\sigma(\mathcal{D})) : \Omega \to \mathbb{R}$,*

$$\mathbb{P}(B|\sigma(\mathcal{D}))(\omega) \equiv \sum_{i \geq 1} \mathbb{P}(B|D_i)\mathbb{I}_{D_i}(\omega).$$

Se deja como ejercicio verificar que la variable aleatoria $\mathbb{P}(B|\sigma(\mathcal{D}))$ es medible respecto de $\sigma(\mathcal{D})$, la σ-álgebra generada por D.

[3] aquí nos referimos a probabilidad en Ω es decir respecto de $\mathbb{P}$

Veamos algunas propiedades de la esperanza condicional. En todos los casos Y es medible respecto de $\mathcal{A}$, $\mathfrak{F} \subset \mathcal{A}$ es una σ-álgebra, y suponemos que están definidas las esperanzas condicionales que aparecen.

Proposición 2.7 (Propiedades de la esperanza condicional)**.**

1. *Si C es una constante e $Y = C$ c.s, entonces $\mathbb{E}(Y|\mathfrak{F}) = C$ c.s.*
2. *Si $Y \leq Z$ c.s, entonces $\mathbb{E}(Y|\mathfrak{F}) \leq \mathbb{E}(Z|\mathfrak{F})$ c.s.*
3. *$|\mathbb{E}(Y|\mathfrak{F})| \leq \mathbb{E}(|Y||\mathfrak{F})$ c.s.*
4. *Si a, b son constantes tal que $a\mathbb{E}(Y) + b\mathbb{E}(Z)$ está definida, entonces*
$$\mathbb{E}(aY + bZ|\mathfrak{F}) = a\mathbb{E}(Y|\mathfrak{F}) + b\mathbb{E}(Y|\mathfrak{F}) \quad c.s. \tag{2.7}$$
5. *Si $\mathfrak{F}_* \equiv \{\emptyset, \Omega\}$ entonces $\mathbb{E}(Y|\mathfrak{F}_*) = \mathbb{E}(Y)$ c.s.*
6. *$\mathbb{E}(Y|\mathcal{A}) = Y$ c.s.*
7. *$\mathbb{E}(\mathbb{E}(Y|\mathfrak{F})) = \mathbb{E}(Y)$ c.s.*
8. *Si $\mathfrak{F}_1 \subset \mathfrak{F}_2$ entonces $\mathbb{E}[\mathbb{E}(Y|\mathfrak{F}_2)|\mathfrak{F}_1] = \mathbb{E}(Y|\mathfrak{F}_1)$ c.s.*
9. *Si $\mathfrak{F}_2 \subset \mathfrak{F}_1$ entonces $\mathbb{E}[\mathbb{E}(Y|\mathfrak{F}_2)|\mathfrak{F}_1] = \mathbb{E}(Y|\mathfrak{F}_2)$ c.s.*
10. *Si Y es independiente de la σ-álgebra $\mathfrak{F}$ [4], tal que $\mathbb{E}(Y)$ está definida, entonces $\mathbb{E}(Y|\mathfrak{F}) = \mathbb{E}(Y)$ c.s.*
11. *Sea Z medible respecto de $\mathfrak{F}$. Supongamos que $\mathbb{E}|Z| < \infty$ y $\mathbb{E}|YZ| < \infty$, entonces $\mathbb{E}(YZ|\mathfrak{F}) = Z\mathbb{E}(Y|\mathfrak{F})$ c.s.*

Demostración.

1. Se sigue de que como $Y = C$ c.s., $\int_F Y d\mathbb{P} = \int_F C d\mathbb{P}$ para todo $F \in \mathfrak{F}$, y de la unicidad c.s. de la esperanza condicional.
2. Si $Y \leq Z$ entonces para todo $F \in \mathfrak{F}$, $\int_F Y d\mathbb{P} \leq \int_F Z d\mathbb{P}$ y de (2.5) se sigue que $\int_F \mathbb{E}(Y|\mathfrak{F}) d\mathbb{P} \leq \int_F \mathbb{E}(Z|\mathfrak{F}) d\mathbb{P}$, como esta igualdad vale para todo $F \in \mathfrak{F}$ se sigue que $\mathbb{E}(Y|\mathfrak{F}) \leq \mathbb{E}(Z|\mathfrak{F})$ como queríamos.
3. Se sigue de que $-|Y| \leq Y \leq |Y|$. De donde, usando la propiedad anterior, $-\mathbb{E}[|Y||\mathfrak{F}] \leq \mathbb{E}(Y|\mathfrak{F}) \leq \mathbb{E}[|Y||\mathfrak{F}]$ c.s., o, lo que es idéntico $|\mathbb{E}(Y|\mathfrak{F})| \leq \mathbb{E}(|Y||\mathfrak{F})$ c.s.
4. De la linealidad de la esperanza, para todo $F \in \mathfrak{F}$
$$\int_F (aY + bZ) d\mathbb{P} = \int_F aY d\mathbb{P} + \int_F bZ d\mathbb{P} \underbrace{=}_{(2.5)} \int_F a\mathbb{E}(Y|\mathfrak{F}) d\mathbb{P} + \int_F b\mathbb{E}(Z|\mathfrak{F}) d\mathbb{P} = \int_F [a\mathbb{E}(Y|\mathfrak{F}) + b\mathbb{E}(Z|\mathfrak{F})] d\mathbb{P}.$$
Como
$$\int_F \mathbb{E}(aY + bZ|\mathfrak{F}) d\mathbb{P} \underbrace{=}_{(2.5)} \int_F (aY + bZ) d\mathbb{P} = \int_F [a\mathbb{E}(Y|\mathfrak{F}) + b\mathbb{E}(Z|\mathfrak{F})] d\mathbb{P}$$
vale para todo $F \in \mathfrak{F}$, de la unicidad de la esperanza condicional se sigue (2.7).
5. Se sigue de que $\mathbb{E}(Y)$ es $\mathfrak{F}_*$-medible y si $F = \emptyset$ o $F = \Omega$, $\int_F Y d\mathbb{P} = \int_F \mathbb{E}(Y) d\mathbb{P}$.
6. Como Y es $\mathcal{A}$-medible, y $\mathbb{E}(Y|\mathcal{A})$ es la única c.s. $\mathcal{A}$-medible que verifica 2.5 para todo $A \in \mathcal{A}$, y Y lo verifica, son iguales c.s.
7. Se sigue del punto siguiente tomando $\mathfrak{F}_1 = \mathfrak{F}_*$ y usando el punto 5. Más precisamente, supongamos que vale que si $\mathfrak{F}_1 \subset \mathfrak{F}_2$ entonces $\mathbb{E}[\mathbb{E}(Y|\mathfrak{F}_2)|\mathfrak{F}_1] = \mathbb{E}(Y|\mathfrak{F}_1)$ *c.s.*, tomemos $\mathfrak{F}_2 = \mathfrak{F}$ y $\mathfrak{F}_1 = \mathfrak{F}_*$, por lo tanto $\mathbb{E}(\mathbb{E}(Y|\mathfrak{F})|\mathfrak{F}_*) = \mathbb{E}(Y|\mathfrak{F}_*)$. Finalmente, por el punto 5, $\mathbb{E}(Y|\mathfrak{F}_*) = \mathbb{E}(Y)$ y $\mathbb{E}(\mathbb{E}(Y|\mathfrak{F})|\mathfrak{F}_*) = \mathbb{E}(\mathbb{E}(Y|\mathfrak{F}))$.
8. Sea $F_1 \in \mathfrak{F}_1$ entonces por (2.5) aplicado a Y y a $\mathbb{E}(Y|\mathfrak{F}_2)$,
$$\int_{F_1} \mathbb{E}(Y|\mathfrak{F}_1) d\mathbb{P} = \int_{F_1} Y d\mathbb{P} \quad \text{y} \quad \int_{F_1} \mathbb{E}[\mathbb{E}(Y|\mathfrak{F}_2)|\mathfrak{F}_1] d\mathbb{P} = \int_{F_1} \mathbb{E}(Y|\mathfrak{F}_2) d\mathbb{P}.$$
Usando que $\mathfrak{F}_1 \subset \mathfrak{F}_2$, $F_1 \in \mathfrak{F}_2$, la integral anterior es $\int_{F_1} Y d\mathbb{P}$. Por lo tanto probamos que para todo $F_1 \in \mathfrak{F}_1$,
$$\int_{F_1} \mathbb{E}(Y|\mathfrak{F}_1) d\mathbb{P} = \int_{F_1} \mathbb{E}[\mathbb{E}(Y|\mathfrak{F}_2)|\mathfrak{F}_1] d\mathbb{P}.$$
Con lo cual $\mathbb{E}(Y|\mathfrak{F}_1) = \mathbb{E}[\mathbb{E}(Y|\mathfrak{F}_2)|\mathfrak{F}_1]$ c.s.

[4] es decir Y es independiente de $\mathbb{I}_B$ para todo $B \in \mathfrak{F}$

9. Sea $F_1 \in \mathfrak{F}_1$, por definición de $\mathbb{E}[\mathbb{E}(Y|\mathfrak{F}_2)|\mathfrak{F}_1]$

$$\int_{F_1} \mathbb{E}[\mathbb{E}(Y|\mathfrak{F}_2)|\mathfrak{F}_1]d\mathbb{P} \underbrace{=}_{(2.5)} \int_{F_1} \mathbb{E}(Y|\mathfrak{F}_2)d\mathbb{P}.$$

Observemos que $\mathbb{E}(Y|\mathfrak{F}_2)$ es $\mathfrak{F}_2$-medible y por lo tanto $\mathfrak{F}_1$-medible, con lo cual es igual c.s. a $\mathbb{E}[\mathbb{E}(Y|\mathfrak{F}_2)|\mathfrak{F}_1]$, por unicidad de $\mathbb{E}[\mathbb{E}(Y|\mathfrak{F}_2)|\mathfrak{F}_1]$

10. Como $\mathbb{E}(Y)$ es $\mathfrak{F}$-medible, hay que probar que para todo $F \in \mathfrak{F}$,

$$\int_F Y d\mathbb{P} = \int_F \mathbb{E}(Y)d\mathbb{P}.$$

Es decir $\mathbb{E}[Y\mathbb{I}_F] = \mathbb{E}(Y)\mathbb{E}(\mathbb{I}_F)$. Esto vale, por la hipótesis de independencia de Y con $\mathfrak{F}$, si $\mathbb{E}(|Y|) < \infty$ [5] Como estamos suponiendo que existe $\mathbb{E}(Y)$, hay que descomponer $Y = Y^+ - Y^-$. Entonces, como $\min\{\mathbb{E}(Y^+\mathbb{I}_F), \mathbb{E}(Y^-\mathbb{I}_F)\} < \infty$, ya que $\min\{\mathbb{E}(Y^+), \mathbb{E}(Y^-)\} < \infty$,

$$\mathbb{E}[Y\mathbb{I}_F] = \mathbb{E}[(Y^+ - Y^-)\mathbb{I}_F] = \mathbb{E}[Y^+\mathbb{I}_F] - \mathbb{E}[Y^-\mathbb{I}_F].$$

Como Y^+ y $\mathbb{I}_F$ son no negativas, e independientes vale que $\mathbb{E}[Y^+\mathbb{I}_F] = \mathbb{E}[Y^+]\mathbb{E}[\mathbb{I}_F]$, y lo mismo para Y^-. Con lo cual,

$$\mathbb{E}[Y\mathbb{I}_F] = [\mathbb{E}(Y^+) - \mathbb{E}(Y^-)]\mathbb{E}(\mathbb{I}_F) = \mathbb{E}(Y)\mathbb{E}(\mathbb{I}_F).$$

11. El punto 11 lo vamos a probar a partir del siguiente teorema.

□

Teorema 2.8. *Sea $\{Y_n\}_n$ una sucesión de variable a valores en $\overline{\mathbb{R}}$, definidas en $(\Omega, \mathcal{A}, \mathbb{P})$, $\mathcal{A}$ medibles, y $\mathfrak{F} \subset \mathcal{A}$ una σ-álgebra. Supongamos $\mathbb{E}(|Y_n|) < \infty$ para todo n.*

1. *Si $|Y_n| \leq Z$ tal que $\mathbb{E}(Z) < \infty$ y $Y_n \to Y$ c.s., entonces $\mathbb{E}(Y_n|\mathfrak{F}) \to \mathbb{E}(Y|\mathfrak{F})$ c.s., y $\mathbb{E}(|Y_n - Y||\mathfrak{F}) \to 0$, c.s.*
2. *Si $Y_n \geq Z$ tal que $\mathbb{E}(Z) > -\infty$ y $Y_n \uparrow Y$ c.s., [6] entonces $\mathbb{E}(Y_n|\mathfrak{F}) \uparrow \mathbb{E}(Y|\mathfrak{F})$ c.s.*
3. *Si $Y_n \leq Z$ tal que $\mathbb{E}(Z) < \infty$ y $Y_n \downarrow Y$ c.s, entonces $\mathbb{E}(Y_n|\mathfrak{F}) \downarrow \mathbb{E}(Y|\mathfrak{F})$ c.s.*
4. *Si $Y_n \geq Z$ y $\mathbb{E}(Z) > -\infty$, entonces $\mathbb{E}(\underline{\lim} Y_n|\mathfrak{F}) \leq \underline{\lim}\mathbb{E}(Y_n|\mathfrak{F})$, c.s.*
5. *Si $Y_n \leq Z$ y $\mathbb{E}(Z) < \infty$ entonces $\overline{\lim}\mathbb{E}(Y_n|\mathfrak{F}) \leq \mathbb{E}(\overline{\lim} Y_n|\mathfrak{F})$ c.s.*
6. *Si $Y_n \geq 0$ entonces $\mathbb{E}(\sum_n Y_n|\mathfrak{F}) = \sum_n \mathbb{E}(Y_n|\mathfrak{F})$ c.s. [7] En particular, si $B_1, B_2, \ldots$ son disjuntos 2 a 2,*

$$\mathbb{P}(\cup_n B_n|\mathfrak{F}) = \sum_n \mathbb{P}(B_n|\mathfrak{F}) \quad c.s.$$

Demostración.

1. Sea $\zeta_n = \sup_{m \geq n} |Y_m - Y| \geq 0$. Como $\mathbb{E}(Y_n)$ y $\mathbb{E}(Y)$ son finitas,

$$|\mathbb{E}(Y_n|\mathfrak{F}) - \mathbb{E}(Y|\mathfrak{F})| \underbrace{=}_{\text{Propiedad 4}} |\mathbb{E}(Y_n - Y|\mathfrak{F})| \underbrace{\leq}_{\text{Propiedad 3}} \mathbb{E}(|Y_n - Y||\mathfrak{F}) \underbrace{\leq}_{\text{Propiedad 2}} \mathbb{E}(\zeta_n|\mathfrak{F}).$$

Como $\zeta_{n+1} \leq \zeta_n$ c.s., tenemos que $\mathbb{E}(\zeta_{n+1}|\mathfrak{F}) \leq \mathbb{E}(\zeta_n|\mathfrak{F})$ c.s. por la propiedad 2, por lo tanto existe $h = \lim_n \mathbb{E}(\zeta_n|\mathfrak{F}) \geq 0$ c.s., entonces

$$0 \leq \int_\Omega h d\mathbb{P} \leq \int_\Omega \mathbb{E}(\zeta_n|\mathfrak{F})d\mathbb{P} = \int_\Omega \zeta_n d\mathbb{P}.$$

La segunda desigualdad es porque $h \leq \mathbb{E}(\zeta_n|\mathfrak{F})$ c.s., para todo n y la tercera por (2.5). Como $0 \leq \zeta_n \leq |Y_n| + |Y| \leq 2Z$ y $\mathbb{E}(Z) < \infty$, por $Y_m \to Y$ c.s. $\zeta_n \to 0$ c.s. , y del Teorema de Convergencia Dominada, $\int_\Omega \zeta_n d\mathbb{P} \to 0$, por lo tanto $\int_\Omega h d\mathbb{P} = 0$ y como $h \geq 0$, $h = 0$ c.s.

2. Supongamos primero $Z \equiv 0$. Como Y_n es creciente, por la propiedad 2, $0 \leq \mathbb{E}(Y_n|\mathfrak{F}) \leq \mathbb{E}(Y_{n+1}|\mathfrak{F})$ c.s., por lo tanto existe $\zeta = \lim_n \mathbb{E}(Y_n|\mathfrak{F})$ c.s.. Por (2.5), para todo $F \in \mathfrak{F}$,

$$\int_F Y_n d\mathbb{P} = \int_F \mathbb{E}(Y_n|\mathfrak{F})d\mathbb{P}.$$

[5] aquí se usa que si Z y Y son v.a. independientes tal que $\mathbb{E}|Y| < \infty$ y $\mathbb{E}|Z| < \infty$, entonces $\mathbb{E}(YZ) = \mathbb{E}(Y)\mathbb{E}(Z)$.
[6] vamos a usar la notación $\uparrow$ para indicar que una sucesión es creciente, y $\downarrow$ que es decreciente, no necesariamente en sentido estricto.
[7] la suma no necesariamente es de una cantidad finita de variables, ni tiene por qué ser convergente

Por el Teorema de Convergencia Monótona,[8] para todo $F \in \mathfrak{F}$,

$$\int_F Y_n d\mathbb{P} \to \int_F Y d\mathbb{P} = \int_F \mathbb{E}(Y|\mathfrak{F}) d\mathbb{P} \qquad y \qquad \int_F \mathbb{E}(Y_n|\mathfrak{F}) d\mathbb{P} \to \int_F \zeta d\mathbb{P}.$$

Por lo tanto

$$\int_F \mathbb{E}(Y|\mathfrak{F}) d\mathbb{P} = \int_F \zeta d\mathbb{P}.$$

Como la identidad anterior vale para todo $F \in \mathfrak{F}$ se sigue que $\zeta = \mathbb{E}(Y|\mathfrak{F})$ c.s. por unicidad de la esperanza condicional. Para el caso general observemos que si $Y_n \uparrow Y$ entonces $0 \leq Y_n^+ \uparrow Y^+$. [9] De lo que ya probamos se sigue que $\mathbb{E}(Y_n^+|\mathfrak{F}) \uparrow \mathbb{E}(Y^+|\mathfrak{F})$, c.s. . Además, es fácil ver que $Y_n^- \downarrow Y^-$ y $0 \leq \mathbb{E}(Y_n^-) \leq \mathbb{E}(Z^-) < \infty$ para todo n, con lo cual, por el punto 1, $\mathbb{E}(Y_n^-|\mathfrak{F}) \downarrow \mathbb{E}(Y^-|\mathfrak{F})$. Esto, junto con $\mathbb{E}(Y_n^+|\mathfrak{F}) \uparrow \mathbb{E}(Y^+|\mathfrak{F})$, termina de probar el punto 2.

3. El punto 3 se sigue del punto 2 tomando $-Y_n$.

4. Sea $\zeta_n = \text{ínf}_{m \geq n} Y_m$, entonces $\zeta_n \uparrow \underline{\text{lím}} Y_n$. Del punto 2, $\mathbb{E}(\zeta_n|\mathfrak{F}) \uparrow \mathbb{E}(\underline{\text{lím}} Y_n|\mathfrak{F})$ c.s., en donde, para aplicar el punto 2, usamos que $Y_n \geq Z$ por lo tanto, tomando ínfimo, $\zeta_n \geq Z$ y $\mathbb{E}(Z) > -\infty$. Por lo tanto

$$\mathbb{E}(\underline{\text{lím}} Y_n|\mathfrak{F}) = \lim_n \mathbb{E}(\zeta_n|\mathfrak{F}) = \underline{\text{lím}}\, \mathbb{E}(\zeta_n|\mathfrak{F}) \leq \underline{\text{lím}}\, \mathbb{E}(Y_n|\mathfrak{F})$$

donde en la última desigualdad usamos que $\zeta_n \leq Y_n$ y la anterior es porque al existir el límite, coincide con su límite inferior.

5. Se sigue del punto anterior tomando $-Y_n$ y usando que $\underline{\text{lím}} a_n = \overline{\text{lím}} - a_n$ para cualquier sucesión a_n.

6. De (2.7)

$$\mathbb{E}\left(\sum_{k=1}^n Y_k \Big| \mathfrak{F}\right) = \sum_{k=1}^n \mathbb{E}(Y_k|\mathfrak{F}), \quad c.s.$$

y se concluye usando el punto 2. El caso particular se sigue de tomar $Y_n = \mathbb{I}_{B_n}$

□

Observación 2.9. *La conclusión que se obtiene para conjuntos en el punto 6 del teorema anterior establece que, fijados $B_1, B_2, \ldots$, para todo ω en Ω excepto en un conjunto de probabilidad nula, las variables aleatorias $\mathbb{P}(\cup_n B_n|\mathfrak{F})$ y $\sum \mathbb{P}(B_n|\mathfrak{F})$ son iguales. No obstante, si cambiamos los conjuntos $B_1, B_2, \ldots$, el subconjunto de Ω donde esta igualdad vale, cambia. Por lo tanto, no podemos afirmar que exista un conjunto de probabilidad 1, para el cual $\mathbb{P}(\cdot|\mathfrak{F})$ sea una medida, ya que tendríamos que excluir todos los conjuntos de probabilidad nula que se obtienen cambiando los $B_1, B_2, \ldots$. Pero esto no queda necesariamente un conjunto con probabilidad nula, porque tenemos, posiblemente, una cantidad no numerable de subconjuntos de Ω a excluir, ya que tenemos una cantidad no numerable de posibles $B_1, B_2, \ldots$ disjuntos. Se puede probar que dado $\mathfrak{F}$, existe una version de $\mathbb{P}(\cdot|\mathfrak{F})$,[10] que se denomina regular, para la cual $\mathbb{P}(\cdot|\mathfrak{F})(\omega)$ es una medida, para todo ω excepto en un conjunto de probabilidad nula.*

Definición 2.10. Una función $\mathbb{P}(\omega; B)$ definieda para todo $\omega \in \Omega$ y $B \in \mathfrak{F}$ es una probabilidad condicional regular respecto de $\mathfrak{F}$ si

1. $\mathbb{P}(\omega; \cdot)$ es una medida de probabilidad en $\mathfrak{F}$ para todo $\omega \in \Omega$.

2. Para cada $B \in \mathfrak{F}$ la función $\mathbb{P}(\omega; B) = \mathbb{P}(B|\mathfrak{F})(\omega)$ c.s.

Prueba de la propiedad 11 de la Proposición 2.7

Sea Z medible respecto de $\mathfrak{F}$ tal que $\mathbb{E}|Z| < \infty$, y $\mathbb{E}|YZ| < \infty$ entonces $\mathbb{E}(YZ|\mathfrak{F}) = Z\mathbb{E}(Y|\mathfrak{F})$ c.s. Supongamos que $Z = \mathbb{I}_B$ con $B \in \mathfrak{F}$, para todo $A \in \mathfrak{F}$,

$$\int_A \mathbb{E}(YZ|\mathfrak{F}) d\mathbb{P} \underbrace{=}_{(2.5)} \int_A YZ d\mathbb{P} = \int_{A \cap B} Y d\mathbb{P} \underbrace{=}_{(2.5)} \int_{A \cap B} \mathbb{E}(Y|\mathfrak{F}) d\mathbb{P} = \int_A \mathbb{I}_B \mathbb{E}(Y|\mathfrak{F}) d\mathbb{P} = \int_A Z\mathbb{E}(Y|\mathfrak{F}) d\mathbb{P},$$

donde en la tercera igualdad usamos que $A \cap B \in \mathfrak{F}$. Como la igualdad anterior vale para todo $A \in \mathfrak{F}$ tenemos que vale la propiedad para $Z = \mathbb{I}_B$ con $B \in \mathfrak{F}$. Por la linealidad de la integral, vale la propiedad 11 para funciones simples: $Z = \sum_{k=1}^n y_k \mathbb{I}_{B_k}$ con $B_k \in \mathfrak{F}$.

[8] observar que aquí usamos que $Y_n \geq 0$ para todo n

[9] esto se sigue de que si ω es tal que $Y_{n-1}^-(\omega) = 0$, $Y_n^+(\omega) \geq Y_n(\omega) \geq Y_{n-1}(\omega) = Y_{n-1}^+(\omega)$, el caso $Y_{n-1}^-(\omega) > 0$ es análogo ya que aquí $Y_{n-1}^+(\omega) = 0$ y siempre $Y_n^+ \geq 0$ con lo cual $Y_n^+(\omega) \geq Y_{n-1}^+(\omega)$

[10] es decir, es igual a $\mathbb{P}(\cdot|\mathfrak{F})$ excepto en un conjunto de probabilidad nula

Sea ahora Z una función $\mathfrak{F}$-medible tal que $\mathbb{E}|Z| < \infty$, y $\{Z_n\}_n$ una sucesión de funciones simples $\mathfrak{F}$-medibles, tal que $|Z_n| \leq Z$ y $Z_n \to Z$ c.s. [11] Por lo tanto $\mathbb{E}(YZ_n|\mathfrak{F}) = Z_n\mathbb{E}(Y|\mathfrak{F})$. Como $|YZ_n| \leq |YZ|$ y estamos suponiendo que $\mathbb{E}|YZ| < \infty$, por el punto 1 del Teorema anterior, $\mathbb{E}(YZ_n|\mathfrak{F}) \to \mathbb{E}(YZ|\mathfrak{F})$ c.s. Observemos que $|\mathbb{E}(Y|\mathfrak{F})| \leq \mathbb{E}(|Y||\mathfrak{F})$ y la variable $\mathbb{E}(|Y||\mathfrak{F})$ es finita c.s. porque $\mathbb{E}|Y| < \infty$. Por lo tanto $Z_n\mathbb{E}(Y|\mathfrak{F}) \to Z\mathbb{E}(Y|\mathfrak{F})$ y esto concluye la demostración. Observar que es esencial que $\mathbb{E}(Y|\mathfrak{F}) < \infty$ ya que si $Z_n = 1/n$, $(1/n)\infty = \infty$ que no converge a 0∞ que es 0.

Ejercicio 2.11. Desigualdad de Cauchy-Schwartz. Probar que si $X, Y \in L^2(\Omega)$,

$$\mathbb{E}[XY|\mathcal{A}] \leq \sqrt{\mathbb{E}[X^2|\mathcal{A}] \cdot \mathbb{E}[Y^2|\mathcal{A}]}, \quad c.s. \tag{2.8}$$

en particular $\mathbb{E}[|X||\mathcal{A}] \leq \sqrt{\mathbb{E}[X^2|\mathcal{A}]}$ c.s.

Teorema 2.12. ***Desigualdad de Jensen.*** *Sean X e Y variables aleatorias y ϕ una función convexa.*[12] *Supongamos que $\mathbb{E}|Y| < \infty$ y $\mathbb{E}|\phi(Y)| < \infty$.*

$$\phi(\mathbb{E}(Y|X)) \leq \mathbb{E}(\phi(Y)|X) \tag{2.9}$$

Observación 2.13. *Consideremos X, Y, Z tal que $X = Y$, $Z = -Y$ e Y con distribución $N(0,1)$ aunque X y Z tienen la misma distribución sus distribuciones conjuntas con Y son diferentes, además*

$$\mathbb{E}[X|Y] = \mathbb{E}[Y|Y] = Y \qquad \textit{mientras que} \qquad \mathbb{E}[Z|Y] = \mathbb{E}[-Y|Y] = -Y$$

En este caso, las esperanzas condicionales son diferentes. Es decir tener la misma distribución no garantiza que la esperanza condicional respecto de una tercera variable sea la misma. Lo que si es cierto, (y se deja como ejercicio verificarlo) es que si tenemos (X, Y) con la misma distribución conjunta que (Z, T) entonces $\mathbb{E}(X|Y) = \mathbb{E}(Z|T)$ c.s.

2.2 Esperanza condicional respecto de una variable

La esperanza condicional respecto de una variable aleatoria está motivada por el siguiente resultado, cuya demostración puede encontrar en [29], p. 174.[13]

Teorema 2.14. *Sea $(\Omega, \mathcal{A}, \mathbb{P})$ un espacio de probabilidad y $X : \Omega \to \mathbb{R}$. Sea $\mathfrak{F}_X \subset \mathcal{A}$ la menor σ-álgebra que hace que X sea medible, una variable Z es medible respecto de $\mathfrak{F}_X$ si y solo si existe $m : \overline{\mathbb{R}} \to \overline{\mathbb{R}}$ una función $\mathcal{B}(\overline{\mathbb{R}})$-medible tal que $Z = m(X)$.*

Observación 2.15. *El teorema no implica que si $m(X) = Z$, la σ-álgebra generada por Z es $\mathfrak{F}_X$, podría ser más chica. Un caso extremo de esto es cuando Z es constante.*

Definición 2.16. Si Y es una variable aleatoria y $\mathfrak{F}_X$ es la σ-álgebra generada por una variable X,[14] se define $\mathbb{E}(Y|X) \equiv \mathbb{E}(Y|\mathfrak{F}_X)$ si esta última está bien definida. De forma análoga se define $\mathbb{P}(B|X)$ como $\mathbb{E}(\mathbb{I}_B|\mathfrak{F}_X)$.

Observación 2.17. *En virtud del Teorema 2.14, por ser $\mathbb{E}(Y|X)$ una función $\mathfrak{F}_X$ medible, $\mathbb{E}(Y|X) = m(X)$, para alguna m. Es decir la esperanza condicional es una función de X. Además, por definición, cualquier otra variable Z cuya σ-álgebra generada sea $\mathfrak{F}_X$ va a cumplir que $\mathbb{E}(Y|Z) = \mathbb{E}(Y|X)$.*

2.3 Esperanza condicional respecto de $X = x$

Veamos dos formas de definir la esperanza condicional de una variable Y respecto de $X = x$.

La primera es usando la definición que ya dimos: como $\mathbb{E}(Y|X)$ es $\mathfrak{F}_X$-medible, siendo $\mathfrak{F}_X$ la menor σ-álgebra que hace que X sea medible, existe (ver Teorema 2.14) $m : \overline{\mathbb{R}} \to \overline{\mathbb{R}}$ una función $\mathcal{B}(\overline{\mathbb{R}})$-medible, tal que

$$m\big(X(\omega)\big) = \mathbb{E}(Y|X)(\omega), \quad \text{para todo } \omega \in \Omega.^{15}.$$

Como $\mathbb{E}(Y|X)$ está definida a menos de conjuntos de probabilidad nula, m está definida a menos de conjuntos de medida P_X nula.

Denotamos $m(x) = \mathbb{E}(Y|X = x)$. Observar que para todo $A \in \mathfrak{F}_X$,

$$\int_A Y d\mathbb{P} = \int_A \mathbb{E}(Y|X) d\mathbb{P} = \int_A m(X) d\mathbb{P}.$$

[11] Aquí estamos usando un resultado clásico de teoría de la medida que establece que cualquier función medible se puede aproximar por funciones simples, c.s. Más aún, si la función a aproximar es positiva, la sucesión de funciones que aproximan se puede tomar no decreciente.

[12] Una función $\phi : \mathbb{R} \to \mathbb{R}$ es convexa cuando $\phi(\lambda x + (1-\lambda)y) \leq \lambda\phi(x) + (1-\lambda)\phi(y)$ para todo $x, y \in \mathbb{R}$ y $\lambda \in (0,1)$.

[13] En [29] se prueba para variables a valores reales, pero la prueba se extiende de manera natural a variables a valores en $\overline{\mathbb{R}}$.

[14] es decir la menor σ-álgebra que hace medible a X

[15] aquí $\mathbb{E}(Y|X)$ refiere a la variable aleatoria introducida en la Definición 2.16

Sea $B \in \mathcal{B}(\overline{\mathbb{R}})$, haciendo un cambio de variable (ver Teorema 10.7 en el apéndice),

$$\int_{\{\omega: X(\omega)\in B\}} m(X) d\mathbb{P} = \int_B m(x) P_X(dx)$$

donde P_X es la distribución de X.

Veamos la otra, cuya existencia se sigue del Teorema de Radón-Nikodym,

Definición 2.18. Sean Y y X dos variables aleatorias a valores en $\overline{\mathbb{R}}$, supongamos que $\mathbb{E}(Y)$ está definida. La esperanza condicional de Y respecto de $X = x$, que denotamos $\mathbb{E}(Y|X = x)$, es cualquier $\mathcal{B}(\overline{\mathbb{R}})$-medible $m : \overline{\mathbb{R}} \to \overline{\mathbb{R}}$ que cumpla que para todo $B \in \mathcal{B}(\overline{\mathbb{R}})$,

$$\int_{\{\omega: X(\omega)\in B\}} Y d\mathbb{P} = \int_B m(x) P_X(dx). \tag{2.10}$$

La existencia y unicidad (a menos de conjuntos de medida cero respecto de P_X) de esta función se sigue de que la medida Q definida en $\mathcal{B}(\overline{\mathbb{R}})$ como

$$Q(B) = \int_{\{\omega: X(\omega)\in B\}} Y d\mathbb{P}$$

es absolutamente continua respecto de P_X la distribución de X.

Veamos que si m es una función que cumple (2.10) entonces $m(X) = \mathbb{E}(Y|X)$. Por el Teorema de cambio de Variable de la integral de Lebesgue (ver Teorema 10.7 en el apéndice)

$$\int_{\{\omega: X(\omega)\in B\}} Y d\mathbb{P} \underbrace{=}_{(2.10)} \int_B m(x) P_X(dx) = \int_{\{\omega: X(\omega)\in B\}} m(X) d\mathbb{P}$$

para todo $B \in \mathcal{B}(\overline{\mathbb{R}})$. Por otra parte, por definición de $\mathbb{E}(Y|X)$,

$$\int_{\{\omega: X(\omega)\in B\}} Y d\mathbb{P} = \int_{\{\omega: X(\omega)\in B\}} \mathbb{E}(Y|X) d\mathbb{P}.$$

Por lo tanto, para todo $B \in \mathcal{B}(\overline{\mathbb{R}})$

$$\int_{\{\omega: X(\omega)\in B\}} \mathbb{E}(Y|X) d\mathbb{P} = \int_{\{\omega: X(\omega)\in B\}} m(X) d\mathbb{P}$$

Como la función $m(X)$ es $\mathfrak{F}_X$-medible, y $\sigma(\{\omega : X(\omega) \in B : B \in \mathcal{B}(\overline{\mathbb{R}})\}) = \mathfrak{F}_X$, tenemos que $m(X) = \mathbb{E}(Y|X)$.

Por lo tanto, si conocemos $\mathbb{E}(Y|X = x)$ podemos obtener $\mathbb{E}(Y|X)$. Recíprocamente, si conocemos $\mathbb{E}(Y|X)$ vimos que queda unívocamente definida m. Las 11 propiedades de la Proposición 2.7 que probamos para $\mathbb{E}(Y|X)$ valen para $\mathbb{E}(Y|X = x)$ cambiando la igualdad c.s, por una igualdad P_X c.s. Y lo mismo para los 6 puntos del Teorema 2.8. Por ejemplo la propiedad 11 se escribe de la siguiente forma: Si $\mathbb{E}|Y| < \infty$ y $\mathbb{E}|Yf(X)| < \infty$, donde f es $\mathcal{B}(\overline{\mathbb{R}})$-medible, entonces

$$\mathbb{E}(Yf(X)|X = x) = f(x)\mathbb{E}(Y|X = x) \quad P_X \text{ c.s.}^{16}$$

Por su parte el análogo de la propiedad 10 establece que si Y y X son independientes, entonces

$$\mathbb{E}(Y|X = x) = \mathbb{E}(Y) \quad P_X \text{ c.s.}$$

En general vale el siguiente resultado

Teorema 2.19.

1. Sean Y y X variables aleatorias y $\varphi = \varphi(x, y)$ $\mathcal{B}(\mathbb{R}^2)$-medible, tal que $\mathbb{E}|\varphi(X, Y)| < \infty$, entonces

$$\mathbb{E}\big(\varphi(X, Y)|X = x\big) = \mathbb{E}\big(\varphi(x, Y)|X = x\big) \quad P_X \text{ c.s.} \tag{2.11}$$

2. Si Y y X son independientes y $\varphi = \varphi(x, y)$ es $\mathcal{B}(\mathbb{R}^2)$-medible, tal que $\mathbb{E}|\varphi(X, Y)| < \infty$, entonces

$$\mathbb{E}\big(\varphi(X, Y)|X = x\big) = \mathbb{E}\big(\varphi(x, Y)\big) \quad P_X \text{ c.s.} \tag{2.12}$$

Demostración.

[16] esto significa que la igualdad vale para todo x excepto en un conjunto de medida cero respecto de P_X

1. Veamos primero que vale para $\varphi(x,y) = \mathbb{I}_B(x,y)$ con $B \in \mathcal{B}(\mathbb{R}^2)$, supongamos primero que $B = B_1 \times B_2$ con $B_i \in \mathcal{B}(\mathbb{R})$ para $i = 1, 2$. Por (2.10) tenemos que probar que para todo $A \in \mathcal{B}(\overline{\mathbb{R}})$

$$\int_{\{\omega: X(\omega)\in A\}} \mathbb{I}_{B_1\times B_2}(X,Y)d\mathbb{P} = \int_A \mathbb{E}\big(\mathbb{I}_{B_1}(x)\mathbb{I}_{B_2}(Y)|X=x\big)P_X(dx). \tag{2.13}$$

La integral de la izquierda es $\mathbb{P}(\{X \in B_1 \cap A\} \cap \{Y \in B_2\})$. La de la derecha es

$$\int_A \mathbb{I}_{B_1}(x)\mathbb{E}\big(\mathbb{I}_{B_2}(Y)|X=x\big)P_X(dx) = \int_{A\cap B_1} \mathbb{E}\big(\mathbb{I}_{B_2}(Y)|X=x\big)P_X(dx).$$

Si ahora usamos (2.10) con $A = A \cap B_1$ y $Y = \mathbb{I}_{B_2}(Y)$ obtenemos que la integral anterior es

$$\int_{\{\omega: X(\omega)\in A\cap B_1\}} \mathbb{I}_{B_2}(Y)d\mathbb{P} = \mathbb{P}\big(\{Y \in B_2\} \cap \{X \in A \cap B_1\}\big).$$

Por lo tanto hemos probado (2.11), es decir el resultado vale para $\varphi(x,y) = \mathbb{I}_B(x,y)$ con $B \in \mathcal{B}(\mathbb{R}^2)$.

De esto se sigue, usando la linealidad de la esperanza, que vale para indicatrices de uniones finitas de rectángulos disjuntos de la forma $B_1 \times B_2$. Estos conjuntos forman un álgebra y por lo tanto vale para indicatrices de conjuntos en un álgebra. Para ver que vale para indicatrices de cualquier conjunto $\mathcal{B}(\mathbb{R}^2)$-medible se prueba que la clase de conjuntos que la cumplen son una clase monótona [17]. Esto último se hace a través del Teorema de Fubini. Por lo tanto se cumple para una clase monótona que contiene a un álgebra cuya σ-álgebra generada es la σ-álgebra de Borel. Es decir se prueba que vale para la indicatriz de cualquier boreliano. La tesis del teorema se sigue de aproximar φ por combinaciones lineales finitas de indicatrices de $\mathbb{I}_B$, siendo B un boreliano.

2. Veamos primero que vale para $\varphi(x,y) = \mathbb{I}_B(x,y)$ con $B \in \mathcal{B}(\mathbb{R}^2)$. Es decir

$$\mathbb{E}\big[\mathbb{I}_B(X,Y)|X=x\big] = \mathbb{E}\big[\varphi(x,Y)\big] \quad P_X \text{ c.s.}$$

Para ver esto supongamos que $B = B_1 \times B_2$, con $B_i \in \mathcal{B}(\mathbb{R})$ para $i = 1, 2$. Por (2.10) tenemos que ver que para todo $A \in \mathcal{B}(\overline{\mathbb{R}})$

$$\int_{\{\omega: X(\omega)\in A\}} \mathbb{I}_{B_1\times B_2}(X,Y)d\mathbb{P} = \int_A \mathbb{E}\big(\mathbb{I}_{B_1\times B_2}(x,Y)\big)P_X(dx).$$

La integral de la izquierda es $\mathbb{P}(Y \in B_2, X \in A \cap B_1)$ y usando la independencia es

$$\mathbb{P}(Y \in B_2)\mathbb{P}(X \in A \cap B_1).$$

La integral de la derecha es

$$\int_A \mathbb{I}_{B_1}(x)\mathbb{E}\big(\mathbb{I}_{B_2}(Y)\big)P_X(dx) = \mathbb{P}(Y \in B_2)\int_A \mathbb{I}_{B_1}(x)P_X(dx) = \mathbb{P}(Y \in B_2)\mathbb{P}(X \in A \cap B_1).$$

Se concluye ahora igual que como hicimos en la parte anterior.

□

Al igual que hicimos antes, podemos definir $\mathbb{P}(A|X=x)$.

Definición 2.20. Dado $A \in \mathfrak{F}$ se define $\mathbb{P}(A|X=x) = \mathbb{E}(\mathbb{I}_A|X=x)$.

Observar que de (2.10) se sigue que, para todo $B \in \mathcal{B}(\overline{\mathbb{R}})$

$$\mathbb{P}\big(A \cap \{X \in B\}\big) = \int_B \mathbb{P}(A|X=x)P_X(dx).$$

Veamos algunos casos particulares.

Ejemplo 2.21. Sea X una variable discreta tal que $\mathbb{P}(X = x_k) > 0$, y $\sum_k \mathbb{P}(X = x_k) = 1$. Entonces, para todo $k \geq 1$,

$$\mathbb{P}(A|X=x_k) = \frac{\mathbb{P}\big(A \cap \{X = x_k\}\big)}{\mathbb{P}(X=x_k)}.$$

Esto es un caso particular de lo siguiente: si Y es tal que $\mathbb{E}(Y)$ existe, entonces

$$\mathbb{E}(Y|X=x_k) = \frac{1}{\mathbb{P}(X=x_k)}\int_{\{\omega: X(\omega)=x_k\}} Y d\mathbb{P}.$$

Esta última igualdad se sigue de forma inmediata de (2.10). Da una idea de qué es $\mathbb{E}[Y|X=x]$ cuando $\mathbb{P}(X=x) > 0$. Para ese x la función $m(x) = \mathbb{E}[Y|X=x]$ calcula la esperanza de la variable Y' definida como Y pero en el espacio muestral $(\Omega', \mathcal{A}', \mathbb{P}')$, donde $\Omega' = \{\omega : X(\omega) = x\}$, $\mathcal{A}'$ es la restricción a Ω' de $\mathcal{A}$ y $\mathbb{P}'$ es la probabilidad $\mathbb{P}$ condicionada a Ω'.

[17] recordar que una clase monótona es cerrada por uniones numerables crecientes y por intersecciones numerables decrecientes, además, la clase monótona generada por un álgebra coincide con la σ álgebra generada por el álgebra

Ejercicio 2.22. Probar que si (X,Y) es un vector aleatorio bidimensional tal que $Rec(X,Y) = \{(x_n, y_m) : n, m \in \mathbb{N}\}$ y definimos la probabilidad condicional en el sentido usual, como

$$\mathbb{P}_{Y|X=x}(y) = \mathbb{P}(Y = y|X = x) = \frac{\mathbb{P}_{Y,X}(y,x)}{\mathbb{P}_X(x)} \quad \forall y \in Rec(Y), \forall x \in Rec(X),$$

entonces, si $\sum_{k,j}^{\infty} |y_j| \mathbb{P}_{Y,X}(y_j, x_k) < \infty$,

$$\mathbb{E}(Y|X) = \sum_{y \in Rec(Y)} y \mathbb{P}_{Y|X}(y), \tag{2.14}$$

donde $\mathbb{P}_{Y|X}(y)$ es la variable aleatoria definida en $\omega \in \Omega$ como $\mathbb{P}_{Y|X}(y)(\omega) = \mathbb{P}_{Y|X=X(\omega)}(y)$.

Proposición 2.23. *Sea (X,Y) un vector aleatorio tal que existe la densidad $f_{X,Y}(x,y)$ de la distribución conjunta. Sean f_Y y f_X las marginales de Y y X respectivamente. Supongamos que $\int_{-\infty}^{\infty} \int_{-\infty}^{\infty} |y| f_{X,Y}(x,y) dxdy < \infty$. Definamos, si $f_X(x) = 0$, $f_{Y|X}(y|x) = 0$ y si $f_X(x) \neq 0$,*

$$f_{Y|X}(y|x) = \frac{f_{X,Y}(x,y)}{f_X(x)},$$

entonces

$$\mathbb{P}(Y \in C|X = x) = \int_C f_{Y|X}(y|x) dy. \tag{2.15}$$

De forma similar, si existe $\mathbb{E}(Y)$,

$$\mathbb{E}(Y|X = x) = \int_{\mathbb{R}} y f_{Y|X}(y|x) dy. \tag{2.16}$$

donde las igualdades (2.15) y (2.16) son para todo x salvo un conjunto de medida P_X nula.

Demostración. Por definición $\mathbb{P}(Y \in C|X = x) = \mathbb{E}(\mathbb{I}_{Y^{-1}(C)}|X = x)$, por lo tanto por (2.10) basta probar que para todo $B \in \mathcal{B}(\overline{\mathbb{R}})$

$$\int_{X \in B} \mathbb{I}_{Y^{-1}(C)} d\mathbb{P} = \int_B \left[\int_C f_{Y|X}(y|x) dy \right] dP_X(x).$$

$$\int_B \int_C f_{Y|X}(y|x) dy dP_X(x) = \int_B \left[\int_C f_{Y|X}(y|x) dy \right] f_X(x) dx = \int_{B \times C} f_{Y,X}(x,y) dxdy.$$

Por otra parte

$$\int_{X \in B} \mathbb{I}_{Y^{-1}(C)} d\mathbb{P} = \mathbb{P}\big(\{Y \in C\} \cap \{X \in B\}\big) = \int_{B \times C} f_{X,Y}(x,y) dxdy.$$

La prueba de (2.16) se sigue de (2.15) aproximando Y por funciones simples, se deja como ejercicio. □

Se puede probar el siguiente teorema (ver Teorema 3 en p. 226 en [29]), que generaliza (2.16) y (2.14)

Teorema 2.24. *Si $\mathbb{P}(\omega; B)$ es una probabilidad condicional regular (ver Definición 2.10) respecto a $\mathfrak{F}$ y $\mathbb{E}|Y| < \infty$,*

$$\mathbb{E}(Y|\mathfrak{F})(\omega) = \int_\Omega Y(\tilde{\omega}) \mathbb{P}(\omega, d\tilde{\omega}),$$

en particular, si denotamos $\mathbb{P}_x(\cdot)$ la medida en la σ-álgebra de Borel, definida como $\mathbb{P}_x(B) = \mathbb{P}(Y \in B|X = x)$,

$$\mathbb{E}(Y|X = x) = \int_{\mathbb{R}} y \mathbb{P}_x(dy), \tag{2.17}$$

ya que, como vimos $\mathbb{E}(Y|X = X(\omega)) = \mathbb{E}(Y|X)(\omega)$.

Observación 2.25. *Para el caso en el cual (X,Y) tiene densidad conjunta, (2.17) es (2.16) (ver, para más detalles, la observación (2.27)), y en el caso discreto (2.17) es (2.14).*

2.4 Distribución Condicional y Varianza Condicional

Definición 2.26. Vamos a introducir la **distribución condicional**, y la **varianza condicional**,

1. La función $F_{Y|X=x}(z) \equiv \mathbb{P}(Y \leq z|X = x)$ se la denomina distribución condicional.

2. Si Y es una variable aleatoria y $A \in \mathcal{A}$ tal que $\mathbb{P}(A) > 0$, $F_{Y|A}(z) \equiv \mathbb{P}(Y \leq z|A)$.

Observación 2.27. *$F_{Y|X=x}(z)$ es, por definición, $\mathbb{E}[\mathbb{I}_{\{Y\leq z\}}|X=x]$, por lo tanto, no es la distribución de la variable aleatoria $\mathbb{E}[Y|X]$. Fijado x no tiene por qué ser la distribución de una variable ya que hereda el problema que se mencionó en la Observación 2.9. No obstante, al igual que antes, se puede probar que hay una versión regular de $F_{Y|X=x}$ (ver Teorema 4 p.227, en [29]), la cual es una función de distribución para todo x excepto en un conjunto de medida nula respecto de P_X. En el caso de tener densidad, está dada por $f_{Y|X}(y|x)$. En lo que sigue vamos a asumir que estamos trabajando con la versión regular de $F_{Y|X=x}$.*

Ejemplo 2.28. Un ejemplo que será de utilidad más adelante es cuando Y toma únicamente los valores 0 o 1, en tal caso,

$$F_{X|Y=1}(x) = \mathbb{P}(X \leq x|Y=1) = \frac{\mathbb{P}(X\leq x, Y=1)}{\mathbb{P}(Y=1)}.$$

La derivada de esta función respecto de x, que denotaremos $f_1(x)$, es la densidad en x, de la variable X, condicional a la clase $Y=1$, mientras que $f_0(x)$ es la densidad en x, de la variable X, condicionada a la clase $Y=0$. Supongamos que X tiene densidad f, veamos que

$$\mathbb{P}(Y=1|X=x)f(x) = f_1(x)\mathbb{P}(Y=1). \tag{2.18}$$

Por definición de $\mathbb{P}(Y=1|X=x) = \mathbb{E}\big[\mathbb{I}_1(Y)|X=x\big]$ y se cumple que para todo boreliano B,

$$\int_B \mathbb{P}(Y=1|X=x)dP_X = \mathbb{E}\big[\mathbb{I}_1(Y)\mathbb{I}_B(X)\big] = \mathbb{P}(Y=1, X\in B)$$

por su parte,

$$\int_B f_1(x)\mathbb{P}(Y=1)dx = \mathbb{P}(X\in B|Y=1)\mathbb{P}(Y=1) = \mathbb{P}(Y=1, X\in B),$$

donde en la primera igualdad hemos usado (2.15) y en la segunda la definición usual de probabilidad condicional entre subconjuntos de Ω. Esto prueba (2.18).

Observación 2.29. *La función f_1 es la densidad de la variable X' definida en el espacio de probabilidad $(\Omega', \mathcal{A}', \mathbb{P}')$ donde que $\Omega' = \{\omega\in\Omega : Y(\omega)=1\}$. La σ-álgebra $\mathcal{A}'$ es la restricción de $\mathcal{A}$ a Ω', es decir $A'\in\mathcal{A}'$ si y solo si existe $A\in\mathcal{A}$ tal que $A' = A\cap\Omega'$, y $\mathbb{P}'$ es la probabilidad condicional, es decir $\mathbb{P}(A') = \mathbb{P}(A'|Y=1)$. En este espacio se define $X'(\omega) = X(\omega)$, para $\omega\in\Omega'$. Lo mismo vale para f_0.*

Veamos la varianza condicional y algunas propiedades.

Definición 2.30. Dada una variable X y otra Y tal que $\mathbb{E}(Y|X)$ está definida, se define la varianza condicional de Y dado X como

$$\mathbb{V}(Y|X) = \mathbb{E}\Big[\big(Y-\mathbb{E}(Y|X)\big)^2|X\Big].$$

Observación 2.31. *La varianza condicional es una nueva variable aleatoria, por lo tanto, no va a ser en general $\mathbb{V}(\mathbb{E}(Y|\mathfrak{F}))$.*

La varianza condicional nos dice cuanta varianza nos queda por explicar si usamos $\mathbb{E}(Y|X)$ para predecir Y, ya que, condicionado a X, $Y-\mathbb{E}(Y|X)$ es el residuo, o error de predicción que tenemos, si usamos como predictor la esperanza condicional.

Por otra parte, el error cuadrático medio si usamos el predictor $f(X)$ es $\mathbb{E}[(Y-f(X))^2]$, si sumamos y restamos $\mathbb{E}(Y|X)$ queda

$$\begin{aligned}
\mathbb{E}[(Y-f(X))^2] =& \ \mathbb{E}\bigg[\Big(Y-\mathbb{E}(Y|X)+\mathbb{E}(Y|X)-f(X)\Big)^2\bigg] \\
=& \ \mathbb{E}\bigg\{\mathbb{E}\Big[\big(Y-\mathbb{E}(Y|X)+\mathbb{E}(Y|X)-f(X)\big)^2\Big|X\Big]\bigg\} \\
=& \ \mathbb{E}\bigg\{\mathbb{E}\Big[\big(Y-\mathbb{E}(Y|X)\big)^2\Big|X\Big]\bigg\} + 2\mathbb{E}\bigg\{\mathbb{E}\Big[\big(Y-\mathbb{E}(Y|X)\big)\big(\mathbb{E}(Y|X)-f(X)\big)\Big|X\Big]\bigg\} \\
&+ \mathbb{E}\bigg\{\mathbb{E}\Big[\big(\mathbb{E}(Y|X)-f(X)\big)^2\Big|X\Big]\bigg\}
\end{aligned} \tag{2.19}$$

Veamos que el segundo término es 0, primero observemos que $\mathbb{E}(Y|X)-f(X)$ es una variable aleatoria medible respecto de X con lo cual sale para afuera, es decir

$$\mathbb{E}\bigg\{\mathbb{E}\Big[\big(Y-\mathbb{E}(Y|X)\big)\big(\mathbb{E}(Y|X)-f(X)\big)\Big|X\Big]\bigg\} = \mathbb{E}\bigg\{\Big(\mathbb{E}(Y|X)-f(X)\Big)\mathbb{E}\Big[\big(Y-\mathbb{E}(Y|X)\big)\Big|X\Big]\bigg\}$$

Por su parte,

$$\mathbb{E}\Big[\big(Y-\mathbb{E}(Y|X)\big)\Big|X\Big]=\mathbb{E}[Y|X]-\mathbb{E}\Big[\mathbb{E}(Y|X)|X\Big]=\mathbb{E}[Y|X]-\mathbb{E}[Y|X]\mathbb{E}[1|X]=0.$$

Si tomamos $f(X)=\mathbb{E}(Y|X)$ se obtiene que que el error cuadrático medio de este predictor es la esperanza de la varianza condicional. Además haciendo una cuenta análoga a (2.19), donde en lugar de tomar esperanza condicional tomamos esperanza, se prueba que

$$\mathbb{E}\big|f(X)-Y\big|^2=\mathbb{E}\big|f(X)-\mathbb{E}(Y|X)\big|^2+\mathbb{E}\big|\mathbb{E}(Y|X)-Y\big|^2.$$

Es decir $\mathbb{E}[Y|X]$ es el mejor predictor de Y usando X (minimiza el error cuadrático medio).

Otra identidad importante es la ley de la varianza total,

$$\mathbb{V}(Y)=\mathbb{E}\big(\mathbb{V}(Y|X)\big)+\mathbb{V}\big(\mathbb{E}(Y|X)\big)$$

Como cada uno de los sumandos en la ecuación anterior son no negativos, obtenemos en particular que $\mathbb{V}(Y)\geq\mathbb{E}\big(\mathbb{V}(Y|X)\big)$ lo cual significa que cuando condicionamos, en promedio reducimos la varianza.

Ejemplo 2.32. Sea Y la altura de una persona elegida al azar en el mundo y X el país de la persona elegida, donde $X=1,2,3,\ldots,n$, y n es el número de países. $\mathrm{Var}(Y\mid X=i)$ es la varianza de Y en el país i. $\mathbb{E}[\mathrm{Var}(Y\mid X)]$ es el promedio de las varianzas en cada país. Por otro lado, $\mathbb{E}[Y\mid X=i]$ es la altura promedio en el país i, por lo tanto $\mathrm{Var}(\mathbb{E}[Y\mid X])$ es la varianza de las alturas medias. Podemos interpretar la ley de la varianza total de la siguiente manera. La varianza de Y se puede descomponer en dos partes: la primera es el promedio de las varianzas en cada país individual, mientras que la segunda es la varianza entre los promedios de altura en cada país.

3 Regresión y clasificación

3.1 Regresión

Predecir o modelar una variable continua Y a partir de otra variable o conjunto de variables X es un problema que tiene diversas aplicaciones. Un ejemplo clásico de modelado de la relación entre X e Y es la regresión lineal, en la cual se asume que tiene validez el modelo $Y = \Gamma(X) + \epsilon$, donde Γ es un funcional lineal y ϵ es un error que usualmente se asume no correlacionado con X. En estos modelos los errores se asumen desconocidos, así como también Γ, y se busca estimar Γ a partir de una muestra de entrenamiento. Según cómo sea el espacio donde viven las X, y la relación Γ, el problema se aborda de diferentes maneras. En el caso de la regresión lineal finito dimensional (es decir X toma valores en $\mathbb{R}^d$), Γ es el producto interno de X con un vector $\beta = (\beta_1, \ldots, \beta_d)$. Si X toma valores en $L^2[0,1]$, Γ es un operador lineal en $L^2[0,1]$, por ejemplo $\Gamma(X) = \int_0^1 \beta(t)X(t)dt$. Si se asume que X toma valores en un espacio de Hilbert con núcleo reproductor K ("Reproducing Kernel Hilbert Space, RKHS"), $\Gamma(X) = \langle X, \beta \rangle_K$ donde el producto interno es el del RKHS.

La pregunta que surge de forma natural es qué criterio elegir para decir que un predictor de Y basado en X es bueno. Si X toma valores en un espacio $\mathcal{X}$ e $Y \in \mathbb{R}$, vimos en el capítulo anterior que la función medible f que minimiza $\mathbb{E}|f(X) - Y|$, es la esperanza condicional de Y dado X, es decir $f(X) = \mathbb{E}[Y|X]$. Vamos a denotar $m^*(x) = \mathbb{E}(Y|X = x)$, y, como vimos, para cualquier f medible,

$$\mathbb{E}\Big|f(X) - Y\Big|^2 = \mathbb{E}\Big|f(X) - m^*(X)\Big|^2 + \mathbb{E}\Big|m^*(X) - Y\Big|^2. \tag{3.1}$$

En general la distribución de (X, Y) es desconocida y, por lo tanto, también m^*. Lo que buscamos es "estimar" m^* a partir de una muestra de entrenamiento $D_n = \{(X_1, Y_1), \ldots, (X_n, Y_n)\}$ iid de $(X, Y) \in \mathcal{X} \times \mathbb{R}$. Dado que el segundo término en la ecuación anterior no se puede achicar, es decir es la cota inferior del error cuadrático medio de cualquier predictor f, lo que queremos es una función m_n que dependerá de la muestra, que minimice $\mathbb{E}|m_n(X) - m^*(X)|^2$. Si $X : (\Omega, \mathcal{A}, \mathbb{P}) \to (\mathcal{X}, \mathfrak{F})$ es una función medible, donde $\mathcal{X}$ es un conjunto en el cual tenemos una σ-álgebra $\mathfrak{F}$, si denotamos μ como la distribución de X, entonces

$$\mathbb{E}\Big|m_n(X) - m^*(X)\Big|^2 = \int_{\mathcal{X}} \Big|m_n(x) - m^*(x)\Big|^2 \mu(dx) = \int_{\mathcal{X}} \mathbb{V}\big(m_n(x)\big)\mu(dx) + \int_{\mathcal{X}} \Big[\text{sesgo}\big(m_n(x)\big)\Big]^2 \mu(dx)$$

donde $\text{sesgo}(m_n(x)) = m^*(x) - \mathbb{E}(m_n(x))$. De acuerdo a la identidad anterior, para achicar el error cuadrático medio hay que achicar la varianza y el sesgo, pero esto, en general, va en direcciones opuestas. Achicar el sesgo conduce a aumentar la varianza y, recíprocamente, achicar la varianza aumenta el sesgo.

Algunos predictores son locales, es decir, fijado x, usan los puntos de la muestra que están "cerca" de x[1]. Entre estos se encuentran los métodos basados en vecinos más cercanos, o basados en núcleos. Otros son globales, usan toda la muestra, como "support vector machine". La consistencia de los métodos locales, para el caso $\mathcal{X} = \mathbb{R}^d$ se suele probar a partir de un general de Teorema de Stone de 1977, cuyas hipótesis se tienen que verificar para cada método de regresión. Este teorema es universal en el sentido de que solo le pide al par (X, Y) que exista $\mathbb{E}[Y|X]$.

3.2 El teorema de Stone (1977)

Un teorema general de Stone (1977), ver [30], nos permitirá deducir la consistencia universal de diversas reglas de clasificación a partir de estimar la función de regresión.

Consideremos un estimador de $m^*(x) = \mathbb{E}(Y|X = x)$ de la forma

$$m_n(x) = \sum_{i=1}^{n} Y_i w_{ni}(x),$$

donde los pesos $w_{ni}(x) = w_{ni}(x, X_1, \ldots, X_n)$ son no negativos y $\sum_{i=1}^n w_{ni}(x) = 1$.

[1] o todos los puntos, pero ponderando por la distancia a x como hace el método basado en núcleos

El siguiente Teorema es una versión un poco mas restrictiva del Teorema demostrado por Stone en 1977 (ver [30]), una prueba con las hipótesis originales puede encontrarse en [15]. Con las hipótesis que lo enunciaremos, Stone prueba que en realidad las mismas son necesarias y suficientes para que se cumpla la tesis del teorema, ver Corolario 1 en [30].

Teorema 3.1. *(Stone (1977)).*
Supongamos que los pesos verifiquen las siguientes tres condiciones

i) Existe $c > 0$ tal que para toda f medible,no negativa con $\mathbb{E}(f(X)) < \infty$,

$$\mathbb{E}\left(\sum_{i=1}^{n} w_{ni}(X) f(X_i)\right) \leq c\mathbb{E}(f(X)).$$

ii)

$$\lim_{n\to\infty} \mathbb{E}\left(\sum_{i=1}^{n} w_{ni}(X)\mathbb{I}_{\{\|X_i - X\| > a\}}\right) = 0, \quad \forall a > 0.$$

iii)

$$\lim_{n\to\infty} \mathbb{E}\left(\max_{1\leq i\leq n} w_{ni}(X)\right) = 0,$$

entonces

$$\lim_{n\to\infty} \mathbb{E}\left(\Big[m^*(X) - m_n(X)\Big]^2\right) = 0,$$

para toda distribución de (X, Y) tal que $\mathbb{E}(Y^2) < \infty$.

Demostración. Denotemos por $\tilde{m}_n(x) = \sum_{i=1}^{n} m^*(X_i) w_{ni}(x)$ (no observable). Luego, tenemos que

$$\begin{aligned}\mathbb{E}\left(\Big[m^*(X) - m_n(X)\Big]^2\right) &= \mathbb{E}\left(\Big[m^*(X) - \tilde{m}_n(X) + \tilde{m}_n(X) - m_n(X)\Big]^2\right) \\ &\leq 2\Bigg(\underbrace{\mathbb{E}\Big[\big(m^*(X) - \tilde{m}_n(X)\big)^2\Big]}_{\mathbf{A}} \quad + \quad \underbrace{\mathbb{E}\Big[\big(\tilde{m}_n(X) - m_n(X)\big)^2\Big]}_{\mathbf{B}}\Bigg). \end{aligned} \tag{3.2}$$

Donde hemos usado $(a+b)^2 \leq 2(a^2+b^2)$. Bastará con probar que ambos términos convergen a 0.

Acotación de A

Como los pesos suman 1 podemos escribir:

$$\mathbf{A} = \mathbb{E}\left(\left\{\sum_{i=1}^{n} w_{ni}(X)\Big[m^*(X) - m^*(X_i)\Big]\right\}^2\right).$$

Como los pesos son no negativos y suman uno, son una probabilidad. Podemos, por lo tanto, acotar el termino anterior, usando la desigualdad de Jensen, por

$$\mathbb{E}\left(\sum_{i=1}^{n} w_{ni}(X)\Big[m^*(X) - m^*(X_i)\Big]^2\right).$$

Si la función m^* es continua de soporte compacto, es uniformemente continua y está acotada. Denotemos L la cota superior de m^*, por tanto dado $\epsilon > 0$ existe $a > 0$ tal que si $\|x_1 - x\| \leq a$, $|m^*(x_1) - m^*(x)| < \epsilon$. Luego, como además $|m^*(x_1) - m^*(x)| \leq L$, tenemos que

$$\mathbb{E}\left(\sum_{i=1}^{n} w_{ni}(X)\Big[m^*(X) - m^*(X_i)\Big]^2\right) \leq \mathbb{E}\left(\sum_{i=1}^{n} w_{ni}(X)\mathbb{I}_{\{\|X - X_i\| \geq a\}} L\right) \quad + \quad \mathbb{E}\left(\sum_{i=1}^{n} w_{ni}(X)\epsilon^2\right) \to \epsilon^2, \tag{3.3}$$

por la hipótesis (ii).

Como el conjunto de funciones continuas con soporte compacto es denso en $L^2(\mu)$, siendo μ la distribución de X, para todo $\epsilon > 0$ podemos elegir η, continua y con soporte compacto, tal que $\mathbb{E}([m^*(X) - \eta(X)]^2) < \epsilon$. Tenemos que

$$\mathbb{E}\left(\sum_{i=1}^{n} w_{ni}(X)\Big[m^*(X) - m^*(X_i)\Big]^2\right) = \mathbb{E}\left(\sum_{i=1}^{n} w_{ni}(X)\Big[m^*(X) - \eta(X) + \eta(X) - \eta(X_i) + \eta(X_i) - m^*(X_i)\Big]^2\right).$$

Luego, usando $(a+b+c)^2 \leq 3(a^2+b^2+c^2)$, la esperanza anterior se acota superiormente por

$$3\mathbb{E}\left(\sum_{i=1}^{n} w_{ni}(X)\left[\Big(m^*(X)-\eta(X)\Big)^2+\Big(\eta(X)-\eta(X_i)\Big)^2+\Big(\eta(X_i)-m^*(X_i)\Big)^2\right]\right)=$$

$$3\mathbb{E}\left(\sum_{i=1}^{n} w_{ni}(X)\left[\Big(m^*(X)-\eta(X)\Big)^2\right]\right)+3\mathbb{E}\left(\sum_{i=1}^{n} w_{ni}(X)\left[\Big(\eta(X)-\eta(X_i)\Big)^2\right]\right)+$$
$$3\mathbb{E}\left(\sum_{i=1}^{n} w_{ni}(X)\left[\Big(\eta(X_i)-m^*(X_i)\Big)^2\right]\right)=\textbf{I+II+III}$$

I: Como los pesos suman 1,

$$\mathrm{I}=3\mathbb{E}\left(\Big(m^*(X)-\eta(X)\Big)^2\sum_{i=1}^{n} w_{ni}(X)\right)=3\mathbb{E}\left(\Big(m^*(X)-\eta(X)\Big)^2\right)\leq 3\epsilon$$

donde la última desigualdad es por la forma en que tomamos η.

II: Para acotar superiormente el término II razonamos como en (3.3): observemos que como η es continua y tiene soporte compacto, es uniformemente continua por tanto dado $\epsilon>0$ existe $a>0$ tal que si $\|x_1-x\|\leq a$, $|\eta(x_1)-\eta(x)|<\epsilon$, y está acotada superiormente por alguna constante positiva, por lo tanto existe $L>0$ tal que,

$$\big(\eta(X)-\eta(X_i)\big)^2\leq L,$$

es decir,

$$\mathbb{E}\left(\sum_{i=1}^{n} w_{ni}(X)\Big[\eta(X)-\eta(X_i)\Big]^2\right)\leq L\mathbb{E}\left(\sum_{i=1}^{n} w_{ni}(X)\mathbb{I}_{\{\|X-X_i\|\geq a\}}\right)\quad+\quad\mathbb{E}\left(\sum_{i=1}^{n} w_{ni}(X)\epsilon^2\right)\to\epsilon^2,$$

III: Para acotar el término III usamos la hipótesis i, y obtenemos

$$\mathrm{III}\leq 3c\mathbb{E}\left(\Big(m^*(X)-\eta(X)\Big)^2\right)\leq 3c\epsilon.$$

siendo c como en i. Por tanto,

$$\limsup_{n\to\infty}\mathbb{E}\left(\Big(m^*(X)-\tilde{m}_n(X)\Big)^2\right)\leq 3\epsilon(1+\epsilon+c).$$

Como esta desigualdad vale para todo ϵ, y el término de la izquierda es ≥ 0, obtenemos que

$$\lim_{n\to\infty}\mathbb{E}\left(\Big(m^*(X)-\tilde{m}_n(X)\Big)^2\right)=0.$$

Acotación de B

$$\mathbf{B}=\mathbb{E}\left[\left(\sum_{i=1}^{n} w_{ni}(X)\big(m^*(X_i)-Y_i\big)\right)^2\right]=\sum_{i=1}^{n}\sum_{j=1}^{n}\mathbb{E}\Big[w_{ni}(X)\big(m^*(X_i)-Y_i\big)w_{nj}(X)\big(m^*(X_j)-Y_j\big)\Big].$$

Primero observemos que

$$\mathbb{E}\Big[\mathbb{E}\Big\{w_{ni}(X)\big(Y_i-m^*(X_i)\big)\big(Y_j-\eta(X_j)\big)\Big|X,X_1,\ldots,X_n,Y_j\Big\}\Big]=$$
$$\mathbb{E}\Big[w_{ni}(X)\big(Y_j-\eta(X_j)\big)\mathbb{E}\Big\{\big(Y_i-m^*(X_i)\big)\Big|X,X_1,\ldots,X_n,Y_j\Big\}\Big].$$

Veamos que la esperanza condicional de adentro es 0, si $i\neq j$,

$$\mathbb{E}\Big(\big(Y_i-m^*(X_i)\big)\Big|X,X_1,\ldots,X_n,Y_j\Big)=\mathbb{E}(Y_i|X_i)-\mathbb{E}\Big[m^*(X_i)\Big|X,X_1,\ldots,X_n,Y_j\Big]=\mathbb{E}(Y_i|X_i)-m^*(X_i)$$

En la primera igualdad usamos que Y_i es independiente de Y_j, X y de X_l para $l \neq j$. En la segunda usamos que $m^*(X_i)$ es una función medible respecto de la σ-álgebra generada por X_i y, por lo tanto, "sale para afuera" de la esperanza condicional. Como (X_i, Y_i) tiene la misma distribución conjunta que (X, Y), $\mathbb{E}(Y_i|X_i) = m^*(X_i)$ (ver Observación 2.13).

Es decir, obtuvimos que

$$\mathbf{B} = \sum_{i=1}^{n} \mathbb{E}\left[w_{ni}^2(X) \Big(m^*(X_i) - Y_i \Big)^2 \right]$$

Si condicionamos a $X, X_1, \ldots, X_n$.

$$\mathbb{E}\left[w_{ni}^2(X) \Big(m^*(X_i) - Y_i \Big)^2 \right] = \mathbb{E}\left[w_{ni}^2(X) \mathbb{E}\Big\{ \big(m^*(X_i) - Y_i \big)^2 \Big| X, X_1, \ldots, X_n \Big\} \right] = \mathbb{E}\Big[w_{ni}^2(X) \sigma^2(X_i) \Big]$$

donde $\sigma^2(x) = \mathbb{E}\big((Y - m^*(X))^2 | X = x\big)$. Si σ^2 está acotada superiormente, basta observar que

$$\mathbb{E}\left(\sum_{i=1}^{n} w_{ni}^2(X) \right) \leq \mathbb{E}\left(\max_{1 \leq i \leq n} w_{ni}(X) \sum_{i=1}^{n} w_{ni}(X) \right) = \mathbb{E}\left(\max_{1 \leq i \leq n} w_{ni}(X) \right) \to 0,$$

por (iii). Si σ^2 no es acotada, se aproxima en $L^2(\mu)$ por una función continua y acotada.

□

3.3 Clasificación

En el problema de clasificación la variable Y es discreta y nos referiremos a sus posibles valores como etiquetas. En el caso en que $Y \in \{0, 1\}$ se llama clasificación binaria. Como mencionamos en la introducción la mejor regla de clasificación es

$$g^*(x) = \mathbb{I}_{\{\mathbb{P}(Y=1|X=x) > 1/2\}} = \mathbb{I}_{\{m^*(x) > 1/2\}}, \tag{3.4}$$

y se conoce como **regla de Bayes**. Observar que

$$g^*(x) = \mathbb{I}_{\{m^*(x) > 1 - m^*(x)\}} = \mathbb{I}_{\{\mathbb{P}(Y=1|X=x) > \mathbb{P}(Y=0|X=x)\}}.$$

La probabilidad de que esta regla se equivoque en un nuevo par (X, Y) es $\mathbb{P}(g^*(X) \neq Y)$. Probaremos que para cualquier otra regla g vale

$$\mathbb{P}(g(X) \neq Y) \geq \mathbb{P}(g^*(X) \neq Y).$$

Usualmente los algoritmos de regresión como k vecinos mas cercanos o núcleos se adaptan de forma inmediata para este contexto. La consistencia universal de los algoritmos locales se deduce de verificar las hipótesis del Teorema de Stone antes mencionado, para cada caso. Intuitivamente un algoritmo local va a funcionar bien si la función m^* se puede aproximar, en un punto x, usando las etiquetas de valores próximos a x. Si m^* es continua esto es inmediato. En dimensión finita estas hipótesis se cumplen sin pedirle nada a m^*, pero en espacios más generales hay que pedir que se cumpla lo que se conoce como la condición de Besicovitch, que no es otra cosa que el teorema de diferenciación de Lebesgue, pero aplicado a la distribución de las X. En dimensión finita este teorema se cumple para casi todo x respecto de la distribución de las X.

Si en lugar de aproximar m^* de forma no paramétrica, a partir de la muestra de entrenamiento, como hace k-nn, buscamos la "mejor" función f en una determinada clase de funciones $\mathcal{C}$ veremos que el desempeño de el mejor clasificador en la clase dependerá de cuán rica sea esta clase de funciones. Es claro que cuanto más pobre sea esta clase peor vamos a poder aproximarnos al ideal $g^*(x)$. El desarrollo de esta teoría, y de buena parte de lo que haremos, se debe a Vapnik, y puede encontrarse en varios libros, una referencia clásica es [9] y [15].

3.3.1 Clasificación Binaria

Vamos a ver primero el caso de clasificación binaria, es decir $Y \in \{0, 1\}$. De forma muy general vamos a suponer que ambas están definidas en algún espacio de probabilidad $(\Omega, \mathcal{A}, \mathbb{P})$ y que X toma valores en un conjunto cualquiera $\mathcal{X}$ en el cual hay una σ-álgebra que hace que X sea medible. Comencemos probando que la distribución del par (X, Y) queda determinada si se conoce P_X (la distribución de X), y $m^*(x) = \mathbb{P}(Y = 1|X = x)$. Para eso observemos que (X, Y) toma valores en $\mathcal{X} \times \{0, 1\}$. Si $C \subset \mathcal{X} \times \{0, 1\}$,

$$C = \Big\{ C \cap (\mathcal{X} \times \{0\}) \Big\} \cup \Big\{ C \cap (\mathcal{X} \times \{1\}) \Big\} = C_0 \times \{0\} \cup C_1 \times \{1\}.$$

Como esta unión es disjunta

$$\begin{aligned}
\mathbb{P}\big((X,Y)\in C\big) =& \mathbb{P}(X\in C_0, Y=0) + \mathbb{P}(X\in C_1, Y=1)\\
=& \int_\Omega \mathbb{I}_{\{\omega: X(\omega)\in C_0\}}\mathbb{I}_0(Y)d\mathbb{P} + \int_\Omega \mathbb{I}_{\{\omega: X(\omega)\in C_1\}}\mathbb{I}_1(Y)d\mathbb{P}\\
=& \int_{\{\omega: X(\omega)\in C_0\}} \mathbb{I}_0(Y)d\mathbb{P} + \int_{\{\omega: X(\omega)\in C_1\}} \mathbb{I}_1(Y)d\mathbb{P}\\
\underbrace{=}_{(2.10)}& \int_{C_0} \mathbb{E}(\mathbb{I}_0(Y)|X=x)P_X(dx) + \int_{C_1} \mathbb{E}(\mathbb{I}_1(Y)|X=x)P_X(dx)\\
=& \int_{C_0} \mathbb{P}(Y=0|X=x)P_X(dx) + \int_{C_1} \mathbb{P}(Y=1|X=x)P_X(dx)\\
=& \int_{C_0} \big(1-m^*(x)\big)P_X(dx) + \int_{C_1} m^*(x)P_X(dx)
\end{aligned}$$

3.3.2 Regla de Bayes

La regla de Bayes, $g^*(x) = \mathbb{I}_{\{m^*(x)>1/2\}}$, minimiza el error de clasificación, más precisamente,

Teorema 3.2. *Siguiendo la notación antes introducida,*

$$\mathbb{P}\big(g^*(X)\neq Y\big) = \min_{g:\mathcal{X}\to\{0,1\}} \mathbb{P}\big(g(X)\neq Y\big) \tag{3.5}$$

donde el mínimo se toma en el conjunto de todas las funciones medibles de $\mathcal{X}$ a $\{0,1\}$.

Demostración. Sea $g:\mathcal{X}\to\{0,1\}$ medible

$$\begin{aligned}
\mathbb{P}\big(g(X)\neq Y\big|X=x\big) =& 1-\mathbb{P}\big(g(X)=Y\big|X=x\big)\\
=& 1-\Big[\mathbb{P}\big(Y=1,g(X)=1\big|X=x\big)+\mathbb{P}\big(Y=0,g(X)=0\big|X=x\big)\Big]
\end{aligned}$$

Si usamos ahora (2.11),

$$\mathbb{P}\big(Y=1,g(X)=1\big|X=x\big) = \mathbb{P}\big(Y=1,g(x)=1\big|X=x\big) = \mathbb{I}_{\{g(x)=1\}}\mathbb{P}(Y=1|X=x),$$

donde las igualdades anteriores valen para todo x excepto en un conjunto de P_X medida nula. Análogamente

$$\mathbb{P}\big(Y=0,g(X)=0\big|X=x\big) = \mathbb{I}_{\{g(x)=0\}}\mathbb{P}(Y=0|X=x).$$

Por lo tanto,

$$\begin{aligned}
\mathbb{P}\big(g(X)\neq Y\big|X=x\big) =& 1-\Big[\mathbb{I}_{\{g(x)=1\}}\mathbb{P}\big(Y=1\big|X=x\big)+\mathbb{I}_{\{g(x)=0\}}\mathbb{P}\big(Y=0\big|X=x\big)\Big]\\
=& 1-\Big[\mathbb{I}_{\{g(x)=1\}}m^*(x)+\mathbb{I}_{\{g(x)=0\}}\big(1-m^*(x)\big)\Big]
\end{aligned}\tag{3.6}$$

De forma totalmente idéntica,

$$\mathbb{P}\big(g^*(X)\neq Y\big|X=x\big) = 1-\Big[\mathbb{I}_{\{g^*(x)=1\}}m^*(x)+\mathbb{I}_{\{g^*(x)=0\}}\big(1-m^*(x)\big)\Big] \tag{3.7}$$

Si restamos (3.6) y (3.7) obtenemos,

$$\begin{aligned}
\mathbb{P}\big(g(X)\neq Y\big|X=x\big) - \mathbb{P}\big(g^*(X)\neq Y\big|X=x\big) =& 1-\Big[\mathbb{I}_{\{g(x)=1\}}m^*(x)+\mathbb{I}_{\{g(x)=0\}}\big(1-m^*(x)\big)\Big]-\\
& \Big(1-\big[\mathbb{I}_{\{g^*(x)=1\}}m^*(x)+\mathbb{I}_{\{g^*(x)=0\}}(1-m^*(x))\big]\Big)\\
=& m^*(x)\Big(\mathbb{I}_{\{g^*(x)=1\}}-\mathbb{I}_{\{g(x)=1\}}\Big)+\big(1-m^*(x)\big)\Big(\mathbb{I}_{\{g^*(x)=0\}}-\mathbb{I}_{\{g(x)=0\}}\Big)
\end{aligned}$$

Si usamos,

$$\mathbb{I}_{\{g^*(x)=0\}}-\mathbb{I}_{\{g(x)=0\}} = \Big(1-\mathbb{I}_{\{g^*(x)=1\}}-(1-\mathbb{I}_{\{g(x)=1\}})\Big) = \mathbb{I}_{\{g(x)=1\}}-\mathbb{I}_{\{g^*(x)=1\}} = -\Big(\mathbb{I}_{\{g^*(x)=1\}}-\mathbb{I}_{\{g(x)=1\}}\Big)$$

obtenemos que

$$\mathbb{P}\big(g(X)\neq Y|X=x\big) - \mathbb{P}\big(g^*(X)\neq Y|X=x\big) = \big(2m^*(x)-1\big)\Big(\mathbb{I}_{\{g^*(x)=1\}}-\mathbb{I}_{\{g(x)=1\}}\Big). \tag{3.8}$$

Por definición de g^*, $2m^*(x)-1>0$ si y solo si $g^*(x)=1$, por lo tanto $\mathbb{I}_{\{g^*(x)=1\}}-\mathbb{I}_{\{g(x)=1\}}\geq 0$, y el producto anterior da mayor o igual que 0. Por otra parte $2m^*(x)-1<0$ si y solo si $g^*(x)=0$, por lo tanto $\mathbb{I}_{\{g^*(x)=1\}}-\mathbb{I}_{\{g(x)=1\}}\leq 0$, y el producto anterior da mayor o igual que 0. En resumen

$$\mathbb{P}\big(g(X)\neq Y|X=x\big)\geq \mathbb{P}\big(g^*(X)\neq Y|X=x\big)$$

□

Definición 3.3. El **error de clasificación de una regla** cualquiera g, $\mathbb{P}(g(X)\neq Y)$, lo vamos a denotar $L(g)$ y es, tomando esperanza en (3.6),

$$L(g)=1-\mathbb{E}\Big[\mathbb{I}_{\{g(X)=1\}}m^*(X)+\mathbb{I}_{\{g(X)=0\}}\big(1-m^*(X)\big)\Big]. \tag{3.9}$$

En el caso particular de la regla de Bayes denotamos $L^*=L(g^*)$, y obtenemos

$$L^*=1-\mathbb{E}\Big[\mathbb{I}_{\{m^*(X)>1/2\}}m^*(X)+\mathbb{I}_{\{m^*(X)\leq 1/2\}}\big(1-m^*(X)\big)\Big]. \tag{3.10}$$

Ejercicio 3.4. Las siguientes propiedades se dejan como ejercicio

1. De (3.10) se sigue que
$$L^*=\mathbb{E}\Big[\text{mín}\big\{m^*(X),1-m^*(X)\big\}\Big].$$
2. De la igualdad anterior, junto con (2.18) se sigue de forma inmediata que si X tiene densidad
$$L^*=\int \text{mín}\big\{(1-p)f_0(x),pf_1(x)\big\}dx,$$
donde $p=\mathbb{P}(Y=1)$. Esto implica que $L^*\leq \text{mín}\{p,(1-p)\}\leq 1/2$ [2]
3. Como la regla de Bayes es equivalente a $g^*(x)=\mathbb{I}_{\{m^*(x)>1-m^*(x)\}}$, si X tiene densidad esta regla es, usando (2.18)
$$g^*(x)=\begin{cases}1 & \text{si } pf_1(x)>(1-p)f_0\\ 0 & \text{en caso contrario}\end{cases} \tag{3.11}$$
donde $p=\mathbb{P}(Y=1)$.
4. También de (3.10) se sigue que $L^*=(1/2)-(1/2)\mathbb{E}|2m^*(X)-1|$.

3.3.3 Reglas plug-in

Si tenemos m una función que aproxima m^*, que luego la construiremos en base a una muestra, es natural proponer como regla de clasificación la función $g(x)=\mathbb{I}_{\{m(x)>1/2\}}$. Ya vimos que la probabilidad de error de esta regla, es decir $L(g)=\mathbb{P}(g(X)\neq Y)$, es mayor o igual que L^*, la probabilidad de error de la regla de Bayes. Veamos una cota superior para la diferencia $L(g)-L^*$, en términos de m y m^*.

Teorema 3.5.

$$\mathbb{P}\big(g(X)\neq Y\big)-L^*=2\int_{\mathcal{X}}\big|m^*(x)-1/2\big|\mathbb{I}_{\{g(x)\neq g^*(x)\}}P_X(dx)$$

además

$$\mathbb{P}\big(g(X)\neq Y\big)-L^*\leq 2\int_{\mathcal{X}}\big|m^*(x)-m(x)\big|P_X(dx) \tag{3.12}$$

Demostración. Vamos a usar (3.8), observemos que

$$\big(2m^*(x)-1\big)\Big(\mathbb{I}_{\{g^*(x)=1\}}-\mathbb{I}_{\{g(x)=1\}}\Big)=\big|2m^*(x)-1\big|\mathbb{I}_{\{g^*(x)\neq g(x)\}}$$

Esta última igualdad se prueba considerando todos los casos posibles, por ejemplo si $\mathbb{I}_{\{g^*(x)=1\}}=1$ entonces $\big(2m^*(x)-1\big)>0$, si $\mathbb{I}_{\{g(x)=1\}}=1$ entonces $\mathbb{I}_{\{g^*(x)\neq g(x)\}}=0$ y vale la igualdad, si $\mathbb{I}_{\{g(x)=1\}}=1$ ambos lados de la igualdad son 0. Los otros casos se razonan de forma análoga. Por lo tanto de (3.8)

$$\mathbb{P}\big(g(X)\neq Y|X=x\big)-\mathbb{P}\big(g^*(X)\neq Y|X=x\big)=\int_{\mathcal{X}}2\big|m^*(x)-1/2\big|\mathbb{I}_{\{g^*(x)\neq g(x)\}}$$

Para probar (3.12) basta probar que $g(x)\neq g^*(x)$ implica que $\big|m^*(x)-1/2\big|\leq\big|m(x)-m^*(x)\big|$. Esto también se prueba considerando los dos casos posibles:

[2] aquí se usa, y se deja como ejercicio verificar, que $\int \text{mín}\{f(x),g(x)\}dx\leq \text{mín}\{\int f(x)dx,\int g(x)dx\}$.

- si $g^*(x) = 1$ y $g(x) = 0$, entonces $m^*(x) \geq 1/2$ y $m(x) < 1/2$, por lo tanto

$$0 \leq m^*(x) - 1/2 \leq m^*(x) - m(x),$$

de donde $|m^*(x) - 1/2| \leq |m(x) - m^*(x)|$.

- Si $g^*(x) = 0$ y $g(x) = 1$ entonces $m^*(x) < 1/2$ y $m(x) \geq 1/2$, y por lo tanto

$$0 \leq 1/2 - m^*(x) \leq m(x) - m^*(x).$$

□

Observar que

$$\int_{\mathcal{X}} \big|m^*(x) - m(x)\big| P_X(dx) = 2\mathbb{E}\big|m^*(X) - m(X)\big| \leq 2\sqrt{\mathbb{E}\big|m^*(X) - m(X)\big|^2} \tag{3.13}$$

De esto y (3.12) se sigue que si m es próxima a m^* en $L^2_{P_X} = \{f : \mathcal{X} \to \mathbb{R} : \int f^2(x) P_X(dx) < \infty\}$, la regla plug-in $g(x) = \mathbb{I}_{\{m(x)>1/2\}}$ va a tener un error de clasificación próximo al error de Bayes.

Ejercicio 3.6.

1. Sea $T : \mathcal{X} \to \mathcal{X}'$ una función medible cualquiera, si L^*_X denota el error de Bayes de (X, Y) y $L^*_{T(X)}$ el de $(T(X), Y)$, probar que $L^*_{T(X)} \geq L^*_X$. Esto muestra que transformando las X se pierde información, ya que el error de Bayes aumenta.

2. Si m y m' son funciones medibles, a valores en $[0, 1]$ y definimos las reglas $g'(x) = \mathbb{I}_{\{m'(x)\geq 1/2\}}$ y $g(x) = \mathbb{I}_{\{m(x)\geq 1/2\}}$, probar que

$$|L(g) - L(g')| \leq \mathbb{P}\big(g'(X) \neq g(X)\big) \quad y \quad |L(g) - L(g')| \leq \mathbb{E}\Big[\big|2m^*(X) - 1\big|\mathbb{I}_{\{g'(X)\neq g(X)\}}\Big]$$

3.3.4 Criterios de minimización del error

Hay reglas de clasificación que se construyen usando un determinado tipo de funciones, por ejemplo, los clasificadores lineales buscan el subespacio que mejor separa los datos, es decir son reglas $g : \mathbb{R}^d \to \{0, 1\}$ del tipo $g(x) = \mathbb{I}_{\langle x,a\rangle + c_0 > 0}$ donde $a \in \mathbb{R}^d$ es un vector fijo, que determina, junto con c_0 el subespacio. Lo mismo sucede con las redes neuronales. Si elegimos el subespacio que mejor clasifica, es claro que la probabilidad de error de esta regla será mayor que si optimizamos entre todas las posibles reglas de clasificación.

Si tenemos una clase de funciones medibles, $\mathcal{C} = \{g : \mathcal{X} \to \{0, 1\}\}$[3], es decir, un subconjunto de todas las posibles reglas de clasificación, la probabilidad de equivocarse de la regla en esta clase que clasifica mejor, es decir $\text{ínf}_{g\in\mathcal{C}} \mathbb{P}(g(X) \neq Y)$, va a ser mayor que si elegimos la regla de Bayes. Es decir

$$\inf_{g\in\mathcal{C}} L(g) \geq L^*$$

La diferencia

$$\inf_{g\in\mathcal{C}} L(g) - L^*$$

obviamente depende de $\mathcal{C}$, y no es aleatoria. Cuanto más funciones contenga $\mathcal{C}$ menor será esta diferencia, pero una vez fijada, no la podemos achicar tomando más datos.

En general $L(g)$ no se conoce; una posibilidad es estimarlo por medio del error empírico, $\hat{L}_n(g)$, basado en una muestra $D_n = \big\{(X_1, Y_1), \ldots, (X_n, Y_n)\big\}$ iid de (X, Y), esto es

$$\hat{L}_n(g) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{I}_{\{g(X_i)\neq Y_i\}} \tag{3.14}$$

o, lo que es lo mismo, $\hat{L}_n(g) = (1/n)\sum_{i=1}^n (g(X_i) - Y_i)^2$. Observar que $\mathbb{E}(\hat{L}_n(g)) = L(g)$. Hasta aquí g es una función que no depende de la muestra, por lo tanto, por la ley fuerte de los grandes números, $\hat{L}_n(g) \to L(g)$, casi seguramente, cuando $n \to \infty$. Vamos a denotar g^*_n una regla de clasificación en $\mathcal{C}$ que minimize el error empírico, es decir

$$\hat{L}_n(g^*_n) \leq \hat{L}_n(g) \quad \forall g \in \mathcal{C}.$$

La probabilidad de error de esta regla (claramente g^*_n depende de la muestra), condicionada a la muestra, es

$$L(g^*_n) = \mathbb{P}\big(g^*_n(X) \neq Y \big| D_n\big).^4.$$

[3]Por ahora no vamos a considerar el caso en que construimos la regla a partir de una muestra. Es decir la clase de funciones $\mathcal{C}$ **no** depende de la muestra

[4]esta probabilidad, que es condicionada a la muestra, mide la capacidad de predicción de la regla que elegimos, en un nuevo dato

Es claro que

$$\big|\hat{L}_n(g_n^*) - L(g_n^*)\big| \leq \sup_{g\in\mathcal{C}} \big|\hat{L}_n(g) - L(g)\big|. \tag{3.15}$$

ya que $g_n^* \in \mathcal{C}$. Esta desigualdad (que vale c.s.) nos dice que si $\sup_{g\in\mathcal{C}} |\hat{L}_n(g) - L(g)|$ es chico, entonces el error de predicción de la regla que elegimos[5], esto es $L(g_n^*)$ (que con la muestra no lo podemos calcular), está cerca del error empírico $\hat{L}_n(g_n^*)$, que si podemos calcular. Acá tenemos el efecto contrario al que teníamos con la diferencia $\text{ínf}_{g\in\mathcal{C}} L(g) - L^*$, ya que cuanto más grande sea la clase de funciones, más grande va a ser $\sup_{g\in\mathcal{C}} |\hat{L}_n(g) - L(g)|$. En el caso extremo de que la clase $\mathcal{C}$ contiene a todas las funciones medibles, podemos tomar una que interpole los datos de la muestra de entrenamiento, y, por lo tanto $\hat{L}_n(g_n^*) = 0$, pero para la mayoría de las distribuciones de pares (X, Y) esta regla va a tener una capacidad de predicción muy baja, es decir $L(g_n^*)$ va a ser alto.

La teoría de Vapnik-Chervonenkis permite acotar superiormente $\sup_{g\in\mathcal{C}} |\hat{L}_n(g) - L(g)|$, en función de que tan rica es la clase $\mathcal{C}$. O, más precisamente, en función de su dimensión de Vapnik-Chervonenkis, un concepto que estudiaremos más adelante en profundidad.

Por otra parte, vale la cota,

$$L(g_n^*) - \inf_{g\in\mathcal{C}} L(g) \leq 2 \sup_{g\in\mathcal{C}} \big|\hat{L}_n(g) - L(g)\big|, \tag{3.16}$$

que nos dice que si acotamos superiormente $\sup_{g\in\mathcal{C}} \big|\hat{L}_n(g) - L(g)\big|$, estamos acotando superiormente $L(g_n^*) - \text{ínf}_{g\in\mathcal{C}} L(g)$, que cuantifica la diferencia entre el poder de predicción de la regla que elegimos, g_n^*, y el poder de predicción de la regla en la clase que tiene mayor poder predictivo $\text{ínf}_{g\in\mathcal{C}} L(g)$.

La cota (3.16) es consecuencia de

$$L(g_n^*) - \inf_{g\in\mathcal{C}} L(g) = \underbrace{L(g_n^*) - \hat{L}_n(g_n^*)}_{\mathbf{A}} + \underbrace{\hat{L}_n(g_n^*) - \inf_{g\in\mathcal{C}} L(g)}_{\mathbf{B}},$$

de que para toda regla $g' \in \mathcal{C}$

$$\mathbf{B} = \hat{L}_n(g_n^*) - \inf_{g\in\mathcal{C}} L(g) \leq \hat{L}_n(g') - \inf_{g\in\mathcal{C}} L(g) = \inf_{g\in\mathcal{C}} \hat{L}_n(g') - L(g) \leq \inf_{g\in\mathcal{C}} \big|\hat{L}_n(g') - L(g)\big| \leq \big|\hat{L}_n(g') - L(g')\big|$$
$$\leq \sup_{g\in\mathcal{C}} \big|L(g) - \hat{L}_n(g)\big|.$$

En la primera desigualdad usamos que, por definición de g_n^*, $\hat{L}_n(g_n^*) \leq \hat{L}_n(g')$ para todo $g' \in \mathcal{C}$. La siguiente igualdad es trivial ya que $\hat{L}_n(g')$ no depende de g. En la penúltima desigualdad usamos que $g' \in \mathcal{C}$. Por otra parte, para acotar $\mathbf{A}$,

$$\mathbf{A} = L(g_n^*) - \hat{L}_n(g_n^*) \leq \big|L(g_n^*) - \hat{L}_n(g_n^*)\big| \leq \sup_{g\in\mathcal{C}} \big|L(g) - \hat{L}_n(g)\big|,$$

donde en la última desigualdad usamos que $g^* \in \mathcal{C}$.

Observación 3.7. *La aplicación más importante de la estimación del error es la selección de una función de clasificación en una clase $\mathcal{C}$ de funciones. Dada una clase $\mathcal{C}$, es tentador elegir la regla que minimice sobre la clase una estimación de la probabilidad de error de dicha regla. Un buen método debería elegir un clasificador con una probabilidad de error que esté cerca de la probabilidad de error mínima en la clase. Aquí requerimos mucho más que cotas para la diferencia $\hat{L}_n(g) - L(g)$ que sean libres de distribución: no es suficiente poder estimar la probabilidad de error de todos los clasificadores en la clase, ya que, dada una clase $\mathcal{C}$ de funciones de decisión de la forma $g : \mathbb{R}^d \to \{0, 1\}$ (es decir, los datos de entrenamiento no juegan ningún papel en la decisión) tal que para cada $\epsilon > 0$*

$$\sup_{g\in\mathcal{C}} \mathbb{P}\Big\{\big|\hat{L}_n(g) - L(g)\big| > \epsilon\Big\} \leq 2e^{-2n\epsilon^2}$$

para cada distribución. Si $\mathcal{F}_n$ es la clase de funciones g_n^* que minimizan $\hat{L}_n(g)$ sobre la clase $\mathcal{C}$, existe una distribución de (X, Y) tal que

$$\mathbb{P}\left\{\sup_{g\in\mathcal{F}_n} L(g) - \inf_{g\in\mathcal{C}} L(g) = 1\right\} = 1$$

para todo n. La teoría de V.C acota $\mathbb{P}\{\sup_{g\in\mathcal{C}} |\widehat{L}_n(g) - L(g)| > \epsilon\}$.

Cuando $\mathcal{C}$ tiene una cantidad finita, N, de funciones, esto se puede hacer por medio de la desigualdad de Hoeffding (ver Teorema 10.6 en el Apéndice), como establece el siguiente Teorema.

Teorema 3.8. *Denotemos $\#\mathcal{C} = N$, el cardinal de $\mathcal{C}$. Para todo $\epsilon > 0$,*

$$\mathbb{P}\left(\sup_{g\in\mathcal{C}} \big|\hat{L}_n(g) - L(g)\big| > \epsilon\right) \leq 2N \exp\big(-2n\epsilon^2\big).$$

La Teoría de Vapnik-Chervonenkis que veremos más adelante aborda el caso en que $\mathcal{C}$ no es finito, en particular el de redes neuronales.

[5] es decir el error que comete la regla que elegimos en un nuevo dato

3.3.5 Reglas construidas a partir de una muestra

Si la regla g que construimos depende de la muestra D_n, es decir $g(x) = g_n(x, D_n)$ la probabilidad de error de la misma se denota

$$L(g_n) = \mathbb{P}\big(g_n(X) \neq Y | D_n\big).$$

Vamos a definir ahora dos tipos de consistencia.

Definición 3.9. Una regla de clasificación $g_n(x, D_n)$ es **consistente** para una distribución (X, Y), si

$$\mathbb{E}\big(L(g_n)\big) = \mathbb{P}\big(g_n(X) \neq Y\big) \to L^*. \tag{3.17}$$

Se dice que es **fuertemente consistente** si $L_n \to L^*$ c.s.

Observemos que como $L^* \leq L(g_n) \leq 1$, (3.17) es equivalente a la convergencia en probabilidad de $L(g_n)$ a L^* es decir, para todo $\varepsilon > 0$,

$$\mathbb{P}\big(L(g_n) - L^* > \varepsilon\big) \to 0 \quad \text{cuando } n \to \infty.$$

Definición 3.10. Una regla de clasificación es **universalmente consistente** si es consistente para todo par (X, Y). Análogamente se define universalmente fuertemente consistente.

Observación 3.11. *Si bien probaremos que existen reglas universalmente consistentes, el teorema que sigue establece que para n fijo,* ***para cualquier regla*** g_n *y cualquier* $\epsilon > 0$*, existe una distribución de* (X, Y) *tal que* $L^* = 0$*, pero* $\mathbb{E}(L_n)$ *es mayor que* $1/2 - \epsilon$*. Es decir, no podemos encontrar una regla que se desempeñe uniformemente bien en cualquier conjunto de datos, no importa que tan grande sea. Dicho se otra forma, el Teorema 3.12 nos dice que para n fijo, para cualquier regla* g_n*,* $L(g_n)$ *como estimador de* L^* *puede hacerse tan malo como se quiera, eligiendo la distribución adecuadamente.*

Teorema 3.12. *Sea* $\epsilon > 0$*. Para todo n y para toda regla de clasificación* g_n*, existe una distribución* (X, Y) *tal que* $L^* = 0$*, y*

$$\mathbb{E}\big(L(g_n)\big) \geq 1/2 - \epsilon.$$

Observación 3.13. *El teorema que sigue complementa el teorema anterior en el sentido de que tampoco es posible encontrar reglas universalmente consistentes que converjan a una determinada tasa (observar que la diferencia es que aquí el tamaño muestral, n, no es fijo). Dicho de otra manera, el Teorema 3.14 afirma que aún para el caso en que* $L(g_n)$ *es consistente, su tasa de convergencia puede ser arbitrariamente lenta.*

Teorema 3.14. *Sea* $\{a_n\}_n$ *tal que* $0 \leq a_1 \leq 1/16$ *y* $a_n \downarrow 0$*. Para toda sucesión* $\{g_n\}_n$ *de reglas de clasificación, existe una distribución de* (X, Y) *tal que* $L^* = 0$*, pero*

$$\mathbb{E}\big(L(g_n)\big) \geq a_n.$$

Teorema 3.15. *Para todo n y para todo estimador* $\hat{L}_n$ *del error de Bayes* L^**, se cumple que para todo* $\epsilon > 0$ *existe una distribución de* (X, Y) *tal que*

$$\mathbb{E}\big|\hat{L}_n - L^*\big| \geq \frac{1}{4} - \epsilon.$$

Observación 3.16. *Otro resultado interesante es que, en términos de orden de convergencia, la clasificación es más rápida que la regresión. Es decir, las reglas de clasificación construidas a partir de estimadores* m_n *de* m^**, consistentes en* $L^2(X)$*, son débilmente consistentes (por* (3.12) *y* (3.13)*), pero el orden de convergencia además es más rápido, como establece el siguiente teorema (ver Teorema 6.5 en [9])*

Teorema 3.17. *Sea* m_n *débilmente consistente, es decir* $\mathbb{E}(m_n(X) - m(X))^2 \to 0$*, si definimos la regla plug-in*

$$g_n(x) = \begin{cases} 0 & \text{si } m_n(x) \leq \frac{1}{2} \\ 1 & \text{de lo contrario,} \end{cases}$$

entonces

$$\lim_{n\to\infty} \frac{\mathbb{E}\big(L(g_n)\big) - L^*}{\sqrt{\mathbb{E}\left\{\big(m_n(X) - m^*(X)\big)^2\right\}}} = 0,$$

es decir, $\mathbb{E}(L(g_n)) - L^*$ *converge a cero más rápido que el error* L_2 *de la estimación de regresión.*

Si nos restringimos a la clase $\mathcal{C}$ de funciones $g_n(x, D_n)$ construidas en base a la muestra de entrenamiento D_n, tenemos el análogo de las ecuaciones (3.15) y (3.16) [6]

$$L(g_n^*) - \inf_{g_n \in \mathcal{C}} L(g_n) \leq 2 \sup_{g_n \in \mathcal{C}} \big|\widehat{L}_n(g_n) - L(g_n)\big|,$$

y

$$\big|\widehat{L}_n(g_n^*) - L(g_n^*)\big| \leq \sup_{g_n \in \mathcal{C}} \big|\widehat{L}_n(g_n) - L(g_n)\big|.$$

[6] pueden ser o bien todas, lo cual es una clase muy grande y de poca utilidad, o un subconjunto de ellas, como por ejemplo las construidas con hiperplanos

Estimación del Error

Si queremos estimar el error $L_n = \mathbb{P}(g_n(X) \neq Y|D_n)$ de un clasificador g_n construido en base a una muestra D_n fija, y tenemos una muestra de testeo $T_m = \{(X_{n+1}, Y_{n+1}), \ldots, (X_{n+m}, Y_{n+m})\}$, vamos a denotar

$$\hat{L}_{n,m} = \frac{1}{m} \sum_{j=1}^{m} \mathbb{I}_{\{g_n(X_{n+j}) \neq Y_{n+j}\}}.$$

Claramente $\mathbb{E}(\hat{L}_{n,m}|D_n) = L_n$.

El siguiente Teorema es consecuencia directa de la desigualdad de Hoeffding (ver Teorema 10.6 en el Apéndice).

Teorema 3.18. *Para todo $\epsilon > 0$,*

$$\mathbb{P}\big(|\hat{L}_{n,m} - L_n| > \epsilon|D_n\big) \leq 2\exp(-2m\epsilon^2)$$

3.4 Clasificación por vecinos mas cercanos, k–NN

Consideremos el vector ordenado de acuerdo a sus distancias a x, $\|X^{(1)} - x\| \leq \ldots \leq \|X^{(n)} - x\|$, y sean $Y^{(i)}$ las correspondientes etiquetas. Sea $k \in \mathbb{N}$, $k \geq 1$, la regla k-NN[7] esta dada por

$$g_n(x) = \begin{cases} 0 & \text{si} \quad \sum_{i=1}^{k} \mathbb{I}_{\{Y^{(i)}=1\}} \leq \sum_{i=1}^{k} \mathbb{I}_{\{Y^{(i)}=0\}} \\ 1 & \text{sino} \end{cases}$$

en el que los empates se rompen al azar. Denotemos $C_n(X) = \{X^{(1)}, \ldots, X^{(k)}\}$ el conjunto de los k vecinos más próximos a X. Consideremos

$$m_n(x) = \frac{1}{k} \sum_{\{i:X_i \in C_n(X)\}} Y_i$$

entonces $g_n(x) = \mathbb{I}_{m_n(x)>1/2}$, es decir, es una regla plug-in, y m_n así definido es un estimador de regresión como los que estudiamos en el Teorema de Stone, con $w_{ni} = 1/k(X)$ si i es tal que $X_i \in C_n(X)$.

El siguiente teorema prueba que esta regla de clasificación es universalmente consistente.

Teorema 3.19. *Stone (1977).*
Si $k = k(n) \to \infty$ y $k/n \to 0$ cuando $n \to \infty$, para toda distribución del par (X, Y), $\mathbb{E}(L_n) \to L^*$.

Demostración. Por (3.12) y (3.13) basta verificar que se cumplen las 3 condiciones del Teorema 3.1. La condición (iii) se sigue de que $k \to \infty$.

Respecto a la condición (ii) ,

$$\mathbb{E}\left(\sum_{i=1}^{n} w_{ni}(X)\mathbb{I}_{\{\|X_i - X\|>a\}}\right) = \mathbb{E}\left(\sum_{i=1}^{n} \frac{1}{k}\mathbb{I}_{\{X_i \in C_n(X)\}}\mathbb{I}_{\{\|X_i - X\|>a\}}\right)$$

Observemos primero que esta sumatoria tiene únicamente k sumandos no nulos, por la primera indicatriz. Por otra parte vale, para todo i,

$$\mathbb{I}_{\{X_i \in C_n(X)\}}\mathbb{I}_{\{\|X_i - X\|>a\}} \leq \mathbb{I}_{\{\|X^{(k)} - X\|>a\}}.$$

Por lo tanto

$$\mathbb{E}\left(\sum_{i=1}^{n} \frac{1}{k}\mathbb{I}_{\{X_i \in C_n(X)\}}\mathbb{I}_{\{\|X_i - X\|>a\}}\right) \leq \mathbb{P}(\|X^{(k)} - X\| > a)$$

Basta probar entonces que $\mathbb{P}(\|X^{(k)} - X\| > a) \to 0$. Denotemos $\aleph_n = \{X_1, \ldots, X_n\}$, observemos que

$$\mathbb{P}\big(\|X^{(k)} - X\| > a\big) = \mathbb{P}\Big(\#\big(\aleph_n \cap B(X, a)\big) < k\Big) = \mathbb{E}\left(\mathbb{P}\Big(\#\big(\aleph_n \cap B(X, a)\big) < k\Big|X\Big)\right)$$

Condicionado a X, la variable aleatoria $\#\big(\aleph_n \cap B(X, a)\big)$ tiene distribución Binomial de parámetros n y $p = P_X\big(B(X, a)\big)$. Observar que para todo X, $p > 0$. Vamos a acotar la probabilidad de adentro, asumiendo que condicionamos a X

$$\mathbb{P}\left(\#\big(\aleph_n \cap B(X, a)\big) < k\Big|X\right) = \mathbb{P}\left(\frac{1}{n}\#\big(\aleph_n \cap B(X, a)\big) < \frac{k}{n}\Big|X\right) = \mathbb{P}\left(\frac{1}{n}\#\big(\aleph_n \cap B(X, a)\big) - p < \frac{k}{n} - p\Big|X\right)$$

[7]NN refiere a su sigla en inglés. Es abreviación de *nearest neighbor*

Si multiplicamos por -1 y tomamos valor absoluto

$$\mathbb{P}\left(\frac{1}{n}\#\big(\aleph_n \cap B(X,a)\big) - p < \frac{k}{n} - p \Big| X\right) \leq \mathbb{P}\left(\Big|\frac{1}{n}\#\big(\aleph_n \cap B(X,a)\big) - p\Big| > p - \frac{k}{n}\Big| X\right)$$

Como $p > 0$ y $k/n \to 0$, para n suficientemente grande $p - \frac{k}{n} > p/2$. Luego se concluye usando la desigualdad de Hoeffding.

Finalmente para verificar la condición (i) hay que probar que para toda f medible, no negativa con $\mathbb{E}(f(X)) < \infty$ se cumple que

$$\mathbb{E}\left(\sum_{i=1}^{n} \frac{1}{k}\mathbb{I}_{\{X_i \text{ esta entre los k-vecinos mas cercanos a } X\}} f(X_i)\right) \leq c\mathbb{E}(f(X)),$$

para alguna constante $c > 0$. No lo probaremos. Ver Lema 5.3 en [9]. □

4 Clasificación lineal y particiones

Veamos ahora una de las reglas más simples de clasificación binaria, para el caso en que $x \in \mathbb{R}^d$, aunque puede formularse de forma idéntica en cualquier espacio de Hilbert. Supongamos que tenemos pesos $a_0, \ldots, a_d$; la regla de clasificación $g(x)$ dada por

$$g(x) = \begin{cases} 1 & \text{si } \sum_{i=1}^d a_i x_i + a_0 > 0 \\ 0 & \text{en caso contrario} \end{cases} \tag{4.1}$$

donde $x = (x_1, \ldots, x_d)$, se conoce como regla de clasificación lineal, y es la base para entender las redes neuronales que veremos más adelante. El vector $a = (a_1, \ldots, a_d)$ y la constante a_0 tendremos que elegirlos a partir de una muestra iid de un par (X, Y).

Observación 4.1. *Volviendo a la discusión que hicimos en la introducción, sobre la cultura de los modelos o de los algoritmos, cabe destacar que, si no hacemos ninguna hipótesis sobre la distribución del par (X, Y) y buscamos a y a_0 que minimicen el riesgo empírico $\hat{L}_n(g)$, entonces estamos ante el abordaje basado en algoritmos, de Rosenblatt de 1962. En este caso el clasificador* (4.1) *se conoce como perceptrón. En este capítulo veremos una cota superior para el error de clasificación de la regla* (4.1)*. En general vamos a estar lejos del error de Bayes L^*, salvo un caso muy particular: Fisher abordó en 1936 el problema asumiendo que las distribuciones de $X|Y = 1$ y $X|Y = 0$ eran normales con medias μ_1 y μ_0 respectivamente, y con matriz de varianzas y covarianzas $\Sigma_1 = \Sigma_0$. Bajo estas hipótesis*[1]*, la regla de Bayes resulta de la forma* (4.1)*, esto es fácil de ver y quedará como ejercicio más adelante. No es el único caso, un ejemplo trivial es cuando las distribuciones condicionales tienen soporte en conjuntos que son separables por un hiperplano, en este caso $L^* = 0$.*

Es intuitivo que, salvo casos muy especiales, esta regla no es consistente universalmente. Una generalización natural de la misma se obtiene al tomar una función $f : \mathbb{R} \to \mathbb{R}$ y clasificar como 1 a x si $f(a^t x + a_0)$ supera un determinado umbral, o 0 en caso contrario. Ejemplos clásicos de estas funciones son $f(x) = \text{signo}(x)$ o $f(x) = \text{máx}\{0, x\}$. Esto lo retomaremos en el capítulo 8.

Vamos a denotar $L(a, a_0) = \mathbb{P}(g(X) \neq Y)$ y

$$L = \inf_{a \in \mathbb{R}^d, a_0 \in \mathbb{R}} L(a, a_0). \tag{4.2}$$

Para estudiar L vamos a ver primero el caso univariado.

4.1 Clasificación Lineal Univariada

Supongamos que X toma valores en $\mathbb{R}$, en este caso la regla de clasificación lineal es

$$g(x) = \begin{cases} y' & \text{si } x \leq x' \\ 1 - y' & \text{en caso contrario} \end{cases}$$

donde $y' \in \{0, 1\}$. Veremos cómo es el error de clasificación de esta regla, como función de x', y'. Vamos a denotar

$$p = \mathbb{P}(Y = 1), \qquad F_1(x) = \mathbb{P}(X \leq x | Y = 1) \qquad \text{y} \qquad F_0(x) = \mathbb{P}(X \leq x | Y = 0).$$

Supongamos que $y' = 1$, en este caso

$$g(x) = \begin{cases} 1 & \text{si } x \leq x' \\ 0 & \text{en caso contrario} \end{cases}$$

$$\begin{aligned} \mathbb{P}(g(X) \neq Y) &= \mathbb{P}(g(X) = 1, Y = 0) + \mathbb{P}(g(X) = 0, Y = 1) \\ &= \mathbb{P}(g(X) = 1 | Y = 0)(1 - p) + \mathbb{P}(g(X) = 0 | Y = 1)p \\ &= \mathbb{P}(X \leq x' | Y = 0)(1 - p) + \mathbb{P}(X > x' | Y = 1)p \\ &= F_0(x')(1 - p) + (1 - F_1(x'))p. \end{aligned} \tag{4.3}$$

[1] observar que estamos en el enfoque basado en modelo, ya que asumimos una distribución para los datos

Por su parte si $y' = 0$

$$g(x) = \begin{cases} 0 & \text{si } x \leq x' \\ 1 & \text{en caso contrario} \end{cases}$$

Razonando como en (4.3)

$$\begin{aligned} \mathbb{P}(g(X) \neq Y) &= \mathbb{P}(g(X) = 1, Y = 0) + \mathbb{P}(g(X) = 0, Y = 1) \\ &= \mathbb{P}(g(X) = 1|Y = 0)(1-p) + \mathbb{P}(g(X) = 0|Y = 1)p \\ &= \mathbb{P}(X > x'|Y = 0)(1-p) + \mathbb{P}(X \leq x'|Y = 1)p \\ &= (1 - F_0(x'))(1-p) + F_1(x')p. \end{aligned} \tag{4.4}$$

Queremos encontrar el par (x^*, y^*) que verifica

$$(x^*, y^*) = \arg\min_{(x',y')} \mathbb{P}(g(X) \neq Y).$$

Este par existe porque estamos permitiendo $x' = -\infty$ o $x' = \infty$. Vamos a denotar L^* el error de la regla correspondiente a (x^*, y^*). Se verifica que

Lema 4.2.

$$L = \frac{1}{2} - \sup_x \Big| pF_1(x) - (1-p)F_0(x) - p + \frac{1}{2} \Big| \tag{4.5}$$

en particular para $p = 1/2$,

$$L = \frac{1}{2} - \frac{1}{2} \sup_x \Big| F_1(x) - F_0(x) \Big|$$

Demostración. Como los casos (4.3) y (4.4) son excluyentes, tenemos que

$$L = \inf_{(x',y')} \left\{ \mathbb{I}_{y'=0} \Big[(1 - F_0(x'))(1-p) + F_1(x')p \Big] + \mathbb{I}_{y'=1} \Big[F_0(x')(1-p) + (1 - F_1(x'))p \Big] \right\}.$$

Si el ínfimo se da cuando $y' = 0$ tenemos que

$$L = \inf_{x'} \Big\{ \big(1 - F_0(x')\big)(1-p) + F_1(x')p \Big\}.$$

Si el ínfimo se da en $y' = 1$,

$$L = \inf_{x'} \Big\{ F_0(x')(1-p) + \big(1 - F_1(x')\big)p \Big\}.$$

Por lo tanto

$$L = \inf_{x'} \text{mín} \Big\{ \underbrace{(1 - F_0(x'))(1-p) + F_1(x')p}_{a} \ , \ \underbrace{F_0(x')(1-p) + (1 - F_1(x'))p}_{b} \Big\}.$$

Si usamos $\text{mín}\{a, b\} = (a + b - |a - b|)/2$ y verificamos que en nuestro caso $a + b = 1$,

$$L = \inf_{x'} \ \frac{1}{2} \left[1 - \Big| (1 - F_0(x'))(1-p) + F_1(x')p - \Big(F_0(x')(1-p) + (1 - F_1(x'))p \Big) \Big| \right].$$

Haciendo cuentas

$$\frac{1}{2} \Big| (1 - F_0(x'))(1-p) + F_1(x')p - \Big(F_0(x')(1-p) + (1 - F_1(x'))p \Big) \Big| = \Big| pF_1(x') - (1-p)F_0(x') - p + \frac{1}{2} \Big|.$$

por lo tanto,

$$L = \frac{1}{2} - \sup_x \Big| pF_1(x) - (1-p)F_0(x) - p + \frac{1}{2} \Big|.$$

□

Denotemos para $i = 0, 1$, $m_i = \mathbb{E}(X|Y = i)$, $\sigma_i^2 = \mathbb{V}(X|Y = i)$. Tenemos la siguiente acotación

Teorema 4.3.

$$L^* \leq L \leq \frac{1}{1 + \left[\frac{m_0 - m_1}{\sigma_0 + \sigma_1} \right]^2}$$

Demostración. Supongamos sin pérdida de generalidad que $m_0 < m_1$. Sea $0 < \Delta_0 < m_1 - m_0$, L es menor o igual que la probabilidad de error de la regla g dada por

$$g(x) = \begin{cases} 0 & \text{si } x \leq m_0 + \Delta_0 \\ 1 & \text{en caso contrario} \end{cases}$$

Denotemos $m_1 - m_0 = \Delta_0 + \Delta_1$, con $\Delta_1, \Delta_2 > 0$. El error de esta regla es

$$L \equiv \mathbb{P}(g(X) \neq Y) = \mathbb{P}(X \leq m_1 - \Delta_1 | Y = 1)p + \mathbb{P}(X > m_0 + \Delta_0 | Y = 0)(1-p). \tag{4.6}$$

Por la desigualdad de Chebyshev-Cantelli (ver Teorema 10.18 en el Apéndice) aplicada a $-X$, tenemos que

$$\mathbb{P}(X \leq m_1 - \Delta_1 | Y = 1) \leq \frac{\sigma_1^2}{\sigma_1^2 + \Delta_1^2} = \frac{1}{1 + \frac{\Delta_1^2}{\sigma_1^2}}$$

y si aplicamos la misma igualdad pero a X, tenemos que

$$\mathbb{P}(X > m_0 + \Delta_0 | Y = 0) \leq \frac{\sigma_0^2}{\sigma_0^2 + \Delta_0^2} = \frac{1}{1 + \frac{\Delta_0^2}{\sigma_0^2}}.$$

Tomemos

$$\Delta_0 = \frac{(m_1 - m_0)\sigma_0}{\sigma_0 + \sigma_1}$$

por lo tanto

$$\Delta_1 = (m_1 - m_2) - \frac{(m_1 - m_0)\sigma_0}{\sigma_0 + \sigma_1} = \frac{\sigma_1(m_1 - m_0)}{\sigma_0 + \sigma_1} = \frac{\sigma_1(m_1 - m_0)\sigma_0}{(\sigma_0 + \sigma_1)} \frac{1}{\sigma_0} = \frac{\sigma_1}{\sigma_0}\Delta_0$$

Usando ahora la cota (4.6)

$$L \leq \frac{p}{1 + \frac{(\frac{\sigma_1}{\sigma_0}\Delta_0)^2}{\sigma_1^2}} + \frac{1-p}{1 + \frac{\Delta_0^2}{\sigma_0^2}} = \frac{p}{1 + \frac{\Delta_0^2}{\sigma_0^2}} + \frac{1-p}{1 + \frac{\Delta_0^2}{\sigma_0^2}}.$$

Finalmente,

$$L \leq \frac{1}{1 + \left[\frac{m_0 - m_1}{\sigma_0 + \sigma_1}\right]^2}$$

□

4.2 Clasificación Lineal Multivariada

Volviendo a la regresión multivariada, fijado $a \in \mathbb{R}^d$, $x' \in \mathbb{R}$, $y' \in \{0, 1\}$, tenemos una regla lineal $g_a : \mathbb{R}^d \to \{0, 1\}$

$$g_a(x) = \begin{cases} y' & \text{si } \langle a, x \rangle \leq x' \\ 1 - y' & \text{en caso contrario} \end{cases}$$

ver Figura 4.1. Por el Lema (4.2), el ínfimo al variar x' e y', del error de g_a es L_a, dado por

$$L_a = \frac{1}{2} - \sup_x \left| pF_{1,a}(x) - (1-p)F_{0,a}(x) - p + \frac{1}{2} \right|.$$

donde $F_{1,a}$ es la distribución de $\sum_{i=1}^d a_i X_i$ condicionada a $Y = 1$ y $F_{0,a}$ es la distribución de $\sum_{i=1}^d a_i X_i$ condicionada a $Y = 0$.

El ínfimo, que denotamos L en (4.2), de la regla lineal (4.1), al variar a, es $\inf_a L_a$, o, lo que es lo mismo

$$L = \frac{1}{2} - \sup_{a \in \mathbb{R}^d} \sup_x \left| pF_{1,a}(x) - (1-p)F_{0,a}(x) - p + \frac{1}{2} \right|.$$

Tenemos, además, la siguiente cota del error, que nos da una idea además, de como elegir a y a_0.

Teorema 4.4. *Sean X_0 y X_1 variables aleatorias distribuidas como X dado $Y = 0$, e $Y = 1$ respectivamente. Sean $m_0 = \mathbb{E}(X_0)$ y $m_1 = \mathbb{E}(X_1)$. Definamos las matrices de covarianzas*

$$\Sigma_1 = \mathbb{E}\Big[(X_1 - m_1)(X_1 - m_1)^T\Big] \quad y \quad \Sigma_0 = \mathbb{E}\Big[(X_0 - m_0)(X_0 - m_0)^T\Big],$$

entonces

$$L^* \leq L \leq \inf_{a \in \mathbb{R}^d} \frac{1}{1 + \frac{\left(a^T(m_1 - m_0)\right)^2}{\left((a^T\Sigma_0 a)^{1/2} + (a^T\Sigma_1 a)^{1/2}\right)^2}}$$

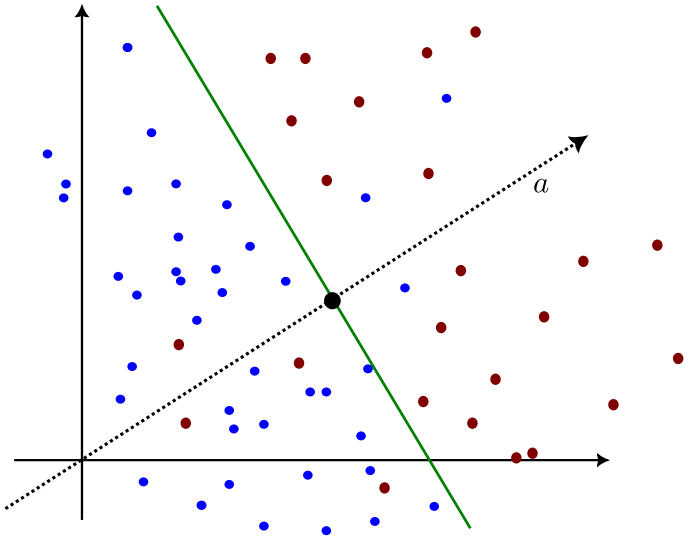

Figura 4.1: Fijado $a \in \mathbb{R}^d$ tal que $\|a\| = 1$, y $x \in \mathbb{R}$, el semiespacio que separa los datos es $\{z : \langle a, z\rangle > x\}$. Es decir $-x$ juega el rol de la constante a_0 en (4.1), y de x' en el clasificador univariado. En la figura, la regla de clasificación lineal para el a y x que se muestran, clasifica mal 3 puntos azules y 4 rojos.

Demostración. La prueba se sigue del Teorema 4.3 aplicada a las variables $a^T X_0$ y $a^T X_1$. □

Observación 4.5. *Si conociéramos m_1, m_0, Σ_0 y Σ_1 el teorema anterior nos dice que tenemos que elegir a de modo de maximizar*

$$\frac{(a^T(m_1 - m_0))^2}{\left((a^T\Sigma_0 a)^{1/2} + (a^T\Sigma_1 a)^{1/2}\right)^2}$$

ya que con esto estaríamos minimizando L. Lo que hace el clásico discriminante lineal de Fisher, que veremos en la sección siguiente, es cambiar estos parámetros por sus estimaciones "plug-in". Esto no pasa con el perceptrón de Rosemblatt que busca el hiperplano que hace que el error de clasificación en cada una de las clases que este hiperplano determina, sea lo más pequeño posible. Este hiperplano se puede hallar por algoritmos de tipo descenso por gradiente, y no requiere de hipótesis sobre la distribución del par (X, Y).

4.2.1 Discriminante lineal de Fisher

Uno de los criterios mas famosos para elegir a y a_0 a partir de los datos fue propuesto por Fisher en 1936. Denotemos para $i = 0, 1$, $\hat{m}_i$ el promedio de las X si $Y = i$, es decir, por ejemplo

$$\hat{m}_1 = \frac{1}{|\{i : Y_i = 1\}|} \sum_{i:Y_i=1} X_i \qquad \text{y} \qquad \hat{m}_0 = \frac{1}{|\{i : Y_i = 0\}|} \sum_{i:Y_i=0} X_i.$$

Si proyectamos los datos $X_1, \ldots, X_n$ ortogonalmente sobre el hiperplano $a^T x = 0$ obtenemos $a^T X_1, \ldots, a^T X_n$. Definimos la dispersión de la proyección de X en la clase $Y = 1$ como

$$\hat{\sigma}_1^2 = \sum_{i:Y_i=1} (a^T X_i - a^T \hat{m}_1)^2 = \sum_{i:Y_i=1} a^T (X_i - \hat{m}_1)(X_i - \hat{m}_1)^T a = a^T S_1 a$$

donde

$$S_1 = \sum_{i:Y_i=1} (X_i - \hat{m}_1)(X_i - \hat{m}_1)^T$$

Y lo mismo para la clase $Y = 0$. El discriminante lineal de Fisher es la función lineal $a^T x$ para la cual a se elije de modo que maximice

$$J(a) = \frac{(a^T \hat{m}_1 - a^T \hat{m}_0)^2}{\hat{\sigma}_1^2 + \hat{\sigma}_0^2} = \frac{\left(a^T(\hat{m}_1 - \hat{m}_0)\right)^2}{a^T(S_1 + S_0)a}.$$

Es decir a es la dirección en la cual las medias empíricas de cada clase, luego de proyectar, están más separadas, ponderando por la dispersión en cada clase.

La solución del problema de maximización se da cuando

$$a = (S_1 + S_0)^{-1}(\hat{m}_1 - \hat{m}_0).$$

La regla de clasificación de Fisher, g_{a_0} clasifica un nuevo x como 1 si $a^T x + a_0 > 0$ y como 0 en caso contrario siendo a el valor antes obtenido, y a_0 una cierta contante. En general su error puede estar lejos del error óptimo que se obtiene con la mejor regla lineal. De hecho dado $\epsilon > 0$ se puede encontrar un par (X, Y) tal que $X \in \mathbb{R}^2$, $\mathbb{E}(\|X\|^2) < \infty$ y que sean linealmente separables, es decir el error L de la mejor regla lineal de clasificación es 0, pero $\inf_{a_0} \mathbb{E}(g_{a_0}) > 1 - \epsilon$.

Ejercicio 4.6. En general la regla de Bayes no tiene por qué ser lineal, un caso muy especial es cuando $X|Y = 1$ y $X|Y = 0$ son normal multivariada con medias μ_1 y μ_0 respectivamente, y matriz de varianzas y covarianzas Σ, que asumimos invertible, y $\mathbb{P}(Y = 1) = 1/2$. Se deja como ejercicio probar que en este caso la regla de Bayes es lineal, y tiene la forma

$$g(x) = \begin{cases} 1 & \text{si } \langle a, x\rangle + a_0 > 0 \\ 0 & \text{en caso contrario} \end{cases}$$

donde $a = \Sigma^{-1}(\mu_1 - \mu_0)$ y $a_0 = -(1/2)a^T(\mu_1 + \mu_0)$. Sugerencia, usar (3.11), con $p = 1/2$. Si $p \neq 1/2$ también queda lineal pero la expresión no es tan limpia.

4.3 Reglas de clasificación basadas en particiones

Un ejemplo de red neuronal de una capa consiste en particionar el espacio en regiones determinadas por hiperplanos, y luego la etiqueta asignada a un punto x es el resultado de hacer voto mayoritario en el simplex de la partición que contiene a x. Para probar su consistencia vamos a ver primero de forma general las reglas basas en particiones. Supongamos que tenemos $A_1, A_2, \ldots$ una partición de $\mathbb{R}^d$, para cada $x \in \mathbb{R}^d$ consideramos $A(x)$ la celda de la partición que contiene a x. Denotamos g_n la regla

$$g_n(x) = \begin{cases} 0 & \text{si } \sum_{i=1}^n \mathbb{I}_1(Y_i)\mathbb{I}_{\{X_i \in A(x)\}} \leq \sum_{i=1}^n \mathbb{I}_0(Y_i)\mathbb{I}_{\{X_i \in A(x)\}} \\ 1 & \text{en caso contrario} \end{cases} \tag{4.7}$$

Para que esta regla sea consistente vamos a pedir que las celdas se vayan achicando pero lo suficientemente lento como para que la cantidad de puntos de la muestra en la celda se vaya a infinito. Es decir si definimos $diam(A) = \sup_{x,y \in A} ||x - y||$ el diámetro de un conjunto, vamos a pedir que $diam(A(X)) \to 0$ en probabilidad y además

$$N(X) = n\mu_n(A(X)) = \sum_{i=1}^n \mathbb{I}_{\{X_i \in A(X)\}} \to \infty \text{ en probabilidad,}$$

donde $A(X)$ es la celda que contiene a X. Es decir tenemos el teorema:

Teorema 4.7. *Sea A un elemento genérico de una partición de $\mathbb{R}^d$. Supongamos que $diam(A(X)) \to 0$ en probabilidad, y $N(X) \to \infty$ en probabilidad, entonces $\mathbb{E}(L_n) \to L^*$.*

Demostración. Recordemos que $m^*(x) = \mathbb{P}(Y = 1|X = x)$. Por la desigualdad (3.12) es suficiente probar que

$$\mathbb{E}\big|\hat{m}_n(X) - m^*(X)\big| \to 0,$$

donde

$$\hat{m}_n(x) = \frac{1}{N(x)} \sum_{i:X_i \in A(x)} Y_i.$$

Denotemos $\overline{m}(x) = \mathbb{E}\big(m^*(X)|X \in A(x)\big)$.[2]

Por la desigualdad triangular

$$\mathbb{E}\big|\hat{m}_n(X) - m^*(X)\big| \leq \underbrace{\mathbb{E}\big|\hat{m}_n(X) - \overline{m}(X)\big|}_{\mathbf{A}} + \underbrace{\mathbb{E}\big|\overline{m}(X) - m^*(X)\big|}_{\mathbf{B}}.$$

Acotación de A

Denotemos

$$\mathcal{A} = \sigma(\{X, \mathbb{I}_{\{X_1 \in A(X)\}}, \ldots, \mathbb{I}_{\{X_n \in A(X)\}}\}).$$

Como $N(X)$ condicionado $\mathcal{A}$, es $N(X)$ c.s y $|\hat{m}_n(X) - \overline{m}(X)| \leq 1$

$$\mathbb{E}\Big[\big|\hat{m}_n(X) - \overline{m}(X)\big|\Big|\mathcal{A}\Big] \leq \mathbb{E}\Big[\big|\hat{m}_n(X) - \overline{m}(X)\big|\mathbb{I}_{\{N(X)>0\}}\Big|\mathcal{A}\Big] + \mathbb{I}_{\{N(X)=0\}} \tag{4.8}$$

[2]en la definición de $\overline{m}(x)$ estamos condicionando al suceso $B_x = \{\omega : X(\omega) \in A(x)\}$ por lo tanto $\overline{m}(x) = \mathbb{E}(m^*(X)\mathbb{I}_{B_x})/\mathbb{P}(B_x)$. La variable aleatoria $\overline{m}(X)$, dado un ω, calcula $x = X(\omega)$, y luego $B_{X(\omega)}$. Observar que podemos suponer que $\mathbb{P}(X \in A(x)) > 0$ ya que $\mathbb{E}\big|\hat{m}_n(X) - m^*(X)\big|$ se puede descomponer en suma de integrales sobre celdas de la partición cuya partición tenga probabilidad positiva respecto de P_X.

Condicionado a $\mathcal{A}$, $\overline{m}(X)$ y $N(X)$ son constantes y $N(X)\hat{m}_n(X)$ es una variable aleatoria con distribución Binomial de parámetros $N(X)$, $\overline{m}(X)$, por lo tanto,

$$\mathbb{E}\left[\left|\hat{m}_n(X)-\overline{m}(X)\right|\mathbb{I}_{\{N(X)>0\}}\middle|\mathcal{A}\right]=\mathbb{E}\left[\left|\frac{B(N(X),\overline{m}(X))}{N(X)}-\overline{m}(X)\right|\mathbb{I}_{\{N(X)>0\}}\middle|\mathcal{A}\right].$$

Definimos la variable,

$$Z=\Big[\frac{B(N(X),\overline{m}(X))}{N(X)}-\overline{m}(X)\Big].$$

Si aplicamos la desigualdad de Cauchy-Schwartz (2.8) con $Y=\mathbb{I}_{\{N(X)>0\}}$ y $X=Z$ obtenemos que

$$\mathbb{E}[|ZY|\ |\mathcal{A}]\le\sqrt{\mathbb{E}[Z^2|\mathcal{A}]}\sqrt{\mathbb{E}[\mathbb{I}_{\{N(X)>0\}}|\mathcal{A}]}=\sqrt{\mathbb{E}[Z^2|\mathcal{A}]}\mathbb{I}_{\{N(X)>0\}}=\mathbb{E}\Big[\sqrt{\mathbb{E}[Z^2|\mathcal{A}]}\mathbb{I}_{\{N(X)>0\}}\Big|\mathcal{A}\Big]$$

$$\mathbb{E}(Z^2|\mathcal{A})=\mathbb{V}\left(\frac{B(N(X),\overline{m}(X))}{N(X)}\Big|\mathcal{A}\right)=\frac{1}{N(X)^2}\mathbb{V}\Big(B\big(N(X),\overline{m}(X)\big)|\mathcal{A}\Big)=\frac{1}{N(X)^2}N(X)\overline{m}(X)(1-\overline{m}(X))$$
$$=\frac{1}{N(X)}\overline{m}(X)(1-\overline{m}(X)).\quad(4.9)$$

De donde se sigue que

$$\mathbb{E}\left[\left|\frac{B(N(X),\overline{m}(X))}{N(X)}-\overline{m}(X)\right|\mathbb{I}_{\{N(X)>0\}}\middle|\mathcal{A}\right]\le\mathbb{E}\left[\sqrt{\frac{\overline{m}(X)(1-\overline{m}(X))}{N(X)}}\mathbb{I}_{\{N(X)>0\}}\middle|\mathcal{A}\right]$$

Como $\overline{m}(X)(1-\overline{m}(X))\le 1/4$ obtenemos de (4.9) que

$$\mathbb{E}\left[\left|\frac{B(N(X),\overline{m}(X))}{N(X)}-\overline{m}(X)\right|\mathbb{I}_{\{N(X)>0\}}\middle|\mathcal{A}\right]\le\mathbb{E}\left[\frac{1}{2\sqrt{N(X)}}\mathbb{I}_{\{N(X)>0\}}\middle|\mathcal{A}\right]$$

Si combinamos esto último con (4.8) y tomamos esperanza obtenemos que

$$\mathbf{A}\le\mathbb{E}\left[\frac{1}{2\sqrt{N(X)}}\mathbb{I}_{\{N(X)>0\}}\right]+\mathbb{P}(N(X)=0)$$

Para todo $k>0$,

$$\mathbb{E}\left[\frac{1}{2\sqrt{N(X)}}\mathbb{I}_{\{N(X)>0\}}\right]=\ \mathbb{E}\left[\frac{1}{2\sqrt{N(X)}}\mathbb{I}_{\{0<N(X)\le k\}}\right]+\mathbb{E}\left[\frac{1}{2\sqrt{N(X)}}\mathbb{I}_{\{N(X)>k\}}\right]\le\frac{1}{2}\mathbb{P}\big(N(X)\le k\big)+\frac{1}{2\sqrt{k}}$$

Fijado $\epsilon>0$ tomamos k suficientemente grande tal que $1/(2\sqrt{k})<\epsilon/3$. Para ese k, como $N(X)\to\infty$ en probabilidad, se puede tomar n suficientemente grande tal que $(1/2)\mathbb{P}(N(X)\le k)<\epsilon/3$ y además $P(N(X)=0)<\epsilon/3$.

Acotación de B

Dado $\epsilon>0$ sea m_ϵ a valores en $[0,1]$, uniformemente continua en un conjunto $C\subset\mathbb{R}^d$ acotado, que se anula en C^c, tal que $\mathbb{E}|m_\epsilon(X)-m^*(X)|<\epsilon$. Denotemos $\overline{m}_\epsilon(x)=\mathbb{E}[m_\epsilon(X)|X\in A(x)]$, entonces

$$\mathbb{E}\big|\overline{m}(X)-m^*(X)\big|\ \le\mathbb{E}\big|\overline{m}(X)-\overline{m}_\epsilon(X)\big|\ +\ \mathbb{E}\big|\overline{m}_\epsilon(X)-m_\epsilon(X)\big|\ +\ \mathbb{E}\big|m_\epsilon(X)-m^*(X)\big|:=\text{I+II+III}$$

Por la forma en que elegimos m_ϵ, III$<\epsilon$. Además

$$\text{I}=\mathbb{E}\big|\overline{m}(X)-\overline{m}_\epsilon(X)\big|=\mathbb{E}\left|\mathbb{E}\Big(m^*(X)-m_\epsilon(X)\Big|X\in A(X)\Big)\right|$$
$$\le\mathbb{E}\left[\mathbb{E}\Big[\big|m^*(X)-m_\epsilon(X)\big|\Big|X\in A(X)\Big]\right]=\text{III}.$$

Para acotar II, como m_ϵ es uniformemente continua existe $\theta=\theta(\epsilon)>0$ tal que si $diam(A(X))<\theta$ entonces $|m_\epsilon(z)-m_\epsilon(t)|<\epsilon$ si $|z-t|<\theta$. Observemos que, por (2.6)

$$\overline{m}_\epsilon(x)=\frac{1}{\mathbb{P}(X\in A(x))}\int_{X\in A(x)}m_\epsilon(X)d\mathbb{P}=\frac{1}{P_X(A(x))}\int_{A(x)}m_\epsilon(z)P_X(dz).$$

De donde,

$$\mathbb{E}\Big[|\overline{m}_\epsilon(X)-m_\epsilon(X)|\mathbb{I}_{\{diam(A(X))<\theta\}}\Big|X\in A(X)\Big]$$
$$\le\mathbb{E}\left[\frac{1}{P_X(A(X))}\int_{A(X)}|m_\epsilon(z)-m_\epsilon(X)|\mathbb{I}_{\{diam(A(X))<\theta\}}P_X(dz)\middle|X\in A(X)\right]$$

Por lo tanto II$\le\epsilon+\mathbb{P}(diam(A(X)>\theta))$. □

4.3.1 Histogramas

Como consecuencia del teorema anterior se prueba el caso particular en que la partición está formada por cubos de longitud h_n, es decir, tomamos conjuntos de la forma

$$\prod_{i=1}^{d}\Big[k_i h_n,(k_i+1)h_n\Big) \quad \text{con } k_i\in\mathbb{Z}$$

Denotamos esta partición como $\mathcal{P}_n=\{A_{n1},A_{n2},\dots\}$. Si $x\in A_{ni}$ denotamos $A_n(x)=A_{ni}$. La regla (4.7) para el caso de la partición $\mathcal{P}$ se llama la regla de histograma. El siguiente teorema establece su consistencia, es decir $\mathbb{E}(L_n)\to L^*$. Para eso vamos a verificar que se cumplen las hipótesis del Teorema (4.7)

Teorema 4.8. *Si $h_n\to 0$ tal que $nh_n^d\to\infty$ la regla de histograma es universalmente consistente, es decir, $\mathbb{E}(L_n)\to L^*$.*

Demostración. El diámetro de cada celda es $\sqrt{d}h_n^d\to 0$. Por lo tanto resta probar que para todo $M>0$, $\mathbb{P}(N(X)<M)\to 0$. Sea $\epsilon>0$ arbitrario, y S una bola centrada en el origen, de radio $R>0$, donde tomamos R de modo que $\mu(S^c)<$ *epsilon*. El número de celdas A_{ni} de la partición $\mathcal{P}_n$ que cortan a S está acotada superiormente por $(2R)^d/h_n^d$, ya que esta es la cantidad de elementos de la partición en el cubo $[-R,R]^d$.

$$\mathbb{P}\big(N(X)<M\big)=\sum_{j=1}^{\infty}\mathbb{P}\big(X\in A_{nj},N(X)<M\big)\le\sum_{j:A_{nj}\cap S\neq\emptyset}\mathbb{P}\big(X\in A_{nj},N(X)<M\big)\ +\\ \sum_{j:A_{nj}\cap S=0}\mathbb{P}\big(X\in A_{nj},N(X)<M\big).$$

Por un lado acotamos

$$\sum_{j:A_{nj}\cap S=0}\mathbb{P}\big(X\in A_{nj},N(X)<M\big)\le\sum_{j:A_{nj}\cap S=0}\mathbb{P}(X\in A_{nj})\le\mathbb{P}(S^c)$$

por otro lado, si denotamos μ a la distribución de X, podemos separar la primer sumatoria en

$$\begin{aligned}\sum_{j:A_{nj}\cap S\neq\emptyset}\mathbb{P}\big(X\in A_{nj},N(X)<M\big)&=\sum_{\substack{j:A_{nj}\cap S\neq\emptyset\\ \mu(A_{nj})\le 2M/n}}\mathbb{P}\big(X\in A_{nj},N(X)<M\big)\\ &\quad+\sum_{\substack{j:A_{nj}\cap S\neq\emptyset\\ \mu(A_{nj})>2M/n}}\mathbb{P}\big(X\in A_{nj},N(X)<M\big)\\ &\le\sum_{\substack{j:A_{nj}\cap S\neq\emptyset\\ \mu(A_{nj})\le 2M/n}}\mathbb{P}(X\in A_{nj})+\sum_{\substack{j:A_{nj}\cap S\neq\emptyset\\ \mu(A_{nj})>2M/n}}\mathbb{P}(n\mu_n(A_{nj})<M)\mathbb{P}(X\in A_{nj})=\text{I+II}\end{aligned}$$

donde $\mu_n(A_{nj})$ es la cantidad de puntos de la muestra en la celda A_{nj} dividido n, es decir, su medida empírica. Acotamos superiormente I$\le[(2R)^d/h^d](2M)/n\to 0$ cuando $n\to\infty$ porque $nh_n^d\to 0$. Para acotar superiormente II escribimos

$$\mathbb{P}\Big(\mu_n(A_{nj})\le M/n\Big)=\mathbb{P}\Big(\mu_n(A_{nj})-\mathbb{E}\big(\mu_n(A_{nj})\big)\le M/n-\mathbb{E}\big(\mu_n(A_{nj})\big)\Big)$$

Observar que $\mathbb{E}(\mu_n(A_{nj}))=\mu(A_{nj})$. Como estamos sumando en los j tal que $M/n<\mu(A_{nj})/2$ tenemos que $M/n-\mu(A_{nj})<-\mu(A_{nj})/2$. Por la desigualdad de Markov podemos acotar

$$\mathbb{P}\left(\mu_n(A_{nj})-\mathbb{E}\big(\mu_n(A_{nj})\big)\le-\frac{\mu(A_{nj})}{2}\right)\le\frac{4\mathbb{V}\big(\mu_n(A_{nj})\big)}{\mu(A_{nj})^2}=\frac{4\mu(A_{nj})\big(1-\mu(A_{nj})\big)}{n^2\mu(A_{nj})^2}\le 4\frac{1}{n^2\mu(A_{nj})}$$

Observemos que si j es de los términos considerados en el segundo sumando

$$4\frac{1}{n^2\mu(A_{nj})}\le\frac{2}{Mn}\to 0\quad\text{cuando } n\to\infty.$$

Juntando las cotas de I y II obtuvimos que

$$\sum_{j:A_{nj}\cap S\neq\emptyset}\mathbb{P}\big(X\in A_{nj},N(X)<M\big)\le\frac{(2R)^d2M}{nh^d}+\frac{2}{Mn}\to 0.$$

Por lo tanto

$$\limsup_{n\to\infty}\mathbb{P}(N(X)<M)\le\mu(S^c)<\epsilon$$

Como ϵ es arbitrario se tiene que $\mathbb{P}(N(X)<M)\to 0$, o, lo que es lo mismo $N(X)\to\infty$, en probabilidad. □

Observación 4.9. *Se puede probar que la regla de clasificación basada en histogramas del teorema anterior es universalmente fuertemente consistente, ver Teorema 9.4, p. 138, en [9]*

5 Modelos lineales

Un modelo lineal tiene la forma[1]

$$Y_i = \sum_{j=1}^{p} x_{ij}\beta_j + \epsilon_i \quad (i = 1, 2, ..., n) \tag{5.1}$$

donde $Y_1, Y_2, ..., Y_n$ son observaciones experimentales de una cierta variable aleatoria con valores reales, $x_{ij} (i = 1, ..., n; j = 1, ..., p)$ son ciertas constantes conocidas, $\beta_1, ..., \beta_p$ son p parámetros desconocidos y $\epsilon_1, \epsilon_2, ..., \epsilon_n$ son los "errores", es decir, las fluctuaciones aleatorias de las observaciones $Y_1, Y_2, ..., Y_n$ en torno a los valores dados por el primer término del segundo miembro de (5.1).

El estudio de un problema mediante un modelo de la forma (5.1), a los efectos de hacer inferencia sobre algún fenómeno, debe ser acompañado de hipótesis sobre la naturaleza estadística de los errores. Diversas hipótesis aparecerán en lo que sigue, pero en todos los casos habremos de suponer que la esperanza matemática de los errores es cero:

$$\mathbb{E}(\epsilon_i) = 0 \quad (i = 1, ..., n).$$

La notación que usaremos es la siguiente:

$$Y = \begin{pmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{pmatrix}, \quad A = \begin{pmatrix} x_{11} & x_{12} & \cdots & x_{1p} \\ x_{21} & x_{22} & \cdots & x_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{np} \end{pmatrix}, \quad \beta = \begin{pmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_p \end{pmatrix}, \quad \epsilon = \begin{pmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{pmatrix}$$

con lo que (5.1) se escribe:

$$Y = A\beta + \epsilon \tag{5.2}$$

Ejemplo 1: magnitud desconocida

$Y_1, Y_2, ..., Y_n$ son una muestra de observaciones de una cierta magnitud μ desconocida, que se mide con error. Entonces:

$$Y_i = \mu + \epsilon_i \quad (i = 1, ..., n)$$

o sea que, con la notación anterior:

$$A = \begin{pmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{pmatrix}, \quad p = 1, \quad \beta = \mu.$$

En un ejemplo de este tipo pueden plantearse diversos problemas, a saber:

- Estimar el parámetro μ a partir de las n observaciones;
- Supongamos que las observaciones se realizan con un aparato de medida, cuya dispersión en torno al verdadero valor μ puede medirse de manera adecuada por la varianza $\sigma^2 = \mathbb{V}(\epsilon_i) \quad (i = 1, ..., n)$, común a todas las observaciones. La estimación de σ^2 dará una idea de la dispersión del aparato.
- Otro problema natural es, a partir de la muestra, hacer una prueba de hipótesis sobre el valor de μ, por ejemplo, decidir si $\mu < \mu_0$, donde μ_0 es un valor dado con alguna significación física. Del mismo modo, podemos estar interesados en realizar una prueba de hipótesis sobre el valor de σ^2, por ejemplo, decidir si la dispersión del aparato no supera un valor dado.

[1] notas basadas en notas de Mario Wschebor

Ejemplo 2. Ajuste de funciones

Supongamos que tenemos un fenómeno representado por una función $\{X(t) : t \in I\}$, I es un intervalo de la recta real, y que $X(t)$ es la superposición de una ley determinística y ciertas fluctuaciones aleatorias. Representamos la parte determinística por funciones de cierta clase, preferentemente sencillas, por ejemplo polinomios de grado k.

Suponemos que tenemos observaciones efectuadas en tiempos fijos y conocidos $t_i, (t_i, X_i)(i = 1, ..., n)$. Entonces:

$$X_i = \beta_0 + \beta_1 t_i + \ldots + \beta_k t_i^k + \epsilon_i \quad (i = 1, ..., n)$$

donde los coeficientes son $k + 1$ parámetros desconocidos. Con la notación convenida es:

$$A = \begin{pmatrix} 1 & t_1 & t_1^2 & \cdots & t_1^k \\ 1 & t_2 & t_2^2 & \cdots & t_2^k \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & t_n & t_n^2 & \cdots & t_n^k \end{pmatrix}, \quad p = k + 1, \quad \beta = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_k \end{pmatrix}.$$

Nuevamente en este ejemplo, se pueden plantear diversos problemas:

- Estimar los coeficientes $\beta_0, \beta_1, \ldots, \beta_k$ mediante funciones de las observaciones (para esas estimaciones emplearemos la notación $\hat{\beta}_0, \hat{\beta}_1, ..., \hat{\beta}_k$);
- Tener una idea de la proximidad entre los estimadores y el verdadero valor de los parámetros, para poder dar intervalos de confianza y poder realizar pruebas de hipótesis sobre el valor de los mismos o sobre relaciones entre ellos;
- Criterios para elegir el grado k de los polinomios, por ejemplo, probar la hipótesis de que $\beta_k = 0$.

En este tipo de ejemplo, observar que, en lugar de los monomios $1, t, t^2, ..., t^k$ podemos utilizar funciones de cualquier otra forma, que sean por alguna razón adecuadas para representar la información contenida en los datos. Habrá que adecuar la matriz A, cuyas columnas serán ahora los valores de las nuevas funciones en los tiempos de observación t_i. Lo que se debe tener en cuenta es que la parte determinística de los segundos miembros sea lineal como función de los parámetros desconocidos.

Por otra parte, el mismo tipo de modelo lineal se puede utilizar para ajustar funciones de más de una variable real, sin que haya ningún cambio esencial en la formulación.

Ejemplo 3. Modelo lineal a efectos fijos

Vamos a suponer que estamos en un modelo del tipo

$$Y_i = \beta_0 + \sum_{j=1}^{p} \beta_j x_{ij} + \epsilon_i \quad (i = 1, ..., n) \tag{5.3}$$

donde asumimos que las x_{ij} no son aleatorias. Esto se denomina **modelo de efectos fijos**. Los errores $\epsilon_1, \ldots, \epsilon_n$ son variables aleatorias que asumimos, por ahora, con esperanza 0 e independientes. En (5.3) $\beta_0, \ldots, \beta_j$ son parámetros que luego vamos a querer estimar. Lo que observamos son las x_{ij} e Y_i, el resto no es observable. No estamos ante un problema de regresión clásico ya que no tenemos una muestra $(X_1, Y_1), \ldots, (X_n, Y_n)$ iid de un par (X, Y). Suponer efectos fijos tiene que ver con el diseño del experimento y el tipo de datos con el que estamos trabajando. Muchas veces es una hipótesis poco razonable y hay que pensar las x_{ij} como aleatorias con cierta distribución (en esos casos se asume generalmente que x_{ij} y ϵ son no correlacionadas). Los modelos de efectos fijos se suelen usar en lo que se denominan datos de panel, es decir, se observan k características, de N personas, a lo largo de T instantes de tiempo, y se plantea

$$Y_{it} = x_{it}^T \beta + \alpha_i + \epsilon_{it} \quad i = 1, \ldots, N \qquad t = 1, \ldots, T$$

donde $x_{it}, \beta \in \mathbb{R}^k$. Observar que α_i depende del individuo pero no del tiempo, y no es observable, como tampoco lo son los errores ϵ_{it} ni el vector β.

El problema en que las x_{ij} son variables aleatorias se denomina de efectos aleatorios y no lo abordaremos en estas notas.

La matriz A correspondiente al (5.3) es

$$A = \begin{pmatrix} 1 & x_{11} & \cdots & x_{1p} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & \cdots & x_{np} \end{pmatrix}$$

Un modelo de este tipo se suele utilizar para estudiar la influencia de las variables $x_1, x_2, ..., x_p$ en el resultado Y.

Ejemplo 4. Clasificación simple (análisis de varianza)

Tenemos q tratamientos que influyen en una magnitud Y. Hacemos observaciones Y_{ij} $(i = 1, ..., q; j = 1, ..., n_i)$, de dicha magnitud, es decir, n_i observaciones aplicando el i-ésimo tratamiento, cuyo efecto medio denotamos por μ_i. El modelo lineal toma la forma:

$$Y_{ij} = \mu_i + \epsilon_{ij} \quad (i = 1, ..., q; j = 1, ..., n_i)$$

y, en notación vectorial, se tiene:

$$n = \sum_{i=1}^{q} n_i, \quad Y = \begin{pmatrix} Y_{11} \\ \vdots \\ Y_{1n_1} \\ Y_{21} \\ \vdots \\ Y_{2n_2} \\ \vdots \\ Y_{q1} \\ \vdots \\ Y_{qn_q} \end{pmatrix}, \quad \epsilon = \begin{pmatrix} \epsilon_{11} \\ \vdots \\ \epsilon_{1n_1} \\ \epsilon_{21} \\ \vdots \\ \epsilon_{2n_2} \\ \vdots \\ \epsilon_{q1} \\ \vdots \\ \epsilon_{qn_q} \end{pmatrix}, \quad \beta = \begin{pmatrix} \mu_1 \\ \mu_2 \\ \vdots \\ \mu_q \end{pmatrix}, \quad A = \begin{pmatrix} 1 & 0 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 1 & 0 & \cdots & 0 \\ 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & \cdots & 1 \\ \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & \cdots & 1 \end{pmatrix}$$

Nuevamente se presenta el problema de estimación de β a partir de las observaciones. Problemas corrientes para este modelo son, a título de ejemplo:

- Realizar una prueba de hipótesis de que $\mu_1 = \mu_2 = ... = \mu_q$, es decir que los efectos de los q tratamientos no difieren significativamente a la luz de las observaciones realizadas;
- Idem de que $\mu_1 = (\mu_2 + \mu_3)/2$, o cualquier otra relación lineal entre las μ's.
- Idem, por ejemplo, de que $\mu_1 > 2\mu_2$, es decir, que el efecto esperado del primer tratamiento es mayor que el doble del segundo.

Ejemplo 5. Clasificación doble

En este caso, la magnitud Y depende de dos tratamientos y se desea estudiar su influencia conjunta. El modelo asume la forma:

$$Y_{ijk} = \mu + \alpha_i + \beta_j + \tau_{ij} + \epsilon_{ijk} \quad (i = 1, ..., m; j = 1, ..., q; k = 1, ..., n_{ij})$$

μ es el efecto medio general de los tratamientos, α_i mide la influencia del primer tratamiento, β_j la del segundo y τ_{ij} la interacción entre ambos. Obsérvese que el modelo es aditivo con respecto a estos elementos. El número total de observaciones es $n = \sum\{n_{ij} : i = 1, .., m; j = 1, ..., q\}$ y el número de parámetros $p = 1 + m + q + mq$.

Queda a cargo del lector la descripción de la matriz A.

Un ejemplo entre muchos de un problema natural en este modelo, es hacer una prueba de hipótesis de que $\tau_{ij} = 0 \forall i, j$, es decir, que no hay interacción entre ambos tratamientos (con este modelo).

Es claro que del mismo modo es posible construir modelos de clasificación múltiple de orden mayor que 2.

5.1 Estimación 1

Introducimos la siguiente notación, en el entendido que están bien definidos los valores esperados que aquí figuran. Si

$$Z = \begin{pmatrix} Z_1 \\ \vdots \\ Z_m \end{pmatrix}$$

es un vector aleatorio en $\mathbb{R}^m$. Definimos:

$$\mathbb{E}(Z) = \begin{pmatrix} \mathbb{E}(Z_1) \\ \vdots \\ \mathbb{E}(Z_m) \end{pmatrix}.$$

Si en lugar de un vector, tenemos una matriz aleatoria, definimos la esperanza del mismo modo, como la matriz que tiene por elementos las esperanzas, respectivamente. Su varianza se define por

$$\mathbb{V}(Z) = \mathbb{E}\Big(\big(Z - \mathbb{E}(Z)\big)\big(Z - \mathbb{E}(Z)\big)^T\Big),$$

de modo que el elemento que ocupa el lugar (i,j) de la matriz $\mathbb{V}(Z)$ es

$$\mathbb{E}\Big(\big(Z_i - \mathbb{E}(Z_i)\big)\big(Z_j - \mathbb{E}(Z_j)\big)\Big) = \mathrm{Cov}(Z_i, Z_j).$$

El lector verificará fácilmente que si c es un vector fijo de m coordenadas, entonces:

$$\mathbb{E}(c^T Z) = c^T \mathbb{E}(Z), \quad \mathbb{V}(c^T Z) = c^T \mathbb{V}(Z) c.$$

5.1.1 Estimación lineal insesgada de mínima varianza [ELIVM]

En lo que sigue, agregamos al modelo (5.1) las hipótesis siguientes sobre los errores:

$$\mathbb{E}(\epsilon) = 0, \quad \mathbb{V}(\epsilon) = \sigma^2 I_n,$$

donde σ^2 es una constante positiva e I_n es la matriz identidad $n \times n$. Esto significa que $\mathbb{E}(\epsilon_i) = 0$ para todo i y que $\mathrm{Cov}(\epsilon_i, \epsilon_j) = 0$ para todo $i \neq j$.

Supondremos además que $p < n$ y que la matriz A tiene rango máximo, es decir p. Esto implica que la matriz $A^T A$, de $p \times p$, es invertible.

Sea $\lambda \in \mathbb{R}^p$ un vector fijo. Bajo esas hipótesis nos proponemos estimar una combinación lineal de los parámetros:

$$\lambda^T \beta = \sum_{j=1}^{p} \lambda_j \beta_j.$$

mediante un estimador que es una combinación lineal de las observaciones:

$$c^T Y = \sum_{i=1}^{n} c_i Y_i.$$

El problema consiste en hallar $c \in \mathbb{R}^n$ de modo que:

1. $\mathbb{E}(c^T Y) = \lambda^T \beta$ para todo $\beta \in \mathbb{R}^p$ (estimación insesgada);
2. $\mathbb{V}(c^T Y)$ sea mínima.

La condición de sesgo nulo se escribe:

$$\mathbb{E}(c^T Y) = c^T \mathbb{E}(Y) = c^T A \beta = \lambda^T \beta \quad \text{para todo} \quad \beta \in \mathbb{R}^p,$$

es decir $(A^T c - \lambda)^T \beta = 0$ para todo β. Como esto vale para todo β, $A^T c - \lambda = 0$.

Por otro lado:

$$\mathbb{V}(c^T Y) = c^T \mathbb{V}(Y) c = c^T \mathbb{V}(\epsilon) c = c^T \sigma^2 I_n c = \sigma^2 \|c\|^2,$$

(donde $\|.\|$ denota la norma euclideana).

En resumen, nuestro problema de ELIVM se reduce a hallar el punto $c_0 \in \mathbb{R}^n$ del hiperplano S - cuya ecuación es $A^T c = \lambda$ - que está a mínima distancia del origen de coordenadas, ver Figura 5.1. Este es un problema sencillo de álgebra lineal, con solución única, ya que (teorema de Pitágoras) se trata de hallar la proyección ortogonal del origen sobre S. La solución es:

$$c_0 = A(A^T A)^{-1} \lambda.$$

En efecto, es inmediato que $c_0 \in S$. Además, c_0 es perpendicular al hiperplano, o lo que es lo mismo, al subespacio $N(A^T) = \{x \in \mathbb{R}^n : A^T x = 0\}$ paralelo a S. En efecto, si $x \in N(A^T)$, entonces

$$c_0^T x = \lambda^T (A^T A)^{-1} A^T x = 0,$$

es decir que $c_0 \perp x$ para todo $x \in N(A^T)$.

Por lo tanto, el estimador buscado es:

$$\hat{\lambda}^T \beta = c_0^T Y = \lambda^T (A^T A)^{-1} A^T Y.$$

En particular, como se indicó más arriba, esto permite estimar las coordenadas de β o el propio vector β mediante:

$$\hat{\beta} = (A^T A)^{-1} A^T Y.$$

Para referencia futura, resumamos que

$$\mathbb{E}(\hat{\beta}) = \beta, \qquad \mathbb{V}(\hat{\beta}) = \sigma^2 (A^T A)^{-1}.$$

La primera igualdad es una condición que hemos utilizado para el cálculo de β. La segunda resulta de, como

$$\hat{\beta} - \beta = (A^T A)^{-1} A^T Y - \beta = (A^T A)^{-1} A^T (A\beta + \epsilon) - \beta = (A^T A)^{-1} A^T \epsilon,$$

$$\mathbb{V}(\hat{\beta}) = \mathbb{E}\Big[\big(\hat{\beta} - \beta\big)\big(\hat{\beta} - \beta\big)^T\Big] = \mathbb{E}\Big[\big((A^T A)^{-1} A^T \epsilon\big)\big((A^T A)^{-1} A^T \epsilon\big)^T\Big] = \mathbb{E}\Big[\big((A^T A)^{-1} A^T \epsilon\big)\big(\epsilon^T A (A^T A)^{-1}\big)\Big]$$

$$= (A^T A)^{-1} A^T \mathbb{V}(\epsilon) A (A^T A)^{-1} = \sigma^2 (A^T A)^{-1},$$

ya que $\mathbb{V}(\epsilon) = \sigma^2 I_n$.

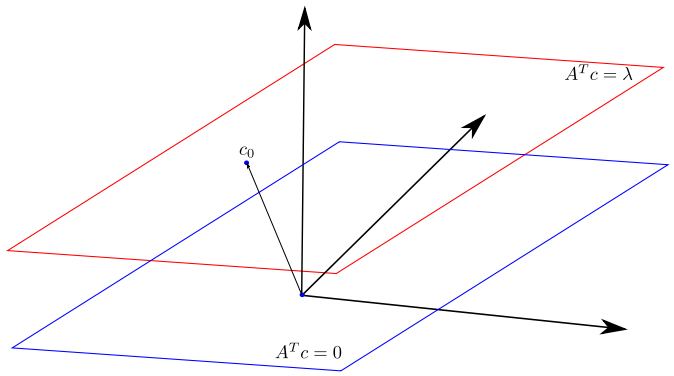

Figura 5.1: En azul el subespacio $N(A^T) = \{c : A^T c = 0\}$ y en rojo el espacio afin $S = \{c : A^T c = \lambda\}$. c_0 es el vector del espacio afin cuya norma es mas chica, es decir, es la proyección ortogonal del vector nulo sobre este espacio.

5.1.2 Conexión con mínimos cuadrados

Consideramos el modelo lineal (5.1), con la hipótesis de que la matriz A es de rango máximo (igual a p). El llamado método de los mínimos cuadrados, consiste en estimar β por el valor $\hat{\beta}$ que hace mínimo el valor de

$$\|\epsilon\|^2 = \sum_{i=1}^{n} \epsilon_i^2, \qquad \text{es decir} \qquad \|Y - A\hat{\beta}\|^2 = \min_{\beta \in \mathbb{R}^p} \|Y - A\beta\|^2.$$

Veamos que la solución a este problema está dada por el mismo $\hat{\beta}$ que encontramos en el párrafo anterior. Ver Figura 5.2

Se verifica sin dificultad que el subespacio

$$R(A) = A(\mathbb{R}^p) = \{x \in \mathbb{R}^n : x = A\beta \quad \text{para algún} \quad \beta \in \mathbb{R}^p\}$$

tiene dimensión p y que su complemento ortogonal es $N(A^T)$ (ver más arriba).

La solución (Pitágoras) tiene que verificar que $Y - A\hat{\beta}$ sea perpendicular al subespacio $R(A)$. Por lo tanto, $Y - A\hat{\beta} \in N(A^T)$, lo que implica que

$$A^T(Y - A\hat{\beta}) = 0 \Rightarrow (A^T A)\hat{\beta} = A^T Y$$

y volvemos a encontrar la solución (5).

5.1.3 Descomposición ortogonal del error. Estimación insesgada de la varianza

Con la misma notación e hipótesis de los dos párrafos previos, tenemos:

$$\epsilon = Y - A\beta = (Y - A\hat{\beta}) + (A\hat{\beta} - A\beta)$$

y sabemos que $Y - A\hat{\beta} \in N(A^T)$, $A\hat{\beta} - A\beta \in R(A)$ siendo ambos sumandos ortogonales, ver Figura 5.2.

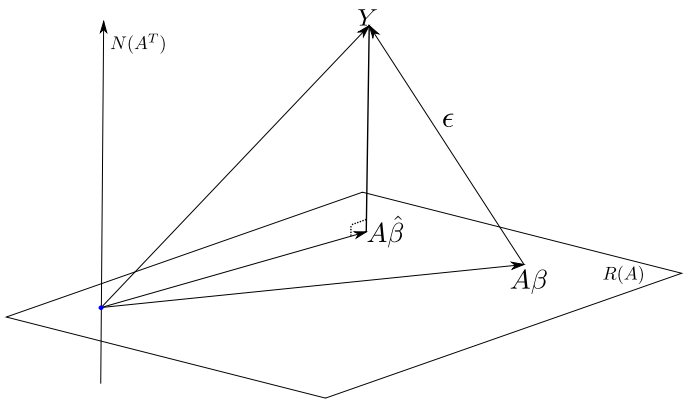

Figura 5.2: En azul el origen de coordenadas. La solución $\hat{\beta}$ se encuentra proyectando Y ortogonalmente en $S = R(A)$.

Consideremos una base ortonormal de $\mathbb{R}^n$, $\{v_1, \ldots, v_p, v_{p+1}, \ldots, v_n\}$ tal que $v_1, \ldots, v_p \in R(A)$; $v_{p+1}, \ldots, v_n \in N(A^T)$ y escribimos el vector de errores ϵ en esta nueva base:

$$\epsilon = \sum_{i=1}^{p} \tilde{\epsilon}_i v_i + \sum_{i=p+1}^{n} \tilde{\epsilon}_i v_i$$

y como la descomposición de un vector como suma de dos pertenecientes a subespacios ortogonales es única, se deduce que:

$$Y - A\hat{\beta} = \sum_{i=p+1}^{n} \tilde{\epsilon}_i v_i, \quad A\hat{\beta} - A\beta = \sum_{i=1}^{p} \tilde{\epsilon}_i v_i.$$

Por otra parte, el vector $\tilde{\epsilon}$ de las coordenadas de ϵ en la nueva base, se obtiene mediante la transformación del cambio de base:

$$\tilde{\epsilon} = U\epsilon$$

donde la matriz U es ortogonal, es decir que $UU^T = I_n$. Por lo tanto:

$$\mathbb{E}(\tilde{\epsilon}) = U\mathbb{E}(\epsilon) = 0, \quad \mathbb{V}(\tilde{\epsilon}) = U\mathbb{V}(\epsilon)U^T = U\sigma^2 I_n U^T = \sigma^2 UU^T = \sigma^2 I_n.$$

Es decir, la matriz de varianza de $\tilde{\epsilon}$ es la misma que la de ϵ.

Además:

$$\|Y - A\hat{\beta}\|^2 = \left\| \sum_{i=p+1}^{n} \tilde{\epsilon}_i v_i \right\|^2 = \sum_{i=p+1}^{n} \tilde{\epsilon}_i^2,$$

y por lo tanto:

$$\mathbb{E}\left(\left\| Y - A\hat{\beta} \right\|^2 \right) = \sum_{i=p+1}^{n} \mathbb{E}\left(\tilde{\epsilon}_i^2\right) = (n-p)\sigma^2.$$

Esto implica que

$$s_n^2 = \frac{1}{n-p} \|Y - A\hat{\beta}\|^2$$

es un estimador insesgado de la varianza σ^2.

En las secciones que siguen veremos como quedan los estimadores en cada uno de los ejemplos de la introducción.

Ejemplo 1: magnitud desconocida

Se tiene, en este caso:

$$R(A) = \left\{ \lambda(1, 1, ..., 1)^T : \lambda \in \mathbb{R} \right\} \quad \text{(dimensión 1)}$$

y por lo tanto, la condición se reduce a:

$$Y - \hat{\mu}(1, 1, ..., 1)^T \perp (1, 1, ..., 1)^T \Rightarrow \sum_{i=1}^{n} Y_i - n\hat{\mu} = 0 \Rightarrow \hat{\mu} = \frac{1}{n} \sum_{i=1}^{n} Y_i = \bar{Y}.$$

Además:

$$s_n^2 = \frac{1}{n-1} \left\| Y - \hat{\mu}(1, 1, ..., 1)^T \right\|^2 = \frac{1}{n-1} \sum_{i=1}^{n} (Y_i - \bar{Y})^2.$$

Ejemplo 2. Ajuste de funciones

Para no escribir fórmulas complicadas aquí, nos limitamos al caso $k = 1$ (ajuste por una recta). De todos modos, si se trata de ajustes por polinomios de grado mayor u otras funciones, el procedimiento es enteramente similar. Se tiene:

$$A = \begin{pmatrix} 1 & t_1 \\ 1 & t_2 \\ \vdots & \vdots \\ 1 & t_n \end{pmatrix}$$

Suponemos además que $\sum_{i=1}^{n} t_i = 0$. Si esta condición no se verifica, el lector pensará cómo reducir el problema al caso en que sí se verifica. También suponemos que los t_i no son todos nulos: si lo fueran, la matriz A no sería de rango máximo (igual a 2).

Se tiene:

$$A^T A = \begin{pmatrix} n & 0 \\ 0 & \sum_{i=1}^{n} t_i^2 \end{pmatrix}$$

y el lector verificará fácilmente que:

$$\hat{a}_0 = \bar{X}, \quad \hat{a}_1 = \frac{\sum_{i=1}^n t_i X_i}{\sum_{i=1}^n t_i^2}, \quad s_n^2 = \frac{1}{n-2}\left[n\bar{X}^2 + \left(\frac{\sum_{i=1}^n t_i X_i}{\sum_{i=1}^n t_i^2}\right)^2 \sum_{i=1}^n t_i^2\right].$$

También se obtiene:

$$\mathbb{V}\begin{pmatrix}\hat{a}_0\\ \hat{a}_1\end{pmatrix} = \sigma^2(A^TA)^{-1} = \sigma^2\begin{pmatrix}\frac{1}{n} & 0\\ 0 & \frac{1}{\sum_{i=1}^n t_i^2}\end{pmatrix},$$

lo que implica que $\hat{a}_0, \hat{a}_1$ son no correlacionadas.

Ejemplo 4. Clasificación simple (análisis de varianza)

En el modelo de clasificación simple, el lector verificará que:

$$A^TA = \begin{pmatrix} n_1 & 0 & \cdots & 0\\ 0 & n_2 & \cdots & 0\\ \vdots & \vdots & \ddots & \vdots\\ 0 & 0 & \cdots & n_q\end{pmatrix}$$

y

$$\hat{\mu}_i = \bar{Y}_i \quad (i = 1, ..., q)$$

donde $\bar{Y}_i$ es el promedio de las observaciones de la magnitud μ_i. Para la estimación insesgada de la varianza obtenemos:

$$s_n^2 = \frac{1}{n-q}\sum_{i=1}^q\sum_{j=1}^{n_i}(Y_{ij} - \bar{Y}_i)^2.$$

5.2 Modelos lineales con errores normales. Distribución de los estimadores

En esta sección consideramos el modelo lineal

$$Y = A\beta + \epsilon$$

con la hipótesis de que el vector de errores tiene coordenadas ϵ_i normales centradas independientes, $\mathbb{V}(\epsilon_i) = \sigma^2$. Suponemos que A tiene rango p.

Aplicando el mismo método que la sección previa, podemos estimar los parámetros β, σ^2. Para β obtenemos nuevamente el estimador de mínimos cuadrados,

$$\hat{\beta} = (A^TA)^{-1}A^TY$$

y para σ^2:

$$\hat{\sigma}^2 = \frac{1}{n}\|Y - A\hat{\beta}\|^2.$$

Nótese que $\hat{\sigma}^2$ no es insesgado, ya que

$$\mathbb{E}(\hat{\sigma}^2) = \frac{n-p}{n}\sigma^2.$$

A continuación, calculamos la distribución de los estimadores básicos. Con ellas, estamos en condiciones de obtener intervalos de confianza (en dimensión 1) o, en general, regiones de confianza (en cualquier dimensión) para los estimadores y también realizar algunas pruebas de hipótesis. En este breve texto no habremos de detallar la manera de hacer esto, que es común en la literatura sobre modelos lineales y nos limitaremos - en la sección siguiente - a presentar la prueba F.

Comenzamos eligiendo una base ortonormal de $\mathbb{R}^n$, $\{v_1, ..., v_n\}$, de modo que

$$v_1, \ldots, v_p \in R(A), \quad v_{p+1}, \ldots, v_n \in N(A).$$

Se tiene:

$$\epsilon = Y - A\beta = (A\hat{\beta} - A\beta) + (Y - A\hat{\beta}) = \sum_{i=1}^p \tilde{\epsilon}_i v_i + \sum_{i=p+1}^n \tilde{\epsilon}_i v_i$$

y como la descomposición de un vector como suma de vectores en subespacios ortogonales es única, se deduce que:

$$\frac{1}{\sigma}(A\hat{\beta} - A\beta) = \sum_{i=1}^p \frac{1}{\sigma}\tilde{\epsilon}_i v_i, \qquad \frac{1}{\sigma}(X - A\hat{\beta}) = \sum_{i=p+1}^n \frac{1}{\sigma}\tilde{\epsilon}_i v_i.$$

Como $\frac{\epsilon}{\sigma}$ tiene distribución normal estándar en $\mathbb{R}^n$, el vector de sus coordenadas en cualquier base ortonormal también es normal estándar. De modo que $\tilde{\epsilon}_1/\sigma, \ldots, \tilde{\epsilon}_n/\sigma$ son independientes, normales estándar en $\mathbb{R}^1$. De esto y de las igualdades anteriores se deduce que:

1.
$$\frac{1}{\sigma}A(\hat{\beta}-\beta)$$
tiene distribución normal estándar de dimensión p.

Por lo tanto
$$\hat{\beta} = \beta + (A^TA)^{-1}A^TA(\hat{\beta}-\beta),$$
también tiene distribución normal con media β y varianza $\sigma^2(A^TA)^{-1}$.

2.
$$(n-p)\frac{s_n^2}{\sigma^2} = \frac{1}{\sigma^2}\|X - A\hat{\beta}\|^2 = \sum_{i=p+1}^{n}\left(\frac{\tilde{\epsilon}_i}{\sigma}\right)^2.$$

Por lo tanto, $(n-p)s_n^2/\sigma^2$ tiene distribución χ^2_{n-p}.

Una observación lateral es que s_n^2 es un estimador consistente de σ^2 cuando $n \to \infty$, en virtud de la ley de los grandes números. En esto no interviene la normalidad de los errores, sino solamente su independencia.

3. $\hat{\beta}$ (o alternativamente $A\hat{\beta}$ y s_n^2) son variables aleatorias independientes, ya que la primera es función de $\tilde{\epsilon}_1, ..., \tilde{\epsilon}_p$ y la segunda de $\tilde{\epsilon}_{p+1}, ..., \tilde{\epsilon}_n$.

4. Sea $\lambda \in \mathbb{R}^p$ y consideremos la combinación lineal $\lambda^T\beta$ de las coordenadas del vector de parámetros β. Entonces,
$$\zeta = \frac{\lambda^T\hat{\beta} - \lambda^T\beta}{s_n\left[\lambda^T(A^TA)^{-1}\lambda\right]^{1/2}}$$
tiene distribución t_{n-p} cualquiera sea λ. En efecto, ζ se escribe bajo la forma:
$$\zeta = \frac{\left(\lambda^T\hat{\beta} - \lambda^T\beta\right)/\left[\sigma\left[\lambda^T(A^TA)^{-1}\lambda\right]^{1/2}\right]}{s_n/\sigma}$$
El numerador es normal estándar, el denominador es
$$\sqrt{\frac{1}{n-p}\sum_{i=p+1}^{n}\left(\frac{\tilde{\epsilon}_i}{\sigma}\right)^2}$$
y ambos son independientes. Por lo tanto, la distribución de ζ es t_{n-p}.

5.3 La prueba F

En esta sección consideramos un modelo lineal de la forma
$$Y = A\beta + \epsilon$$
con las hipótesis previas, donde además el vector de errores ϵ tiene distribución normal en $\mathbb{R}^n$, centrada y con varianza igual a $\sigma^2 I_n$. Es decir que las coordenadas del error son centradas, independientes con varianza común σ^2.

Sea R_0 un subespacio propio del espacio de parámetros $\mathbb{R}^p$. Exponemos una prueba para la hipótesis
$$H_0 : \beta \in R_0$$
versus la alternativa de que $\beta \notin R_0$.

Algunos ejemplos:

- En el Ejemplo 1 de la Introducción, la hipótesis de que $\mu = 0$ (o de que $\mu = \mu_0$ teniendo la precaución de reemplazar las observaciones por $X_i - \mu$).
- En el Ejemplo 2, si se ajusta por un polinomio, probar la hipótesis de que algunos de sus coeficientes son nulos.
- En el Ejemplo 4, probar la hipótesis de que $\mu_1 = \mu_2 = \cdots = \mu_q$, es decir que no hay diferencia significativa entre los tratamientos, que corresponde a tomar
$$R_0 = \left\{(\mu_1, \mu_2, ..., \mu_q) : \mu_1 = \mu_2 = \cdots = \mu_q\right\}$$
que es un subespacio de dimensión 1.
- También uno podría tener interés en probar otras hipótesis lineales, como por ejemplo, $\mu_1 + \mu_2 = \mu_3$.

Volvamos al problema inicial. Sea $S_0 = \{A\beta : \beta \in R_0\}$ la imagen de R_0 por A.

- $p = \dim(R(A))$.
- S_0 es un subespacio de $R(A) \subset \mathbb{R}^n$, $p_0 = \dim(S_0)$.
- Es obvio que H_0 es equivalente a $A\beta \in S_0$ y que el problema tiene interés si $p_0 < p$.

Vayamos ahora a la construcción de la región crítica de la prueba F.

Sea $A\beta \in R(A)$, donde β es el verdadero valor (desconocido) del parámetro y sea $A\beta_0$ su proyección ortogonal sobre S_0 (ver figura).

Denotamos con c el vector $c = A\beta_0 - A\beta$. La norma de c es la distancia de $A\beta$ a S_0. Que se cumpla la hipótesis nula quiere decir que $c = 0$.

Sea $A\hat{\beta}_0$ la proyección ortogonal del vector de observaciones X sobre el subespacio S_0. Es decir, $A\hat{\beta}_0$ se obtiene del mismo modo que $A\hat{\beta}$ - mínimos cuadrados - sólo que en un modelo lineal en el que los parámetros verifican la hipótesis nula.

En el modelo original, tenemos la estimación insesgada de la varianza:

$$s^2 = \frac{1}{n-p}\|Y - A\hat{\beta}\|^2$$

y en el modelo bajo H_0:

$$s_0^2 = \frac{1}{n-p_0}\|Y - A\hat{\beta}_0\|^2.$$

El lector verá que $A\hat{\beta} - A\hat{\beta}_0 \perp S_0$ (Pitágoras). Consideramos la siguiente descomposición del error como suma de 3 términos:

$$\epsilon = Y - A\beta = (Y - A\hat{\beta}) + (A\hat{\beta}_0 - A\beta_0) + (A\hat{\beta} - A\hat{\beta}_0 + c).$$

Sabemos que el primer término pertenece al subespacio $N(A^T) = \{x \in \mathbb{R}^n : A^T x = 0\}$ (recordar quién es $\hat{\beta}$). Ello implica que es ortogonal a los otros dos sumandos. Pero además, éstos son ortogonales entre sí, porque $A\hat{\beta}_0 - A\beta_0 \in S_0$ y $A\hat{\beta} - A\hat{\beta}_0 \perp S_0$, $c \perp S_0$.

Tomamos ahora una base ortonormal de $\mathbb{R}^n$:

$$\{v_1, ..., v_{p_0}, v_{p_0+1}, ..., v_p, v_{p+1}, ..., v_n\}$$

de tal modo que

$$v_1, ..., v_{p_0} \in S_0; v_{p_0+1}, ..., v_p \in T_0; v_{p+1}, ..., v_n \in N(A^T),$$

en que hemos denotado con T_0 el complemento ortogonal de S_0 en $R(A)$.

Se tiene:

$$\frac{1}{\sigma}\epsilon = \sum_{i=1}^{n} \frac{1}{\sigma}\tilde{\epsilon}_i v_i.$$

Ahora, tenemos en cuenta que el vector aleatorio $\frac{1}{\sigma}\epsilon$ tiene distribución normal estándar. Entonces, dada la invariancia de esta distribución bajo transformaciones ortogonales, si lo expresamos como combinación lineal de cualquier base ortonormal, el vector de coeficientes tiene la misma distribución. Es decir, que:

$$\frac{1}{\sigma}\tilde{\epsilon} = \begin{pmatrix} \frac{1}{\sigma}\tilde{\epsilon}_1 \\ \frac{1}{\sigma}\tilde{\epsilon}_2 \\ \vdots \\ \frac{1}{\sigma}\tilde{\epsilon}_n \end{pmatrix}$$

es normal estándar. Dado que la descomposición de un vector como suma de sus proyecciones sobre tres subespacios ortogonales dos a dos es única, se deduce que:

$$\frac{1}{\sigma}(A\hat{\beta} - A\hat{\beta}_0 + c) = \sum_{i=p_0+1}^{p} \frac{1}{\sigma}\tilde{\epsilon}_i v_i, \qquad \frac{1}{\sigma}(X - A\hat{\beta}) = \sum_{i=p+1}^{n} \frac{1}{\sigma}\tilde{\epsilon}_i v_i.$$

De aquí, obtenemos las siguientes conclusiones:

1. $$\frac{1}{\sigma^2}\|A\hat{\beta} - A\hat{\beta}_0\|^2$$

   tiene distribución $\chi^2_{p-p_0, \|\frac{c}{\sigma}\|}$ (es decir, χ^2 excéntrica con $p - p_0$ grados de libertad y excentricidad $\|\frac{c}{\sigma}\|$).

2.
$$\frac{1}{\sigma^2}\|Y - A\hat{\beta}\|^2$$
tiene distribución χ^2_{n-p}.

3.
$$\frac{1}{\sigma^2}\|A\hat{\beta} - A\hat{\beta}_0\|^2 \qquad \text{y} \qquad \frac{1}{\sigma^2}\|Y - A\hat{\beta}\|^2$$
son variables aleatorias independientes, ya que la primera es función de $\tilde{\epsilon}_{p_0+1}, ..., \tilde{\epsilon}_p$ y la segunda de $\tilde{\epsilon}_{p+1}, ..., \tilde{\epsilon}_n$.

Denotamos
$$\tilde{s}^2 = \frac{1}{p-p_0}\|A\hat{\beta} - A\hat{\beta}_0\|^2 = \frac{1}{p-p_0}\left[\|Y - A\hat{\beta}_0\|^2 - \|Y - A\hat{\beta}\|^2\right].$$

Si se cumple la hipótesis nula, $c = 0$ y $\frac{1}{\sigma^2}\|A\hat{\beta} - A\hat{\beta}_0\|^2$ tiene distribución $\chi^2_{p-p_0}$. Entonces,
$$\frac{\tilde{s}^2}{\sigma^2} = \frac{1}{p-p_0}\|A\hat{\beta} - A\hat{\beta}_0\|^2$$
tiene distribución $\chi^2_{p-p_0}$ y
$$F = \frac{\tilde{s}^2/(p-p_0)}{s^2/(n-p)}$$
tiene distribución $F_{p-p_0,n-p}$.

Finalmente, la región crítica para la prueba H_0 de significación α está dada por
$$F \geq F_{p-p_0,n-p}(\alpha),$$
donde $F_{p-p_0,n-p}(\alpha)$ es el valor de la tabla F-Fisher-Snedecor para $p-p_0, n-p$ grados de libertad y una cola α.

5.4 Regresión Logística

La regresión logística es un modelo de aprendizaje supervisado utilizado para la **clasificación binaria** (los datos de que disponemos son pares (X_i, Y_i) con $X_i \in \mathbb{R}^d$ y $Y_i \in \{0, 1\}$. Se utiliza para predecir la probabilidad de que una observación pertenezca a una de dos clases posibles. Utiliza la **función sigmoide**, $\sigma : \mathbb{R} \to [0, 1]$,

$$\sigma(z) = \frac{1}{1+e^{-z}} \tag{5.4}$$

Aquí z es una combinación lineal de las características de entrada:

$$z = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_d x_d \tag{5.5}$$

La salida de la función sigmoide se interpreta como la probabilidad de que la observación pertenezca a la clase 1. Si la probabilidad es mayor a 0.5, clasificamos la observación como 1; de lo contrario, como 0. Es decir, si $X_i = (X_{1,i}, \ldots, X_{d,i})$, es un imput en $\mathbb{R}^d$, tomamos el vector $\tilde{X}_i = (1, X_i)$ y $\beta = (\beta_0, \ldots, \beta_d)$

$$\mathbb{P}(Y_i = 1|X_i) = \sigma(\beta^T \tilde{X}_i) = \frac{1}{1+e^{-\beta^T \tilde{X}_i}} = \frac{e^{\beta^T \tilde{X}_i}}{1+e^{\beta^T \tilde{X}_i}}$$

por lo tanto

$$\mathbb{P}(Y_i = 0|X_i) = 1 - \sigma(\beta^T \tilde{X}_i) = 1 - \frac{1}{1+e^{-\beta^T \tilde{X}_i}} = \frac{1}{1+e^{\beta^T \tilde{X}_i}}$$

El vector de parámetros w lo vamos a estimar por el método de máxima verosimilitud y una vez que tenemos el estimador $\hat{\beta}$, clasificamos un nuevo dato X como 1 si $\sigma(\hat{\beta}^T X) > 1/2$. Es inmediato verificar que esto pasa si y sólo si $\mathbb{P}(Y = 1|X) > \mathbb{P}(Y = 0|X)$, si y solo si $\hat{\beta}^T \tilde{X} > 0$, por lo tanto es una regla de clasificación lineal.

Se generaliza a $k \geq 2$ clases tomando la **función softmax**:

$$\mathbb{P}(Y = j|X) = \frac{\exp(t_j)}{\sum_{j=1}^k \exp(t_j)} \qquad j = 1, \ldots, k \tag{5.6}$$

donde $t_j = f(\beta_j, X)$ es una función de $(1, X)$, y de parámetros $\beta_j \in \mathbb{R}^{d+1}$ que hay que encontrar. En el caso de la regresión logística esta función es simplemente un producto interno. Se deja como ejercicio verificar que para $k = 2$ la función softmax (con $t_j = f(\beta_j, X) = \beta^t \tilde{X}$) es la función logística con $\beta = \beta_2 - \beta_1$.

Ejercicio 5.1. Verificar que

$$f(x;\mu,s) = \frac{e^{(x-\mu)/s}}{s(1+e^{-(x-\mu)/s})^2}$$

define una función de densidad cuya función de distribución es

$$F(x;\mu,s) = \frac{1}{1+e^{-(x-\mu)/s}}$$

y cuya función cuantil es

$$Q(p;\mu,s) = \mu + s\log(p(1-p))$$

En particular esto implica que (5.5) se puede escribir como $Q(\sigma(z);0,1) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_n x_n$, es decir sigue un modelo lineal.

5.4.1 Ajuste de β por máxima verosimilitud

Para un conjunto de datos $(X_1, Y_1), \ldots, (X_n, Y_n) \in \mathbb{R}^d \times \{0,1\}$ la función de verosimilitud es

$$V(\beta) = \prod_{i=1}^{n} \mathbb{P}(Y = Y_i | X_i)$$

Dado que Y_i puede ser 0 o 1,

$$V(\beta) = \prod_{i=1}^{n} \left[\sigma(\beta^T \tilde{X}_i)\right]^{Y_i} \left[1 - \sigma(\beta^T \tilde{X}_i)\right]^{1-Y_i}$$

Y la log-verosimilitud:

$$\log V(\beta) = \sum_{i=1}^{n} \left[Y_i \log\left(\sigma(\beta^T \tilde{X}_i)\right) + (1-Y_i)\log\left(1 - \sigma(\beta^T \tilde{X}_i)\right)\right]$$

El objetivo es encontrar los parámetros β que maximicen la log-verosimilitud. Sin embargo, en la práctica, en lugar de maximizar la log-verosimilitud, minimizamos la función de pérdida, que es el negativo de la log-verosimilitud. Esto se debe a que las técnicas de optimización estándar están formuladas para la minimización de funciones.

Tomando el negativo de la log-verosimilitud, obtenemos la **función de pérdida de entropía cruzada**

$$J(\beta) = -\log V(\beta) = -\sum_{i=1}^{n} \left[Y_i \log\left(\sigma(\beta^T \tilde{X}_i)\right) + (1-Y_i)\log\left(1 - \sigma(\beta^T \tilde{X}_i)\right)\right] \tag{5.7}$$

En el caso de tener $k \geq 2$ clases, si denotamos $p_i = \frac{\exp(t_j)}{\sum_{j=1}^{k}\exp(t_j)}$, la entropía cruzada del dato (X_i, Y_i) es

$$-\sum_{j=1}^{k} \mathbb{I}_{\{Y_i=j\}} \log(p_i)$$

y por lo tanto la función que queremos minimizar es

$$J(\beta) = -\sum_{i=1}^{n}\sum_{j=1}^{k} \mathbb{I}_{\{Y_i=j\}} \log(p_i),$$

Esto se sigue igual que antes de plantear la log-verosimilitud, se deja como ejercicio verificarlo.

Para encontrar β se procede por el método de descenso por gradiente que veremos más adelante.

6 Teoría de Vapnik-Chervonenkis

En este capítulo vamos a retomar lo que vimos en la subsección 3.3.4, es decir, estudiar las reglas que se eligen en una determinada familia de funciones. Recordemos la notación que introdujimos, vamos a denotar $\mathcal{C} = \{g : \mathcal{X} \to \{0,1\}\}$, $L(g) = \mathbb{P}(g(X) \neq Y)$ y se verifica que $L(g) \geq L^*$ para todo g. Por lo tanto $\inf_{g\in\mathcal{C}} L(g) \geq L^*$. La diferencia $\inf_{g\in\mathcal{C}} L(g) - L^*$ va a ser más chica cuanto más grande sea la clase de funciones, no obstante, una vez que elegimos $\mathcal{C}$, no se puede achicar este error.

Cuando tenemos una muestra $D_n = \{(X_1, Y_1), \ldots, (X_n, Y_n)\}$ un criterio para elegir una regla cuyo error sea próximo a $\inf_{g\in\mathcal{C}}$ es elegir una regla en $\mathcal{C}$ que minimice el error empírico

$$\hat{L}_n(g) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{I}_{\{g(X_i)\neq Y_i\}}.$$

Vamos a denotar

$$g_n^* = \arg\min_{g\in\mathcal{C}} \hat{L}_n(g).$$

Observación 6.1. *La regla g^* no tiene por qué ser única, pero denotaremos como g^* cualquier regla que minimice el riesgo empírico $\hat{L}_n$. Como $\hat{L}_n(g)$ toma una cantidad finita de valores, siempre vamos poder encontrar una.*

Es claro que

$$L(g_n^*) = \mathbb{P}\big(g_n^*(X) \neq Y | D_n\big) \geq \inf_{g\in\mathcal{C}} L(g).$$

En este caso, si la clase $\mathcal{C}$ es muy grande, puede pasar que la diferencia $L(g_n^*) - \inf_{g\in\mathcal{C}} L(g)$ sea grande. No obstante, la diferencia anterior se puede controlar, y hacer chica, si n es suficientemente grande. Usualmente el error de aproximación $\inf_{g\in\mathcal{C}} L(g) - L^*$ es más grande que el error de estimación $L(g_n^*) - \inf_{g\in\mathcal{C}} L(g)$.

Supongamos que $\mathcal{C}$ tiene una cantidad finita de funciones, que denotamos $|\mathcal{C}|$. Si $\min_{g\in\mathcal{C}} L(g) = 0$ veamos que con probabilidad 1, $\hat{L}_n(g_n^*) = 0$. Sea $g' \in \mathcal{C}$ tal que $\mathbb{P}(g'(X) \neq Y) = 0$ entonces, condicionado a D_n, $\hat{L}_n(g_n^*) \leq \hat{L}_n(g')$, por lo tanto

$$0 \leq \mathbb{P}(\hat{L}_n(g_n^*) > 0) = \mathbb{E}(\mathbb{P}(\hat{L}_n(g_n^*) > 0|D_n)) \leq \mathbb{E}(\mathbb{P}(\hat{L}_n(g') > 0|D_n)) = \mathbb{E}(\mathbb{P}(\exists i : g'(X_i) \neq Y_i|D_n)) \leq n\mathbb{P}(g'(X) \neq Y) = 0.$$

Para este caso particular tenemos el siguiente resultado de Vapnik y Chervonenkis de 1974.

Teorema 6.2. *Supongamos que $|\mathcal{C}| < \infty$ y $\min_{g\in\mathcal{C}} L(g) = 0$. Para todo $n > 0$ y $\epsilon > 0$,*

$$\mathbb{P}\big(L(g_n^*) > \epsilon\big) \leq |\mathcal{C}| \exp(-n\epsilon), \tag{6.1}$$

y

$$\mathbb{E}\big(L(g_n^*)\big) \leq \frac{1 + \log|\mathcal{C}|}{n}. \tag{6.2}$$

Demostración. Como $\hat{L}_n(g_n^*) = 0$ con probabilidad 1, tenemos que

$$\mathbb{P}\big(L(g_n^*) > \epsilon\big) \leq \mathbb{P}\Big(\max_{g\in\mathcal{C}:\hat{L}_n(g)=0} L(g) > \epsilon\Big) = \mathbb{E}\Big(\mathbb{I}_{\left\{\max_{g\in\mathcal{C}:\hat{L}_n(g)=0} L(g)>\epsilon\right\}}\Big).$$

De

$$\mathbb{I}_{\left\{\max_{g\in\mathcal{C}:\hat{L}_n(g)=0} L(g)>\epsilon\right\}} = \max_{g\in\mathcal{C}} \mathbb{I}_{\{\hat{L}_n(g)=0\}}\mathbb{I}_{\{L(g)>\epsilon\}} = \max_{g\in\mathcal{C}:L(g)>\epsilon} \mathbb{I}_{\{\hat{L}_n(g)=0\}}.$$

se sigue, tomando esperanza,

$$\mathbb{E}\Big(\max_{g\in\mathcal{C}:L(g)>\epsilon} \mathbb{I}_{\{\hat{L}_n(g)=0\}}\Big) \leq \sum_{g\in\mathcal{C}:L(g)>\epsilon} \mathbb{E}(\mathbb{I}_{\{\hat{L}_n(g)=0\}}) = \sum_{g\in\mathcal{C}:L(g)>\epsilon} \mathbb{P}(\hat{L}_n(g) = 0)$$

De $L(g) > \epsilon$, y usando la independencia de los (X_i, Y_i)

$$\mathbb{P}\big(\hat{L}_n(g) = 0\big) = \mathbb{P}\big(\forall i = 1, \ldots, n : g(X_i) = Y_i\big) = \Big[\mathbb{P}\big(g(X) = Y\big)\Big]^n = \Big[1 - L(g)\Big]^n \leq (1-\varepsilon)^n.$$

Finalmente, de $(1-x) \leq \exp(-x)$ se sigue (6.1). Para probar (6.2), para todo $u > 0$,

$$\begin{aligned}
\mathbb{E}\big(L(g_n^*)\big) &= \int_0^\infty \mathbb{P}\big(L(g_n^*) > t\big)dt \\
&\leq \int_0^u 1dt + \int_u^\infty \mathbb{P}\big(L(g_n^*) > t\big)dt \\
&= u + \int_u^\infty \mathbb{P}\big(L(g_n^*) > t\big)dt \\
&\leq u + |\mathcal{C}| \int_u^\infty \exp(-nt)dt \\
&= u + |\mathcal{C}| \exp(-nu)/n
\end{aligned}$$

Como u es arbitrario podemos elegir el que minimiza $u + |\mathcal{C}| \exp(-nu)/n$. Derivando esto se da en $u = \log(|\mathcal{C}|)/n$ de donde se obtiene (6.2). □

La hipótesis $\text{mín}_{g\in\mathcal{C}} L(g) = 0$ es muy restrictiva. Recordemos la cota (3.16)

$$L(g_n^*) - \inf_{g\in\mathcal{C}} L(g) \leq 2 \sup_{g\in\mathcal{C}} \big|\hat{L}_n(g) - L(g)\big|. \tag{6.3}$$

Acotar $\sup_{g\in\mathcal{C}} \big|\hat{L}_n(g) - L(g)\big|$ conduce a acotar uniformemente en una familia de conjuntos, la diferencia entre la medida empírica y la teórica de estos conjuntos. Esto se sigue de que, si ν es la medida de probabilidad en $\mathbb{R}^d \times \{0,1\}$ del par (X,Y), es decir $\nu(A) = \mathbb{P}\big((X,Y) \in A\big)$ con $A \subset \mathbb{R}^d \times \{0,1\}$, y ν_n es la medida empírica basada en una muestra D_n de pares iid de (X,Y), es decir

$$\nu_n(A) = \frac{1}{n} \sum_{i=1}^n \mathbb{I}_{\{(X_i,Y_i)\in A\}}.$$

Entonces

$$L(g) = \nu\Big(\big\{(x,y) : g(x) \neq y\big\}\Big).$$

Es decir, $L(g)$ es la ν-medida del conjunto

$$\big\{\{x : g(x) = 1\} \times \{0\}\big\} \bigcup \big\{\{x : g(x) = 0\} \times \{1\}\big\}.$$

De forma análoga

$$\hat{L}_n(g) = \nu_n\Big(\big\{(x,y) : g(x) \neq y\big\}\Big).$$

Por lo tanto

$$\sup_{g\in\mathcal{C}} \big|\hat{L}_n(g) - L(g)\big| = \sup_{A\in\mathcal{A}} \big|\nu_n(A) - \nu(A)\big|,$$

donde $\mathcal{A}$ es la familia de conjuntos

$$\big\{\{x : g(x) = 1\} \times \{0\}\big\} \bigcup \big\{\{x : g(x) = 0\} \times \{1\}\big\}, \quad g \in \mathcal{C}.$$

Es decir a cada $g \in \mathcal{C}$ le asociamos $A \in \mathcal{A}$, dado por

$$A = \big\{\{x : g(x) = 1\} \times \{0\}\big\} \bigcup \big\{\{x : g(x) = 0\} \times \{1\}\big\}.$$

Por la desigualdad de Hoeffding,

$$\mathbb{P}\big(|\nu_n(A) - \nu(A)| > \epsilon\big) \leq 2\exp(-2n\varepsilon^2).$$

Por lo tanto si $|\mathcal{A}| < \infty$,

$$\mathbb{P}\Big(\sup_{A\in\mathcal{A}} \big|\nu_n(A) - \nu(A)\big| > \epsilon\Big) \leq 2|\mathcal{A}| \exp(-2n\varepsilon^2).$$

La Teória de Vapnik-Chervonenkis permite acotar la probabilidad anterior para el caso en que $|\mathcal{A}|$ no necesariamente es finito. Para eso vamos a empezar mostrando las ideas fundamentales en la prueba del Teorema de Glivenko-Cantelli, en la próxima sección.

6.1 Glivenko-Cantelli

Teorema 6.3. *Sean $Z_1, Z_2, \ldots, Z_n$ variables aleatorias independientes idénticamente distribuidas, a valores reales, con función de distribución $F(z) = \mathbb{P}(Z_1 \leq z)$. Denotemos la distribución empírica como:*

$$F_n(z) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{I}_{\{Z_i \leq z\}},$$

entonces

$$\mathbb{P}\left\{\sup_{z\in\mathbb{R}} \left|F(z) - F_n(z)\right| > \varepsilon\right\} \leq 8(n+1)e^{-n\varepsilon^2/32},$$

en particular, por el lema de Borel-Cantelli

$$\lim_{n\to+\infty} \sup_{z\in\mathbb{R}} \left|F(z) - F_n(z)\right| = 0 \quad \textit{con probabilidad 1.}$$

Demostración. Vamos a introducir algo de notación que usaremos a lo largo de esta sección.

$$\nu(A) = \mathbb{P}\{Z_1 \in A\} \qquad \text{y} \qquad \nu_n(A) = \frac{1}{n}\sum_{j=1}^{n} \mathbb{I}_{\{Z_j \in A\}} \quad \text{para todo conjunto medible } A \subset \mathbb{R}.$$

Denotemos $\mathcal{A}$ la clase de los conjuntos de la forma $(-\infty, z]$ con $z \in \mathbb{R}$. Con esta notación

$$\sup_{z\in\mathbb{R}} \left|F(z) - F_n(z)\right| = \sup_{A\in\mathcal{A}} \left|\nu_n(A) - \nu(A)\right|.$$

Vamos a demostrar el teorema en varias etapas, siguiendo las ideas de simetrización de Dudley (1978) y Pollard (1984). Asumiremos que $n\varepsilon^2 > 2$ en caso contrario la cota es trivial.

PASO 1. Simetrización respecto de un remuestreo de Z. Definimos las variables $Z_1', \ldots, Z_n' \in \mathbb{R}$ tal que $Z_1, \ldots, Z_n$, $Z_1', \ldots, Z_n'$ son independientes e idénticamente distribuidas. Denotemos como ν_n' la medida empírica correspondiente a la nueva muestra:

$$\nu_n'(A) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{I}_{\{Z_i' \in A\}}.$$

Vamos a probar primero que para $n\varepsilon^2 > 2$,

$$\mathbb{P}\left\{\sup_{A\in\mathcal{A}} \left|\nu_n(A) - \nu(A)\right| > \varepsilon\right\} \leq 2\mathbb{P}\left\{\sup_{A\in\mathcal{A}} \left|\nu_n(A) - \nu_n'(A)\right| > \frac{\varepsilon}{2}\right\}.$$

Para ver esto, sea $A^* \in \mathcal{A}$ para el cual $\left|\nu_n(A^*) - \nu(A^*)\right| > \varepsilon$ si tal conjunto existe. En caso contrario tomamos cualquier $A^* \in \mathcal{A}$ fijo. Entonces

$$\mathbb{P}\left\{\sup_{A\in\mathcal{A}} \left|\nu_n(A) - \nu_n'(A)\right| > \frac{\varepsilon}{2}\right\} \geq \ \mathbb{P}\left\{\left|\nu_n(A^*) - \nu_n'(A^*)\right| > \frac{\varepsilon}{2}\right\}.$$

Vamos a acotar por abajo esta probabilidad, para eso observemos que

$$\left|\nu_n(A^*) - \nu(A^*)\right| \leq \left|\nu_n(A^*) - \nu_n'(A^*)\right| + \left|\nu_n'(A^*) - \nu(A^*)\right|.$$

Por lo tanto

$$\left\{\left\{\left|\nu_n(A^*) - \nu(A^*)\right| > \varepsilon\right\} \cap \left\{\left|\nu_n'(A^*) - \nu(A^*)\right| < \frac{\varepsilon}{2}\right\}\right\} \subset \left\{\left|\nu_n(A^*) - \nu_n'(A^*)\right| > \frac{\varepsilon}{2}\right\}.$$

De esto se sigue que

$$\begin{aligned}\mathbb{P}\left\{\left|\nu_n(A^*) - \nu_n'(A^*)\right| > \frac{\varepsilon}{2}\right\} &\geq \mathbb{P}\left\{\left|\nu_n(A^*) - \nu(A^*)\right| > \varepsilon, \left|\nu_n'(A^*) - \nu(A^*)\right| < \frac{\varepsilon}{2}\right\} \\ &= \mathbb{E}\left\{\mathbb{I}_{\left\{|\nu_n(A^*) - \nu(A^*)| > \varepsilon\right\}} \mathbb{P}\left\{\left|\nu_n'(A^*) - \nu(A^*)\right| < \frac{\varepsilon}{2}\middle| Z_1, \ldots, Z_n\right\}\right\}.\end{aligned}$$

Donde en la última igualdad se usó

$$\mathbb{I}_{\left\{|\nu_n(A^*) - \nu(A^*)| > \varepsilon\right\}},$$

depende únicamente de $Z_1, \ldots, Z_n$. [1] Condicionado a $Z_1, \ldots, Z_n$, la variable $n\nu'_n(A^*)$ tiene distribución Binomial de parámetros n y $\nu(A^*)$ La probabilidad dentro de la esperanza puede ser acotada usando la desigualdad de Chebyshev:

$$\begin{aligned}\mathbb{P}\left\{\left|\nu'_n(A^*) - \nu(A^*)\right| < \frac{\varepsilon}{2}\middle| Z_1, \ldots, Z_n\right\} &\geq 1 - \frac{\nu(A^*)(1-\nu(A^*))}{n\varepsilon^2/4} \\ &\geq 1 - \frac{1}{n\varepsilon^2} > \frac{1}{2},\end{aligned}$$

ya que hemos supuesto que $n\varepsilon^2 > 2$, donde las desigualdades anteriores son c.s. Obtuvimos entonces

$$\begin{aligned}\mathbb{P}\left\{\sup_{A\in\mathcal{A}}\left|\nu_n(A) - \nu'_n(A)\right| > \frac{\varepsilon}{2}\right\} &\geq \frac{1}{2}\mathbb{P}\Big\{\left|\nu_n(A^*) - \nu(A^*)\right| > \varepsilon\Big\} \\ &= \frac{1}{2}\mathbb{P}\left\{\sup_{A\in\mathcal{A}}\left|\nu_n(A) - \nu(A)\right| > \varepsilon\right\}.\end{aligned}$$

PASO 2. Simetrización por signos aleatorios. Sean $\sigma_1, \ldots, \sigma_n$ variables aleatorias independientes idénticamente distribuidas, con $\mathbb{P}(\sigma_1 = -1) = \mathbb{P}(\sigma_1 = 1) = 1/2$, independientes de $Z_1, \ldots, Z_n, Z'_1, \ldots, Z'_n$. Como las $Z_1, Z'_1, \ldots, Z_n, Z'_n$ son independientes idénticamente distribuidas, la distribución de

$$\sup_{A\in\mathcal{A}}\left|\sum_{i=1}^n \big(\mathbb{I}_A(Z_i) - \mathbb{I}_A(Z'_i)\big)\right|,$$

es la misma que la de[2]

$$\sup_{A\in\mathcal{A}}\left|\sum_{i=1}^n \sigma_i\big(\mathbb{I}_A(Z_i) - \mathbb{I}_A(Z'_i)\big)\right|.$$

Usando el Paso 1,

$$\begin{aligned}\mathbb{P}\left\{\sup_{A\in\mathcal{A}}\left|\nu_n(A) - \nu(A)\right| > \varepsilon\right\} &\leq 2\mathbb{P}\left\{\sup_{A\in\mathcal{A}}\frac{1}{n}\left|\sum_{i=1}^n \big(\mathbb{I}_A(Z_i) - \mathbb{I}_A(Z'_i)\big)\right| > \frac{\varepsilon}{2}\right\} \\ &= 2\mathbb{P}\left\{\sup_{A\in\mathcal{A}}\frac{1}{n}\left|\sum_{i=1}^n \sigma_i\big(\mathbb{I}_A(Z_i) - \mathbb{I}_A(Z'_i)\big)\right| > \frac{\varepsilon}{2}\right\}\end{aligned}$$

Si acotamos

$$\left|\sum_{i=1}^n \sigma_i\big(\mathbb{I}_A(Z_i) - \mathbb{I}_A(Z'_i)\big)\right| \leq \left|\sum_{i=1}^n \sigma_i\mathbb{I}_A(Z_i)\right| + \left|\sum_{i=1}^n \sigma_i\mathbb{I}_A(Z'_i)\right|$$

obtenemos

$$\begin{aligned}\mathbb{P}\left\{\sup_{A\in\mathcal{A}}\left|\nu_n(A) - \nu(A)\right| > \varepsilon\right\} &\leq 2\mathbb{P}\left\{\sup_{A\in\mathcal{A}}\frac{1}{n}\left|\sum_{i=1}^n \sigma_i\mathbb{I}_A(Z_i)\right| > \frac{\varepsilon}{4}\right\} \quad + \quad 2\mathbb{P}\left\{\sup_{A\in\mathcal{A}}\frac{1}{n}\left|\sum_{i=1}^n \sigma_i\mathbb{I}_A(Z'_i)\right| > \frac{\varepsilon}{4}\right\} \\ &= 4\mathbb{P}\left\{\sup_{A\in\mathcal{A}}\frac{1}{n}\left|\sum_{i=1}^n \sigma_i\mathbb{I}_A(Z_i)\right| > \frac{\varepsilon}{4}\right\}.\end{aligned}$$

PASO 3. Condicionar a $Z_1, \ldots, Z_n$. Para acotar la probabilidad anterior vamos a condicionar a $Z_1, \ldots, Z_n$. Fijemos $Z_1, \ldots, Z_n \in \mathbb{R}$, al variar z en $\mathbb{R}$ el número de vectores diferentes $\big(\mathbb{I}_{\{Z_1\leq z\}}, \ldots, \mathbb{I}_{\{Z_n\leq z\}}\big)$ es a lo sumo $n+1$. Para ver esto, supongamos que $z < \text{mín}(Z_1, \ldots, Z_n)$ entonces $\big(\mathbb{I}_{\{Z_1\leq z\}}, \ldots, \mathbb{I}_{\{Z_n\leq z\}}\big)$ es el vector nulo. Por su parte si $z \geq \text{máx}(Z_1, \ldots, Z_n)$ es el vector donde todas las coordenadas son 1. En general si $z \in [Z^{(i)}, Z^{(i+1)})$ siendo $Z^{(i)}$ el i-ésimo estadístico de orden, entonces el vector $\big(\mathbb{I}_{\{Z_1\leq z\}}, \ldots, \mathbb{I}_{\{Z_n\leq z\}}\big)$ tiene por lo menos i unos (podrían ser más). Sea $\mathcal{I}$ el conjunto de estos a lo sumo $n+1$ posibles vectores de n coordenadas.

[1] observar que A^* depende de $Z_1, \ldots, Z_n$

[2] para ver esto basta considerar los tres valores, -1,0,1 que toman $X_i = \sigma_i\big(\mathbb{I}_A(Z_i) - \mathbb{I}_A(Z'_i)\big)$ y $Y_i = \mathbb{I}_A(Z_i) - \mathbb{I}_A(Z'_i)$, y ver que estas dos variables tienen la misma distribución

Condicionando a $Z_1, \ldots, Z_n$ podemos escribir

$$\begin{aligned}\mathbb{P}\left\{\sup_{A\in\mathcal{A}}\ \frac{1}{n}\left|\sum_{i=1}^{n}\sigma_i\mathbb{I}_A(Z_i)\right| > \frac{\varepsilon}{4}\Big|Z_1,\ldots,Z_n\right\} &= \mathbb{P}\left\{\max_{v\in\mathcal{I}}\ \frac{1}{n}\left|\sum_{i=1}^{n}\sigma_i\mathbb{I}_A(Z_i)\right| > \frac{\varepsilon}{4}\Big|Z_1,\ldots,Z_n\right\}\\ &\leq \sum_{v\in\mathcal{I}}\mathbb{P}\left\{\frac{1}{n}\left|\sum_{i=1}^{n}\sigma_i\mathbb{I}_A(Z_i)\right| > \frac{\varepsilon}{4}\Big|Z_1,\ldots,Z_n\right\}\\ &\leq \sum_{v\in\mathcal{I}}\sup_{A\in\mathcal{A}}\mathbb{P}\left\{\frac{1}{n}\left|\sum_{i=1}^{n}\sigma_i\mathbb{I}_A(Z_i)\right| > \frac{\varepsilon}{4}\Big|Z_1,\ldots,Z_n\right\}\\ &= |\mathcal{I}|\sup_{A\in\mathcal{A}}\mathbb{P}\left\{\frac{1}{n}\left|\sum_{i=1}^{n}\sigma_i\mathbb{I}_A(Z_i)\right| > \frac{\varepsilon}{4}\Big|Z_1,\ldots,Z_n\right\}. \qquad (6.4)\end{aligned}$$

PASO 4. Desigualdad de Hoeffding. Con $z_1, \ldots, z_n$ fijos, $\sum_{i=1}^{n}\sigma_i\mathbb{I}_A(z_i)$ es la suma de n variables aleatorias independientes con media 0, a valores entre -1 y 1, por lo tanto si aplicamos la desigualdad de Hoeffding obtenemos

$$\mathbb{P}\left\{\frac{1}{n}\left|\sum_{i=1}^{n}\sigma_i\mathbb{I}_A(Z_i)\right| > \frac{\varepsilon}{4}\Big|Z_1,\ldots,Z_n\right\} \leq 2e^{-n\varepsilon^2/32}.$$

Entonces

$$\mathbb{P}\left\{\sup_{A\in\mathcal{A}}\ \frac{1}{n}\left|\sum_{i=1}^{n}\sigma_i\mathbb{I}_A(Z_i)\right| > \frac{\varepsilon}{4}\Big|Z_1,\ldots,Z_n\right\} \leq 2(n+1)e^{-n\varepsilon^2/32}.$$

Si tomamos valor esperado de ambos lados

$$\mathbb{P}\left\{\sup_{A\in\mathcal{A}}\ \frac{1}{n}\left|\sum_{i=1}^{n}\sigma_i\mathbb{I}_A(Z_i)\right| > \frac{\varepsilon}{4}\right\} \leq 2(n+1)e^{-n\varepsilon^2/32}.$$

□

Observación 6.4. *La cota que se obtiene en el teorema anterior es peor que la famosa cota DKW (Dvoretzky-Kiefer-Wolfowitz 1956) que establece que*

$$\mathbb{P}\left\{\sup_{z\in\mathbb{R}}\big|F(z)-F_n(z)\big| > \epsilon\right\} \leq 2e^{-2n\epsilon^2}.$$

6.2 Coeficiente de Fragmentación

Veamos ahora como generalizar el resultado anterior para el caso en que las variables Z_i toman valores en $\mathbb{R}^d$, para eso veamos primero algunas definiciones.

Definición 6.5. Sea $\mathcal{A}$ una familia de subconjuntos medibles. Para $(z_1, \ldots, z_n) \in (\mathbb{R}^d)^n$, sea $N_{\mathcal{A}}(z_1, \ldots, z_n)$ el número de conjuntos distintos[3], de la forma

$$\big\{\{z_1, \ldots, z_n\} \cap A : A \in \mathcal{A}\big\},$$

definimos el número

$$s(\mathcal{A}, n) := \max_{(z_1,\ldots,z_n)\in(\mathbb{R}^d)^n} N_{\mathcal{A}}(z_1, \ldots, z_n).$$

Esto se denomina el n-ésimo **coeficiente de fragmentación** de la familia.

Observación 6.6.

1. *Como función de $(\mathbb{R}^d)^n \to \mathbb{R}$, $N_{\mathcal{A}}(z_1, \ldots, z_n)$ toma una cantidad finita de valores posibles: $1, \ldots, 2^n$, por lo tanto se puede definir $s(\mathcal{A}, n)$ como un máximo y no como un supremo.*
2. *Si $\mathcal{A}' \subset \mathcal{A}$ es una subfamilia de conjuntos de la familia $\mathcal{A}$, es inmediato que $s(\mathcal{A}', n) \leq s(\mathcal{A}, n)$.*

Ejemplo 6.7. Para entender qué es el coeficiente de fragmentación veamos un ejemplo simple. Supongamos que $\mathcal{A}$ es la familia de todos los intervalos $(-\infty, t)$ con $t \in \mathbb{R}$, si tenemos $\mathcal{X}_n = \{X_1, \ldots, X_n\}$ un conjunto de n números reales que suponemos ordenados $X_1 < X_2 < \ldots < X_n$, es claro que $N_{\mathcal{A}}(X_1, \ldots, X_n) = n + 1$. Es decir a lo sumo podemos elegir $n+1$ subconjuntos de $\mathcal{X}_n$ usando $\mathcal{A}$. Un resultado importante que veremos más adelante dice que la cantidad de subconjuntos que podemos elegir con $\mathcal{A}$ (es decir conjuntos de la forma $A \cap \mathcal{X}_n$ con $A \in \mathcal{A}$), es la misma que la cantidad de subconjuntos de pares (X_i, Y_i) con $Y_i \in \{0, 1\}$ que podemos elegir con la familia de la forma $A \times \{0\} \cup A^c \times \{1\}$.

En $\mathbb{R}^2$, si la familia $\mathcal{A}$ es el conjunto de todos los semisespacios, es claro que si $n > 2$ no vamos a poder obtener todos los subconjuntos de un conjunto de n puntos. Claramente $s(\mathcal{A}, n) \leq 2^n$ y si $s(\mathcal{A}, k) < 2^k$ para algún entero k entonces $s(\mathcal{A}, n) < 2^n$ para todo $n > k$.

[3] contando el conjunto vacío

Definición 6.8. Sea $\mathcal{A}$ una familia de conjuntos tal que $|\mathcal{A}| \geq 2$. El mayor entero $k \geq 1$ para el cual $s(\mathcal{A}, k) = 2^k$ se denotará $V_{\mathcal{A}}$ y se denomina **dimensión de Vapnik-Chernonenkis** de $\mathcal{A}$. [4] Si $s(\mathcal{A}, n) = 2^n$ para todo n entonces $V_{\mathcal{A}} = \infty$. En el caso en que $S(\mathcal{A}, k) = 2^k$ decimos que $\mathcal{A}$ fragmenta completamente a $\{z_1, \ldots, z_n\}$.

Si, por ejemplo, tomamos $\mathcal{A}$ como los subconjuntos de la forma $(-\infty, x]$ con $x \in \mathbb{R}$ entonces $s(\mathcal{A}, 2) = 3 < 2^2$ y $V_{\mathcal{A}} = 1$.

Teorema 6.9. *(Vapnik-Chernonenkis (1971)). Sea ν una probabilidad, y $\mathcal{A}$ una familia de conjuntos, entonces, para todo n y para todo $\varepsilon > 0$ tenemos que*

$$\mathbb{P}\left\{\sup_{A\in\mathcal{A}} \big|\nu_n(A) - \nu(A)\big| > \varepsilon\right\} \leq 8s(\mathcal{A}, n)e^{-n\varepsilon^2/32}.$$

Demostración. La demostración sigue las ideas de la prueba del teorema anterior. De forma análoga asumimos que $n\varepsilon^2 \geq 2$, en los primeros 2 pasos demostramos que

$$\mathbb{P}\left\{\sup_{A\in\mathcal{A}} \big|\nu_n(A) - \nu(A)\big| > \varepsilon\right\} \leq 4\mathbb{P}\left\{\sup_{A\in\mathcal{A}} \frac{1}{n}\left|\sum_{i=1}^{n} \sigma_i \mathbb{I}_A(Z_i)\right| > \frac{\varepsilon}{4}\right\}.$$

Esto puede hacerse de manera totalmente análoga a lo hecho en el teorema anterior. Veamos el paso 3.
PASO 3. Condicionar. Para acotar la probabilidad

$$\mathbb{P}\left\{\sup_{A\in\mathcal{A}} \frac{1}{n}\left|\sum_{i=1}^{n} \sigma_i \mathbb{I}_A(Z_i)\right| > \frac{\varepsilon}{4}\right\},$$

nuevamente condicionamos a $Z_1, \ldots, Z_n$. Fijemos $z_1, \ldots, z_n \in \mathbb{R}^d$ y observemos que al variar $A \in \mathcal{A}$ el número de vectores distintos $\big(\mathbb{I}_A(z_1), \ldots, \mathbb{I}_A(z_n)\big)$ es justamente el número de subconjuntos distintos, de $\{z_1, \ldots, z_n\}$, que se producen al intersectar con elementos de $\mathcal{A}$. Por definición este número no excede $s(\mathcal{A}, n)$. Razonando como en (6.4)

$$\mathbb{P}\left\{\sup_{A\in\mathcal{A}} \frac{1}{n}\left|\sum_{i=1}^{n} \sigma_i \mathbb{I}_A(Z_i)\right| > \frac{\varepsilon}{4}\Big| Z_1, \ldots, Z_n\right\} \leq s(\mathcal{A}, n) \sup_{A\in\mathcal{A}} P\left\{\frac{1}{n}\left|\sum_{i=1}^{n} \sigma_i \mathbb{I}_A(Z_i)\right| > \frac{\varepsilon}{4}\Big| Z_1, \ldots, Z_n\right\}.$$

Por lo tanto es suficiente acotar la probabilidad condicional

$$\mathbb{P}\left\{\frac{1}{n}\left|\sum_{i=1}^{n} \sigma_i \mathbb{I}_A(Z_i)\right| > \frac{\varepsilon}{4}\Big| Z_1, \ldots, Z_n\right\}.$$

Esto se hace igual que en el Teorema anterior, aplicamos la desigualdad de Hoeffding y obtenemos

$$\mathbb{P}\left\{\sup_{A\in\mathcal{A}} \big|\nu_n(A) - \nu(A)\big| > \varepsilon\right\} \leq 8s(\mathcal{A}, n)e^{-n\varepsilon^2/32}.$$

□

La cota anterior sirve si $s(\mathcal{A}, n)$ no crece muy rápido con n. Si la clase $\mathcal{A}$ contiene, por ejemplo, a todos los borelianos $s(\mathcal{A}, n) = 2^n$, y, por lo tanto, no es de utilidad.

En general una cota mejor que la que se obtiene en el Teorema 6.9 es [5]

$$\mathbb{P}\left\{\sup_{A\in\mathcal{A}} \big|\nu_n(A) - \nu(A)\big| > \epsilon\right\} \leq cs(\mathcal{A}, n^2)e^{-2n\epsilon^2}, \tag{6.5}$$

La prueba de esta cota se puede ver como ejercicio guiado en [9].

6.2.1 Condición necesaria y suficiente

De la demostración anterior se puede probar que

$$\mathbb{P}\left\{\sup_{A\in\mathcal{A}} \big|\nu_n(A) - \nu(A)\big| > \varepsilon\right\} \leq 8\mathbb{E}\big(N_{\mathcal{A}}(Z_1, \ldots, Z_n)\big)e^{-n\varepsilon^2/32}.$$

De esto se sigue que se puede obtener la ley fuerte de los grandes números, uniforme:

$$\sup_{A\in\mathcal{A}} |\nu_n(A) - \nu(A)| \to 0 \quad \text{en probabilidad}$$

[4] observar que como estamos pidiendo que $|\mathcal{A}| \geq 2$, $s(\mathcal{A}, 1) = 2$ y por lo tanto $V_{\mathcal{A}} \geq 1$

[5] observar que $s(\mathcal{A}, n) \leq s(\mathcal{A}, n^2)$, la mejora está en el exponente de la exponencial, cuyo impacto es mayor ya que, en los casos donde estas desigualdades son usadas, el coeficiente de fragmentación es a lo sumo polinomial en n.

si

$$\frac{\mathbb{E}\Big(\log\big(N_{\mathcal{A}}(Z_1,\ldots,Z_n)\big)\Big)}{n}\to 0.$$

Ya que

$$\mathbb{E}\big(N_{\mathcal{A}}(Z_1,\ldots,Z_n)\big)e^{-n\varepsilon^2/32}=\exp\left(-n\Big(\frac{\epsilon^2}{32}+\frac{\log(\mathbb{E}[N_{\mathcal{A}}(Z_1,\ldots,Z_n)])}{n}\Big)\right).$$

Y, por la desigualdad de Jensen, $\log\Big(\mathbb{E}\big[N_{\mathcal{A}}(Z_1,\ldots,Z_n)\big]\Big)\geq\mathbb{E}\Big(\log\big(N_{\mathcal{A}}(Z_1,\ldots,Z_n)\big)\Big)$.

Vapnik y Chervonenkis probaron luego que esta condición es necesaria para obtener una ley fuerte uniforme.

6.2.2 Elección de clasificadores

Definición 6.10. Si $\mathcal{C}$ es una familia de clasificadores $g:\mathbb{R}^d\to\{0,1\}$, definimos $\mathcal{A}$ como la familia de subconjuntos de $\mathbb{R}^d\times\{0,1\}$ de la forma

$$\big\{\{x:g(x)=1\}\times\{0\}\big\}\bigcup\big\{\{x:g(x)=0\}\times\{1\}\big\},\quad g\in\mathcal{C}.$$

El n-ésimo coeficiente de fragmentación, $s(\mathcal{C},n)$, de la familia de clasificadores $\mathcal{C}$ se define como

$$s(\mathcal{C},n)=s(\mathcal{A},n).$$

Análogamente se define su dimensión de Vapnik-Chervonenkis como

$$V_{\mathcal{C}}=V_{\mathcal{A}}.$$

Como consecuencia del Teorema 6.9, de (3.16), y de la definición anterior, tenemos el siguiente teorema.

Teorema 6.11. *Sea $\mathcal{C}$ una familia de clasificadores $g:\mathbb{R}^d\to\{0,1\}$. Si denotamos $\hat{L}_n(g)=\frac{1}{n}\sum_{i=1}^n\mathbb{I}_{g(X_i)\neq Y_i}$ y $L(g)=\mathbb{P}(g(X)\neq Y)$.*

$$\mathbb{P}\Big\{\sup_{g\in\mathcal{C}}\big|\hat{L}_n(g)-L(g)\big|>\varepsilon\Big\}\leq 8s(\mathcal{C},n)\exp(-n\varepsilon^2/32),$$

y

$$\mathbb{P}\Big\{L(g_n^*)-\inf_{g\in\mathcal{C}}L(g)>\varepsilon\Big\}\leq 8s(\mathcal{C},n)\exp(-n\varepsilon^2/128),\tag{6.6}$$

donde g_n^* denota cualquier clasificador en la clase $\mathcal{C}$ que minimice $\hat{L}_n(g)$.

El error del clasificador g_n^*, que es óptimo empírico en la familia de funciones $\mathcal{C}$, va a estar cerca del error teórico del mejor en $\mathcal{C}$, por (6.6). Muchas veces elegir g_n^* es computacionalmente muy costoso y lo que se hace es elegir un clasificador g_n tal que

$$\mathbb{P}\Big\{\hat{L}_n(g_n)\leq\inf_{g\in\mathcal{C}}\hat{L}_n(g)+\varepsilon_n\Big\}\geq 1-\delta_n,$$

para algún par de sucesiones $\varepsilon_n,\delta_n\to 0$. En este caso se prueba lo siguiente

Ejercicio 6.12.

$$\mathbb{P}\Big\{L(g_n)-\inf_{g\in\mathcal{C}}L(g)>\varepsilon\Big\}\leq\delta_n+\mathbb{P}\Big\{2\sup_{g\in\mathcal{C}}\big|\hat{L}_n(g)-L(g)\big|>\varepsilon-\varepsilon_n\Big\}.$$

El siguiente corolario (que no demostraremos, puede verse como ejercicio guiado en [9]), muestra que minimizando el error empírico en una cierta clase $\mathcal{C}$ (es decir tomando g_n^*), la regla que se obtiene tiene un error que está a menos de $C\sqrt{\log(s(C,n))/n}$ del error del mejor en la clase. Más adelante veremos que podemos acotar $s(C,n)\leq C'n^{V_{\mathcal{C}}}$ si $V_{\mathcal{C}}>2$ con lo cual si $V_{\mathcal{C}}<\infty$, vemos que se puede obtener una cota del tipo $C\sqrt{\log(n)/n}$ donde C depende de $V_{\mathcal{C}}$.

Corolario 6.13. *Usando la notación del Teorema 6.11*

$$\mathbb{E}\big(L(g_n^*)\big)-\inf_{g\in\mathcal{C}}L(g)\leq 16\sqrt{\frac{\log(e\cdot 8\cdot s(\mathcal{C},n))}{2n}}.$$

Una pregunta que surge naturalmente es si la elección que estamos haciendo del clasificador (minimizando L_n en la clase $\mathcal{C}$) es la mejor que se puede hacer, es decir, si no es posible encontrar una regla g_n basada en la muestra, que mejore la cota del corolario anterior. Se puede probar que la cota del corolario anterior son óptimas, ver Teoremas 14.1 y 14.5 de [9].

6.3 Aspectos combinatorios de la teoría de Vapnik-Chervonenkis

Una propiedad importante para calcular coeficientes de fragmentación establece que la cantidad de formas de fragmentar n puntos $X_1, \dots, X_n$ de $\mathbb{R}^d$ (elegir subconjuntos), con una familia $\mathcal{A}$ de subconjuntos de $\mathbb{R}^d$ is igual a la cantidad de formas de fragmentar pares de puntos $(X_1, Y_1), \dots, (X_n, Y_n)$ (con $Y_i \in \{0, 1\}$) tomando conjuntos de la familia

$$\overline{\mathcal{A}} = \Big\{ A \times \{0\} \bigcup A^c \times \{1\} : A \in \mathcal{A} \Big\}.$$

Como mencionamos antes, $\mathcal{A}$ va a ser la familia de conjuntos de la forma $\{x : g(x) = 1\}$, y, por lo tanto, $\overline{\mathcal{A}}$ es la familia de los pares (x, y) tal que $g(x) \neq y$. Como vimos, nos interesa $s(\overline{\mathcal{A}}, n)$.

Teorema 6.14. *Para todo n, $s(\overline{\mathcal{A}}, n) = s(\mathcal{A}, n)$.*

Demostración. Denotemos $\mathcal{X}_n = \{X_1, \dots, X_n\}$ y $D_n = \{(X_1, Y_1), \dots, (X_n, Y_n)\}$. Veremos que $s(\mathcal{A}, n) \leq s(\overline{\mathcal{A}}, n)$ y luego $s(\overline{\mathcal{A}}, n) \leq s(\mathcal{A}, n)$.

$s(\mathcal{A}, n) \leq s(\overline{\mathcal{A}}, n)$:

Veremos que a cualquier partición de $\mathcal{X}_n$ hecha con un $A \in \mathcal{A}$, le podemos asignar una de los pares (X_i, Y_i) formada por un elemento $\overline{A} \in \overline{\mathcal{A}}$, y luego que esta correspondencia es inyectiva. Sea $A \in \mathcal{A}$, denotemos $A \cap \mathcal{X}_n := \{X_{i_1}, \dots, X_{i_k}\}$. Consideremos $\overline{A} \in \overline{\mathcal{A}}$ definido como $\overline{A} = A \times \{0\} \cup A^c \times \{1\}$. Entonces $D_n \cap \overline{A}$ corresponde a los pares (X_i, Y_i) donde $X_i \in \{X_{i_1}, \dots, X_{i_k}\}$, y su etiqueta $Y_i = 0$, o a los pares (X_i, Y_i) tal que $X_i \notin \{X_{i_1}, \dots, X_{i_k}\}$ e $Y_i = 1$. Esto prueba $D_n \cap \overline{A}$ es una partición formada por un elemento $\overline{A} \in \overline{\mathcal{A}}$. Ver Figura 6.1.

Veamos que si $A_1, A_2 \in \mathcal{A}$ dan particiones distintas de $\mathcal{X}_n$, los correspondientes $\overline{A}_1, \overline{A}_2 \in \overline{\mathcal{A}}$ dan particiones distintas de D_n. Si $A_1, A_2 \in \mathcal{A}$ dan particiones distintas de $\mathcal{X}_n$, existe $X_{i_j} \in A_1 \cap \mathcal{X}_n$ pero $X_{i_j} \notin A_2 \cap \mathcal{X}_n$ (el caso en que $X_{i_j} \notin A_1 \cap \mathcal{X}_n$ pero $X_{i_j} \in A_2 \cap \mathcal{X}_n$ es análogo). Recordemos que

$$\overline{A}_1 = A_1 \times \{0\} \cup A_1^c \times \{1\} \qquad \text{y} \qquad \overline{A}_2 = A_2 \times \{0\} \cup A_2^c \times \{1\}.$$

- Si $Y_{i_j} = 0$, como $X_{i_j} \in A_1$, $(X_{i_j}, 0) \in A_1 \times \{0\} \in \overline{A}_1$ pero $(X_{i_j}, 0) \notin \overline{A}_2$, ya que $X_{i_j} \notin A_2$. Por lo tanto $\overline{A}_1$ y $\overline{A}_2$ dan particiones distintas de D_n.
- Si $Y_{i_j} = 1$, $(X_{i_j}, 1) \notin \overline{A}_1$ ya que $(X_{i_j}, 1) \notin A_1 \times \{0\}$ y $(X_{i_j}, 1) \notin A_1^c \times \{1\}$ porque $X_{i_j} \in A_1$. Pero $(X_{i_j}, 1) \in \overline{A}_2$ ya que $(X_{i_j}, 1) \in A_2^c \times \{1\}$. Nuevamente esto significa que $\overline{A}_1$ y $\overline{A}_2$ dan particiones distintas de D_n.

Esto concluye la prueba de que $s(\mathcal{A}, n) \leq s(\overline{\mathcal{A}}, n)$.

$s(\overline{\mathcal{A}}, n) \leq s(\mathcal{A}, n)$:

Veremos que a cualquier partición de D_n le corresponde una de $\mathcal{X}_n$ y que esta correspondencia es inyectiva. Supongamos, sin pérdida de generalidad, que D_n fue dado de la siguiente manera

$$(X_1, 0), \dots, (X_m, 0), (X_{m+1}, 1), \dots, (X_n, 1).$$

Supongamos que un conjunto $A \in \mathcal{A}$ es tal que el correspondiente $\overline{A} = A \times \{0\} \cup A^c \times \{1\} \in \overline{\mathcal{A}}$ separa los pares $N_{k+l} = \{(X_{i_1}, 0), \dots, (X_{i_k}, 0), (X_{j_1}, 1), \dots, (X_{j_l}, 1)\}$, es decir, por definición $\overline{A} \cap D_n = N_{k+l}$. Esto significa que A elige del conjunto $\mathcal{X}_n$ los k puntos $X_{i_1}, \dots, X_{i_k}$ entre $X_1, \dots, X_m$,. Además $A^c \times \{1\}$ elige los pares $(X_{j_1}, 1), \dots, (X_{j_l}, 1)$ por lo tanto A tuvo que haber elegido los puntos $\{X_{m+1}, \dots, X_n\} \setminus \{X_{j_1}, \dots, X_{j_l}\}$. Por lo tanto, probamos que

$$A \cap \mathcal{X}_n = \{X_{i_1}, \dots, X_{i_k}\} \bigcup \Big(\{X_{m+1}, \dots, X_n\} \setminus \{X_{j_1}, \dots, X_{j_l}\} \Big)$$

análogamente,

$$A^c \cap \mathcal{X}_n = \{X_{j_1}, \dots, X_{j_l}\} \bigcup \Big(\{X_1, \dots, X_m\} \setminus \{X_{i_1}, \dots, X_{i_k}\} \Big).$$

Es decir, A y A^c definen una partición de $\mathcal{X}_n$.

Veamos que si $\overline{A}_1, \overline{A}_2 \in \overline{\mathcal{A}}$ eligen pares distintos, entonces $A_1, A_2 \in \mathcal{A}$ eligen puntos distintos. Supongamos que existe $(X_i, Y_i) \in \overline{A}_1 \cap D_n$ pero $(X_i, Y_i) \notin \overline{A}_2 \cap D_n$.

- Si $Y_i = 0$ esto significa que $X_i \in A_1 \cap \mathcal{X}_n$. Que $(X_i, 0) \notin \overline{A}_2 \cap D_n$ quiere decir que $X_i \notin A_2 \cap \mathcal{X}_n$. Por lo tanto A_1 elije X_i pero A_2 no.
- Si $Y_i = 1$ entonces $X_i \in A_1^c \cap \mathcal{X}_n$. Que $(X_i, 1) \notin \overline{A}_2 \cap D_n$ quiere decir que $X_i \notin A_2^c \cap \mathcal{X}_n$ y por lo tanto $X_i \in A_2 \cap \mathcal{X}_n$. Es decir, A_2 elije X_i pero A_1 no.

Esto prueba que $s(\overline{\mathcal{A}}, n) \leq s(\mathcal{A}, n)$.

□

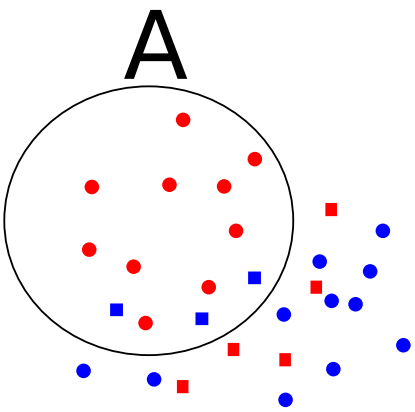

Figura 6.1: En la figura se muestra: con círculos y cuadrados rojos los puntos $(X_i, 0)$. Con círculos y cuadrados azules se representan los pares $(X_i, 1)$. Los cuadrados rojos son los pares $(X_i, 0)$ que $\overline{A}$ no logra agarrar, y los cuadrados azules son los pares $(X_i, 1)$ que $\overline{A}$ no logra agarrar.

Veamos ahora un teorema que nos permitirá acotar superiormente $s(\mathcal{A}, n)$ en términos de la dimensión de Vapnik-Chervonenkis de la familia $\mathcal{A}$.

Teorema 6.15. *Si $\mathcal{A}$ es una familia de conjuntos con dimensión de Vapnik-Chervonenkis $V_\mathcal{A}$ entonces para todo n*

$$s(\mathcal{A}, n) \leq \sum_{i=0}^{\text{mín}\{n, V_\mathcal{A}\}} \binom{n}{i}. \tag{6.7}$$

Demostración. Si $V_\mathcal{A} = \infty$ entonces, por definición de $V_\mathcal{A}$, $s(\mathcal{A}, n) = 2^n$ y $\text{mín}\{n, V_\mathcal{A}\} = n$, para todo n. Por lo tanto la desigualdad (6.7) es, para todo n, una igualdad, por el teorema binomial.

Supongamos que $V_\mathcal{A} < \infty$. Vamos a probar (6.7) por inducción en n y $V_\mathcal{A}$. Por definición de $V_\mathcal{A}$, (6.7) vale si sumamos hasta $V_\mathcal{A}$, sin explicitar que estamos sumando hasta $\text{mín}\{n, V_\mathcal{A}\}$, ya que si $n \leq V_\mathcal{A}$, $s(\mathcal{A}, n) = 2^n$.

Es suficiente probar que para todo $\mathcal{X}_n = \{x_1, \ldots, x_n\}$,

$$N_\mathcal{A}(x_1, \ldots, x_n) \leq \sum_{i=0}^{V_\mathcal{A}} \binom{n}{i}.$$

Caso base: (6.7) es la igualdad $2 = 2$ para $n = 1$ para cualquier familia $\mathcal{A}$ con $|\mathcal{A}| > 1$, ya que $V_\mathcal{A} \geq 1$, y $s(\mathcal{A}, 1) = 2$. Si $n = 2$ y $V_\mathcal{A} = 1$, $s(\mathcal{A}, 2) < 2^2$ y el lado derecho de (6.7) vale 3, por lo tanto se verifica (6.7). Si $V_\mathcal{A} = 1$ y $n > 2$ el lado izquierdo queda menor o igual a 3, y el derecho mayor o igual a 3 y por lo tanto también vale la desigualdad, para cualquier n. En la Figura 6.2 representamos como puntos verde los pares $(n, V_\mathcal{A})$ donde vale la desigualdad en el caso base.

Paso inductivo A los efectos de contar cantidad de fragmentaciones de $\mathcal{X}_n$ con elementos de $\mathcal{A}$ podemos asumir que estamos contando subconjuntos de $\mathcal{X}_n$. Es decir, $\mathcal{A}$ lo podemos tomar como una familiar de subconjuntos de $\mathcal{X}_n$ tal que $s(\mathcal{A}, n) = |\mathcal{A}|$, en particular $|\mathcal{A}| < \infty$.

La hipótesis de inducción es que (6.7) es cierto para todo $k < n$, para toda familia de subconjuntos de $\{x_1, \ldots, x_k\}$ de dimensión menor o igual que $V_\mathcal{A}$, y para n y toda familia de dimensión menor que $V_\mathcal{A}$. En la Figura 6.2 estos puntos se representan en azul. Queremos probarlo para el punto en rojo.

Definimos las siguientes clases de subconjuntos de $\{x_, \ldots, x_n\}$.

$$\mathcal{A}' = \{A - \{x_n\} : A \in \mathcal{A}\},$$

y

$$\hat{\mathcal{A}} = \{\hat{A} \in \mathcal{A} : x_n \notin \hat{A}, \hat{A} \cup \{x_n\} \in \mathcal{A}\}.$$

Observemos que tanto $\mathcal{A}'$ como $\hat{\mathcal{A}}$ están formados por subconjuntos de $\{x_1, \ldots, x_{n-1}\}$. Veamos que $|\mathcal{A}| = |\mathcal{A}'| + |\hat{\mathcal{A}}|$, escribimos

$$\mathcal{A}' = \underbrace{\{A - \{x_n\} : x_n \in A, A \in \mathcal{A}\}}_{B_1} \cup \underbrace{\{A - \{x_n\} : x_n \notin A, A \in \mathcal{A}\}}_{B_2}.$$

Es decir B_1 son todos los subconjuntos (que no necesariamente están en $\mathcal{A}$) que se obtienen quitando x_n de subconjuntos de $\mathcal{A}$ que lo contenían. Mientras que B_2 son los subconjuntos de $\mathcal{A}$ que no tenían a x_n. Por lo tanto $B_1 \cap B_2 = \hat{\mathcal{A}}$.

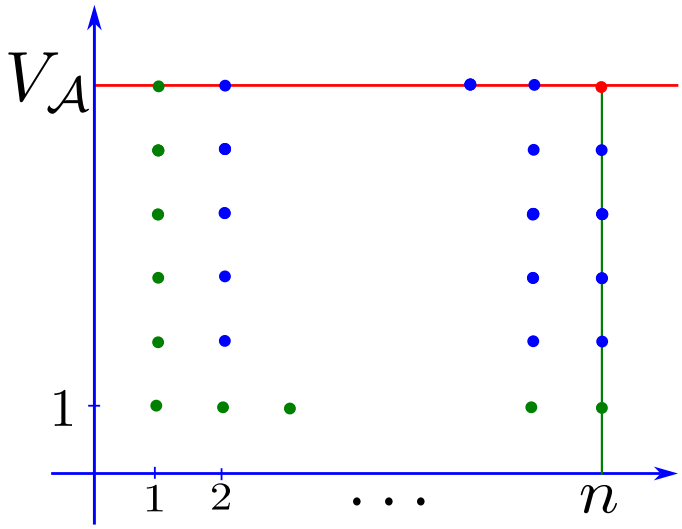

Figura 6.2: En verde se representan los puntos cuyas coordenadas cumplen la desigualdad (6.7) para el caso base. En azul se representan los puntos para los cuales se asume que vale la cota en el paso inductivo, y, finalmente, el punto rojo es el punto cuyas coordenadas queremos probar que verifican (6.7).

Entonces

$$\begin{aligned}
|\mathcal{A}'| &= |B_1| + |B_2| - |B_1 \cap B_2| \\
&= \left|\{A - \{x_n\} : x_n \in A, A \in \mathcal{A}\}\right| + \left|\{A - \{x_n\} : x_n \notin \mathcal{A}, A \in \mathcal{A}\}\right| - \left|\hat{\mathcal{A}}\right| \\
&= \left|\{A : x_n \in A, A \in \mathcal{A}\}\right| + \left|\{A : x_n \notin A, A \in \mathcal{A}\}\right| - \left|\hat{\mathcal{A}}\right| \\
&= |\mathcal{A}| - |\hat{\mathcal{A}}|.
\end{aligned}$$

Como $\mathcal{A}'$ es una familia de subconjuntos de $\{x_1, \ldots, x_{n-1}\}$ (no necesariamente es una subfamilia de $\mathcal{A}$), y $|\mathcal{A}'| \leq |\mathcal{A}|$ podemos aplicar la hipótesis de inducción y obtenemos que [6]

$$|\mathcal{A}'| = s(\mathcal{A}', n-1) = \sum_{i=1}^{V_{\mathcal{A}'}} \binom{n-1}{i} \leq \sum_{i=1}^{V_{\mathcal{A}}} \binom{n-1}{i}.$$

Veremos que $V_{\hat{\mathcal{A}}} \leq V_{\mathcal{A}} - 1$, lo cual implica que

$$|\hat{\mathcal{A}}| = s(\hat{\mathcal{A}}, n-1) = \sum_{i=1}^{V_{\hat{\mathcal{A}}}} \binom{n-1}{i} \leq \sum_{i=1}^{V_{\mathcal{A}}-1} \binom{n-1}{i},$$

por hipótesis de inducción. Para ver ésto consideremos un conjunto cualquiera $S \subset \{x_1, \ldots, x_{n-1}\}$ tal que $\hat{\mathcal{A}}$ lo divide completamente. Si probamos que $S \cup \{x_n\}$ es divido completamente por $\mathcal{A}$ entonces $|S \cup \{x_n\}| \leq V_{\mathcal{A}}$ por definición de $V_{\mathcal{A}}$. Además, como S no contiene a x_n, $|S \cup \{x_n\}| = |S| + 1$. Pero, como S es un elemento arbitrario, dividido completamente por $\hat{\mathcal{A}}$ obtenemos que $V_{\hat{\mathcal{A}}} \leq V_{\mathcal{A}} - 1$.

Veamos que $S \cup \{x_n\}$ es divido completamente por $\mathcal{A}$. Si $S' \subset S \cup \{x_n\}$ y S' no contiene a x_n entonces $S' \subset S$ es dividido completamente por $\mathcal{A}$, ya que S era dividido completamente por $\hat{\mathcal{A}} \subset \mathcal{A}$.

Consideremos $S' \subset S$ cualquiera y veamos que $S' \cup \{x_n\}$ es la intersección de $S \cup \{x_n\}$ y un elemento de $\mathcal{A}$. Como S es dividido completamente por $\hat{\mathcal{A}}$, si $S' \subset S$ entonces existe $\hat{A} \in \hat{\mathcal{A}}$ tal que $S' = S \cap \hat{A}$. Pero como por definición $x_n \notin \hat{A}$, tenemos que

$$S' = (S \cup \{x_n\}) \cap \hat{A} \qquad \text{y, por lo tanto} \qquad S' \cup \{x_n\} = (S \cup \{x_n\}) \cap (\hat{A} \cup \{x_n\}).$$

Por definición de $\hat{\mathcal{A}}$, si $\hat{A} \in \hat{\mathcal{A}}$ entonces $\hat{A} \cup \{x_n\} \in \mathcal{A}$ y por lo tanto $S \cup \{x_n\}$ es dividido completamente por $\mathcal{A}$. Por lo tanto hemos demostrado que

$$s(\mathcal{A}, n) = |\mathcal{A}| = |\mathcal{A}'| + |\hat{\mathcal{A}}| \leq \sum_{i=0}^{V_{\mathcal{A}}} \binom{n-1}{i} + \sum_{i=0}^{V_{\mathcal{A}}-1} \binom{n-1}{i},$$

finalmente la conclusión se sigue de la identidad $\binom{n}{i} = \binom{n-1}{i} + \binom{n-1}{i-1}$ para todo $i > 0$, $i \in \mathbb{N}$. □

[6] aquí estamos usando que nuestra hipótesis de inducción establece que vale (6.7) para cualquier familia de conjuntos de cardinal menor que n (en particular para $\mathcal{A}'$), y que $V_{\mathcal{A}'} \leq V_{\mathcal{A}}$.

El siguiente teorema facilita el cálculo del coeficiente de fragmentación de varias clases de conjuntos.

Teorema 6.16.

1. *Si* $\mathcal{A} = \mathcal{A}_1 \cup \mathcal{A}_2$*, entonces* $s(\mathcal{A}, n) \leq s(\mathcal{A}_1, n) + s(\mathcal{A}_2, n)$.
2. *Dada una clase* $\mathcal{A}$ *definimos* $\mathcal{A}^c = \{A^c : A \in \mathcal{A}\}$*. Entonces* $s(\mathcal{A}^c, n) = s(\mathcal{A}, n)$.
3. *Si* $\mathcal{A} = \left\{\hat{A} \cap \tilde{A} : \hat{A} \in \hat{\mathcal{A}}, \tilde{A} \in \tilde{\mathcal{A}}\right\}$*, entonces* $s(\mathcal{A}, n) \leq s(\hat{\mathcal{A}}, n)s(\tilde{\mathcal{A}}, n)$.
4. *Si* $\mathcal{A} = \left\{\hat{A} \cup \tilde{A} : \hat{A} \in \hat{\mathcal{A}}, \tilde{A} \in \tilde{\mathcal{A}}\right\}$*, entonces* $s(\mathcal{A}, n) \leq s(\hat{\mathcal{A}}, n)s(\tilde{\mathcal{A}}, n)$.
5. *Si* $\mathcal{A} = \left\{\hat{A} \times \tilde{A} : \hat{A} \in \hat{\mathcal{A}}, \tilde{A} \in \tilde{\mathcal{A}}\right\}$*, entonces* $s(\mathcal{A}, n) \leq s(\hat{\mathcal{A}}, n)s(\tilde{\mathcal{A}}, n)$.

Demostración. Los puntos 1,2 y 5 quedan como ejercicio. Para probar 3 fijamos n puntos $x_1, \ldots, x_n$. Supongamos que con conjuntos $C_1, \ldots, C_N \in \hat{\mathcal{A}}$ elegimos $N \leq s(\hat{\mathcal{A}}, n)$ subconjuntos distintos de $x_1, \ldots, x_n$, esto significa que

$$\{A \cap \{X_1, \ldots, X_n\} : A \in \mathcal{A}\} = C_1 \cup \cdots \cup C_N.$$

Con subconjuntos de $\tilde{\mathcal{A}}$ podemos elegir a lo sumo $s(\tilde{\mathcal{A}}, |C_i|)$ subconjuntos de tamaño $|C_i|$. Cada elección que hacemos de $x_1, \ldots, x_n$ con elementos de la forma $\hat{A} \cap \tilde{A}$, corresponde a una elección de dichos puntos con $\hat{A}$ (y por lo tanto va a ser alguno de los C_i de antes). Para ese C_i podemos elegir a lo sumo $s(\tilde{\mathcal{A}}, |C_i|)$ subconjuntos. Por lo tanto, la cantidad de subconjuntos distintos, de $x_1, \ldots, x_n$ que podemos agarrar, con elementos de $\hat{\mathcal{A}} \cap \tilde{\mathcal{A}}$ está acotada superiormente por

$$\sum_{i=1}^{N} s(\tilde{\mathcal{A}}, |C_i|) \leq \sum_{i=1}^{N} s(\tilde{\mathcal{A}}, n) = N s(\tilde{\mathcal{A}}, n) \leq s(\hat{\mathcal{A}}, n)s(\tilde{\mathcal{A}}, n)$$

donde hemos usado que en general $s(\mathcal{A}, n) \leq s(\mathcal{A}, n + m)$. Esto prueba 3. Para probar el punto 4 observemos que

$$\mathcal{A} = \left\{\hat{A} \cup \tilde{A} : \hat{A} \in \hat{\mathcal{A}}, \tilde{A} \in \tilde{\mathcal{A}}\right\} = \left\{\left(\hat{A}^c \cap \tilde{A}^c\right)^c : \hat{A} \in \hat{\mathcal{A}}, \tilde{A} \in \tilde{\mathcal{A}}\right\}.$$

Si usamos el punto 2,

$$s(\mathcal{A}, n) = s\left(\left\{\hat{A}^c \cap \tilde{A}^c : \hat{A} \in \hat{\mathcal{A}}, \tilde{A} \in \tilde{\mathcal{A}}\right\}, n\right).$$

Si ahora aplicamos el punto 3,

$$s\left(\left\{\hat{A}^c \cap \tilde{A}^c : \hat{A} \in \hat{\mathcal{A}}, \tilde{A} \in \tilde{\mathcal{A}}\right\}, n\right) \leq s(\hat{\mathcal{A}}^c, n)s(\tilde{\mathcal{A}}^c, n).$$

Y ahora volvemos a aplicar el punto 2. □

El siguiente teorema se sigue de forma inmediata a partir de la definición del coeficiente de fragmentación.

Teorema 6.17. *Si* $\mathcal{A}$ *contiene finitos subconjuntos,* $s(\mathcal{A}, n) \leq |\mathcal{A}|$ *para todo* n*, y por lo tanto* $V_{\mathcal{A}} \leq \log_2 |\mathcal{A}|$ *y*

Veamos ahora dos ejemplos que muestran que la cota del Teorema 6.15 es óptima.

Teorema 6.18.

1. *Si* $\mathcal{A} = \left\{(-\infty, x] : x \in \mathbb{R}\right\}$*, entonces* $V_{\mathcal{A}} = 1$ *y*

$$s(\mathcal{A}, n) = n + 1 = \binom{n}{0} + \binom{n}{1}$$

2. *Si* $\mathcal{A}$ *es la clase de todos los intervalos de* $\mathbb{R}$*, entonces* $V_{\mathcal{A}} = 2$ *y*

$$s(\mathcal{A}, n) = \frac{n(n+1)}{2} + 1 = \binom{n}{0} + \binom{n}{1} + \binom{n}{2}$$

Demostración. El punto 1 es trivial. Para ver el punto 2, observemos primero que si tenemos 3 puntos, $X_1 < X_2 < X_3$, por conexidad, no se pueden elegir X_1 y X_3 sin elegir X_2. Por lo tanto $V_{\mathcal{A}} = 2$, ya que 2 puntos se pueden fragmentar completamente con intervalos. Para calcular $s(\mathcal{A}, n)$ observemos que hay $n - k + 1$ subconjuntos de k puntos de $X_1, \ldots, X_n$ de la forma $[a, b] \cap \{X_1, \ldots, X_n\}$ (que son $\{X_1, \ldots, X_k\}, \{X_2, \ldots, X_{k+1}\}, \ldots, \{X_{n-k+1}, \ldots, X_n\}$. Además hay que sumar el conjunto vacío. Por lo tanto tenemos

$$s(\mathcal{A}, n) = 1 + \sum_{k=1}^{n} (n - k + 1) = \frac{n(n+1)}{2} + 1.$$

□

En $\mathbb{R}^d$ tenemos el siguiente resultado.

Teorema 6.19.

1. *Si* $\mathcal{A} = \big\{(-\infty, x_1] \times \cdots \times (-\infty, x_d]\big\}$, *entonces* $V_{\mathcal{A}} = d$.

2. *Si* $\mathcal{A}$ *es la clase de todos los rectángulos* $V_{\mathcal{A}} = 2d$.

Demostración. Para probar el punto 1 basta observar que $\{e_1, \ldots, e_d\}$ los vectores de la base canónica, unión el origen $(0, \ldots, 0)$ no se pueden fragmentar completamente, pero si se puede fragmentar completamente $\{e_1, \ldots, e_d\}$.

Para probar el punto 2 vamos a ver primero que $V_{\mathcal{A}} \geq 2d$. Esto se sigue de que los $2d$ puntos

$$(1, 0, \ldots, 0), (0, 1, 0, \ldots, 0), \ldots, (0, \ldots, 0, 1), (-1, 0, \ldots, 0), (0, -1, 0, \ldots, 0), \ldots, (0, \ldots, 0, -1)$$

se pueden fragmentar completamente con rectángulos. Para ver que no se pueden fragmentar $2d + 1$ puntos. Dados $2d + 1$ puntos cualesquiera, consideremos el conjunto $\mathcal{X}$ de puntos formado por: el punto con primera coordenada más chica, el punto con primera coordenada más grande, el punto con segunda coordenada más chica, el punto con segunda coordenada más grande, y así sucesivamente. Este conjunto tiene a lo sumo $2d$ puntos, es claro que el punto de los $2d + 1$ que resta, no se puede separar por rectángulos de $\mathcal{X}$. □

Teorema 6.20. *Sea* $\mathscr{G}$ *un espacio vectorial de dimensión* $r < \infty$, *de funciones de* $\mathbb{R}^d$ *a valores reales. La familia de conjuntos*

$$\mathcal{A} = \Big\{\big\{x : g(x) \geq 0\big\} : g \in \mathscr{G}\Big\},$$

tiene dimensión $V_{\mathcal{A}} \leq r$.

Demostración. Tenemos que probar que un conjunto de $m = r + 1$ no se puede fragmentar con conjuntos de la forma $\{x : g(x) \geq 0\}$. Fijemos m puntos $X_1, \ldots, X_m$, supongamos que podemos fragmentarlo completamente. Consideremos el mapa lineal $T : \mathscr{G} \to \mathbb{R}^m$

$$T(g) = \Big(g(X_1), \ldots, g(X_m)\Big).$$

Como $\mathscr{G}$ es un espacio vectorial de dimensión $r < \infty$ tenemos que $\dim(T(\mathscr{G})) \leq r = m - 1$. Por lo tanto existe $\gamma = (\gamma_1, \ldots, \gamma_m) \in \mathbb{R}^m$, ortogonal a $T(\mathscr{G})$, es decir para todo $g \in \mathscr{G}$,

$$\gamma_1 g(X_1) + \cdots + \gamma_m g(X_m) = 0.$$

Podemos suponer que alguno de los γ_i es negativo, ya que si todos son positivos se toma $-\gamma$. Podría pasar que todos los i sean negativos. Escribimos la ecuación anterior como

$$\sum_{i:\gamma_i \geq 0} \gamma_i g(X_i) = \sum_{i:\gamma_i < 0} -\gamma_i g(X_i).$$

Donde si todos los g_i son negativos a la izquierda la suma da 0. Si suponemos que el conjunto $X_1, \ldots, X_m$ se puede fragmentar completamente, entonces existe $g \in \mathscr{G}$ que elige exactamente los X_i con $i : \gamma_i \geq 0$. Por lo tanto el lado derecho de la ecuación es estrictamente negativo, ya que, no elegir un X_i con g significa que $g(X_i) < 0$. Por otra parte, el lado izquierdo es mayor o igual que 0. Esto es una contradicción por lo tanto no se pueden fragmentar m puntos. □

Como corolarios de este teorema se obtiene una cota para la dimensión de Vapnik-Chervonenkis de dos clases importantes de conjuntos.

Corolario 6.21.

1. *Sea* $\mathcal{A} = \Big\{x : a^T x \geq b,\ a \in \mathbb{R}^d,\ b \in \mathbb{R}\Big\}$. *Entonces* $V_{\mathcal{A}} \leq d + 1$. [7]

2. *Sea* $\mathcal{A}$ *la familia de todas las bolas cerradas de* $\mathbb{R}^d$. *Entonces* $V_{\mathcal{A}} \leq d + 2$.

Demostración.

1. Se aplica el Teorema 6.20, al espacio $d + 1$ dimensional generado por las funciones $g_i(x) = x_i$ para $i = 1, \ldots, d$ y $g_{d+1}(x) = 1$.

2. Se aplica el Teorema 6.20, al espacio $d + 2$ dimensional generado por las funciones

$$g_1(x) = \sum_{i=1}^{d} |x_i|^2,\ g_2(x) = x_1, \ldots, g_{d+1}(x) = x_d,\ g_{d+2}(x) = 1,$$

ya que

$$\sum_{i=1}^{d} |x_i - a_i|^2 - b = g_1(x) - 2\sum_{i=1}^{d} g_{i+1}(x) a_i + \sum_{i=1}^{d} a_i^2 - b.$$

□

[7] se puede probar que en este caso la dimensión es exactamente $d + 1$.

En el caso de $\mathcal{A}$ sea la clase de todos los polígonos convexos, $V_{\mathcal{A}} = \infty$, y esto vale en cualquier dimensión. Esto se sigue de que cualquier subconjunto de un conjunto de n puntos en el círculo, se pueden elegir con polígonos.

Teorema 6.22. *Para todos $n \geq 1$ y $V_{\mathcal{A}} < n/2$,*

$$s(\mathcal{A}, n) \leq e^{n\mathcal{H}\left(\frac{V_{\mathcal{A}}}{n}\right)},$$

donde $\mathcal{H}(x) = -x\log x - (1-x)\log(1-x)$ para $x \in (0,1)$, y $\mathcal{H}(0) = \mathcal{H}(1) = 0$.

La prueba del Teorema 6.22 se sigue de forma inmediata de (6.7) y del siguiente Lema:

Lema 6.23. *Para $k < n/2$,*

$$\sum_{i=0}^{k} \binom{n}{i} \leq e^{n\mathcal{H}\left(\frac{k}{n}\right)}.$$

Demostración. Introducimos $\lambda = k/n \leq 1/2$. Por el teorema binomial,

$$1 = \big(\lambda + (1-\lambda)\big)^n = \sum_{i=0}^{n} \binom{n}{i} \lambda^i (1-\lambda)^{n-i} \underbrace{\geq}_{\lambda \leq 1/2} \sum_{i=0}^{\lambda n} \binom{n}{i} \lambda^i (1-\lambda)^{n-i} = \sum_{i=0}^{\lambda n} \binom{n}{i} \frac{\lambda^i}{(1-\lambda)^i} (1-\lambda)^n$$

$$\geq \sum_{i=0}^{\lambda n} \binom{n}{i} \left(\frac{\lambda}{1-\lambda}\right)^{\lambda n} (1-\lambda)^n \quad \text{(ya que } \lambda/(1-\lambda) \leq 1) = \lambda^{n\lambda}(1-\lambda)^{n(1-\lambda)} \sum_{i=0}^{\lambda n} \binom{n}{i} = e^{-n\mathcal{H}(\lambda)} \sum_{i=0}^{k} \binom{n}{i}.$$

□

Una consecuencia del Teorema 6.22 es el siguiente Teorema

Teorema 6.24. *Para todo $n > 2V_{\mathcal{A}}$,*

$$s(\mathcal{A}, n) \leq \sum_{i=0}^{V_{\mathcal{A}}} \binom{n}{i} \leq \left(\frac{e \cdot n}{V_{\mathcal{A}}}\right)^{V_{\mathcal{A}}}.$$

Observar que si $V_{\mathcal{A}} > 2$ $(e/V_{\mathcal{A}}) < 1$ y por lo tanto

$$\boxed{s(\mathcal{A}, n) \leq n^{V_{\mathcal{A}}}.}$$

Demostración. Usando el Teorema 6.22, para $V_{\mathcal{A}} < n/2$,

$$s(\mathcal{A}, n) \leq \exp\left(n\left[-\frac{V_{\mathcal{A}}}{n}\log\left(\frac{V_{\mathcal{A}}}{n}\right) - \left(1-\frac{V_{\mathcal{A}}}{n}\right)\log\left(1-\frac{V_{\mathcal{A}}}{n}\right)\right]\right)$$

$$= \left(\frac{n}{V_{\mathcal{A}}}\right)^{V_{\mathcal{A}}} \underbrace{\exp\left[-n\Big(1-\frac{V_{\mathcal{A}}}{n}\Big)\log\Big(1-\frac{V_{\mathcal{A}}}{n}\Big)\right]}_{\leq e^{V_{\mathcal{A}}}}$$

donde en la desigualdad

$$\exp\left[-n\Big(1-\frac{V_{\mathcal{A}}}{n}\Big)\log\Big(1-\frac{V_{\mathcal{A}}}{n}\Big)\right] \leq e^{V_{\mathcal{A}}}$$

usamos que, para todo $x \in [0,1]$, $-(1-x)\log(1-x) \leq x$. □

6.4 Error de resustitución

En la sección 3.3.5 vimos que si tenemos una muestra de testeo T_m y una de entrenamiento D_n podemos estimar la probabilidad de error, L_n, que comete una regla g_n basada en D_n. En algunos casos se puede estimar dicho error en la propia muestra, por medio del error de resustitución que definimos como

$$L_n^{(R)} = \frac{1}{n}\sum_{i=1}^{n} \mathbb{I}_{\{g_n(X_i) \neq Y_i\}}$$

Donde g_n es una regla basada en $D_n = \{(X_1, Y_1), \ldots, (X_n, Y_n)\}$. Es decir usamos la muestra D_n para entrenar el clasificador, y para estimar su error. Es claro, en general

$$L_n^{(R)} \leq L(g_n) = \mathbb{P}(g_n(X) \neq Y | D_n)$$

porque usamos la misma muestra para entrenar y calcular su error. En un caso extremo, si g_n es la regla del 1 vecino mas cercano es claro que $L_n^{(R)} = 0$ ya que el 1 vecino mas cercano de cada X_i, en D_n es X_i. Se puede demostrar que si k es suficientemente grande $L_n^{(R)}$ estima bien $L(g_n)$. Este es el caso de las reglas que se construyen por medio de la optimización en una clase con dimensión VC finita, pero deja afuera a los clasificadores por vecinos más cercanos, y a los clasificadores basados en histogramas, cuya dimensión VC es infinita.

Veremos primero un teorema que nos dice que si partimos $\mathbb{R}^d$ en k celdas fijas, y usamos una regla de clasificador g_n que es constante en cada celda, el error de resustitución $L_n^{(R)}$ estará próximo al error $L(g_n)$ del clasificador.

Teorema 6.25. *Sea g_n una regla de clasificación constante en las celdas de una partición de $\mathbb{R}^d$ en k celdas. Entonces*

$$\mathbb{P}\big(|L_n^{(R)} - L(g_n)| > \epsilon\big) \leq 8 \cdot 2^k \exp\big(-n\varepsilon^2/32\big)$$

Demostración. Definamos $A_n \subset \mathbb{R}^d \times \{0,1\}$ como el conjunto (x,y) donde g_n se equivoca, es decir

$$A_n = \big\{(x,y) : g_n(x) \neq y\big\}$$

Por lo tanto $L(g_n) = \mathbb{P}\big((X,Y) \in A_n | D_n\big)$ y

$$L_n^{(R)} = \frac{1}{n}\sum_{i=1}^{n} \mathbb{I}_{\{(X_i,Y_i)\in A_n\}}.$$

Si denotamos ν la medida asociada a (X,Y) y ν_n la empírica, $L(g_n) = \nu(A_n)$ y $L_n^{(R)} = \nu_n(A_n)$. Observar que aquí primero sorteamos A_n y luego le calculamos su medida ν y ν_n respectivamente. Como A_n depende de la muestra, por ejemplo $\mathbb{E}(\nu_n(A_n)) \neq \mathbb{E}(\nu(A_n))$. Si llamamos $\mathcal{C}$ a la clase de todos los conjuntos de la forma $\{(x,y) : g(x) \neq y\}$ donde $g : \mathbb{R}^d \to \{0,1\}$ es cualquier regla de clasificación constante en las k celdas. Entonces

$$\big|L(g_n) - L_n^{(R)}\big| \leq \sup_{C\in\mathcal{C}} \big|\nu(C) - \nu_n(C)\big|.$$

Tenemos que calcular el coeficiente de fragmentación de la clase $\mathcal{C}$. Por el Teorema 6.14, hay que calcular el coeficiente de fragmentación de la clase $\mathcal{C}'$ de todos los subconjuntos de $\mathbb{R}^d$, que se obtienen como unión de celdas de una partición fija. Como son k celdas, tenemos 2^k elementos en $\mathcal{C}'$. Por el Teorema 6.17 tenemos entonces que $s(\mathcal{C}', n) \leq 2^k$. □

Observación 6.26.

1. *Es importante tener en cuenta que en el teorema anterior la partición es fija. Si permitimos variar la partición el coeficiente 2^k aumenta. Para entender esto supongamos que $d = 1$. Una partición de $\mathbb{R}$ en k celdas es un conjunto de k intervalos disjuntos. Por cada partición fija tenemos 2^k fragmentaciones del conjunto de puntos, ya que podemos elegir o no cada intervalo. Por otra parte la cantidad de fragmentaciones de n puntos de la recta que podemos hacer con k intervalos disjuntos es*

$$\binom{n+k}{k}.$$

   Es decir la cantidad de fragmentaciones de n pares (X_i, Y_i) tal que $X_i \in \mathbb{R}$ y $Y_i \in \{0,1\}$, para $i = 1,\ldots,n$, con subconjuntos de la forma

$$\Big\{\{x : g_{P_k}(x) = 0\} \times \{1\}\Big\} \bigcup \Big\{\{x : g_{P_k}(x) = 1\} \times \{0\}\Big\}$$

   donde g_{P_k} es cualquier regla de clasificación constante en celdas de una partición P_k, está acotada superiormente por $2^k\binom{n+k}{k}$.

2. *Si en $\mathbb{R}^d$ tenemos un hiperplano, podemos fragmentar n puntos $X_1,\ldots,X_n$ de a lo sumo n^{d+1} formas como se sigue de aplicar el punto 1 del Corolario 6.21, y el Teorema 6.24 Por lo tanto la cantidad de formas de fragmentar n pares (X_i, Y_i) con reglas de clasificación constantes en las celdas de una partición de $\mathbb{R}^d$ por 1 hiperplano está acotada superiormente por $4n^{d+1}$.*

3. *Si tenemos a lo sumo k hiperplanos, podemos fragmentar n puntos $X_1,\ldots,X_n$ de a lo sumo $n^{(d+1)k}$ formas como se sigue de aplicar el punto 3 del Teorema 6.16 y la cota n^{d+1} que vimos antes. Por lo tanto la cantidad de formas de fragmentar n pares (X_i, Y_i) con reglas de clasificación constantes en las celdas de una partición de $\mathbb{R}^d$ por a lo sumo k hiperplanos está acotada superiormente por $2^k n^{(d+1)k}$.*

Una consecuencia importante de la observación anterior es que elegir la regla que minimiza $L_n^{(R)}$ es en algunos casos cercano a lo óptimo, para ciertas reglas basadas en particiones, posiblemente dependientes de la muestra. Para eso denotamos $\mathcal{C}_n$ a la clase de todas las reglas de clasificación construidas en base a particiones de $\mathbb{R}^d$ dependientes eventualmente de la muestra, usando a lo sumo k hiperplanos, como en el punto 3 de la observación anterior. Denotemos g_n^* a la regla en $\mathcal{C}_n$ que minimiza $L_n^{(R)}$ en $\mathcal{C}_n$ es decir

$$L_n^{(R)}(g_n^*) \leq L_n^{(R)}(g_n) \quad g_n \in \mathcal{C}_n. \tag{6.8}$$

De la cota (3.16) sabemos que

$$L(g_n^*) - \inf_{g_n \in \mathcal{C}_n} L(g) \leq 2 \sup_{g_n \in \mathcal{C}_n} |L_n^{(R)}(g_n) - L(g)|.$$

Por lo tanto procediendo igual que en el Teorema 6.25, si usamos el punto 3 de la observación anterior obtenemos el siguiente resultado

Corolario 6.27. *Sea $\mathcal{C}_n$ la clase de todas reglas de clasificación binaria basadas en particiones de $\mathbb{R}^d$* [8] *en a lo sumo k celdas, que son constantes en las celdas. Sea g_n^* como en* (6.8)*.*

$$\mathbb{P}\left(L(g_n^*) - \inf_{g_n \in \mathcal{C}_n} L(g_n) > \epsilon \right) \leq 8 \cdot 2^k (n+1)^{(d+1)k} \exp\left(\frac{-n\epsilon^2}{128} \right). \tag{6.9}$$

En particular si $k = o(n/\log(n))$, $L(g_n^*) - \inf_{g_n \in \mathcal{C}_n} L(g_n) \to 0$ en probabilidad.

[8] la partición puede depender de la muestra

7 Regresión por mínimos cuadrados

En este capítulo vamos a mostrar algunos resultados que son el equivalente de la teoría V.C. que vimos en el capítulo anterior, pero para el caso de regresión en lugar de clasificación. No vamos a ver demostraciones, las mismas pueden encontrarse en [15]. Consideremos $D_n = \{(X_1, Y_1), \ldots, (X_n, Y_n)\}$ iid de (X, Y). En el problema de estimación por mínimos cuadrados, minimizamos el riesgo empírico L_2:

$$\frac{1}{n}\sum_{i=1}^{n} |f(X_i) - Y_i|^2 , \tag{7.1}$$

sobre un conjunto de funciones $\mathcal{F}_n$ que depende de n, el tamaño de la muestra.

Si $X_1, \ldots, X_n$ son todos distintos (lo que ocurre con probabilidad 1 si X tiene densidad), entonces minimizar (7.1) sobre el conjunto de todas las funciones medibles, lleva a una estimación que interpola los datos $(X_1, Y_1), \ldots, (X_n, Y_n)$, y tiene riesgo empírico L_2 igual a 0. Obviamente, no será consistente en general.

Por lo tanto, primero se elige una clase "adecuada"de funciones $\mathcal{F}_n$ (quizás dependiendo de los datos, pero al menos dependiendo del tamaño de la muestra n) y luego se selecciona una función de esta clase que minimiza el riesgo empírico L_2, es decir, se define la estimación m_n por

$$m_n(\cdot) = \arg \min_{f \in \mathcal{F}_n} \frac{1}{n}\sum_{j=1}^{n} |f(X_j) - Y_j|^2,$$

lo que significa, por definición,

$$m_n \in \mathcal{F}_n \qquad \text{y} \qquad \frac{1}{n}\sum_{j=1}^{n} |m_n(X_j) - Y_j|^2 = \min_{f \in \mathcal{F}_n} \frac{1}{n}\sum_{j=1}^{n} |f(X_j) - Y_j|^2. \tag{7.2}$$

Aquí asumimos la existencia de funciones minimizantes, aunque no necesariamente su unicidad. En casos donde los mínimos no existen, el mismo análisis puede llevarse a cabo con funciones cuyo error es arbitrariamente cercano al ínfimo, pero, por simplicidad, mantenemos la suposición de existencia. Se puede ver que en la mayoría de las aplicaciones los mínimos de hecho existen.

La clase de funciones candidatas crece a medida que el tamaño de la muestra n crece. La elección de $\mathcal{F}_n$ tiene dos efectos sobre el error de la estimación. Por un lado, si $\mathcal{F}_n$ no es demasiado grande, entonces el riesgo empírico L_2 está próximo al riesgo L_2 teórico, uniformemente sobre $\mathcal{F}_n$. Es decir, el error introducido al minimizar el riesgo empírico L_2 en lugar del riesgo L_2 será pequeño. Por otro lado, debido al requisito de que nuestra estimación esté contenida en $\mathcal{F}_n$, no puede ser mejor (con respecto al error L_2) que la mejor función en $\mathcal{F}_n$. Esto se formula en el siguiente lema:

Lema 7.1. *Sea $\mathcal{F}_n = \mathcal{F}_n(D_n)$ una clase de funciones $f : \mathbb{R}^d \to \mathbb{R}$ dependiendo de los datos $D_n = \{(X_1, Y_1), \ldots, (X_n, Y_n)\}$. Si m_n satisface (7.2) entonces*

$$\int |m_n(x) - m^*(x)|^2 \mu(dx) \le 2 \sup_{f \in \mathcal{F}_n} \left| \frac{1}{n}\sum_{j=1}^{n} |f(X_j) - Y_j|^2 - \mathbb{E}\left\{|f(X) - Y|^2\right\} \right| \; + \; \inf_{f \in \mathcal{F}_n} \int |f(x) - m^*(x)|^2 \mu(dx), \tag{7.3}$$

donde $m^*(X) = \mathbb{E}(Y|X)$.

El lema anterior establece, al igual que en la teoría de V.C., que para demostrar la consistencia de un estimador que minimice (7.1), basta con probar dos cosas

1. el riesgo empírico L_2 es uniformemente (sobre $\mathcal{F}_n$) cercano al riesgo L_2. Es decir, el primer término de la ecuación (7.3) converge a 0.

2. el segundo término de (7.3) se va a 0, es decir, la familia $\mathcal{F}_n$ se hace densa en $L^2(\mu)$.

Para ver el primer objetivo sea $Z = (X, Y)$, $Z_i = (X_i, Y_i)$ $(i = 1, \ldots, n)$, $g_f(x, y) = |f(x) - y|^2$ para $f \in \mathcal{F}_n$ y $\mathcal{G}_n = \{g_f : f \in \mathcal{F}_n\}$. Entonces, la convergencia del primer término en (7.3) es equivalente a probar

$$\sup_{g\in\mathcal{G}_n}\left|\frac{1}{n}\sum_{i=1}^{n}g(Z_i)-\mathbb{E}\{g(Z)\}\right|\to 0\quad(n\to\infty)\quad\text{c.s.}\tag{7.4}$$

Sea $Z, Z_1, Z_2, \ldots$ una secuencia de variables aleatorias independientes e idénticamente distribuidas a valores en $\mathbb{R}^d$, y para $n\in\mathbb{N}$, sea $\mathcal{G}_n$ una clase de funciones $g:\mathbb{R}^d\to\mathbb{R}$. Veremos condiciones suficientes para (7.4).

Si $g:\mathbb{R}^d\to[0,B]$ con $B<\infty$, entonces, por la desigualdad de Hoeffding

$$\mathbb{P}\left\{\left|\frac{1}{n}\sum_{j=1}^{n}g(Z_j)-\mathbb{E}\{g(Z)\}\right|>\epsilon\right\}\le 2e^{-\frac{2n\epsilon^2}{B^2}},$$

por lo tanto, si $|\mathcal{G}_n|<\infty$,

$$\mathbb{P}\left\{\sup_{g\in\mathcal{G}_n}\left|\frac{1}{n}\sum_{j=1}^{n}g(Z_j)-\mathbb{E}\{g(Z)\}\right|>\epsilon\right\}\le 2|\mathcal{G}_n|e^{-\frac{2n\epsilon^2}{B^2}}.\tag{7.5}$$

Para clases finitas $\mathcal{G}_n$ que satisfacen

$$\sum_{n=1}^{\infty}|\mathcal{G}_n|e^{-\frac{2n\epsilon^2}{B^2}}<\infty\tag{7.6}$$

para todo $\epsilon>0$, ya que (7.4) se sigue de (7.5) y el lema de Borel-Cantelli. En las aplicaciones, (7.6) no se cumple porque $\mathcal{G}_n$ tiene cardinal infinito. Pero a veces es posible elegir conjuntos finitos $\mathcal{G}_{n,\epsilon}$ que satisfacen (7.6) y

$$\left\{\sup_{g\in\mathcal{G}_n}\left|\frac{1}{n}\sum_{j=1}^{n}g(Z_j)-\mathbb{E}\{g(Z)\}\right|>\epsilon\right\}\subset\left\{\sup_{g\in\mathcal{G}_{n,\epsilon}}\left|\frac{1}{n}\sum_{j=1}^{n}g(Z_j)-\mathbb{E}\{g(Z)\}\right|>\epsilon'\right\}.$$

para algún ϵ' dependiente de ϵ pero no de n. De la inclusión anterior se sigue que si la clase $\mathcal{G}_{n,\epsilon}$ verifica (7.6) se sigue (7.4). Para construir clases $\mathcal{G}_{n,\epsilon}$ adecuadas, vamos a introducir las siguientes definiciones

Definición 7.2. Sea $\epsilon>0$ y $\mathcal{G}$ un conjunto de funciones de $\mathbb{R}^d$ a $\mathbb{R}$. Un conjunto finito de funciones $g_1,\ldots,g_N:\mathbb{R}^d\to\mathbb{R}$ con la propiedad de que para cada $g\in\mathcal{G}$ existe un $j=j(g)\in\{1,\ldots,N\}$ tal que

$$\|g-g_j\|_\infty:=\sup_z|g(z)-g_j(z)|<\epsilon,$$

se llama un ϵ-cubrimiento de $\mathcal{G}$ con respecto a $\|\cdot\|_\infty$.

Definición 7.3. Sean $\epsilon>0$ y $\mathcal{G}$ un conjunto de funciones de $\mathbb{R}^d$ a $\mathbb{R}$. Definimos $\mathcal{N}(\epsilon,\mathcal{G},\|\cdot\|_\infty)$ la menor cantidad de funciones de un ϵ-cubrimiento de $\mathcal{G}$ con respecto a $\|\cdot\|_\infty$. Si no existe tal cubrimiento definimos $\mathcal{N}(\epsilon,\mathcal{G},\|\cdot\|_\infty)=\infty$. El número entero $\mathcal{N}(\epsilon,\mathcal{G},\|\cdot\|_\infty)$ se denomina **número de ϵ-cubrimiento** de $\mathcal{G}$ con respecto a $\|\cdot\|_\infty$ y se abrevia como $\mathcal{N}_\infty(\epsilon,\mathcal{G})$.

Observación 7.4. *Es claro que una definición análoga se podría haber hecho si en lugar de tomar la norma infinito se toma la norma $L^p(\nu)$, donde ν es una medida de probabilidad en $\mathbb{R}^d$. En este caso el ϵ-cubrimiento se denota $\mathcal{N}(\epsilon,\mathcal{G},\|\cdot\|_{L^p(\nu)})$.*

En virtud de estas definiciones, nuestras clases de funciones $\mathcal{G}_{n,\epsilon}$ van a ser las funciones de un $\epsilon/3$-cubrimiento de $\mathcal{G}_n$, como establece el siguiente resultado (ver Lema 9.1 en [15])

Lema 7.5. *Para $n\in\mathbb{N}$, sea $\mathcal{G}_n$ un conjunto de funciones $g:\mathbb{R}^d\to[0,B]$ y $\epsilon>0$. Entonces*

$$\mathbb{P}\left\{\sup_{g\in\mathcal{G}_n}\left|\frac{1}{n}\sum_{j=1}^{n}g(Z_j)-\mathbb{E}\{g(Z)\}\right|>\epsilon\right\}\le 2\mathcal{N}_\infty(\epsilon/3,\mathcal{G}_n)e^{-\frac{2n\epsilon^2}{9B^2}}.$$

Dado que en la probabilidad anterior el supremo se toma sobre un conjunto posiblemente no numerable, puede haber algunos problemas de medibilidad. En el libro de van der Vaart y Wellner (1996) [35], este problema aborda usando la noción de probabilidad exterior. En la mayoría de nuestras aplicaciones, bastará con considerar conjuntos numerables de funciones; por lo tanto, aquí y en lo sucesivo, ignoraremos completamente este problema.

Observación 7.6. *Si la clase de funciones del Lema anterior verifica*

$$\sum_{n=1}^{\infty}\mathcal{N}_\infty(\epsilon/3,\mathcal{G}_n)e^{-\frac{2n\epsilon^2}{9B^2}}<\infty,$$

obtenemos (7.4)*, como aplicación directa del Lema de Borel-Cantelli.*

El siguiente teorema relaciona la posibilidad de obtener una ley de los grandes números uniforme para toda función de una familia $\mathcal{G}$ de funciones, como en (7.4), con la dimensión V.C. de los conjuntos de nivel determinados por las funciones de $\mathcal{G}$. Si $\mathcal{G}$ es una clase de funciones $g : \mathbb{R}^d \to \mathbb{R}$ vamos a denotar estos conjuntos como

$$\mathcal{G}^+ = \Big\{ \big\{ (z,t) \in \mathbb{R}^d \times \mathbb{R} : t \leq g(z) \big\} : g \in \mathcal{G} \Big\}.$$

Del Teorema 6.20 se sigue que si $\mathcal{G}$ es un espacio vectorial de funciones de dimensión r, $V_{\mathcal{G}^+} = r + 1$.

El coeficiente de fragmentación de los conjuntos de nivel de una familia de funciones está relacionado con el ϵ-cubrimiento respecto de la norma L^p, como establece el siguiente Teorema

Teorema 7.7. *Sea $\mathcal{G}$ una clase de funciones $g : \mathbb{R}^d \to [0, B]$ con $V_{\mathcal{G}^+} \geq 2$, ν una medida de probabilidad en $\mathbb{R}^d$, y $0 < \epsilon < \frac{B}{4}$. Entonces*

$$\mathcal{N}\left(\epsilon, \mathcal{G}, \|\cdot\|_{L_p(\nu)}\right) \leq 3 \left(\frac{2eB^p}{\epsilon^p} \log \frac{3eB^p}{\epsilon^p} \right)^{V_{\mathcal{G}^+}},$$

donde $p \geq 1$.

Teorema 7.8. *Sea $\mathcal{G}$ una clase de funciones $g : \mathbb{R}^d \to \mathbb{R}$ y $G(x) := \sup_{g \in \mathcal{G}} |g(x)|$. Supongamos que $\mathbb{E}G(Z) < \infty$ y $V_{\mathcal{G}^+} < \infty$. Entonces*

$$\sup_{g \in \mathcal{G}} \left| \frac{1}{n} \sum_{i=1}^{n} g(Z_i) - \mathbb{E}g(Z) \right| \to 0 \quad (n \to \infty) \quad \textit{c.s.}$$

8 Redes Neuronales

8.1 El perceptrón de Rosenblatt

Vamos a empezar recordando la regla de clasificación binaria, g, basada en un hiperplano, que vimos en el capítulo 4,

$$g(x) = \begin{cases} 0 & \text{si } \psi(x) \leq 1/2 \\ 1 & \text{en caso contrario} \end{cases} \tag{8.1}$$

donde

$$\psi(x) = c_0 + \sum_{i=1}^{d} c_i x_i = c_0 + c^T x,$$

$x = (x_1, \ldots, x_d)$, y $c = (c_1, \ldots, c_d)$. Este clasificador (denominado perceptrón) es un ejemplo de red neuronal sin capas ocultas. Como mencionamos en el capítulo 4, si no hacemos hipótesis adicionales sobre la distribución de (X, Y) y hallamos los parámetros c_0 y c por medio de la minimización de $\hat{L}_n = \sum_i \mathbb{I}_{g_n(X_i) \neq Y_i}$ estamos siguiendo lo que en la introducción llamamos la cultura de los algoritmos.

Como vimos antes, no es universalmente consistente, como será el caso de las redes neuronales de una capa.

Antes de definir las redes con una capa oculta, vamos a llamar sigmoide a una función, que usualmente denotamos como $\sigma : \mathbb{R} \to \mathbb{R}$, y en general se toma no decreciente. Algunos ejemplos clásicos de sigmoides son

1. El sigmoide umbral
$$\sigma(x) = \begin{cases} -1 & \text{si } x \leq 0 \\ 1 & \text{si } x > 0 \end{cases}$$
2. El sigmoide logístico
$$\sigma(x) = \frac{1 - \exp(-x)}{1 + \exp(-x)}$$
3. La función ReLu (rectified linear unit), $\sigma_R(x) = \text{máx}\{0, x\}$.

Definición 8.1. Una **red neuronal** con una capa oculta y k neuronas es el clasificador g que se obtiene tomando en (8.1) la función

$$\psi(x) = c_0 + \sum_{i=1}^{k} c_i \sigma\big(\psi_i(x)\big). \tag{8.2}$$

Aquí las funciones ψ_i son combinaciones lineales que escribimos como

$$\psi_i(x) = b_i + \sum_{j=1}^{d} a_{ij} x_j \qquad i = 1, \ldots, k$$

La matriz que contiene los pesos a_{ij} (que usualmente se denotan a^1_{ij} para indicar que corresponden a los pesos de la primera capa), la denotamos como W^1. Tiene, por lo tanto, dimensiones $k \times d$.

$$\big(\psi_1(x), \ldots, \psi_k(x)\big)^T = b^1 + W^1 x$$

donde $b^1 = (b_1, \ldots, b_k)^T$. Las funciones $\sigma\big(\psi_i(x)\big)$ se llaman **neuronas**. Ver Figura 8.1

Ejemplo 8.2. Supongamos que tomamos en dimensión 2, $k = 3$, como en la figura 8.2. El plano queda dividido en, a lo sumo, 7 regiones. Si usamos el sigmoide umbral cada una de esas regiones se corresponde con una terna de 1 y -1, a la cual se le puede asociar un vértice en el cubo $[-1, 1]^3$. No necesariamente todo vértice tiene asociado una región, como se ve en la figura 8.2, el punto en negro tiene coordenadas $(1, -1, 1)$ y no tiene una región asociada. Si tomamos en (8.2) $c_0 = 1/2$, $c_1 = c_2 = c_3 = 1$, la regla $g(x) = \mathbb{I}_{\psi(x)>0}$ asignará la etiqueta 1 a las 3 regiones del plano cuyas ternas tienen 2 o mas unos (es decir 2 o mas clasificadores lineales ψ_i lo clasifican como 1). Si hay 2 o mas -1 lo clasifica como 0. Esto corresponde a separar en el hipercubo, con el hiperplano ortogonal al vector $(1, 1, 1)$. Otros valores de c_0, c_1, c_2 y c_3 corresponden a particiones de puntos en el hiperplano con otro plano.

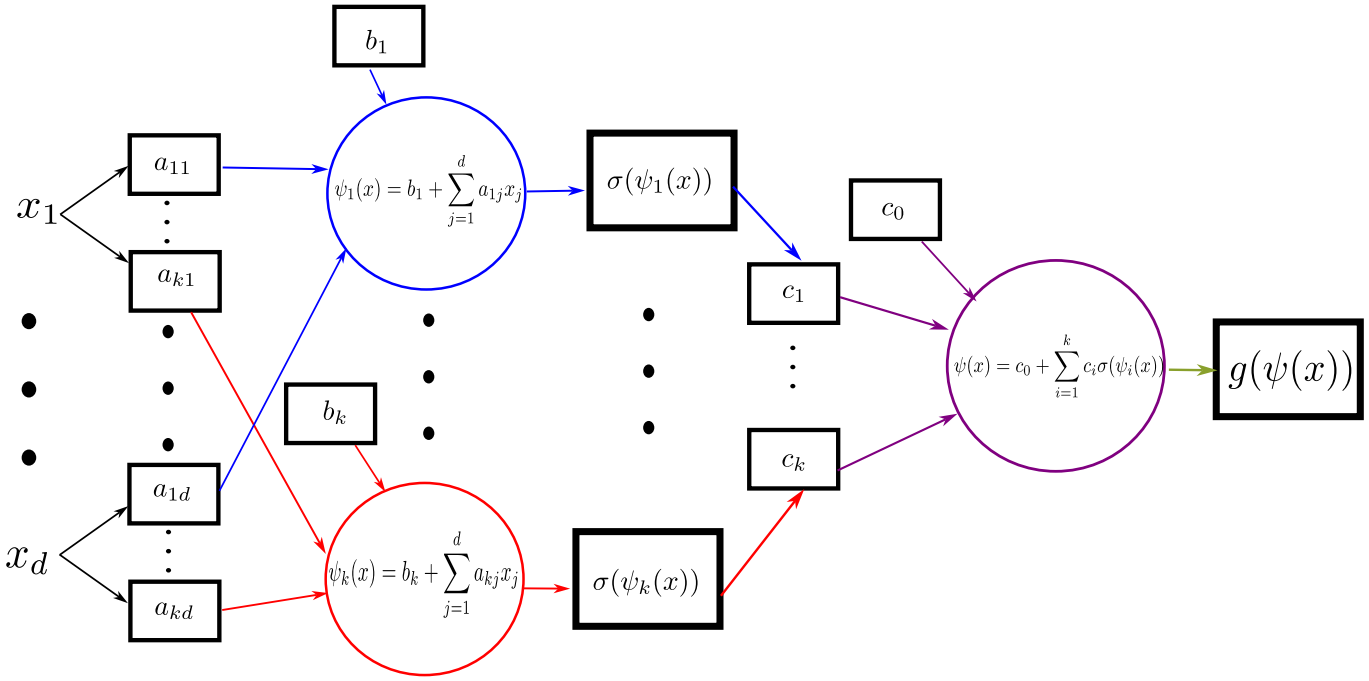

Figura 8.1: Red neuronal con una capa oculta, ver Definición 8.1

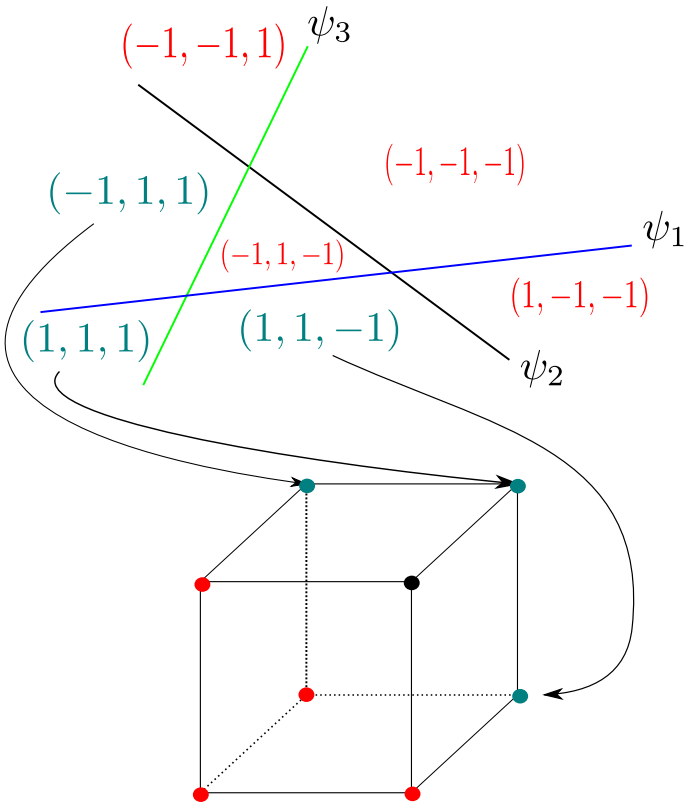

Figura 8.2: Los puntos en rojo serán etiquetados como 0 con la regla g si tomamos $c_0 = 1/2, c_1 = c_2 = c_3 = 1$. El punto en negro corresponde al $(1, -1, 1)$ y no tiene una región del plano asociada.

8.2 Redes con L capas

El procedimiento anterior se puede iterar, veamos como quedan las fórmulas para el caso de 2 capas. Para la regla ψ en (8.2) usamos

$$\psi\big((z_1,\dots,z_l)\big) = c_0 + \sum_{i=1}^{l} c_i z_i. \tag{8.3}$$

donde

$$z_i = \sigma\left(b_i^2 + \sum_{j=1}^{k} a_{ij}^2 u_j\right), \quad 1 \le i \le l,$$

y

$$u_i = \sigma\left(b_i^1 + \sum_{j=1}^{d} a_{ij}^1 x_j\right), \quad 1 \le i \le k.$$

Aquí decimos que tenemos k neuronas en la primera capa oculta, y l neuronas en la segunda capa oculta. Vamos a escribir esto como composición de funciones.

Supongamos que $x \in \mathbb{R}^d$, consideremos $\Psi_1 : \mathbb{R}^d \to \mathbb{R}^k$ tal que para $i = 1,\dots,k$, $(\Psi_1(x))_i = b_i^1 + \sum_{j=1}^{d} a_{ij}^1 x_j$. Entonces $u = \sigma(\Psi_1(x)) \in \mathbb{R}^k$, donde aplicar σ al vector $\Psi_1(x)$ significa que la aplicamos por coordenadas. La función $\sigma(\Psi_1)$ es la primera capa oculta.

Definimos $\Psi_2 : \mathbb{R}^k \to \mathbb{R}^l$ tal que para todo $i = 1,\dots,l$ $(\Psi_2(u))_i = b_i^2 + \sum_{j=1}^{k} a_{ij}^2 u_j$. Entonces

$$z = \sigma\Big(\Psi_2\big(\sigma(\Psi_1(x))\big)\Big) \in \mathbb{R}^l$$

La función $\sigma(\Psi_2)$ es la segunda capa oculta. Finalmente, aplicamos la función linea $\psi : \mathbb{R}^l \to \mathbb{R}$ dada en (8.3) y la regla de clasificación (8.1). Ver Figura 8.3.

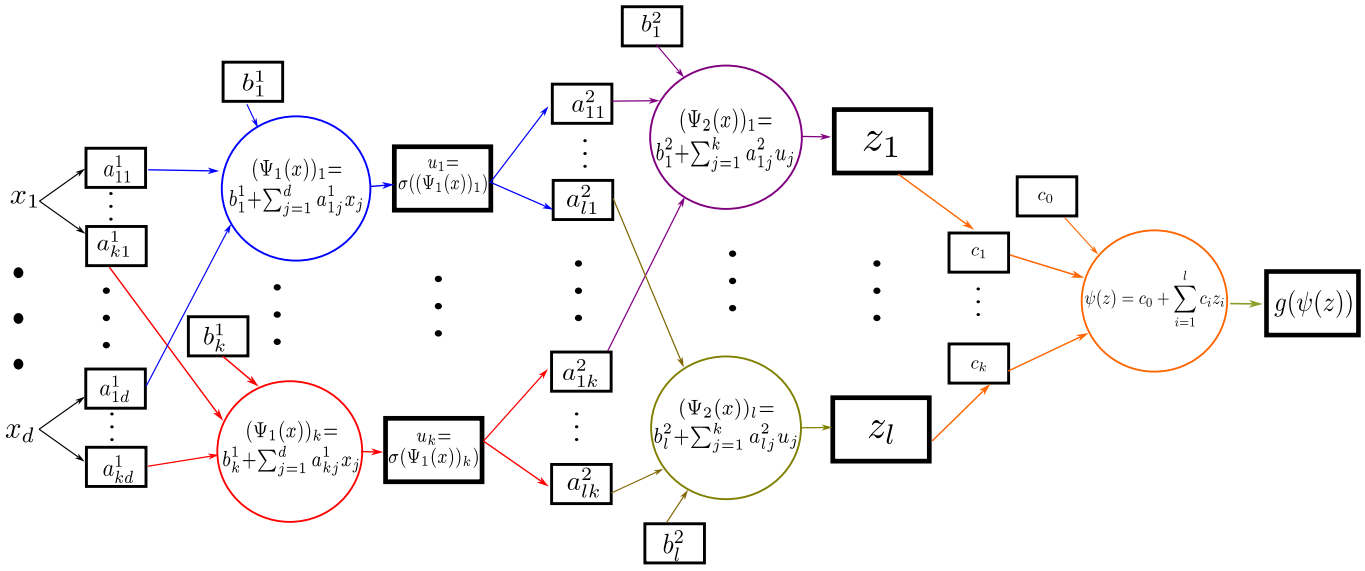

Figura 8.3: Red neuronal con dos capas ocultas.

Las redes así construidas se llaman en inglés "feed-fordward" (propagación hacia adelante), más adelante veremos el caso general de este tipo de redes. El nombre proviene del hecho de que el "output" $y = g(x)$ se obtiene como composición sucesiva de funciones, por lo tanto en el paso siguiente se usa el valor anterior. Si estamos en el caso de clasificación binaria, y tenemos m capas, la red neuronal es la función g como en (8.1), que se obtiene tomando $\psi = c_0 + \sum_{i=1}^{r} c_i u_i$ donde r es la cantidad de neuronas de la última capa, y $u = (u_1,\dots,u_r)$ se escribe de manera compacta como

$$u(x) = f^m\Big(f^{m-1}\Big(f^{m-2}\cdots f^1(x)\Big)\Big),$$

donde $f^m,\dots,f^1$ son m funciones que dependen de parámetros. La función f^1 es la primera capa oculta, y de forma genérica se escribe como $\sigma(b^1 + W^1 x)$ para cierta matriz de pesos W^1 y vector b^1. f^2 es la segunda capa oculta, cuya expresión es de la forma $\sigma(b^i + W^2 z)$ donde z es el vector que contiene las salidas de la capa anterior (es decir la salida de f^1), y así sucesivamente.

Es claro que la cantidad de parámetros a determinar es $(d+1)\times k$ para la primera capa, $(k+1)\times l$ para la segunda capa, y finalmente, $k+1$ para la ψ dada en (8.3). En general la cantidad de parámetros va a ser mucho más grande

que la cantidad de datos. Más adelante veremos como estimar los parámetros y un resultado que establece que las redes neuronales con una sola capa oculta son consistentes, cuando k crece de manera adecuada.

8.3 La función XOR y el sigmoide ReLu

Los sigmoides acotados tienen el problema de que valores grandes de x los transforman en valores muy próximos de $\sigma(x)$. Esto se conoce como fenómeno de saturación, y conlleva a problemas a la hora de estimar los parámetros. Intuitivamente, lo que sucede es que, como usualmente los parámetros se estiman por el método del gradiente estocástico, el cual computa ∇L_n, entonces, si los valores de σ está muy próximos entre si el gradiente de L_n va a ser próximo a 0 en estos casos. Para evitar esto se usa el sigmoide ReLu, que denotaremos como σ_R, y es $\sigma_R(x) = \max\{x, 0\}$. Veamos a modo de ejemplo como queda una red neuronal de una capa, para aproximar la función XOR. Denotemos $\mathbb{X} = \big\{(0,0),(1,1),(0,1),(1,0)\big\}$, recordemos que la función $\mathrm{XOR} : \mathbb{X} \to \{0,1\}$ y vale $\mathrm{XOR}((x,y)) = 1$ si $x \neq y$, y 0 en otro caso. Quisieramos minimizar en θ,

$$\frac{1}{4}\sum_{x\in\mathbb{X}} \big(\mathrm{XOR}(x) - f(x,\theta)\big)^2.$$

Es fácil ver que no se puede aproximar XOR con error 0 usando en (8.1) una función ψ lineal. Vamos a tomar una ψ como en (8.2) con σ la función ReLu y $k = 2$. Tomamos $\psi_1((x_1,x_2)) = x_1 + x_2$, $\psi_2((x_1,x_2)) = x_1 + x_2 - 1$, $c_0 = 0$, $c_1 = 1$ y $c_2 = -2$. Por lo tanto nos queda

$$\psi\big((x_1,x_2)\big) = \max\big\{0, x_1 + x_2\big\} - 2\max\big\{0, x_1 + x_2 - 1\big\}.$$

8.3.1 Otras redes que no veremos

Existen muchas variantes de las redes neuronales que veremos aquí (que son las que se denominan "feed forward"). Las redes feewforward (ver [16]) son la forma más simple de red neuronal. Su nombre se debe a que la información se mueve en una sola dirección, desde la entrada hasta la salida, sin ciclos ni retroalimentación. Son adecuadas para tareas de clasificación y regresión. Vamos a mencionar brevemente otras.

Redes Neuronales Recurrentes (RNN)

Están diseñadas para trabajar con datos secuenciales. Tienen conexiones hacia atrás, lo que permite que la información persista y se use en futuras etapas. Comúnmente usadas en el procesamiento de lenguaje natural (NLP), series temporales y reconocimiento de voz. Variantes incluyen LSTM (Long Short-Term Memory) y GRU (Gated Recurrent Unit), que mejoran la capacidad de las RNN para manejar dependencias a largo plazo. Ver, por ejemplo, [10].

Redes Neuronales de Retroalimentación (Feedback Neural Networks)

Incluyen una red completamente conectada donde las salidas pueden circular de nuevo como entradas. Utilizadas en aplicaciones como redes autoasociativas y modelos de memoria. Ver, por ejemplo, [17]

Redes de Crecimiento y Competición (GNG)

Son redes adaptativas que crecen dinámicamente para ajustar su estructura a los datos de entrada. Utilizadas en clustering y análisis de datos no supervisados. Ver, por ejemplo, [12].

Redes de Kohonen (Mapas Autoorganizados, SOM)

Redes neuronales no supervisadas que proyectan datos de alta dimensión en un mapa de menor dimensión. Útiles para la visualización de datos y reducción de dimensionalidad. Ver, por ejemplo, [20]

Redes de Boltzmann (Boltzmann Machines)

Redes estocásticas que pueden aprender una distribución de probabilidad sobre sus entradas. Incluyen variantes como las Restricted Boltzmann Machines (RBM) y las Deep Belief Networks (DBN). Ver, por ejemplo, [1]

Redes Neuronales de Función de Base Radial (RBF)

Utilizan funciones de base radial como funciones de activación. Adecuadas para problemas de clasificación, regresión y control. Ver, por ejemplo, [7]

Redes Neuronales de Memoria Asociativa

Diseñadas para almacenar y recuperar patrones basados en su similitud con patrones de entrada. Ejemplos incluyen redes de Hopfield y redes de contenido direccional. Ver, por ejemplo, [18]

Redes Generativas Adversariales (GAN)

Compuestas por dos redes (generador y discriminador) que compiten entre sí. Utilizadas para la generación de datos sintéticos, incluyendo imágenes y audio. Ver, por ejemplo, [14].

Redes Neuronales Modulares

Consisten en varios módulos que pueden ser entrenados de manera independiente y luego combinados. Útiles para problemas complejos que pueden ser descompuestos en subproblemas más simples. Ver, por ejemplo, [19].

8.4 Particiones aleatorias con hiperplanos y redes neuronales

La regla de clasificación de la Figura 8.2 le asigna a un x en la partición la etiqueta 0 si más de la mitad de los clasificadores ψ_i lo clasifican como 0, y asigna la etiqueta 1 en caso contrario. Otra posibilidad que vimos en el capítulo anterior, si tenemos una muestra $D_n = \{(X_1, Y_1), \ldots, (X_n, Y_n)\}$, es hacer voto mayoritario en la celda que contiene a x. Después veremos como escribir esta regla como una red neuronal de dos capas. Veamos que esta regla es consistente.

8.4.1 Consistencia de las redes con 2 capas

Consideremos $\mathcal{C}_n$ el conjunto de todas las reglas de clasificación definidas en $\mathbb{R}^d$ que se obtienen al variar k hiperplanos y que son constantes en cada partición que queda determinada. El Corolario (6.27) muestra que si tomamos $k = o(n/\log(n))$, $L(g_n^*) - \text{ínf}_{g_n \in \mathcal{C}_n} L(g_n) \to 0$ en probabilidad, siendo g_n^* la regla que se obtiene al minimizar en $\mathcal{C}_n$ el error de resustitución, que recordemos que es

$$L_n^{(R)} = \frac{1}{n} \sum_{i=1}^{n} \mathbb{I}_{\{g_n(X_i) \neq Y_i\}}.$$

Como la clase $\mathcal{C}_n$ contiene en particular la regla de histogramas cúbicos de la subsección 4.3.1, g_n^H, que se obtiene si hacemos una grilla con los k hiperplanos paralelos a los ejes [1], y por el Teorema 4.8 esta regla es universalmente consistente, es decir $\mathbb{E}(L_n(g_n^H)) = L(g_n^H) \to L^*$, tenemos que $\text{ínf}_{g_n \in \mathcal{C}_n} L(g_n) \to L^*$. Con lo cual $L(g_n^*) \to L^*$ en probabilidad. Observar que las redes neuronales con el sigmoide umbra son constantes en cada elemento de la partición (ver (8.2)).

8.4.2 De particiones a redes neuronales de 2 capas

Veamos que la regla de clasificación que, dada una partición de $\mathbb{R}^d$ en k hiperplanos, constante en cada elemento de la partición que determinan estos hiperplanos, y cuya constante se elije por voto mayoritario de las etiquetas Y_i de los puntos de la muestra que cayeron en ese elemento de la partición, se puede escribir como una red neuronal con 2 capas.

Si tenemos k hiperplanos en $\mathbb{R}^d$, definidos por k funciones lineales $\psi_1, \ldots, \psi_k$ como antes, podemos tomar

$$b(x) = \left(\mathbb{I}_{\{\psi_1(x)>0\}}, \ldots, \mathbb{I}_{\{\psi_k(x)>0\}}\right) \in \{0,1\}^k.$$

A cada región i le podemos asignar el vector $b^i = b(x)$. Vamos a denotar r al número de regiones. Observar que r es menor o igual que 2^k. Para cada b^i veamos que podemos definir un vector $c^i = (c_1^i, \ldots, c_k^i)$ y una constante c_0^i, que cumplan

$$\langle c^i, b^i \rangle + c_0^i > 0 \qquad \text{y} \qquad \langle c^i, b^j \rangle + c_0^i < 0 \quad \text{para todo } \ j \neq i. \tag{8.4}$$

Una forma geométrica de ver que para cada b^i el vector c^i y la constante c_0^i existen, es observar que cada punto b^i es un vértice del hipercubo $[0,1]^k$ (no necesariamente a todos los vértices de dicho hipercubo le corresponde un b^i). Para cada b^i se puede tomar un espacio afín [2], tal que uno de los dos semiespacios que este espacio afín define contienen a b^i y no contienen a los otros vértices del hipercubo. Estos espacios afines determinan ciertos c^i y c_0^i (para cada b^i hay infinitos posibles vectores c^i y valores c_0^i).

Tomemos σ la función sigmoide umbral. Definimos para cada región i, el vector $z \in \mathbb{R}^r$,

$$z_i = \sigma\Big(\langle c^i, b^i \rangle + c_0^i\Big) \quad \text{y} \quad z_j = \sigma\Big(\langle c^i, b^j \rangle + c_0^i\Big) \text{ si } j \neq i$$

Es decir, por cada región i tenemos un vector de r coordenadas $z = (z_1, \ldots, z_r)$ donde $z_i = 1$ y $z_j = -1$ en las otras coordenadas.

[1] podemos tomar por ejemplo $k = n/\log(n)^2$ y partir $[-\log(n), \log(n)]^d$ en k hiperplanos

[2] una traslación de un espacio vectorial $k-1$ dimensional

El objetivo ahora es definir un vector de pesos $w = (w_1, \dots, w_r) \in \{-1, 1\}^r$ y $w_0 \in \{-1, 1\}$, tal que para todos los r vectores z que definimos antes, se verifique

$$\sigma\big(\langle z, w\rangle + w_0\big) = \begin{cases} 1 & \text{si la etiqueta de la región } i \text{ fue } 1 \\ -1 & \text{si la etiqueta de la región } i \text{ fue } 0 \end{cases} \tag{8.5}$$

Es decir, si tenemos x en la primera capa de la red calculamos $b^i(x)$ y luego $z = z(b^i(x))$ como antes. En la segunda capa asignamos una etiqueta aplicando la función $\sigma(\langle z, w\rangle + w_0)$.

Supongamos que en la partición hay s regiones que votaron 1 y t regiones que votaron 0. Definimos $w_0 = 1 + s - t$. Si en la región l el voto fue 1 definimos $w_l = 1$, si en la región l el voto fue 0, definimos $w_l = -1$. Sea ahora x en la región i y $z \in \mathbb{R}^r$ el vector correspondiente que definimos antes, es decir, $z_i = 1$ y $z_j = -1$ para todo $i \neq j$ con lo cual

$$\langle z, w\rangle + w_0 = w_i - \sum_{j \neq i} w_j + 1 + s - t.$$

Si $w_i = 1$ (el voto mayoritario en la región es 1) la ecuación anterior es $1 - ((s-1) - t) + s - t + 1 = 3 > 0$ por lo tanto vale (8.5) mientras que si $w_i = -1$ la ecuación anterior es $-1 - (s - (t-1)) + s - t + 1 = -1$ y también vale (8.5).

Esto prueba que regla de clasificación basada en una partición se puede ver como una red neuronal de dos capas con sigmoides de tipo umbral. Observar que esta regla tiene en total $k + 2^k$ neuronas, ya que son k en la primera capa para determinar la posición del punto en la partición y 2^k para la etiqueta final que se obtiene con (8.5). En la sección que sigue veremo su que las redes neuronales con 1 capa oculta son universalmente consistentes.

8.4.3 Consistencia de las redes con 1 capa

Llamemos $\mathcal{C}_2^k$ a las reglas de clasificación que se obtienen usando (8.1) con (8.3) y en total k neuronas. Si tenemos m hiperplanos, vimos que hacer voto mayoritario en las celdas da como resultado una regla en $\mathcal{C}_2^k$ si $k = m + 2^m$ (ya que en la primera capa de la red aplicamos la función b). Esta regla es consistente si $m \to \infty$ adecuadamente como vimos en la subsección 8.4.1. Estas reglas están basadas en redes neuronales con dos capas ocultas. Veremos que lo mismo vale si tenemos k neuronas con una sola capa oculta.

Denotemos $\mathcal{C}_1^k$ las reglas de clasificación (8.1), es decir $g(x) = \mathbb{I}_{\{m(x)>1/2\}}$ y usando (8.2), es decir k neuronas y 1 capa oculta. Por (3.12) tenemos que

$$L(g) - L^* \leq 2\mathbb{E}\big|m(X) - m^*(X)\big|,$$

donde $m^*(x) = \mathbb{P}(Y = 1 | X = x)$. Por lo tanto $\inf_{g \in \mathcal{C}_1^k} L(g) \to L^*$ cuando $k \to \infty$ si existe $\{m_k\}_k$ con $g_k(x) = \mathbb{I}_{\{m_k(x)>1/2\}} \in \mathcal{C}_1^k$ tal que

$$\mathbb{E}\big|m_k(X) - m^*(X)\big| \to 0 \text{ cuando } k \to \infty.$$

Es decir tenemos que probar que la familia de funciones (8.2), al variar k, es densa en $L_1(\mu)$ para toda μ. El siguiente lema prueba que una condición más fuerte es aproximar m^* uniformemente en cubos $[a, b]^d \subset \mathbb{R}^d$ para todo $a, b \in \mathbb{R}$.

Lema 8.3. *Supongamos que $\mathcal{M}_k$ es una familia de funciones tal que para toda f continua, para todo $a, b \in \mathbb{R}$,*

$$\lim_{k\to\infty} \inf_{m \in \mathcal{M}_k} \sup_{x \in [a,b]^d} \big|m(x) - f(x)\big| = 0.$$

Entonces, para toda distribución de (X, Y),

$$\lim_{k\to\infty} \inf_{g \in \mathcal{C}^k} L(g) - L^* = 0,$$

donde $\mathcal{C}^k$ es la clase de reglas de clasificación de la forma $g(x) = \mathbb{I}_{\{m(x)>1/2\}}$ con $m \in \mathcal{M}_k$

Demostración. Sea $\varepsilon > 0$ fijo, tomemos a, b suficientemente grandes tal que $\mu([a, b]^d) \geq 1 - \varepsilon/3$ siendo μ la distribución de X. Sea $\hat{m}$ continua tal que $\hat{m}(x) = 0$ si $x \notin [a, b]^d$, y

$$\mathbb{E}|m^*(X) - \hat{m}(X)| < \varepsilon/6$$

Sea k y $m \in \mathcal{M}_k$ tal que

$$\sup_{x \in [a,b]^d} |m(x) - \hat{m}(x)| < \varepsilon/6.$$

Por lo tanto si $g(x) = \mathbb{I}_{\{m(x)>1/2\}}$, y denotamos, como siempre $g^*(x) = \mathbb{I}_{\{m^*(x)>1/2\}}$, si acotamos $2|m^*(x) - 1/2| \leq 1$,

$$\begin{aligned}
L(g) - L^* &= 2\int_{\mathcal{X}} \big|m^*(x) - 1/2\big| \mathbb{I}_{\{g(x)\neq g^*(x)\}} P_X(dx) \\
&\leq 2\int_{[a,b]^d} \big|m^*(x) - 1/2\big| \mathbb{I}_{\{g(x)\neq g^*(x)\}} P_X(dx) + \underbrace{\int_{([a,b]^d)^c} \mathbb{I}_{\{g(x)\neq g^*(x)\}} P_X(dx)}_{\leq \mu(([a,b]^d)^c) \leq \varepsilon/3} \\
&\leq 2\int_{[a,b]^d} \big|m^*(x) - 1/2\big| \mathbb{I}_{\{g(x)\neq g^*(x)\}} P_X(dx) + \varepsilon/3 \\
&\leq 2\int_{[a,b]^d} \big|m^*(x) - m(x)\big| P_X(dx) + \varepsilon/3 \\
&\leq 2\mathbb{E}\big|m^*(X) - \hat{m}(X)\big| + 2\mathbb{E}\Big[\big|\hat{m}(X) - m(X)\big| \mathbb{I}_{X\in[a,b]^d}\Big] + \varepsilon/3 \\
&\leq 2 \sup_{x\in[a,b]^d} \big|m(x) - \hat{m}(x)\big| + 2\varepsilon/3 < \varepsilon
\end{aligned}$$

en la cuarta desigualdad usamos, al igual que en la prueba del Teorema 3.12, que $g(x) \neq g^*(x)$ implica que $\big|m^*(x) - 1/2\big| \leq \big|f(x) - m^*(x)\big|$. □

Teorema 8.4. *Para toda $f : [a,b]^d \to \mathbb{R}$ continua, para todo $\varepsilon > 0$, existe una red neuronal con una capa oculta, y ψ como en* (8.2) *tal que*

$$\sup_{x\in[a,b]^d} \big|f(x) - \psi(x)\big| < \varepsilon. \tag{8.6}$$

Demostración. Vamos a probar el resultado para el sigmoide umbral,

$$\sigma(x) = \begin{cases} -1 & \text{si } x \leq 0 \\ 1 & \text{si } x > 0 \end{cases}.$$

La extensión a cualquier sigmoide queda como ejercicio. La idea es primero aproximar f por una combinación finita de sin y cos. Usando la aproximación de Fourier existen un entero M, números reales no nulos $\alpha_1, \ldots, \alpha_M$, $\beta_1, \ldots, \beta_M$, y vectores de $\mathbb{R}^d$, $m_1, \ldots, m_M$ tal que

$$\sup_{x\in[a,b]^d} \left| \sum_{i=1}^{M} \left\{ \alpha_i \cos\left(\frac{\pi}{a} m_i^T x\right) + \beta_i \sin\left(\frac{\pi}{a} m_i^T x\right) \right\} - f(x) \right| < \frac{\varepsilon}{2}. \tag{8.7}$$

Consideremos las $2M$ funciones de variable real $\alpha_i \cos(\pi u/a)$ y $\beta_i \sin(\pi u/a)$. Primero vamos a aproximar cada una de ellas por con redes neuronales de una capa, definidas en $\mathbb{R}$, es decir con funciones de la forma

$$\sum_{i=1}^{k} c_i \sigma(a_i u + b_i) + c_0 \quad u \in \mathbb{R} \tag{8.8}$$

Para eso, observemos que, para $u \in \mathbb{R}$,

$$\mathbb{I}_{[b,c]}(u) = \frac{1}{2}\big(\sigma(u - b) + \sigma(-u + c)\big).$$

Por lo tanto, cualquier combinación lineal finita $\sum a_i \mathbb{I}_{C_i}$ para C_i intervalos de $\mathbb{R}$ se puede escribir en la forma (8.8). Luego aproximamos las $2M$ funciones de variable real $\alpha_i \cos(\pi u/a)$ y $\beta_i \sin(\pi u/a)$ por $2M$ combinaciones lineales finitas de la forma $\sum a_i \mathbb{I}_{C_i}$ para C_i intervalos de $\mathbb{R}$ y escribimos estas combinaciones como redes neuronales de una capa de la forma (8.8). Finalmente, cambiamos en cada una de las $2M$ redes de la forma (8.8) la variable u por la correspondiente variable $m_i^T x$ con $x \in \mathbb{R}^d$. Con lo cual obtenemos redes neuronales en $\mathbb{R}^d$. Llamemos $u_1, \ldots, u_M$ y $v_1, \ldots, v_M$ estas funciones. Podemos tomarlas de modo que

$$\sup_{x\in[a,b]^d} \left| u_i(m_i^T x) - \cos\left(\frac{\pi}{a} m_i^T x\right) \right| \leq \frac{\varepsilon}{4M|\alpha_i|} \qquad i = 1, \ldots, M$$

y

$$\sup_{x\in[a,b]^d} \left| v_i(m_i^T x) - \sin\left(\frac{\pi}{a} m_i^T x\right) \right| \leq \frac{\varepsilon}{4M|\beta_i|} \qquad i = 1, \ldots, M.$$

Por lo tanto de la desigualdad triangular,

$$\sup_{x\in[a,b]^d} \left| \sum_{i=1}^{M} \left\{ \alpha_i \cos\left(\frac{\pi}{a} m_i^T x\right) + \beta_i \sin\left(\frac{\pi}{a} m_i^T x\right) \right\} - \sum_{i=1}^{M} \alpha_i u_i(m_i^T x) + \beta_i v_i(m_i^T x) \right| < \frac{\varepsilon}{2}. \tag{8.9}$$

Definimos

$$\psi(x) = \sum_{i=1}^{M} \alpha_i u_i(m_i^T x) + \beta_i v_i(m_i^T x).$$

De (8.7) y (8.9) y de la desigualdad triangular se sigue (8.6) □

Como corolario inmediato del lema y teorema anterior tenemos que las reglas $g(x) = \mathbb{I}_{\psi(x)>0}$ con ψ una red neuronal con una capa, como en (8.2), son universalmente consistentes si $k \to \infty$.

Corolario 8.5. *Denotemos $\mathcal{C}^k$ las reglas de clasificación* (8.1) *con ψ una red neuronal con una capa oculta, y k nodos, como* (8.2)*. Entonces, para toda distribución (X, Y),*

$$\lim_{k\to\infty} \inf_{g\in\mathcal{C}^k} L(g) - L^* = 0.$$

8.4.4 Dimensión de VC de las redes neuronales

Denotemos $\mathcal{C}^k$ las reglas de clasificación (8.1) con ψ una red neuronal con una capa oculta, y k nodos, como (8.2). Por el Corolario 6.13, y usando la cota de $s(\mathcal{A}, n)$ del Teorema 6.24 sabemos que

$$\mathbb{E}\big(L(g_n^*)\big) - \inf_{g\in\mathcal{C}^k} L(g) \leq \sqrt{\frac{V_{\mathcal{C}^k}\log(n)+4}{n}}$$

para $n > 2V_{\mathcal{C}^k}$, donde acotamos $\log(8e) < 4$.

Vamos a ver algunos cotas de $V_{\mathcal{C}^k}$.

Teorema 8.6. *Sea $\mathcal{C}^k$ como antes*

$$V_{\mathcal{C}^k} \geq 2\left\lfloor \frac{k}{2} \right\rfloor d.$$

Demostración. Vamos a probarlo para el caso del sigmoide umbral, el caso general queda como ejercicio. Tenemos que probar que podemos fragmentar completamente $n = 2\lfloor k/2 \rfloor d$ puntos con conjuntos de la forma $\{x : \psi(x) > 1/2\}$. Observar que como las redes con una capa y k nodos incluyen como caso particular las de una capa y $k-1$ nodos[3] tenemos que $V_{\mathcal{C}^k} \geq V_{\mathcal{C}^{k-1}}$ para todo k. Veamos que de esto se sigue que podemos suponer que k es par. Supongamos que lo hemos probado para los números pares y ahora tenemos un k es impar, $k-1$ es par, con lo cual,

$$V_{\mathcal{C}^k} \geq V_{\mathcal{C}^{k-1}} \geq (k-1)d = 2\left\lfloor \frac{k}{2} \right\rfloor d$$

Por lo tanto lo tendríamos probado para k impar.

Supondremos entonces que k es par, en cuyo caso $n = kd$. Veremos que podemos fragmentar completamente cualquier conjunto $\aleph_n = \{x_1, \ldots, x_n\}$ con $n = kd$ puntos, que tengan la propiedad de que no hay $d+1$ puntos en el mismo hiperplano de dimensión $d-1$.[4] Para esto basta encontrar para cada subconjunto $S \subset \aleph_n$, una función ψ, que dependerá de S, tal que $\psi(x_i) > 1/2$ si y solo si $x_i \in S$.

Veamos que podemos suponer que el cardinal de S es a lo sumo $n/2$. Supongamos que probamos que podemos elegir subconjuntos con cardinal menor o igual que $n/2$, si ahora S tiene cardinal mayor que $n/2$, tomamos ψ que elija los del complemento (es decir $\psi(x_i) > 1/2$ si y solo si $x_i \in S^c$), pero entonces la red neuronal $1/2 - \psi$ elije los de S ya que $1/2 - \psi(x_i) > 0$ para todo $x_i \in S$.

Partimos los puntos de S en subconjuntos de a lo sumo d puntos (en total tenemos a lo sumo $k/2$ grupos ya que $n/2 = (k/2)d$). Para cada uno de estos grupos de a lo sumo d puntos podemos tomar un hiperplano de la forma $a^T x + b = 0$ que los contenga, ver Figura 8.4. Además existe h tal que $a^T x_i + b \in (-h, h)$ si y solo si x_i está en ese subgrupo (ya que estamos suponiendo que no hay $d+1$ puntos en el mismo hiperplano). Es decir la red

$$\sigma(a^T x + b + h) + \sigma(-a^T x - b + h)$$

vale 2 si x_i está en el grupo y 0 en caso contrario.

Para cada uno de estos (a lo sumo) $k/2$ subgrupos denotemos $a_1, \ldots, a_{k/2}$, $b_1, \ldots, b_{k/2}$ y $h_1, \ldots, h_{k/2}$ los parámetros asociados y $h = \min_{j\leq k/2} h_j$. La función

$$\psi(x) = \sum_{j=1}^{k/2} \left(\sigma\Big(a_j^T x + b_j + h\Big) + \sigma\big(-a_j^T x - b_j + h\big) \right) - 1/2$$

es 2 en exactamente los x_i del subconjunto S y $-1/2$ en los $x_i \in S^c$. Esta red tiene a lo sumo k neuronas y una capa oculta. □

[3] basta con hacer 0 los coeficientes de uno de los nodos

[4] si elegimos n puntos al azar que vengan de una distribución con densidad, esto ocurre con probabilidad 1

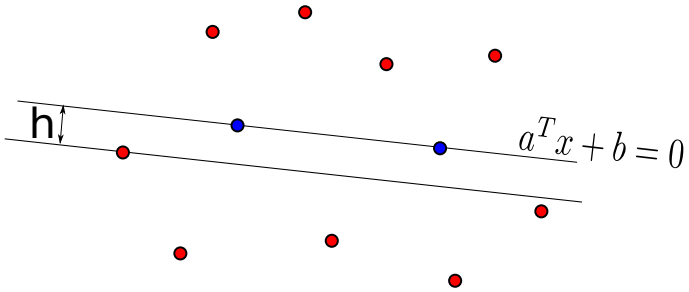

Figura 8.4: Los puntos del subconjunto se representan en azul, y la distancia h que permite separar esos puntos del resto.

El siguiente resultado da una cota superior para $V_{\mathcal{C}^k}$.

Teorema 8.7. *Sea σ el sigmoide umbral y $\mathcal{C}^k$ las reglas de clasificación* (8.1) *con ψ una red neuronal con una capa oculta, y k neuronas, como* (8.2)*. Entonces*

$$\begin{aligned} s(\mathcal{C}^k, n) \leq & \left(\sum_{i=1}^{d+1} \binom{n}{i} \right)^k \left(\sum_{i=1}^{k+1} \binom{n}{i} \right) \\ \leq & \left(\frac{ne}{d+1} \right)^{k(d+1)} \left(\frac{ne}{k+1} \right)^{k+1} \leq (ne)^{kd+2k+1}. \end{aligned} \tag{8.10}$$

Demostración. La segunda desigualdad se sigue de la segunda desigualdad en el Teorema 6.24. Veamos la primera, fijemos $x_1, \ldots, x_n \in \mathbb{R}^d$, tenemos que acotar el número de diferentes d-uplas $(g(x_1), \ldots, g(x_n))$ cuando $g \in \mathcal{C}^k$. Usando el Corolario 6.21, la dimensión de V.C. de 1 hiperplano en $\mathbb{R}^d$ es $d+1$ [5]. Por el Teorema 6.15, tenemos entonces que la cantidad de formas en que podemos fragmentar n puntos con un hiperplano esta acotada por

$$\sum_{i=1}^{d+1} \binom{n}{i}.$$

Como tenemos k hiperplanos, usando el punto 3 del Teorema 6.16, podemos fragmentar esos n puntos de

$$\left(\sum_{i=1}^{d+1} \binom{n}{i} \right)^k$$

formas. A lo sumo podemos fragmentar los n puntos. Finalmente nos queda la combinación lineal $c_0 + \sum_{i=1}^k c_i \sigma(\psi_i(x))$ para fragmentarlos, nuevamente usando el Teorema 6.15, se pueden fragmentar de a lo sumo.

$$\sum_{i=0}^{k+1} \binom{n}{i}$$

formas. En total obtenemos a lo sumo

$$\left(\sum_{i=1}^{d+1} \binom{n}{i} \right)^k \left(\sum_{i=1}^{k+1} \binom{n}{i} \right)$$

formas de fragmentar los n puntos. □

El siguiente corolario es consecuencia inmediata del teorema anterior, en particular prueba que a menos de un factor logarítmico, $(\log(kd))$ la cota inferior que obtuvimos en el Teorema 8.6 es óptima.

Corolario 8.8. *En las hipótesis del Teorema anterior*

$$V_{\mathcal{C}^k} \leq 2(kd + 2k + 1) \log_2 \left(e(kd + 2k + 1) \right)$$

Demostración. Se sigue de que $V_{\mathcal{C}^k} \leq n$ si $s(\mathcal{C}^k, n) \leq 2^n$ sustituyendo n por $2(kd + 2k + 1) \log_2(e(kd + 2k + 1))$ y usando $s(\mathcal{C}^k, n) \leq (ne)^{kd+2k+1}$. □

Teorema 8.9. *Sea σ el sigmoide umbral y $\mathcal{C}^k$ las reglas de clasificación* (8.1) *con ψ una red neuronal con una capa oculta, y k neuronas, como* (8.2)*. Sea $g_n^* \in \mathcal{C}^k$ que minimiza*

$$\hat{L}_n(g) = \sum_{i=1}^n \mathbb{I}_{\{g(X_i) \neq Y_i\}}$$

[5] es decir considerar las fragmentaciones posibles con conjuntos de la forma $\{x : \langle a, x \rangle + b > 0, a \in \mathbb{R}^d, b \in \mathbb{R}\}$

sobre $g \in \mathcal{C}^k$. Si $k \to \infty$ tal que $k\log(n)/n \to 0$ cuando $n \to \infty$, entonces g_n es fuertemente universalmente consistente, es decir

$$\lim_{n\to\infty} L(g_n^*) = L^*,$$

con probabilidad 1, para toda distribución de (X, Y).

Demostración. Descomponemos el error como antes,

$$L(g_n^*) - L^* = \left(L(g_n^*) - \inf_{g\in\mathcal{C}^k} L(g)\right) + \left(\inf_{g\in\mathcal{C}^k} L(g) - L^*\right).$$

El segundo sumando tiende a 0 por el Corolario 8.5. Por (6.6) y la cota obtenida en (8.10)

$$\mathbb{P}\Big\{L(g_n^*) - \inf_{g\in\mathcal{C}^k} L(g) > \varepsilon\Big\} \le 8s(\mathcal{C}^k, n)\exp\left(\frac{-n\varepsilon^2}{128}\right) \le 8(ne)^{kd+2k+1}\exp\left(\frac{-n\varepsilon^2}{128}\right), \tag{8.11}$$

esta serie es convergente porque $k = o(n/\log(n))$. □

8.5 Sobre la cantidad de neuronas y de capas

El teorema que sigue muestra que no es lo mismo incrementar la cantidad de neuronas en una capa que aumentar la cantidad de capas. La demostración del mismo se puede encontrar en [32]. Sea $\mathcal{R}(\sigma_R; m, l)$ el conjunto de todas las redes neuronales con l capas, y a lo sumo m neuronas en cada capa, donde en cada capa usamos el sigmoide ReLu, $\hat{L}_n(g) = (1/n)\sum_{i=1}^n \mathbb{I}_{g(X_i)\neq Y_i}$ y $g(X_i) = \mathbb{I}_{\{f(X_i)>1/2\}}$.

Teorema 8.10 (Telgarski 2015)**.** *Sea k un entero positivo, l el número de capas, y m el número de nodos por capa, tal que $m \le 2^{(k-3)/l-1}$. Entonces, existe una colección de $n := 2^k$ puntos $((x_i, y_i)_{i=1}^n$ con $x_i \in [0, 1]$ e $y \in \{0, 1\}$ tal que*

$$\min_{g\in\mathcal{R}(\sigma_R;2,2k)} \hat{L}_n(g) = 0 \qquad y \qquad \min_{g\in\mathcal{R}(\sigma_R;m,l)} \hat{L}_n(g) \ge \frac{1}{6}.$$

Esto nos dice que, por ejemplo, para obtener el mismo error $\hat{L}_n$ de una red con $2k$ capas pero usando 2 capas, se requiere al menos $2^{(k-3)/2-1}$ neuronas, y con $\sqrt{k-3}$ capas necesita al menos $2^{\sqrt{k-3}-1}$ neuronas.

Otro resultado en esta dirección, más dificil de enunciar con precisión, también de Telgarsky (ver [33]), establece que, para todo entero k existe una red neuronal con sigmoide ReLu, con $2k^3 + 8$ capas, con $3k^3 + 12$ neuronas en total, y con $4 + d$ parámetros, que no se puede aproximar arbitrariamente (en sentido L^2) por redes que **no** tengan una cantidad exponencial de capas (en k) de neuronas, si tienen $\le k$ capas. Este resultado vale aún si permitimos que los sigmoides se cambien por funciones mucho mas generales, que se denominan compuertas semi-algebraicas. Ver Definición 2.1 en [33].

9 Métodos de descenso por gradiente y gradiente estocástico

De forma muy general, entrenar un algoritmo paramétrico de aprendizaje automático consiste en elegir a partir de ciertos datos $(X_1, Y_1), \ldots, (X_n, Y_n)$ el parámetro $\theta \in \mathbb{R}^p$ que minimiza la discrepancia promedio entre el valor Y_i y la predicción $\phi_n(X_i; \theta)$ que hace el clasificador (o la función de regresión, según sea el problema). La discrepancia es cuantificada en términos de una función que denotaremos de forma genérica $\mathcal{L}\big(\phi_n(X_i;\theta), Y_i\big) : \mathbb{R}^2 \to \mathbb{R}^+$. Es decir, queremos resolver el siguiente problema de optimización,

$$\arg\min_{\theta} \quad \sum_{i=1}^{n} \mathcal{L}\big(\phi_n(X_i;\theta), Y_i\big). \tag{9.1}$$

La predicción $\phi_n(X_i;\theta)$ está basada en los datos, y es calculado en un determinado dato X_i, usando el parámetro θ. La función $\mathcal{L}$ puede ser, por ejemplo, $\mathcal{L}(a,b) = (b-a)^2$ o $\mathcal{L}(a,b) = |b-a|$.

Los datos sobre los que se optimiza no son necesariamente todos los datos de los que se dispone. En el caso de redes neuronales con muchas capas en general no se toma toda la muestra, ya que esto es computacionalmente imposible. Se toman subconjuntos de la muestra (que se denominan lotes o "batch" en inglés) y luego se promedian las estimaciones en estos subconjuntos. Idealmente, si pudieramos calcular de forma exacta el gradiente de $\mathcal{L}\big(\phi_n(X_i;\theta), Y_i\big)$ respecto de θ, el problema anterior, fijada la muestra, se vuelve un problema de minimización de una función real. Esto se ataca por medio del método (o métodos, ya que existen variantes) clásico de **descenso por gradiente**. Estos métodos tienen hipótesis de convergencia que, para el caso de redes neuronales, o bien no se pueden verificar o directamente no se cumplen (por ejemplo, convexidad). Es por esto que, suelen encontrar mínimos locales y no globales.

Es claro que (9.1) es una aproximación, basada en la muestra (o un subconjunto de la misma), del error que quisiéramos poder minimizar que es

$$\arg\min_{\theta} \quad \mathbb{E}\Big(\mathcal{L}\big(\phi_n(X;\theta), Y\big)\Big). \tag{9.2}$$

Minimizar (9.2) da un estimador ϕ_n que funciona bien no solo en la muestra en que entrenamos, sino en promedio en cualquier otra de tamaño n que se tome del par (X, Y).

Es importante tener en cuenta que la esperanza en (9.2), cuando únicamente disponemos de una muestra de entrenamiento, no se puede calcular, o no tiene sentido, porque no imponemos hipótesis sobre la distribución del par (X, Y). Aún en el caso en que se pudiera calcular, los métodos clásicos como descenso por gradiente o el método de Newton-Raphson, no tienen por qué encontrar mínimos globales ya que tienen hipótesis que en general no se cumplen.

Para pasar de (9.1) a (9.2), es decir, que minimizando en (9.1) vamos a estar cerca de un mínimo en (9.2), existe una versión estocástica del método clásico de descenso por gradiente, denominada **algoritmo de Robbins-Monro** (ver [24]). Veremos este algoritmo y una variante, que se usa para atacar problemas inherentes al mismo, cuando se aplican a redes neuronales con muchas capas. Uno de estos problemas, es que los gradientes se van haciendo cada vez más pequeños a medida que se agregan capas a la red, por efecto de la regla de la cadena. Otros problemas son los mínimos locales o puntos silla. En la sección 9.1.2 mencionamos algunos problemas y el método SGD con momento que intenta dar solución a algunos de ellos.

Fijado n el error esperado $\mathbb{E}\mathcal{L}(\phi_n(X;\theta), Y)$, como función de θ, suele tener un comportamiento tipo U entorno al mínimo: crece al alejarnos de el. Esto no sucede con la versión empírica, es decir, podríamos seguir achicando el error empírico pero que el esperado aumente. Estos fenómenos se conocen como sobreajuste y se observan en particular en redes neuronales. Son aún más pronunciado si permitimos ir aumentando la cantidad de parámetros (agregando, por ejemplo, más capas o más neuronas).

Vimos que, para el caso en que los clasificadores se obtiene de una familia $\mathcal{C}$ con dimensión de Vapnik-Chervonenkis finita, un criterio asintóticamente consistente de elección de la regla de clasificación es tomar el clasificador $g_n^* \in \mathcal{C}$ que minimiza,

$$\hat{L}_n(g) = \frac{1}{n} \sum_{i=1}^{n} \mathbb{I}_{\{g(X_i) \neq Y_i\}}.$$

Cuando la familia $\mathcal{C}$ depende de ciertos parámetros θ (vamos a asumir $\theta \in \mathbb{R}^p$ para algún $p > 0$) podemos escribir

de forma general el error cuadrático medio empírico como

$$\hat{L}_n^2\big(\phi_n,\theta\big) = \frac{1}{n}\sum_{i=1}^{n}\big(Y_i - \phi_n(X_i;\theta)\big)^2.$$

En el caso de regresión tiene sentido tomar también el error L^1,

$$\hat{L}_n^1\big(\phi_n,\theta\big) = \frac{1}{n}\sum_{i=1}^{n}\big|Y_i - \phi_n(X_i;\theta)\big|.$$

Otra opción es minimizar el error $\hat{L}_n^2$ o $\hat{L}_n^1$ pero agregar además un término con restricciones sobre los parámetros, como se hace en la regresión Lasso. Cada una de ellas da lugar en general a métodos de elección de θ distintos. Una ventaja de usar $\hat{L}_n^2$ es que la función x^2 es derivable.

Fijada la muestra, $\hat{L}_n^2\big(\phi_n,\theta\big)$ y $\hat{L}_n^1\big(\phi_n,\theta\big)$ son funciones determinísticas, que se minimizan usando descenso por gradiente. A grandes rasgos, se parte de un punto θ_0 inicial, y se itera para $k = 1, 2, \ldots,$

$$\theta_k = \theta_{k-1} - \alpha_k \nabla \hat{L}_n^i\big(\phi_n,\theta_{k-1}\big) \qquad i = 1, 2,$$

para cierta sucesión de números reales α_k. Es decir nos movemos en la dirección opuesta del gradiente. Bajo ciertas hipótesis (por ejemplo convexidad de L^2 como función de θ) y eligiendo α_k adecuadamente, se obtiene una sucesión θ_k que converge al mínimo de L^2.

Naturalmente, calcular ∇L^2 conlleva dificultades que luego estudiaremos, para esto se usa el algoritmo de propagación hacia atrás "backpropagation". Vamos a empezar con el método (determinístico) de descenso por gradiente.

La elección de las tasas de aprendizaje α_k es muy importante y el comportamiento del algoritmo puede variar drásticamente según sea esta elección. En `https://fa.bianp.net/teaching/2018/COMP-652/gradient_descent.html` se muestra de manera interactiva cómo cambia el comportamiento del algoritmo según la elección de estas tasas.

9.1 Descenso por gradiente, SGD

Vamos a presentar muy brevemente las ideas fundamentales del método de descenso por gradiente. Para una lectura más profunda, ver, por ejemplo, [2]. Es inmediato que si $f : \mathbb{R}^p \to \mathbb{R}$, ∇f apunta en la dirección mayor crecimiento de f.

Escribimos, para $\alpha > 0$,

$$\theta_\alpha = \theta - \alpha \nabla f(\theta).$$

Haciendo un desarrollo de Taylor de primer orden en θ,

$$f(\theta_\alpha) = f(\theta) + \nabla f(\theta)^T\big(\theta_\alpha - \theta\big) + o\big(\|\theta_\alpha - \theta\|\big) = f(\theta) - \alpha\|\nabla f(\theta)\|^2 + o\big(\alpha\|\nabla f(\theta)\|\big).$$

Por lo tanto, si α es suficientemente chico $f(\theta_\alpha) \leq f(\theta)$. También se verifica si $\theta_\alpha = \theta - \alpha\ell$, donde ℓ es una dirección tal que $\nabla f(\theta)^T\ell < 0$, siendo $\ell = -\nabla f(\theta)$ la de mayor decrecimiento, por la desigualdad de Cauchy-Schwartz. No obstante no siempre es la mejor dirección para moverse. En general se plantea

$$\theta_{k+1} = \theta_k - \alpha_k D_k \nabla f(\theta_k) \tag{9.3}$$

donde D_k es una matriz simétrica definida positiva. La condición de descenso es

$$\nabla f(\theta_k)^T D_k \nabla f(\theta_k) > 0,$$

la cual se cumple porque D_k es definida positiva. En el caso del método de Newton-Raphson se toma $D_k = (\nabla^2 f(\theta_k))^{-1}$ donde $\nabla^2 f(\theta_k)$ denota la matriz Hessiana. Existen diferentes métodos para elegir el tamaño α del paso. Vamos a enunciar una proposición que garantiza que si $\alpha_k \to 0$ tal que $\sum \alpha_k \to \infty$, y la sucesión de vector ℓ_k se elije de forma adecuada, si el algoritmo converge a un cierto θ^*, este es un punto estacionario, es decir $\nabla f(\theta^*) = 0$.

Proposición 9.1. *Sean $f : \mathbb{R}^p \to \mathbb{R}$ diferenciable, y $\{\theta_k\}$, $\{\ell_k\}$ dos sucesiones de vector de $\mathbb{R}^p$ tal que $\theta_{k+1} = \theta_k + \alpha_k\ell_k$. Supongamos que ∇f es Lipschitz en todo $\mathbb{R}^p$ y existen c_1, c_2 constantes positivas tal que*

$$c_1\|\nabla f(\theta_k)\|^2 \leq -\nabla f(\theta_k)^T\ell_k \qquad y \qquad \|\ell_k\|^2 \leq c_2\|\nabla f(\theta_k)\|^2.^1$$

Supongamos que

$$\alpha_k \to 0 \qquad y \qquad \sum_k \alpha_k = \infty.$$

Entonces $f(\theta_k) \to -\infty$ o $f(\theta_k) \to L < \infty$, y $\nabla f(\theta_k) \to 0$. En particular si $\theta_k \to \theta^*$, $\nabla f(\theta^*) = 0$.

[1] la primera condición establece que nos movemos en una dirección ℓ_k que no se va haciendo perpendicular respecto de ∇f y la otra controla la magnitud de este vector

La condición de Lipschitz sobre ∇ se cumple si f es C^2 y $\nabla^2 f$ es acotada en todo $\mathbb{R}^d$. Todas las condiciones sobre ℓ_k se cumplen en el caso particular en que $\ell_k = \nabla f(\theta_k)$.

9.1.1 Algoritmo de Robbins-Monro

En el paper de Robbins-Monro se tiene como objetivo encontrar la raíz $\theta^* \in \mathbb{R}$, que se asume única, de la función (desconocida) $h(\theta)$ a valores reales, utilizando observaciones de $H(\theta, D_n)$ tal que $\mathbb{E}(H(\theta, D_n)) = h(\theta)$, donde $D_n = \{(X_1, Y_1), \ldots, (X_n, Y_n)\}$ son los datos sobre los que se entrena. Este algoritmo se utiliza para encontrar la raíz de una función cuando solo se pueden obtener observaciones ruidosas de la misma, es decir, cuando la función no puede ser evaluada directamente sino a través de mediciones que incluyen ruido o incertidumbre. En el caso del SGD se aplica a

$$h(\theta) = \nabla_\theta \mathbb{E}\Big(\mathcal{L}(\phi_n(X;\theta), Y)\Big)\ ^2$$

y

$$H(\theta, D_n) = \nabla_\theta \left[\frac{1}{n}\sum_{i=1}^{n} \mathcal{L}\big(\phi_n(X_i;\theta), Y_i\big)\right]. \tag{9.4}$$

En [24] se prueba que si H es acotada y

$$\exists \delta > 0: \quad h(\theta) \leq \delta \text{ si } \theta \leq \theta^* \qquad y \qquad h(\theta) \geq \delta \text{ si } \theta \geq \theta^* \tag{9.5}$$

el algoritmo definido como,

$$\theta_{k+1} = \theta_k - \alpha_k H(\theta_k, D_n) \qquad k = 1, \ldots, M,^3 \tag{9.6}$$

cumple que $\mathbb{E}(\theta_k - \theta^*)^2 \to 0$, si las tasas de aprendizaje α_j verifican

$$\sum_{j=1}^{\infty} \alpha_j = \infty \qquad \text{y} \qquad \sum_{j=1}^{\infty} \alpha_j^2 < \infty.$$

Muchas veces como el cálculo de (9.4) es computacionalmente costoso, se toma:

$$\theta_{k+1} = \theta_k - \alpha_k \nabla_{\theta_k} \mathcal{L}\big(\phi_n(X_{i_k};\theta_k), Y_{i_k}\big) \tag{9.7}$$

donde (X_{i_k}, Y_{i_k}) es un punto de la muestra de entrenamiento, elegido al azar. Claramente tiene mas varianza que (9.4). Por otra parte, el cálculo de los gradientes en (9.4) es paralelizable.

9.1.2 SGD con Momento

La condición (9.5) de Robbins-Monro no se cumple en el caso de redes neuronales sino que se tienen mínimos y máximos locales. Ademas, como dijimos, es imposible calcular los gradientes en toda la muestra cuando la cantidad de datos y parámetros es muy grande (redes con muchas capas y muchas neuronas), por lo cual se calculan sobre subconjuntos de la muestra. Otro problema que surge es que como las redes neuronales son una composición de funciones, calcular su gradiente requiere de aplicar la regla de la cadena sucesivamente, esto hace que los gradientes se vayan haciendo cero. El método de SGD con momento se aplica para:

- Superación de mínimos locales y valles poco profundos: funciones de pérdida con múltiples mínimos locales o mesetas. El momento ayuda a superar estos obstáculos al mantener el impulso en la dirección general del descenso. Este impulso que le da el momento permite no quedarse en mínimos locales subóptimos.
- Reducción de oscilaciones en direcciones transversales. En casos donde hay altas curvaturas o gradientes en direcciones perpendiculares, el algoritmo estándar puede oscilar. El momento suaviza estas oscilaciones, permitiendo un descenso más directo hacia el mínimo.

El uso del momento puede acelerar significativamente la convergencia, reduciendo el tiempo de entrenamiento, especialmente en redes neuronales profundas con muchos parámetros. Es importante tener en cuenta que la mayoría de estas sugerencias y variantes del SGD **no tienen un respaldo teórico sólido atrás**, sino que se apoyan en determinados resultados empíricos, ver [31].

Formalmente, se define

$$\begin{aligned}\theta_{k+1} = &\ \theta_k - \alpha_k H(\theta_k, D_n) + \rho \cdot (\theta_k - \theta_{k-1}) \\ = &\ \theta_k - \alpha_k \nabla_{\theta_k} \left[\frac{1}{n}\sum_{i=1}^{n} \mathcal{L}\big(\phi_n(X_i;\theta), Y_i\big)\right] + \rho \cdot \big(\theta_k - \theta_{k-1}\big),\end{aligned}$$

[2] aquí denotamos ∇_θ para indicar que estamos calculando el gradiente en θ.

donde $\rho \in \mathbb{R}^+$ es el factor de momento, típicamente entre 0 y 1, que determina la contribución del momento pasado. Existen muchas variantes del SGD, para los lectores interesados en profundizar, se puede ver, por ejemplo: [5][4].

10 Algunos conceptos básicos de teoría de la medida

En este capítulo vamos a dar algunos conceptos básicos de teoría de la medida necesarios para leer, fundamentalmente, el capítulo de esperanza condicional de estas notas. Dado que únicamente trabajaremos con medidas de probabilidad, vamos a asumir que tenemos siempre una terna $(\Omega, \mathcal{A}, \mathbb{P})$ donde Ω es un conjunto (que pueden ser los reales o $\mathbb{R}^d$), $\mathcal{A}$ es una σ-álgebra en Ω, y $\mathbb{P}$ es una probabilidad definida en $\mathcal{A}$. Una función medible X es una variable aleatoria, es decir está definida entre dos espacios de probabilidad $(\Omega, \mathcal{A}, \mathbb{P})$ y $(\Omega', \mathcal{A}', \mathbb{P}')$, y cumple que $X^{-1}(A') \in \mathcal{A}$ para todo $A' \in \mathcal{A}'$.

Definición 10.1. Función Simple. Dados $E_1, \ldots, E_N$ subconjuntos de Ω, pertenecientes a $\mathcal{A}$, una función X, a valores reales, se dice que es una función simple si existe $a_1, \ldots, a_N$ números reales tal que

$$X(\omega) = \sum_{k=1}^{N} a_k \mathbb{I}_{E_k}(\omega) \tag{10.1}$$

donde $\mathbb{I}_{E_k}(\omega)$ denota la función indicatriz o función característica de E_k, que vale 1 si $\omega \in E_k$ y 0 en otro caso. Observar que si imponemos la restricción de que los $a_1, \ldots, a_N$ sean distintos y no nulos y los $E_1, \ldots, E_N$ disjuntos 2 a 2, es fácil ver que hay una única descomposición de la forma (10.1). Además, cualquier función simple se puede llevar a una que cumpla esas dos propiedades.

Definición 10.2. Si $-\infty \leq X \leq \infty$, diremos que X es medible si además de ser medible en el sentido que dimos antes se cumple que $X^{-1}(-\infty)$ y $X^{-1}(\infty)$ son medibles.

Un teorema importante que usaremos es el siguiente

Teorema 10.3. *Sea X medible definida en $(\Omega, \mathcal{A}, \mathbb{P})$ a valores reales, entonces existe una sucesión de funciones simples $\{\varphi_k\}_{k=1}^{\infty}$ tal que para todo $k > 0$, $|\varphi_k(\omega)| \leq |\varphi_{k+1}(\omega)|$ y*

$$\lim_{k\to\infty} \varphi_k(\omega) = X(\omega) \quad \forall \omega \in \Omega$$

Consideremos una función simple $\varphi(\omega) = \sum_{k=1}^{n} a_k \mathbb{I}_{E_k}(\omega)$, donde los $E_k \in \mathcal{A}$. Definimos

$$\mathbb{E}(\varphi) \equiv \int_\Omega \varphi(\omega) d\mathbb{P}(\omega) = \sum_{k=1}^{n} a_k \mathbb{P}(E_k) \qquad y \qquad \mathbb{E}(\varphi) \equiv \int_E \varphi(\omega) d\mathbb{P}(\omega) \equiv \int_\Omega \varphi(\omega) \mathbb{I}_E(x) d\mathbb{P}(\omega), \tag{10.2}$$

observar que $\varphi(\omega)\mathbb{I}_E(\omega)$ es también una función simple. El siguiente lema prueba que la integral de una función simple está bien definida, es decir (10.2) no depende de la descomposición de φ.

Definición 10.4. El **soporte** de $X : \Omega \to \mathbb{R}$ medible es el conjunto $sop(X) = \{\omega : X(\omega) \neq 0\}$. Observar que $sop(X)$ es medible ya que es igual a $(X^{-1}(0))^c$.

Definición 10.5. Sea X acotada con soporte E, definimos

$$\mathbb{E}(X) \equiv \int_\Omega X(\omega) d\mathbb{P}(\omega) \equiv \lim_{n\to\infty} \int_\Omega \varphi_n(\omega) d\mathbb{P}(\omega)$$

donde φ_n es cualquier sucesión de funciones simples uniformemente acotadas con soporte E.

Vamos a extender la integral a $X : E \subset \Omega \to \mathbb{R} \cup \{\infty\}$ medible tal que $X \geq 0$. Recordar que esto quiere decir que para todo $a \in \mathbb{R}$, $\{X < a\}$ es medible, y además $X^{-1}(\infty)$ es medible. Definimos

$$\mathbb{E}(X) \equiv \int_\Omega X(\omega) d\mathbb{P}(\omega) \equiv \sup_{g \in G_X} \int g(\omega) d\mathbb{P}(\omega)$$

donde

$$G_X = \Big\{ g : 0 \leq g \leq X, \ g \ \text{ es medible, acotada } \Big\}.$$

Se dice que X tiene esperanza si $\mathbb{E}(X) \equiv \int_\Omega X(\omega)d\mathbb{P}(\omega) < \infty$. Definimos, para $E \in \mathcal{A}$,

$$\int_E X(\omega)d\mathbb{P}(\omega) = \int_\Omega X(x)\mathbb{I}_E(\omega)d\mathbb{P}(\omega).$$

Si X toma valores negativos se descompone como $X = X^+ - X^-$, suma de su parte positiva y negativa (que son funciones positivas) y se define su esperanza para cada una de las funciones positivas X^+ y X^- siempre que mín$\{\mathbb{E}(X^+), \mathbb{E}(X^-)\} < \infty$. Finalmente $\mathbb{E}(X) = \mathbb{E}(X^+) - \mathbb{E}(X^-)$.

Teorema 10.6. ***Teorema de convergencia dominada*** *. Sea $\{X_n\}_n$ una sucesión de variables aleatorias tal que $X_n \xrightarrow{c.s.} X$. Supongamos que existe Z integrable tal que para todo n, $|X_n(\omega)| \leq Z(\omega)$ c.s x. Entonces*

$$\lim_{n\to\infty} \int_\Omega |X_n(\omega) - X(\omega)|d\mathbb{P}(\omega) = 0.$$

10.1 Teorema de cambio de variable

El siguiente Teorema va a jugar un rol importante en el capítulo sobre la esperanza condicional, una prueba de el puede encontrarse en la página 196 de [29], o en cualquier libro de medida, por ejemplo [11] o [42].

Teorema 10.7. *Sea $(\Omega, \mathcal{F})$ y $(E, \mathcal{E})$ espacios medibles y $X = X(\omega)$ una función medible $\mathcal{F}/\mathcal{E}$ con valores en E. Sea $\mathbb{P}$ una medida de probabilidad en $(\Omega, \mathcal{F})$ y P_X la medida de probabilidad en $(E, \mathcal{E})$ inducida por $X = X(\omega)$:*

$$P_X(A) = \mathbb{P}\{\omega : X(\omega) \in A\}, \quad A \in \mathcal{E}.$$

Entonces

$$\int_A g(x)P_X(dx) = \int_{X^{-1}(A)} g(X(\omega))\mathbb{P}(d\omega), \quad A \in \mathcal{E},$$

para toda función $\mathcal{E}$-medible $g = g(x)$, $x \in E$ (en el sentido de que si una integral existe, la otra está bien definida, y las dos son iguales).

10.2 Integrales iteradas en $\mathbb{R}^d$.

Consideremos la partición $\mathbb{R}^d = \mathbb{R}^{d_1} \times \mathbb{R}^{d_2}$, tal que $d_1 + d_2 = d$, $d_1, d_2 \geq 1$. Si $f : \mathbb{R}^d \to \mathbb{R}$ es medible definimos las funciones $f^y : \mathbb{R}^{d_1} \to \mathbb{R}$ y $f_x : \mathbb{R}^{d_2} \to \mathbb{R}$ como

$$f^y(x) = f(x, y) \qquad \text{y} \qquad f_x(y) = f(x, y).$$

Más adelante veremos algunos ejemplos de que f medible no implica f^y o f_x medible. Definimos para $E \subset \mathbb{R}^d$,

$$E^y = \{x \in \mathbb{R}^{d_1} : (x, y) \in E\} \qquad y \qquad E_x = \{y \in \mathbb{R}^{d_2} : (x, y) \in E\}.$$

Nuevamente que E sea medible no implica que lo sea E^y o E_x, basta definir en $\mathbb{R}^2$, el conjunto que es en $y = 0$ un conjunto no medible. Este conjunto en $\mathbb{R}^2$ tiene medida 0 y por lo tanto es medible, pero E^y no es medible para $y = 0$ (luego veremos que si E es medible entonces para casi todo E^y es medible).

Teorema 10.8. ***Teorema de Fubini.*** *Sea $f : \mathbb{R}^d \to \mathbb{R}$ medible e integrable, para casi todo $y \in \mathbb{R}^{d_2}$*

1. *f^y es integrable en $\mathbb{R}^{d_1}$*
2. *La función*

$$\int_{\mathbb{R}^{d_1}} f^y(x)dx = F(y)$$

es integrable en $\mathbb{R}^{d_2}$ y

3. $$\int_{\mathbb{R}^{d_2}} \Big(\int_{\mathbb{R}^{d_1}} f^y(x)dx \Big) dy = \int_{\mathbb{R}^d} f(x, y)dxdy. \tag{10.3}$$

10.3 Clases monótonas y Teorema de Radon-Nikodym.

Definición 10.9. Un **álgebra de conjuntos** en X es una familia $\mathcal{A} \subset 2^X$ de conjuntos cerrada por complementos y por uniones finitas.

Definición 10.10. Clase monótona. Una clase monótona en un conjunto X es un subcojunto $\mathcal{C}$ de las partes de X cerrado por uniones numerables crecientes y por intersecciones numerables decrecientes y contiene al vacío. Es inmediato que una σ-álgebra en X es una clase monótona. Además, la intersección de cualquier familia de clases monótonas es una clase monótona, esto permite definir para cualquier $\mathcal{E} \subset 2^X$, la clase monótona generada por $\mathcal{E}$, como la intersección de todas las clases monótonas que contienen a $\mathcal{E}$.

Lema 10.11. *Si $\mathcal{A}$ es un álgebra de conjuntos, la clase monótona $\mathcal{C}$ generada por $\mathcal{A}$ coincide con la σ-álgebra $\mathcal{M}$ generada por $\mathcal{A}$.*

Definición 10.12. Dada una medida signada ν y μ una probabilidad, definidos en el mismo espacio de probabilidad $(\Omega, \mathcal{A}, \mathbb{P})$, ν es **absolutamente continua** respecto de μ, y se denota $\nu \ll \mu$ si $\nu(E) = 0$ para todo $E \in \mathcal{A}$ tal que $\mu(E) = 0$.

Ejercicio 10.13. Se deja como ejercicio verificar que si $\int |Y| d\mathbb{P} < \infty$, $\nu(E) = \int_E Y(\omega) d\mathbb{P}(\omega)$ es una medida y además es absolutamente continua respecto de $\mathbb{P}$. Aquí no se precisa que $Y(\omega) < \infty$ en E ya que asumimos $0\infty = 0$.

Teorema 10.14. ***Teorema de Radon-Nikodym.*** *Sean ν una medida σ-finita, signada, y μ una medida de probabilidad tal que $\nu \ll \mu$ entonces existe X medible, integrable respecto de μ tal que, para todo $A \in \mathcal{A}$,*

$$\nu(\mathcal{A}) = \int_A X(\omega) d\mu(\omega) = \mathbb{E}(X \mathbb{I}_A)$$

Observación 10.15. *Si $\mathfrak{F} \subset \mathcal{A}$ es una σ-álgebra y ν se define en $\mathfrak{F}$ como $\nu(E) = \int_E Y d\mathbb{P}(\omega)$, obtenemos que la X del teorema anterior (tomando μ como la restricción de $\mathbb{P}$ a $\mathfrak{F}$) es $X = \mathbb{E}(Y|\mathfrak{F})$*

Desigualdades de concentración

Veremos varias desigualdades de concentración de variables aleatorias alrededor de su media.[1]

10.4 Desigualdad de Hoeffding

Dada una v.a. X y $s > 0$, por la desigualdad de Markov tenemos que

$$\mathbb{P}(X \geq t) = \mathbb{P}(e^{sX} \geq e^{st}) \leq \mathbb{E}(e^{sX})e^{-st}, \tag{10.4}$$

que puede no ser finita.

El método de Chernov consiste en hallar $s > 0$ que minimize la cota superior. Para el caso de suma de variables independientes nos queda

$$\mathbb{P}(S_n - \mathbb{E}(S_n) \geq t) \leq e^{-st}\mathbb{E}(e^{s\sum_{i=1}^n (X_i - \mathbb{E}(X_i))}) = e^{-st}\prod_{i=1}^n \mathbb{E}(e^{s(X_i - \mathbb{E}(X_i))}), \tag{10.5}$$

por la independencia.

Luego el problema se reduce a hallar buenas cotas para la función generadora de momentos de las variables $X_i - \mathbb{E}(X_i)$, $i = 1, \ldots, n$.

Lema 10.16. *Hoeffding (1983)*
Sea X una variable aleatoria con $\mathbb{E}(X) = 0$, $a \leq X \leq b$. Entonces para todo $s > 0$,

$$\mathbb{E}(e^{sX}) \leq e^{s^2(b-a)^2/8}.$$

Demostración. Sketch. Por la convexidad de la función exponencial tenemos que para $a \leq x \leq b$,

$$e^{sx} \leq \frac{x-a}{b-a}e^{sb} + \frac{b-x}{b-a}e^{sa}.$$

Como $\mathbb{E}(X) = 0$, si $p = \frac{-a}{b-a}$ se tiene que

$$\mathbb{E}(e^{sX}) \leq \frac{-a}{b-a}e^{sb} + \frac{b}{b-a}e^{sa} = \left(1 - p + pe^{s(b-a)}\right)e^{-ps(b-a)}.$$

Sea $u = s(b-a)$. Entonces

$$\left(1 - p + pe^{s(b-a)}\right)e^{-ps(b-a)} := e^{h(u)},$$

con $h(u) = -pu + \ln(1 - p + pe^u)$. Se verifica que $h(0) = h'(0) = 0$ y $h''(u) \leq \frac{1}{4}$. Finalmente desarrollando Taylor tenemos que

$$h(u) = h(0) + h'(0)u + h''(\theta)u^2/2 \leq u^2/8 = s^2(b-a)^2/8,$$

lo que demuestra el lema. □

Ahora podemos reemplazar esta cota en la ecuación (10.5) y obtenemos que

$$\mathbb{P}(S_n - \mathbb{E}(S_n) \geq \epsilon) \leq e^{-s\epsilon}\prod_{i=1}^n \mathbb{E}(e^{s(X_i - \mathbb{E}(X_i))}) \leq e^{-s\epsilon}\prod_{i=1}^n e^{s^2(b_i-a_i)^2/8} = e^{-s\epsilon}e^{s^2\sum_{i=1}^n (b_i-a_i)^2/8}$$

$$= e^{-2\epsilon^2/\sum_{i=1}^n (b_i-a_i)^2},$$

eligiendo $s = \frac{4\epsilon}{\sum_{i=1}^n (b_i-a_i)^2}$.

En resumen, tenemos que

[1] Apuntes tomados de las notas de Ricardo Fraiman

Teorema 10.17. *Sean $X_1, \ldots, X_n$ v.a. independientes, acotadas $a_i \leq X_i \leq b_i$ con probabilidad uno para $i = 1, \ldots, n$. Sea $S_n = \sum_{i=1}^n X_i$. Entonces para todo $\epsilon > 0$ tenemos que*

$$\mathbb{P}\Big(S_n - \mathbb{E}(S_n) \geq \epsilon\Big) \leq e^{-2\epsilon^2 / \sum_{i=1}^n (b_i - a_i)^2},$$

$$\mathbb{P}\Big(S_n - \mathbb{E}(S_n) \leq -\epsilon\Big) \leq e^{-2\epsilon^2 / \sum_{i=1}^n (b_i - a_i)^2}.$$

Y por lo tanto

$$\mathbb{P}\Big(\big|S_n - \mathbb{E}(S_n)\big| \geq \epsilon\Big) \leq \exp\left(\frac{-2\epsilon^2}{\sum_{i=1}^n (b_i - a_i)^2}\right) \tag{10.6}$$

En particular, si consideramos la Binomial, donde X_i son Bernoulli(p) obtenemos que

$$\mathbb{P}\left(\frac{S_n}{n} - p \geq \epsilon\right) \leq e^{-2n\epsilon^2}.$$

Teorema 10.18. Chebyshev-Cantelli. *Sea $t \geq 0$,*

$$\mathbb{P}\big(X - \mathbb{E}(X) \geq t\big) \leq \frac{\mathbb{V}(X)}{\mathbb{V}(X) + t^2}$$

Demostración. Supongamos sin pérdida de generalidad $\mathbb{E}(X) = 0$ entonces

$$t = \mathbb{E}(t - X) \leq \mathbb{E}[(t - X)\mathbb{I}_{X \leq t}]$$

por lo tanto, por la desigualdad de Cauchy-Schwartz

$$t^2 \leq \mathbb{E}[(t - X)^2]\mathbb{P}(X \leq t) = (\mathbb{V}(X) + t^2)\mathbb{P}(X \leq t)$$

de donde se sigue la desigualdad. □

Bibliografía

[1] David H Ackley, Geoffrey E Hinton, and Terrence J Sejnowski. A learning algorithm for boltzmann machines. *Cognitive Science*, 9(1):147–169, 1985.

[2] D. P. Bertsekas. Nonlinear programming. *Journal of the Operational Research Society*, 48(3):334–334, 1997.

[3] David Blackwell and M. A. Girshick. *Theory of Games and Statistical Decisions.* John Wiley and Sons, 1954.

[4] L Bottou. Online learning and stochastic approximations. *Online learning in neural networks*, 17(9):142, 1998.

[5] Léon Bottou, Frank E. Curtis, and Jorge Nocedal. Optimization methods for large-scale machine learning. *SIAM Review*, 60(2):223–311, 2018.

[6] Leo Breiman. Statistical modeling: The two cultures (with comments and a rejoinder by the author). *Statistical science*, 16(3):199–231, 2001.

[7] David S Broomhead and David Lowe. *Radial Basis Functions, Multi-variable Functional Interpolation and Adaptive Networks.* Royal Signals and Radar Establishment Malvern (United Kingdom), 1988.

[8] Stephen Cook. The complexity of theorem-proving procedures. In *Proceedings of the third annual ACM symposium on Theory of computing*, pages 151–158. ACM, 1971.

[9] L. Devroye, L. Györfi, and G. Lugosi. *A Probabilistic Theory of Pattern Recognition.* Springer-Verlag, New York, 1996.

[10] Jeffrey L Elman. Finding structure in time. *Cognitive Science*, 14(2):179–211, 1990.

[11] G. Folland. *Real Analysis: Modern techniques and their applications.* Wiley, 1999.

[12] Bernd Fritzke. A growing neural gas network learns topologies. *Advances in Neural Information Processing Systems*, 7:625–632, 1995.

[13] Francis Galton. Regression towards mediocrity in hereditary stature. *The Journal of the Anthropological Institute of Great Britain and Ireland*, 15:246–263, 1886.

[14] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. *Advances in Neural Information Processing Systems*, 27:2672–2680, 2014.

[15] László Györfi, Michael Kohler, Adam Krzyżak, and Harro Walk. *A Distribution-Free Theory of Nonparametric Regression.* Springer, New York, 2002.

[16] Simon Haykin. *Neural Networks: A Comprehensive Foundation.* Prentice Hall PTR, 1994.

[17] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural Computation*, 9(8):1735–1780, 1997.

[18] John J Hopfield. Neural networks and physical systems with emergent collective computational abilities. *Proceedings of the National Academy of Sciences*, 79(8):2554–2558, 1982.

[19] Robert A Jacobs, Michael I Jordan, Steven J Nowlan, and Geoffrey E Hinton. *Methods for combining experts' probability assessments*, volume 7. MIT Press, 1995.

[20] Teuvo Kohonen. The self-organizing map. *Proceedings of the IEEE*, 78(9):1464–1480, 1990.

[21] E. A. Nadaraya. On estimating regression. *Theory of Probability and Its Applications*, 9(1):141–142, 1964.

[22] Karl Pearson. Contributions to the mathematical theory of evolution. iii. regression, heredity, and panmixia. *Philosophical Transactions of the Royal Society of London. Series A, Containing Papers of a Mathematical or Physical Character*, 187:253–318, 1896.

[23] J.J. Protter. *Probability Essentials.* Springer-Verlag, 2004.

[24] Herbert Robbins and Sutton Monro. A stochastic approximation method. *The Annals of Mathematical Statistics*, 22(3):400–407, 1951.

[25] Frank Rosenblatt. The perceptron: A probabilistic model for information storage and organization in the brain. *Psychological Review*, 65(6):386–408, 1958.

[26] W. Rudin. *Real and Complex Analysis. 3d ed.* 1986.

[27] Claude E. Shannon. A mathematical theory of communication. *The Bell System Technical Journal*, 27(3):379–423, 1948.

[28] Claude E. Shannon. Programming a computer for playing chess. *Philosophical Magazine*, 41(314):256–275, 1950.

[29] A.N. Shiryaev. *Probability.* Springer-Verlag, 1984.

[30] Charles J. Stone. Consistent nonparametric regression. *The Annals of Statistics*, 5(4):595–620, 1977.

[31] Ilya Sutskever, James Martens, George Dahl, and Geoffrey Hinton. On the importance of initialization and momentum in deep learning. In *International conference on machine learning*, pages 1139–1147. PMLR, 2013.

[32] Matus Telgarsky. Representation benefits of deep feedforward networks. *arXiv preprint arXiv:1509.08101*, 2015.

[33] Matus Telgarsky. Benefits of depth in neural networks. In *Conference on learning theory*, pages 1517–1539. PMLR, 2016.

[34] Alan M Turing. Computing machinery and intelligence. *Mind*, 59(236):433–460, 1950.

[35] Aad W. van der Vaart and Jon A. Wellner. *Weak Convergence and Empirical Processes: With Applications to Statistics*. Springer Series in Statistics. Springer, New York, 1996.

[36] Vladimir N Vapnik. An overview of statistical learning theory. *IEEE transactions on neural networks*, 10(5):988–999, 1999.

[37] Vladimir N. Vapnik and Alexey Y. Chervonenkis. A note on one class of perceptrons. *Automation and Remote Control*, 25:103–106, 1964.

[38] Vladimir N. Vapnik and Alexey Y. Chervonenkis. On the uniform convergence of relative frequencies of events to their probabilities. *Theory of Probability and Its Applications*, 16(2):264–280, 1971.

[39] John von Neumann. Zur theorie der gesellschaftsspiele. *Mathematische Annalen*, 100(1):295–320, 1928.

[40] Vladimir Vovk, Alex Gammerman, and Glenn Shafer. *Algorithmic learning in a random world*. Springer Science & Business Media, 2005.

[41] Geoffrey S. Watson. Smooth regression analysis. *Sankhya: The Indian Journal of Statistics, Series A*, 26(4):359–372, 1964.

[42] Stein y Shakarchi. *Real analysis, measure theory, integration, and Hilbert Spaces.* Universitext. Springer-Verlag, 2001.