Introducción a la Teoría del Aprendizaje Estadístico

Nicolas Silva Nash
Departamento de Matemática
Universidad Nacional del Comahue

6 de marzo de 2025

Índice general

Capítulo 1

Un acercamiento amistoso a la Teoría del Aprendizaje Estadístico

1.1. Introducción

La Teoría del Aprendizaje Estadístico proporciona la base teórica para muchos de los algoritmos de aprendizaje automático actuales y, sin lugar a dudas, es una de las ramas más bellamente desarrolladas de la inteligencia artificial en general. Nació con el perceptrón de Rossenblat y la escuela matemática de la Unión Soviética en la década de 1960, y ganó amplia popularidad en la década de 1990 tras el desarrollo de las llamadas Máquinas de Vectores de Soporte (*SVM*, por sus siglas en inglés), que se han convertido en una herramienta estándar para el reconocimiento de patrones en muchas disciplinas, que van desde la visión por computadora hasta la biología computacional.

Proporcionar la base para nuevos algoritmos de aprendizaje no ha sido la única motivación para desarrollar la Teoría del Aprendizaje Estadístico. También ha sido una gesta de caracter filosófico, en el intento de responder a la pregunta de qué nos permite extraer conclusiones válidas a partir de datos empíricos.

1.2. El aprendizaje

En este contexto, el *aprendizaje* es el proceso a través del cual pueden inferirse reglas generales a partir de ejemplos. Nos interesa entender como una máquina -una computadora- puede resolver ciertos problemas sin conocer las reglas de antemano, solo a partir de ejemplos y a través de un *algoritmo de aprendizaje*. El objetivo es que la máquina pueda no solo aprender a reconocer las reglas que rigen a los ejemplos dados, si no que también pueden generalizar dichas reglas para ejemplos que le serán presentados con posterioridad.

Llamamos a esta disciplina *aprendizaje automático* (en inglés, *Machine Learning*, literalmente "aprendizaje de máquinas") y reconocemos sus raíces en otras disciplinas: Estadística Matemática, Ciencias de la Computación e Inteligencia Artificial. Si bien el aprendizaje suele ser una parte fundamental de la mayoría de los esfuerzos en materia de Inteligencia Artificial, el objetivo del Machine Learning es más acotado que el de su rama madre: En vez de intentar definir, explicar o generar comportamiento inteligente o *inteligencia*, aquí nos interesa solamente descubrir los mecanismos a través de los cuales las computadoras pueden resolver algunas tareas acotadas y bien definidas, y que en general escapan a soluciones que pueden ser especificadas con una cantidad finita de código de programación (reglas determinísticas).

Con fines ilustrativos, nos centraremos primero en el más conocido de los problemas del Machine Learning, el de clasificación. Consideremos dos espacios de variables: X, llamado *espacio de entrada*, e Y, el *espacio de etiquetas*. En un problema de clasificación, deseamos poder etiquetar correctamente elementos de X con los valores de Y. Por ejemplo, podríamos querer clasificar un conjunto de datos, en alguna representación fija, de distintos objetos en una cantidad de etiquetas como: silla, cama, microondas, perro, gato. Si esta representación es en imágenes de $N \times M$ pixeles en blanco y negro (realmente, matrices de orden $N \times M$ con coeficientes reales en el interavalo $[0, 1]$ que representan la intensidad de cada pixel, del negro al blanco), el espacio X es el conjunto de dichas matrices y el espacio Y las categorías distintas que corresponden a lo que las imágenes muestran. Con el fin de aprender, a un algoritmo se le muestran ejemplos de imágenes y sus respectivas etiquetas (X_1, Y_1), (X_2, Y_2), $\ldots$, (X_n, Y_n), a partir de los cuales este debe encontrar una función $f : X \to Y$, que comete la menor cantidad de errores posibles. A esta función f la llamamos *clasificador*.

1.3. La historia del aprendizaje automático

El primer modelo de aprendizaje automático, según Vapnik en [2], fue sugerido por F. Rosenblatt, un psicólogo estadounidense, al que llamó *perceptrón*, y su introducción constituye el comienzo del análisis matemático del aprendizaje. Conceptualmente, la idea del perceptrón no era nueva, estando presente en la litetura de Neurofisiología durante varios años. Rosenblatt, sin embargo, se aventuró en describir el modelo como un programa para computadoras y demostró con simples experimentos que dicho modelo era generalizable. El perceptrón fue construído como una solución a un problema particular dentro del aprendizaje automático, el reconocimiento de patrones. En el caso más sencillo, este problema consiste en hallar una regla para separar datos en dos categorías distintas a partir de ejemplos.

Para construir la regla de separación, el perceptrón sigue el modelo más sencillo de neurona, propuesto previamente por McCulloch y Pitts, de acuerdo al cual una neurona recibe n valores (o *inputs*) en la forma de un vector $x = (x^1, \ldots, x^n) \in X \subset \mathbb{R}^n$ y genera una etiqueta (*output*) $y \in \{-1, +1\}$ a través

de una dependecia funcional dada por

$$y = sgn\{(w \cdot x) - b\}$$

con $\cdot$ el producto interno de vectores en $\mathbb{R}^n$, b un valor de límite y un vector w que se genera en el proceso de aprendizaje. Geométricamente, una neurona divide el espacio X en dos regiones: en una la etiqueta y vale $+1$ y en la otra -1. Las dos regiones son separadas por el hiperplano:

$$(w \cdot x) - b = 0$$

El vector w y el escalar b determinan la posición del hiperplano y sus valores son aprendidos por el perceptrón. Cuando combinamos varias neuronas, el perceptrón separa el espacio de entrada en en dos regiones lineales a trozos y no necesariamente conexas. En 1960 no era claro como elegir todos los parametros $(w_1, \ldots, w_k)$ y $(b_1, \ldots, b_k)$ de todas las neuronas, por lo que se fijaban los valores de las primeras $k-1$ y se intentaba encontrar los valores deseables para la última de ellas. Geométricamente, se transformaba el espacio de entrada X en un nuevo espacio Z (eligiendo coeficientes apropiados para las primeras $k-1$ neuronas) y luego se utilizaban los datos de entrenamiento para construir un hiperplano que separe el plano Z. Tomando prestados de la fisiología los conceptos de aprendizaje con estímulos de premios y castigos, Rosenblat propuso un simple algoritmo para hallar estos coeficientes de manera iterativa, el cual describiremos a continuación.

DESCRIBIR perceptrón.

En 1962 Novikoff demostró el primer teorema relacionado al perceptrón. Podemos decir que este teorema inció propiamente la teoría del aprendizaje.

Teorema 1.3.0.1. *Dado un conjunto de datos de entrenamiento como el descrito previamente, de manera que*
(1) La norma de los vectores de entrenamiento z_i está acotada por una constante R:

$$|z_i| \leq R, \qquad \forall i = 1, 2, \ldots, k$$

(2) Los datos de entrenamiento pueden separados con un margen ρ:

$$\sup_w \min_i y_i(z_i \cdot w) > \rho$$

(3) Los datos son alimentandos al perceptrón una cantidad suficiente de veces.

Entonces el algoritmo encuentra el hiperplano que separa los datos de entrenamiento, luego de a lo sumo N correcciones, con N verificando:

$$N \leq \left[\frac{R^2}{\rho^2}\right]$$

Resaltamos este teorema porque jugó un papel fundamental en la creación de la teoría del aprendizaje, conectando el principio de minimización de errores en el conjunto de datos de entrenamiento con la capacidad de generalización de los algoritmos de clasificación y su causa.

1.4. Conceptos básicos del aprendizaje

Volviendo al caso de clasificación binaria en aprendizaje supervisado, partimos de ejemplos (datos de entrenamiento) en un espacio de entrada X con alguna de las dos posibles etiquetas del espacio $Y = \{-1, +1\}$. Aquí, *aprender* se reduce a estimar una relación funcional $f : X \to Y$, el clasificador. Un algoritmo de aprendizaje es aquel que a partir de los datos de entrenamiento construye una función f. Nos interesa construir una teoría que no asuma de manera estricta nada acerca de X e Y, pero nos permitimos asumir ciertas cosas del mecanismo que genera los datos de entrenamiento. En particular, asumiremos que existe una *distribución de probabilidad conjunta* $P = P(X, Y)$ sobre $X \times Y$ y que las muestras son tomadas de forma independiente de esta distribución de forma *iid* -independiente e idénticamente distribuída-. Notemos lo siguiente:

1. *No imponemos condiciones a la distribución de probabilidad P.* La gran diferencia que encontramos entre la Estadística tradicional y la teoría del aprendizaje estadístico es que en esta última trabajamos de manera agnóstica a la distribución que genera las muestras y deseamos llegar a conclusiones generales.

2. *Consideramos a las etiquetas de manera no determinística.* Consideramos a P como una distribución de probabilidad no solo sobre las instancias de X, si no también sobre las propias etiquetas de Y. Por lo tanto, estas últimas no son solo funciones tradicionales de los datos en X, si no que ellas mismas pueden ser aleatorias. Tenemos al menos dos buenas razones para tomar esta consideración: por un lado, el proceso de generación de datos puede tener ruido al asignar etiquetas (por ejemplo tomemos el caso de un detector de spam basado en la opinión de etiquetadores humanos que clasifican emails con un porcentaje de error; incluso los humanos pueden clasificar incorrectament algunos de esos emails), y por otro, ciertos problemas se prestan a que existan clases que se solapan (pensemos en la dificultad en diferenciar a un perro de un gato en una fotografías que los captura desde una gran distancia o con baja resolución).
   En la práctica, en vez de asignar etiquetas a los elementos en X de manera determinística, daremos la probabilidad condicional de la etiqueta y dado el valor x. En el caso de clasificación binaria, basta solo dar la probabilidad $P(Y = 1|X = x)$ de que la etiqueta tenga valor $Y = 1$, dado que la restante es complementaria:

$$P(Y = -1|X = x) \quad = \quad 1 - P(Y = 1|X = x)$$

Ciertos problemas que hagan uso de datos con etiquetas con poco ruido nos llevaran naturalmente a probabilidades condicionales cercanas a 0 y 1, dejando un margen de error pequeño, pero cuando tratemos con solapamiento de clases, las probabilidades condicionales pueden acercarse a $\frac{1}{2}$ para cada etiqueta. Independientemente de la causa, que las probabilidades condicionales sobre las etiquetas se acerquen a $\frac{1}{2}$ vuelve más dificil el aprendizaje, dado que crece el número de errores del clasificador.

3. *Muestro independiente*. Una de las condiciones más fuertes que imponemos en la teoría del aprendizaje estadístico es que asumimos que las muestras son tomadas de forma independiente. En muchas aplicaciones, esta suposición está justificada, pero hay ramas muy importantes de la disciplina en donde esto no se cumple, por ejemplo en el análisis de series de tiempo, en donde la secuencialidad de los datos viola la condición de iid (cada valor depende en alguna medida de los anteriores). Esto es también cierto para las aplicaciones a lenguaje natural, y constituye una de las razones principales por las cuales esta rama más moderna del aprendizaje tiene bases teóricas menos fuertes que el aprendizaje automático tradicional.

4. *La distribución P es fija*. Al no considerar al tiempo como un parámetro, ni existir un orden en las muestras, asumimos que la distribución que las origina es siempre la misma. Esto, como en el punto anterior, no se cumple en las series de tiempo. Otro caso donde se viola esta suposición es en aquellos problemas en donde la distribución de probabilidad de los datos de entrenamiento no coincide con el de los datos posteriores, llamado *covariate shift*, por ejemplo en un sistema de *scoring* de usuarios de una empresa que crece súbitamente y que suma a personas que no se corresponden a los perfiles que existían originalmente en su base de datos (e.g. se admite que inmigrantes no bancarizados y sobre los que no hay datos previos accedan a préstamos).

5. *La distribución P es desconocida al momento de aprender*. Si conociéramos de antemano la probabilidad condicional P, el problema del aprendizaje sería trivial pues podríamos siempre determinar el mejor clasificador posible (aunque no sea perfecto, dada la naturaleza aleatoria de las etiquetas). Solo tenemos acceso a P de manera indirecta, a través de las muestras. Intuitivamente, esto nos hace pensar que, consiguiendo un número lo suficientemente grande de muestras, podemos aproximar las propiedades de la distribución P, pero con errores. Uno de los principales logros de la teoría del aprendizaje estadístico es brindarnos un marco teórico para acotar este error.

1.4.1. Pérdida y Riesgo

Para saber qué tan bien se comporta un clasificador f, necesitamos medir sus equivocaciones. Para esto, definiremos una *función de pérdida*, ℓ, que le asigne un valor al hecho de que f clasifique a cierto $x \in X$ con la etiqueta $y \in Y$.

Llamermos *costo* a dicho valor, dado que más adelante penalizaremos al clasificador en base a los errores que cometa a través del alogritmo de aprendizaje. El ejemplo más sencillo de función de pérdida es la "pérdida-0-1", que le asigna un costo de 0 a una instancia de clasificación correcta y un costo 1 a una incorrecta, es decir:

$$\ell(X, Y, f(X)) := \begin{cases} 1 & \text{si } f(X) \neq Y \\ 0 & \text{si } f(X) = Y \end{cases}$$

En problemas de regresión, la función de pérdida más conocida es el *error cuadrático*, dado por

$$\ell(X, Y, f(X)) := (Y - f(X))^2$$

Por convención, una pérdida igual a 0 implica una clasificación perfecta y valores mayores implican peor clasificación. Es decir, el aprendizaje suele implicar un desafío de optimización en dónde deseamos hallar el mínimo de la función de perdida.

Mientras que la función de pérdida mide el error del clasificador en un punto individual $x \in X$, llamamos *riesgo*, $\mathcal{R}$, del clasificador a la pérdida esperada sobre todos los datos generados por la distribución de probabilidad P. Es decir

$$\mathcal{R}(f) := \mathbb{E}\left(\ell(X, Y, f(X))\right)$$

Desde luego que otra función g es un mejor clasificador que f para un problema dado si su riesgo es más bajo, es decir si $\mathcal{R}(g) < \mathcal{R}(f)$, por lo que el mejor clasificador de todos es aquel con el riesgo más bajo.

Algo que no hemos considerado aún es si los clasificadores f tienen alguna característica especial. Para formalizarlo, tomaremos funciones de un espacio de funciones $\mathcal{F}$ que aplican X en Y. En un principio, parecería aceptable tomar el espacio de todas las funciones posibles, o más precisamente, el conjunto de todas las funciones *medibles* que aplican X en Y, $\mathcal{F}_{all} = \{f \text{ medibles} \mid f : X \to Y\}$. En este caso, podemos señalar cuál es el clasificador ideal, dada la distribución P, al que llamamos *clasificador Bayesiano*, f_{Bayes}, y al que definimos como:

$$f_{\text{Bayes}} := \begin{cases} 1 & \text{si } P(Y = 1|X = x) \geq \frac{1}{2} \\ -1 & \text{en otro caso} \end{cases}$$

Observermos que, en caso de que las etiquetas fueran determinísticas, es decir donde $P(Y = y|X = x) = 1$ para cada $x \in X$ con su respectiva etiqueta y, f_{Bayes} eligiría correctamente en todos los casos. De existir un ligero solapamiento de clases de tal manera que para un cierto x tengamos $P(Y = 1|X = x) = 0{,}9$, entonces tendríamos que en la mayoría de los casos la etiqueta de x es $+1$ y, siendo este el valor elegido por f_{Bayes}, el clasificador Bayesiano estaría en lo correcto.

En la práctica no es posible computar directamente el clasificador Bayesiano dado que, como dijimos, la distribución de probabilidad conjunta P es desconocida. Sin embargo, este clasificador es una herramienta teórica que nos permite formular el problema estandar de la clasificación binaria, el cual es:

Dado un conjunto de datos de entrenamiento $\{(X_1, Y_1), \ldots, (X_n, Y_m)\}$ *obtenidos iid de una distribución* P*, y dada una función de pérdida* ℓ*, deseamos construir una funión clasificadora* $f : X \to Y$ *cuyo riesgo* $\mathcal{R}(f)$ *sea lo más cercano posible al riesgo de* f_{Bayes}*.*

Notemos que no solo es imposible computar el error del clasificador Bayesiano, sino también el propio riesgo de cualquier clasificador f. Es decir, dado un problema definido (minimizar el riesgo del clasificador), con una solución ideal que podemos escribir (el propio clasificador Bayesiano), no tenemos manera de computar ninguna cosa de utilidad. Aquí es donde la teoría de aprendizaje estadístico nos permite llegar a resultados y obtener garantías de la utilidad de esas soluciones.

1.4.2. Generalización

Dado que no conocemos la distribución de probabilidad $P(X, Y)$, no podemos calcular la esperanza de la pérdida de un clasificador cualquiera f, es decir su riesgo. Lo que si podemos hacer, dado un conjunto de entrenamiento, es "contar" (o medir, en general, para problemas que no son de clasificación binaria) el número de errores del clasificador sobre los datos de entrenamiento. A esta cantidad le daremos el nombre de *riesgo empírico* y lo veremos presente en la literatura también como *error de entrenamiento*. Lo definimos como

$$R_{emp}(f) := \frac{1}{n} \sum_{i=1}^{n} \ell(X_i, Y_i, f(X_i))$$

Por ejemplo, de la función de pérdida llamada error cuadrático se deriva el riesgo empírico *error cuadrático medio*, que es ampliamente utilizado en la práctica. Usualmente, un algoritmo de aprendizaje aceptable es capaz de producir un clasificador f que performa aceptablemente bien en un conjunto de datos conocido, es decir tal que el riesgo empírico del clasificador es bajo. Nos interesa que un clasificador f tenga riesgo bajo en todo el espacio de entrada X, no solo en los datos de entrenamiento. Decimos que un clasificador *generaliza* bien si la diferencia entre su riesgo y su riesgo empírico es baja.

DEFINICIÓN Generalización

Por supuesto que una buena generalización no asegura que el riesgo, ni el riesgo empírico, sean bajos, si no que dichas cantidades son cercanas. Pero lo que nos interesa es que el riesgo empírico sea un buen estimativo del riesgo del clasificador, y esto es lo que nos permitirá hacer afirmaciones acerca del error del clasificador en la práctica.

1.4.3. Consistencia

Intuitivamente, parece razonable pedirle a un algoritmo de aprendizaje que, al ser presentado con más y más ejemplos, converja a una solución óptima. En

Estadística, la noción de *consistencia* se relaciona con la capacidad de hacer afirmaciones con respecto a lo que sucede en el límite de una cantidad infinita de muestras y, a diferencia de la generalización, que es acerca de una función en particular, es una propiedad de un conjunto de funciones. Para ilustrar este concepto, denotemos como f_n al clasificador construido por un algoritmo de aprendizaje luego de ser presentado con n puntos de entrenamiento. Por el momento no repararemos en cómo el algoritmo construye esta función, pero podemos estar seguro que la elige de un cierto espacio funcional $\mathcal{F}$. Más adelante veremos que dicho espacio puede estar explícitamente dado, como en el caso de la regresión lineal, o no, y existir implícitamente en el mecanismo del propio algoritmo, como es el caso de las redes neuronales. Más allá de si $\mathcal{F}$ es explícito o no, el algoritmo debe elegir a la mejor función en dicho espacio basándose en los puntos de entrenamiento. Por otro lado, sabemos precisamente cuál es en teoría el mejor clasificador en $\mathcal{F}$: el que tiene el menor riesgo. Por simplicidad, supongamos que este es único (no tiene por qué serlo, puede no haber un solo mínimo global para la función de riesgo). Definimos a este clasificador óptimo como:

$$f_{\mathcal{F}} = \arg\min_{f \in \mathcal{F}} \mathcal{R}(f)$$

Considerando que el clasificador bayesiano introducido en la sección anterior es el mejor clasificador posible, podríamos denotarlo, siguiendo la misma convención, como $f_{\mathcal{F}_{all}}$. Desde luego que el espacio $\mathcal{F}$ que elegimos puede no contenerlo, por lo que $\mathcal{R}(f_{\text{Bayes}}) < \mathcal{R}(f_{\mathcal{F}})$. Con estas ideas, podemos definir los distintos de convergencia que trataremos al considerar el concepto de consistencia.

Definición 1.4.3.1. *Sea $(X_i, Y_i)_{i \in \mathbb{N}}$ una sucesión infinita de puntos de entrenamiento que han sido tomados iid de una distribución P. Sea ℓ una función de pérdida. Por cada $n \in \mathbb{N}$, sea f_n un clasificador construído por algún algoritmo de aprendizaje usando los primeros n puntos de entrenamiento. Sea $\mathcal{F}$ el espacio funcional de todos los posibles clasificadores según el mecanismo de construcción del algoritmo. Entonces*

1. *El algoritmo de aprendizaje se dice* consistente con respecto a $\mathcal{F}$ y a P *si el riesgo $\mathcal{R}(f_n)$ converge en probabilidad al riesgo $\mathcal{R}(f_{\mathcal{F}})$ del mejor clasificador en $\mathcal{F}$. Esto es, que para todo $\epsilon > 0$:*

$$P\left(\mathcal{R}(f_n) - \mathcal{R}(f_{\mathcal{F}}) > \epsilon\right) \to 0 \qquad \textit{cuando} \qquad n \to \infty$$

2. *El algoritmo de aprendizaje se dice* Bayes-consistente con respecto a P *si el riesgo $(R)(f_n)$ converge en probabilidad al riesgo $(R)(f_{Bayes})$ del clasificador bayesiano. Esto es, para todo $\epsilon > 0$*

$$P\left(\mathcal{R}(f_n) - \mathcal{R}(f_{Bayes}) > \epsilon\right) \to 0 \qquad \textit{cuando} \qquad n \to \infty$$

3. *El algoritmo de aprendizaje se dice* universalmente consistente con respecto a $\mathcal{F}$ *si es consistente con respecto a $\mathcal{F}$ para toda distribución de probabilidad P.*

4. *El algoritmo de aprendizaje se dice* universalmente Bayes-consistente *si es Bayes-consistente para toda distribución de probabilidad P.*

Incurrimos en un abuso del lenguaje al decir que el clasificador f_n es consistente para referirnos más precisamente a que el alogritmo de aprendizaje que produce a f_n en base a las primeras n muestras es consistente. Analicemos un poco más el significado de las definiciones anteriores. La primer definición nos habla del caso donde, a medida que n crece, el riesgo del clasificador f_n converge al riesgo del mejor clasificador $f_{\mathcal{F}}$ en el espacio funcional $\mathcal{F}$, con alta probabilidad. En la práctica, esto significa que alguna instancia de f_n puede no tener un riesgo cercano al riesgo de $f_{\mathcal{F}}$, pero que esto es poco probable, es decir que de repetir muchas veces el experimento, la mayoría de las veces el riesgo de f_n será cercano al riesgo de $f_{\mathcal{F}}$ a medida que n crece.

La segunda definición es similar, pero en vez de comparar el riesgo de f_n con el riesgo del mejor clasificador en $\mathcal{F}$, lo hace con el riesgo del clasificador bayesiano f_{Bayes}. La diferencia es clara, la primera defínción se encarga de comparar el riesgo de f_n con el mejor clasificador posible dadas las condiciones del aprendizaje, que son las que definen al espacio funcional $\mathcal{F}$ (por ejemplo, el mejor clasificador lineal si estamos hablando de regresión lineal), mientras que la segunda lo hace con el mejor clasificador posible, independientemente de las condiciones del aprendizaje (la función clasificadora ideal podría no ser lineal para un problema dado).

La tercera y cuarta definición son más fuertes, pues piden que el algoritmo sea consistente para cualquier distribución de probabilidad P. Dado que en la práctica no sabemos cuál es la distribución que genera los datos, estas son las defincciones que nos interesan a la hora de enunciar resultados.

Observamos que la consistencia como está aquí enunciada suele llamarse *consistencia débil*. Existe una noción más fuerte, la de *consistencia fuerte*, que pide que el riesgo empírico de f_n converja al riesgo del mejor clasificador en $\mathcal{F}$, no solo en probabilidad, si no con probabilidad 1, lo que también llamamos *convergencia casi segura*. En la práctica, la consistencia fuerte es una propiedad muy fuerte y no es comúnmente alcanzada por los algoritmos de aprendizaje.

Notemos que en estas definiciones no se hace mención del riesgo empírico $\mathcal{R}_{\text{emp}}(f_n)$, si no del riesgo real $\mathcal{R}(f_n)$, lo que se debe a que la única medida de calidad de un clasificador es su riesgo real y es sobre el cual deseamos obtener resultados. El problema, claro, es que no podemos medirlo, como si podemos hacer con el riesgo empírico. Es natural entonces pensar que, además de la convergencia que hemos anunciado de tipo $\mathcal{R}(f_n) \to \mathcal{R}(f_{\text{Bayes}})$, intentemos buscar las condiciones bajo las cuales el riesgo empírico converge al riesgo real, es decir donde $\mathcal{R}_{\text{emp}}(f_n) \to \mathcal{R}(f_{\text{Bayes}})$. Esta es una de las metas de la teoría del aprendizaje estadístico, y es lo que nos permitirá hacer afirmaciones acerca de la calidad

de los clasificadores en la práctica.

1.4.4. Sobreajuste y subajuste

Consideremos un ejemplo de regresión. Se nos da un conjunto de observaciones empíricas $(x_1, y_1), \ldots, (x_m, y_m) \in X \times Y$, donde, simplemente, tomamos $X = Y = \mathbb{R}$. Por ejemplo, los datos podrían haberse recopilado en un experimento físico donde X representa el peso de un objeto e Y la fuerza necesaria para arrastrar este objeto sobre una superficie rugosa.

La Figura 1.1 muestra una representación gráfica de dicho conjunto de datos (indicados por los puntos redondos), junto con dos posibles dependencias funcionales que podrían explicar los datos. La línea discontinua, $f_{\text{no lineal}}$, representa un modelo bastante complejo y ajusta perfectamente los datos de entrenamiento, es decir, tiene un error de entrenamiento igual a 0. Por otro lado, la línea recta, f_{lineal}, no "explica" completamente los datos de entrenamiento, ya que hay algunos errores residuales, lo que genera un error de entrenamiento pequeño pero positivo (por ejemplo, medido mediante la función de pérdida cuadrática).

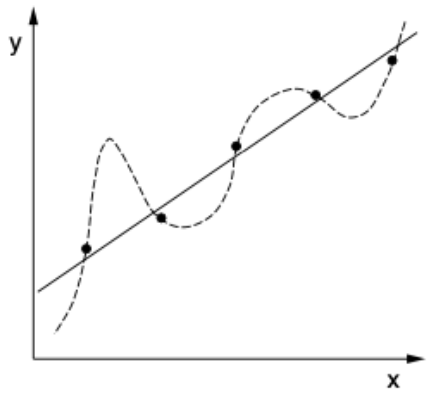

Figura 1.1: Dos posibles modelos de regresión, uno lineal y otro no, para un conjunto de datos dado.

¿Pero qué sucede con los riesgos verdaderos $R(f_{\text{no lineal}})$ y $R(f_{\text{lineal}})$? El problema es que no podemos calcular estos riesgos a partir de los datos de entrenamiento. Además, las funciones $f_{\text{no lineal}}$ y f_{lineal} tienen comportamientos muy diferentes. Por ejemplo, si la línea recta f_{lineal} fuera la verdadera función subyacente, entonces la función discontinua $f_{\text{no lineal}}$ tendría un riesgo verdadero elevado, ya que la "distancia" entre la función verdadera y la estimada es muy grande. Lo mismo ocurre en sentido contrario. En ambos casos, el riesgo verdadero sería mucho mayor que el riesgo empírico.

Este ejemplo resalta una decisión importante que debemos tomar: ¿preferimos ajustar los datos de entrenamiento con una función relativamente compleja, lo que conduce a un error de entrenamiento muy pequeño, o preferimos ajustarlos con una función simple a costa de un error de entrenamiento ligeramente mayor? En el ejemplo anterior, un físico que mida estos puntos de datos podría argumentar que no puede ser coincidencia que las mediciones estén casi alineadas y preferiría atribuir los residuos a errores de medición en lugar de a un modelo erróneo. Pero, ¿es posible caracterizar en qué sentido la línea recta es más simple y por qué esto debería implicar que está, de alguna manera, más cerca de la verdadera dependencia subyacente? Es decir, nos interesa saber cuál es el aumento en el error de entrenamiento que deberíamos estar dispuestos a tolerar para ajustar un modelo más simple.

De una forma u otra, esta cuestión ha ocupado durante mucho tiempo las mentes de los investigadores que estudian el problema del aprendizaje. En la estadística clásica, se ha estudiado como el dilema *sesgo-varianza* (*bias-variance tradeoff*). Si ajustamos un modelo lineal para cada conjunto de datos que encontramos, podríamos pensar que toda dependencia funcional es lineal. Pero no sería por la naturaleza de los procesos que generan los datos, sino un sesgo impuesto por nosotros. Por otro lado, si tomamos un polinomio de grado suficientemente alto para cualquier muestra, siempre podríamos ajustar perfectamente los datos, pero el modelo exacto que obtendríamos estaría sujeto a grandes fluctuaciones, dependiendo de qué tan precisas fueran nuestras mediciones en primer lugar. Esto implicaría que el modelo sufra de una gran varianza.

Una dicotomía relacionada es la existente entre el error de estimación y el error de aproximación. Si usamos una clase pequeña de funciones, incluso la mejor solución posible aproximará pobremente la dependencia real, mientras que una clase grande de funciones llevará a un error de estimación estadística alto. En la terminología del aprendizaje estádistico aplicado, el modelo complejo muestra **sobreajuste** (overfitting), mientras que el modelo lineal simple es más proclive a sufrir de **subajuste** (underfitting).

1.4.5. Los dilemas sesgo-varianza y estimación-aproximación

El ejemplo ilustrado en la Figura 1.1 ya señaló de manera intuitiva el problema de la complejidad del modelo: ¿cuándo un modelo es "más simple" que otro? ¿Es bueno que un modelo sea simple? ¿Qué tan simple? Ya hemos mencionado anteriormente que el objetivo de la clasificación es lograr un riesgo tan bueno como el del clasificador de Bayes. ¿Podríamos simplemente elegir $\mathcal{F}$ como el espacio $\mathcal{F}_{\text{all}}$ de todas las funciones, definir el clasificador $f_n := \arg\min_{f \in \mathcal{F}_{\text{all}}} (\mathcal{R}_{\text{emp}}(f))$, y obtener consistencia? Desafortunadamente, la respuesta es no. Luego veremos que, si optimizamos sobre clases de funciones $\mathcal{F}$ demasiado grandes, y en particular si hacemos $\mathcal{F}$ tan grande que contenga todos los clasificadores de Bayes

para todas las distribuciones de probabilidad P, esto conduce a la inconsistencia. Por lo tanto, si queremos aprender con éxito, necesitamos trabajar con una clase de funciones $\mathcal{F}$ más pequeña. Para investigar las propiedades contrapuestas de la complejidad del modelo y la generalización, queremos introducir algunas nociones que serán útiles más adelante.

Recordemos las definiciones f_n, $f_{\mathcal{F}}$ y f_{Bayes} introducidas anteriormente. Hemos visto que la consistencia de Bayes trata sobre la convergencia del término $R(f_n) - R(f_{\text{Bayes}})$. Es importante notar que podemos descomponer esta cantidad de la siguiente manera:

$$\mathcal{R}(f_n) - \mathcal{R}(f_{\text{Bayes}}) = \underbrace{\mathcal{R}(f_n) - \mathcal{R}(f_{\mathcal{F}})}_{\text{error de generalización}} + \underbrace{\mathcal{R}(f_{\mathcal{F}}) - \mathcal{R}(f_{\text{Bayes}})}_{\text{error de aproximación}}$$

Los dos términos en el lado derecho tienen nombres particulares: el primero se denomina **error de estimación** y el segundo **error de aproximación**. En la Figura 1.2 tenemos una ilustración. El primer término aborda la incertidumbre introducida por el proceso de muestreo aleatorio. Dado un conjunto de datos finitos, necesitamos estimar la mejor función en $\mathcal{F}$. Por supuesto, en este proceso cometeremos errores. Este error se denomina error de estimación. El segundo término no está influido por cantidades aleatorias. Trata sobre el error que cometemos al buscar la mejor función en un espacio de funciones $\mathcal{F}$ pequeño, en lugar de buscar la mejor función en el espacio F_{all} de todas las funciones posibles. La pregunta fundamental en este contexto es qué tan bien las funciones en $\mathcal{F}$ pueden aproximar a las funciones en F_{all} y de allí proviene el nombre error de aproximación.

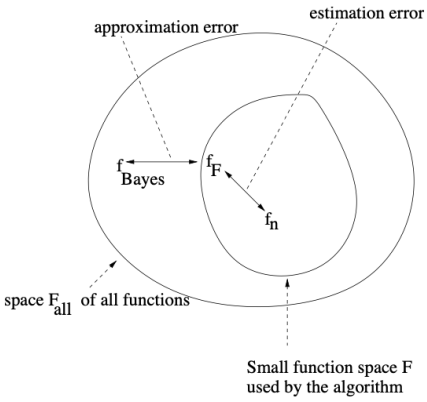

Figura 1.2: Error de estimación y de aproximación.

En estadística, el error de estimación también se llama **varianza**, y el error de aproximación se llama **sesgo** de un estimador. Originalmente, estos térmi-

nos se acuñaron para la situación especial de regresión con función de pérdida cuadrática, pero ahora se usan en contextos más generales, como el que se describe aquí. Su significado intuitivo es el mismo: el primer término mide la variación del riesgo de la función f_n estimada en la muestra, mientras que el segundo mide el sesgo introducido en el modelo al elegir una clase de funciones demasiado pequeña.

En este punto, ya podemos señalar que el espacio $\mathcal{F}$ es el medio para equilibrar el compromiso entre el error de estimación y el error de aproximación. Podemos hacernos una idea gráfica viendo la Figura . Si elegimos un espacio $\mathcal{F}$ muy grande, el término de aproximación será pequeño (el clasificador de Bayes podría incluso estar contenido en $\mathcal{F}$ o ser aproximado de manera cercana por algún elemento en $\mathcal{F}$). Sin embargo, el error de estimación será bastante grande en este caso: el espacio $\mathcal{F}$ contendrá funciones complejas que conducirán al **sobreajuste** (overfitting). El efecto opuesto ocurrirá si la clase de funciones $\mathcal{F}$ es muy pequeña.

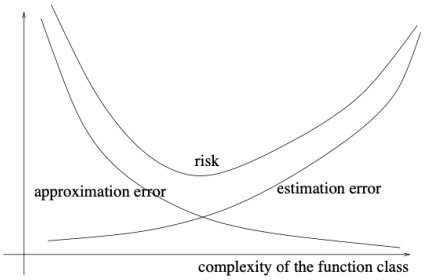

Figura 1.3: Compromiso entre error de estimación y de aproximación. Si el espacio funcional $\mathcal{F}$ que define el algoritmo de aprendizaje es pequeño, es decir es poco complejo, el error de estimación es bajo, pero el de aproximación es alto, y estamos en una situación proclive al subajuste. Si, por otro lado, $\mathcal{F}$ es complejo, el error de estimación es alto y el de aproximación bajo, y tendemos al sobreajuste. El error ideal es usualmente obtenido con una complejidad moderada.

En lo que sigue, trataremos el error de estimación y el error de aproximación por separado. Veremos que tienen comportamientos bastante diferentes y que se necesitan métodos distintos para controlar cada uno.

1.5. El clasificador de los k vecinos más cercanos

Hasta 1977, no se sabía si existía un clasificador universalmente consistente. Esta pregunta fue resuelta positivamente por Stone (1977), quien demostró, mediante una elegante prueba, que un clasificador particular, el denominado clasificador de los k-vecinos más cercanos (*k-nearest neighbors classifier*, abreviado como k-NN), es universalmente consistente. Como el clasificador de los

k-vecinos más cercanos es uno de los clasificadores más simples y todavía se usa ampliamente en la práctica, dedicaremos esta sección a ilustrar las nociones introducidas en la sección anterior, tales como generalización, sobreajuste, subajuste y consistencia, usando el ejemplo del clasificador k-NN.

Consideremos una muestra de puntos con etiquetas $(X_1, Y_1), \ldots, (X_n, Y_n)$ que pertenecen a un espacio métrico. De manera general, el paradigma del aprendizaje consiste en asignar salidas similares a entradas similares. Es decir, creemos que los puntos que están cercanos en el espacio de entrada tienden a tener la misma etiqueta en el espacio de salida. Nótese que si esta afirmación no se cumple, el aprendizaje se vuelve muy difícil o incluso imposible. Para un aprendizaje exitoso, debe existir alguna forma de relacionar las etiquetas de los puntos de entrenamiento con las de los puntos de prueba, y esto siempre implica suposiciones previas sobre las relaciones entre los puntos de entrada. La relación más simple es una distancia entre puntos, pero existen otras formas de medir la similitud, como los *kernels*, que forman la base de algunos de los algoritmos de aprendizaje más populares (Schölkopf y Smola, 2002).

Supongamos entonces que existe una función de distancia en el espacio de entrada, es decir, una función $d : X \times X \to \mathbb{R}$, que asigna un valor de distancia $d(X, X')$ a cada par de puntos de entrenamiento X, X'. Dados algunos puntos de entrenamiento, ahora queremos predecir una buena etiqueta para un nuevo punto X que no se halla en el conjunto de entrenamiento. Una idea simple es buscar el punto de entrenamiento X_i que tenga la distancia más pequeña a X y asignar a X la etiqueta correspondiente Y_i de ese punto. Para definir esto de manera más formal, denotamos por $\mathrm{NN}(X)$ al vecino más cercano de X entre todos los puntos de entrenamiento, es decir:

$$\mathrm{NN}(X) = \arg\min\{X' \in \{X_1, \ldots, X_n\} \mid d(X, X') \leq d(X, Z_2^{),}$$

para todo $Z_2^1 n\{X_1, \ldots, X_n\}\}$.

Luego, podemos definir el clasificador f_n basado en la muestra de n puntos como:

$$f_n(X) = Y_i \quad \text{donde} \quad X_i = \mathrm{NN}(X).$$

Este clasificador se denomina clasificador de un vecino más cercano (*1-nearest neighbor*, 1NN). Podemos generalizarlo en el clasificador de los k-vecinos más cercanos (*kNN*), considerando los k puntos de entrenamiento más cercanos, con $k > 1$ en este caso, y tomando el promedio de todas sus etiquetas.

Definición 1.5.0.1. *Dados un espacio métrico X junto a una función de distancia $d : X \times X \to \mathbb{R}$, un conjunto de puntos de entrenamiento con sus etiquetas $(X_1, Y_1), \ldots, (X_n, Y_n)$ y un entero $k \geq 1$, definimos los* ***k-vecinos más cercanos*** *de un punto X como el conjunto de los k puntos de entrenamiento más cercanos a X, es decir:*

$$kNN(X) = \{X_{i_1}, \ldots, X_{i_k}\} \quad \textit{donde} \quad i_1, \ldots, i_k = \arg \min_{1 \leq j \leq n} d(X, X_j).$$

Es decir, definimos los k-vecinos más cercanos de X, kNN(X), como el conjunto de los k puntos de entrenamiento más cercanos a X. Luego, el clasificador k-NN se define como:

Definición 1.5.0.2. *Con las mismas condiciones de la definición anterior, definimos el* ***clasificador de los k-vecinos más cercanos*** *como la función $f_n : X \to Y$ que asigna a un punto X la etiqueta que resulta de una votación mayoritaria entre las etiquetas de los puntos de entrenamiento en la vecindad de k-vecinos más cercanos de X:*

$$f_n(X) = \begin{cases} +1 & si \sum_{X_i \in kNN(X)} Y_i > 0, \\ -1 & en\ otro\ caso. \end{cases}$$

Es decir, decidimos la etiqueta de X mediante una votación mayoritaria entre las etiquetas de los puntos de entrenamiento en la vecindad de k-vecinos más cercanos de X. Para evitar empates, generalmente se elige k como un número impar.

Teorema 1.5.0.1. *El clasificador de un vecino más cercano (1NN) no es Bayes-consistente.*

Demostración. Consideremos el intervalo real $X = [0, 1]$ y la distribución de probabilidad $P([0, 1])$, que asigna etiquetas de manera uniforme a todos los puntos $X \in [0, 1]$ con ruido, de modo que $P(Y = 1 \mid X = x) = 0{,}9$ para todo $x \in X$. Es decir, la etiqueta correcta (la que asigna el clasificador de Bayes) es $+1$ para todos los puntos $x \in X$. Ya hemos mencionado este ejemplo cuando introdujimos el clasificador de Bayes. En este caso, el clasificador de Bayes es simplemente la función que devuelve 1 para todos los puntos en X, y su riesgo Bayesiano con respecto a la pérdida 0-1 será

$$\begin{aligned} \mathcal{R}(f_{\text{Bayes}}) &= \mathbb{E}(\ell(X, Y, f_{\text{Bayes}}(X))) \\ &= \mathbb{E}(1 \cdot P(f_{\text{Bayes}}(X) \neq Y) + 0 \cdot P(f_{\text{Bayes}}(X) = Y)) \\ &= \mathbb{E}(P(f_{\text{Bayes}}(X) \neq Y)) \\ &= \int_0^1 P(Y = 1 \mid X = x) \cdot \mathbb{I}_{\{f_{\text{Bayes}}(x) \neq 1\}}\, dx \\ &\qquad + \int_0^1 P(Y = 0 \mid X = x) \cdot \mathbb{I}_{\{f_{\text{Bayes}}(x) = 1\}} \\ &= \int_0^1 (0{,}9 \cdot 0 + 0{,}1 \cdot 1)\, dx = 0{,}1. \end{aligned}$$

Ahora investiguemos el comportamiento del clasificador 1NN en este caso. Al tomar puntos de entrenamiento $(X_i, Y_i)_{i=1,\dots,n}$ de acuerdo con la distribución subyacente, estos estarán aproximadamente uniformemente distribuidos en el intervalo $[0, 1]$. En promedio, cada décimo punto tendrá una etiqueta de entrenamiento $Y = -1$, y todos los demás tendrán etiqueta $Y = +1$. Si ahora consideramos el comportamiento del clasificador f_n, podemos escribir la probabilidad de que el clasificador 1NN cometa un error al etiqueta un punto como:

$$\begin{aligned} P(Y \neq f_n(X)) &= P(Y = 1 \mid f_n(X) = 0) + P(Y = 0 \mid f_n(X) = 1) \\ &= 0{,}1 \cdot 0{,}9 + 0{,}9 \cdot 0{,}1 \\ &= 2 \cdot 0{,}1 \cdot 0{,}9 = 0{,}18. \end{aligned}$$

Podemos ver que el riesgo $R(f_n)$ del clasificador f_n es también $0, 18$, independientemente del tamaño de la muestra n. En efecto:

$$\begin{aligned} \mathcal{R}(f_n) &= \mathbb{E}(\ell(X, Y, f_n(X))) \\ &= \int_0^1 1 \cdot P(Y \neq f_n(X = x)) + 0 \cdot P(Y = f_n(X = x))\, dx \\ &= \int_0^1 P(Y \neq f_n(X = x))\, dx \\ &= \int_0^1 0{,}18\, dx = 0{,}18. \end{aligned}$$

Por otro lado el riesgo Bayesiano es 0,1. Por lo tanto, el clasificador 1NN no es consistente, ya que $R(f_n) \not\to R(f_{\text{Bayes}})$.

□

Consideremos por un momento el clasificador de los 100 vecinos más cercanos, 100NN. En este caso, el clasificador cometería muchos menos errores que su primo de un vecino: es muy poco probable tener una vecindad de 100 puntos donde la mayoría de los votos sean $Y = -1$. Así, el clasificador de 100 vecinos más cercanos, aunque sigue sin ser consistente, comete un error menor que el clasificador 1NN.

El truco para lograr consistencia está relacionado con esta observación. Esencialmente, se debe permitir que el tamaño k de la vecindad bajo consideración crezca con el tamaño de la muestra n. Formalmente, se puede demostrar el siguiente teorema:

Teorema 1.5.0.2 (Stone, 1977)**.** *Sea f_n el clasificador de los k-vecinos más cercanos construido a partir de una muestra de n puntos. Si $n \to \infty$ y $k \to \infty$ de modo que $k/n \to 0$, entonces $R(f_n) \to R(f_{Bayes})$ para todas las distribuciones de probabilidad P. Es decir, la regla de clasificación k-NN es universalmente consistente con Bayes.*

Este teorema esencialmente nos dice que si elegimos el parámetro de vecindad k de forma que crezca lentamente con n, por ejemplo $k \approx \log(n)$, entonces la regla de clasificación k-NN es universalmente Bayes-consistente.

En las secciones anteriores mencionamos que la clase de funciones $\mathcal{F}$ de la cual se elige el clasificador es un componente importante para la teoría del aprendizaje estadístico. En el caso del clasificador k-NN, esto no es tan obvio como lo será para los clasificadores que estudiaremos en secciones posteriores. Intuitivamente, se puede decir que, para un parámetro fijo k, la clase de funciones $\mathcal{F}_k$ es un espacio de funciones constantes por partes. Cuanto mayor sea k, más grandes serán las vecindades de los k-vecinos y, por lo tanto, mayores serán los segmentos donde las funciones deben ser constantes. Esto significa que, para valores muy grandes de k, la clase de funciones $\mathcal{F}_k$ es relativamente pequeña (las funciones no pueden oscilar mucho). En el caso extremo de $k = n$, la vecindad de los k-vecinos simplemente incluye todos los puntos de entrenamiento, por lo que el clasificador k-NN no puede cambiar su signo en absoluto; debe ser constante en todo el espacio de entrada X. En este caso, la clase de funciones $\mathcal{F}_k$ contiene solo dos elementos: la función que es constantemente $+1$ y la función que es constantemente -1.

Por otro lado, si k es pequeño, entonces $\mathcal{F}_k$ se vuelve bastante grande (las funciones pueden cambiar sus etiquetas con mucha frecuencia y de manera abrupta). En los términos explicados en las secciones anteriores, podemos decir que si elegimos k demasiado pequeño, entonces la clase de funciones sobreajusta; por ejemplo, esto ocurre en el caso extremo del clasificador 1NN. Por el contrario, si k es demasiado grande, la clase de funciones subajusta, ya que simplemente no contiene funciones capaces de modelar los datos de entrenamiento.

1.6. Minimización del riesgo empírico

En la sección anterior encontramos nuestro primer clasificador simple: el clasificador k-NN. En esta sección, queremos abordar una forma más poderosa de clasificar datos, el llamado principio de **minimización del riesgo empírico**. Recordemos la suposición de que los datos son generados de manera *iid* (independientes e idénticamente distribuidos) a partir de una distribución subyacente desconocida $P(X, Y)$. Como ya hemos visto, el problema de aprendizaje consiste en minimizar el riesgo (o pérdida esperada sobre los datos de prueba):

$$\mathcal{R}(f) = \mathbb{E}(\ell(X, Y, f(X))),$$

donde f es una función que mapea el espacio de entrada X al espacio de etiquetas Y, y ℓ es la función de pérdida.

La dificultad de esta tarea radica en el hecho de que estamos intentando minimizar una cantidad que no podemos evaluar directamente: dado que no conocemos la distribución de probabilidad subyacente P, no podemos calcular el

riesgo $\mathcal{R}(f)$. Sin embargo, lo que sí conocemos son los datos de entrenamiento, muestreados a partir de P. Por lo tanto, podemos intentar inferir una función f a partir de la muestra de entrenamiento cuyo riesgo esté cercano al mejor riesgo posible. Para ello, necesitamos lo que se llama un *principio de inducción.*

Quizás la forma más directa de proceder sea aproximar el riesgo verdadero mediante el **riesgo empírico** calculado sobre los datos de entrenamiento. En lugar de buscar una función que minimice el riesgo verdadero $\mathcal{R}(f)$, intentamos encontrar aquella que minimice el riesgo empírico:

$$\mathcal{R}_{\text{emp}}(f) = \frac{1}{n} \sum_{i=1}^{n} \ell(X_i, Y_i, f(X_i)).$$

Es decir, dados algunos datos de entrenamiento $(X_1, Y_1), \ldots, (X_n, Y_n)$, un espacio de funciones F con el cual trabajar, y una función de pérdida ℓ, definimos el clasificador f_n como la función:

$$f_n := \arg \min_{f \in \mathcal{F}} \mathcal{R}_{\text{emp}}(f).$$

Este enfoque se denomina **principio de inducción de minimización del riesgo empírico**, abreviado como ERM (*Empirical Risk Minimization*). La motivación para este principio está dada por la ley de los grandes números, como explicaremos a continuación.

1.6.1. La ley de los grandes números

Recordemos que, en su forma más simple, la ley de los grandes números establece que, bajo condiciones suaves, la media de variables aleatorias ξ_i que han sido extraídas de manera *iid* (independiente e idénticamente distribuida) a partir de una distribución de probabilidad P, converge a la media de la distribución subyacente cuando el tamaño de la muestra tiende a infinito:

$$\frac{1}{n} \sum_{i=1}^{n} \xi_i \to \mathbb{E}(\xi) \quad \text{cuando } n \to \infty.$$

Aquí, la notación supone que la secuencia $\xi_1, \xi_2, \ldots$ ha sido muestreada de manera *iid* a partir de P y que ξ también está distribuida según P. Este teorema puede aplicarse al caso del riesgo empírico y el riesgo verdadero. Para ver esto, notemos que el riesgo empírico se define como la media de la pérdida $\ell(X_i, Y_i, f(X_i))$ en puntos de muestra individuales, y el riesgo verdadero es la media de esta pérdida sobre toda la distribución. Es decir, a partir de la ley de los grandes números, podemos concluir que, para una función fija f, el riesgo empírico converge al riesgo verdadero a medida que el tamaño de la muestra tiende a infinito:

$$\mathcal{R}_{\text{emp}}(f) = \frac{1}{n} \sum_{i=1}^{n} \ell(X_i, Y_i, f(X_i)) \to \mathbb{E}(\ell(X, Y, f(X))) \quad \text{cuando } n \to \infty.$$

Aquí, la función de pérdida $\ell(X, Y, f(X))$ desempeña el papel de la variable aleatoria ξ. Para una muestra finita dada, esto significa que podemos aproximar el riesgo verdadero (el que nos interesa) de manera bastante precisa mediante el riesgo empírico (el que podemos calcular sobre la muestra).

Una desigualdad famosa, atribuida a Chernoff (1952) y luego generalizada por Hoeffding (1963), caracteriza qué tan bien la media empírica aproxima el valor esperado. Específicamente, si las variables aleatorias ξ_i toman valores solo en el intervalo $[0, 1]$, entonces:

$$P\left(\left|\frac{1}{n}\sum_{i=1}^{n}\xi_i - \mathbb{E}(\xi)\right| \geq \epsilon\right) \leq 2\exp(-2n\epsilon^2). \tag{1.1}$$

Este teorema establece que la probabilidad de que la media de la muestra se desvíe más de ϵ del valor esperado de la distribución está acotada por una cantidad muy pequeña, específicamente $2\exp(-2n\epsilon^2)$. Nótese que, cuanto mayor sea n, más pequeña será esta cantidad; es decir, la probabilidad de desviaciones grandes disminuye rápidamente con n. Una vez más, podemos aplicar este teorema al contexto del riesgo empírico y verdadero. Esto conduce a una cota que establece qué tan probable es que el riesgo empírico esté cerca del riesgo verdadero para una función dada f:

$$P(|\mathcal{R}_{\text{emp}}(f) - \mathcal{R}(f)| \geq \epsilon) \leq 2\exp(-2n\epsilon^2). \tag{1.2}$$

Para cualquier función fija (y un n suficientemente grande), es muy probable que el error de entrenamiento proporcione una buena estimación del error de prueba.

Existen algunos hechos importantes relacionados con la cota de Chernoff (1.2). Primero, una propiedad crucial de la cota de Chernoff es que es de naturaleza probabilística. Establece que la probabilidad de una gran desviación entre el error de prueba y el error de entrenamiento de f es pequeña; cuanto mayor sea el tamaño de la muestra n, menor será esta probabilidad. Por lo tanto, no descarta la presencia de casos en los que la desviación sea grande; simplemente dice que, para una función fija f, esto es muy poco probable. La razón de esto radica en la generación aleatoria de los puntos de entrenamiento. Podría ocurrir que, en algunos casos desafortunados, nuestros datos de entrenamiento sean tan engañosos que sea imposible construir un buen clasificador a partir de ellos. Sin embargo, a medida que el tamaño de la muestra aumenta, tales casos desafortunados se vuelven muy raros. En este sentido, cualquier garantía de consistencia solo puede ser de la forma: el riesgo empírico está cerca del riesgo verdadero, con alta probabilidad.

A primera vista, parece que la cota de Chernoff (1.2) es suficiente para probar la consistencia de la minimización del riesgo empírico. Sin embargo, hay una advertencia importante: la cota de Chernoff solo se cumple para una función fija

f que no depende de los datos de entrenamiento. Sin embargo, el clasificador f_n, por supuesto, depende de los datos de entrenamiento (usamos los datos de entrenamiento para seleccionar f_n). Aunque esto pueda parecer una diferencia matemática sutil, aquí es donde la minimización del riesgo empírico puede fallar por completo. A continuación, discutiremos este problema en detalle y veremos cómo adaptar la ley fuerte de los grandes números para poder tratar funciones que dependen de los datos.

1.6.2. Inconsistencia en la minimización del riesgo empírico

Supongamos que nuestro espacio de datos subyacente es $X = [0, 1]$. Elegimos la distribución uniforme sobre X como la distribución de probabilidad y definimos la etiqueta Y para un punto de entrada X de manera determinista como sigue:

$$Y = \begin{cases} -1 & \text{si } X < 0{,}5, \\ 1 & \text{si } X \geq 0{,}5. \end{cases}$$

Ahora supongamos que se nos da un conjunto de puntos de entrenamiento $(X_i, Y_i)_{i=1,\ldots,n}$ y consideremos el siguiente clasificador:

$$f_n(X) = \begin{cases} Y_i & \text{si } X = X_i \text{ para algún } i = 1, \ldots, n, \\ 1 & \text{en otro caso.} \end{cases}$$

Este clasificador f_n clasifica perfectamente todos los puntos de entrenamiento. Es decir, tiene un riesgo empírico $\mathcal{R}_{\text{emp}}(f_n) = 0$. En consecuencia, dado que el riesgo empírico no puede ser negativo, f_n es un minimizador del riesgo empírico. Sin embargo, f_n claramente no ha aprendido nada; el clasificador simplemente memoriza las etiquetas de entrenamiento y, en otros casos, predice simplemente la etiqueta 1.

Formalmente, esto significa que el clasificador f_n no será consistente. Para ver esto, supongamos que se nos da un punto de prueba (X, Y) extraído de la distribución subyacente. Usualmente, este punto de prueba no será idéntico a ninguno de los puntos de entrenamiento, y en este caso el clasificador simplemente predice la etiqueta 1. Si X resulta ser mayor que 0,5, esta es la etiqueta correcta, pero si $X < 0{,}5$, es la etiqueta incorrecta. Por lo tanto, el clasificador f_n cometerá errores en la mitad de todos los puntos de prueba, lo que implica que su error de prueba es $R(f_n) = 1/2$. Este es el mismo error que se obtendría con una predicción aleatoria. De hecho, este es un buen ejemplo de **sobreajuste**: el clasificador f_n se ajusta perfectamente a los datos de entrenamiento, pero no aprende nada sobre los nuevos datos de prueba.

Es fácil ver que el clasificador f_n es inconsistente. Nótese que, como las etiquetas son una función determinista de los puntos de entrada, el clasificador de Bayes tiene un riesgo igual a 0. Así, tenemos:

$$\frac{1}{2} = R(f_n) \not\to R(f_{\text{Bayes}}) = 0.$$

Hemos construido un ejemplo donde la minimización del riesgo empírico falla de manera estrepitosa. ¿Existe alguna manera de rescatar el principio de ERM? Afortunadamente, la respuesta es sí. El objeto principal al que debemos prestar atención es la clase de funciones $\mathcal{F}$ de la cual extraemos nuestro clasificador. Si permitimos que nuestra clase de funciones contenga funciones que simplemente memorizan los datos de entrenamiento, entonces el principio de ERM no puede funcionar. En particular, si elegimos el minimizador del riesgo empírico del espacio F_{all} de todas las funciones entre X e Y, entonces los valores de f_n en los puntos de entrenamiento $X_1, \ldots, X_n$ no necesariamente contienen ninguna información sobre los valores en otros puntos. Por lo tanto, a menos que impongamos restricciones en el espacio de funciones del cual elegimos nuestra estimación f, no podemos esperar aprender nada.

1.6.3. Convergencia uniforme

Resulta que las condiciones necesarias para que la minimización del riesgo empírico sea consistente implican restringir el conjunto de funciones admisibles. La idea principal de la teoría de Vapnik-Chervonenkis (VC) es que la consistencia de la minimización del riesgo empírico está determinada por el comportamiento en el peor caso sobre todas las funciones $f \in \mathcal{F}$ que la máquina de aprendizaje podría elegir. Veremos que, en lugar de la ley estándar de los grandes números introducida anteriormente, este caso extremo corresponde a una versión de la ley de los grandes números que es uniforme sobre todas las funciones en $\mathcal{F}$.

La Figura 1.4 presenta una representación simplificada de la ley uniforme de los grandes números y la cuestión de la consistencia. Tanto el riesgo empírico como el riesgo verdadero se grafican como funciones de f. Para simplificar, hemos resumido todas las funciones posibles f en un solo eje del gráfico. La minimización del riesgo empírico consiste en elegir la función f que minimiza $\mathcal{R}_{\text{emp}}$. Es consistente si el mínimo de $\mathcal{R}_{\text{emp}}$ converge al de R a medida que el tamaño de la muestra aumenta. Una forma de garantizar la convergencia del mínimo para todas las funciones en $\mathcal{F}$ es la **convergencia uniforme** sobre $\mathcal{F}$: requerimos que para todas las funciones $f \in \mathcal{F}$, la diferencia entre $\mathcal{R}(f)$ y $\mathcal{R}_{\text{emp}}(f)$ se vuelva pequeña simultáneamente. Es decir, requerimos que exista un valor grande de n tal que, para un tamaño de muestra al menos n:

$$\sup_{f \in \mathcal{F}} |\mathcal{R}(f) - \mathcal{R}_{\text{emp}}(f)| \leq \epsilon.$$

Luego

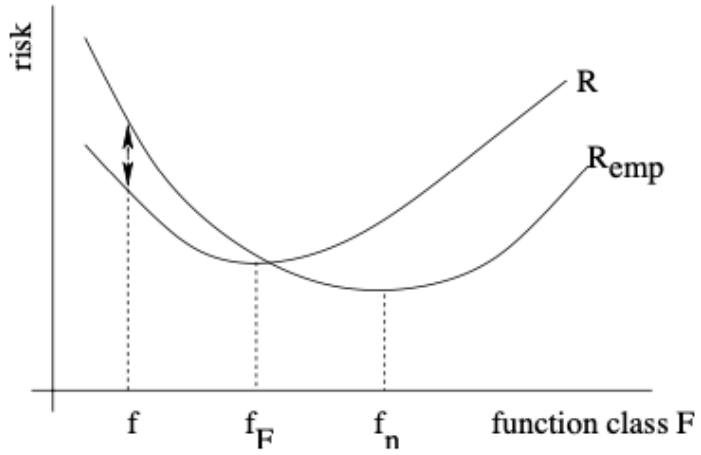

Figura 1.4: Representación simplificada de la convergencia del riesgo empírico al riesgo verdadero. El eje x representa una dimensión de la clase de funciones F, mientras que el eje y denota el riesgo. Para cada función fija f, la ley de los grandes números nos dice que, a medida que el tamaño de la muestra tiende a infinito, el riesgo empírico $\mathcal{R}_{\text{emp}}(f)$ converge al riesgo verdadero $\mathcal{R}(f)$ (indicado por la flecha). Sin embargo, esto no implica que, en el límite de tamaños de muestra infinitos, el minimizador del riesgo empírico, f_n, lleve a un valor del riesgo tan bueno como el del mejor f_F en la clase de funciones. Para que esto sea cierto, requerimos la convergencia uniforme de $\mathcal{R}_{\text{emp}}(f)$ a $\mathcal{R}(f)$ sobre todas las funciones en F. (Adaptado de Schölkopf y Smola, 2002).

$$|\mathcal{R}(f) - \mathcal{R}_{\text{emp}}(f)| \leq \sup_{f \in \mathcal{F}} |\mathcal{R}(f) - \mathcal{R}_{\text{emp}}(f)|.$$

En particular, esto también se cumple para una función f_n que se haya elegido en base a una muestra finita de puntos de entrenamiento

$$P(|R(f_n) - \mathcal{R}_{\text{emp}}(f_n)| \geq \epsilon) \leq P\left(\sup_{f \in \mathcal{F}} |\mathcal{R}(f) - \mathcal{R}_{\text{emp}}(f)| \geq \epsilon\right). \tag{1.3}$$

La cantidad en el lado derecho de la ecuación (1.3) es ahora el objeto de la ley uniforme de los grandes números. Decimos que la ley de los grandes números se cumple de manera uniforme sobre una clase de funciones $\mathcal{F}$ si, para todo $\epsilon > 0$,

$$P\left(\sup_{f \in \mathcal{F}} |\mathcal{R}(f) - \mathcal{R}_{\text{emp}}(f)| \geq \epsilon\right) \to 0 \quad \text{cuando } n \to \infty.$$

Ahora podemos usar la ecuación (1.3) para mostrar que, si la ley uniforme de los grandes números se cumple para alguna clase de funciones $\mathcal{F}$, entonces la minimización del riesgo empírico es consistente con respecto a $\mathcal{F}$. Para verlo, consideremos:

$$|R(f_n) - R(f_{\mathcal{F}})| = R(f_n) - R(f_{\mathcal{F}})$$

Puesto que $R(f_n) - R(f_{\mathcal{F}}) \geq 0$, por definición de $f_{\mathcal{F}}$. Sumando y restando cantidades idénticas, esta expresión se puede reescribir como:

$$\begin{aligned} &R(f_n) - \mathcal{R}_{\text{emp}}(f_n) + \mathcal{R}_{\text{emp}}(f_n) - \mathcal{R}_{\text{emp}}(f_{\mathcal{F}}) + \mathcal{R}_{\text{emp}}(f_F) - R(f_{\mathcal{F}}) \\ &\leq R(f_n) - \mathcal{R}_{\text{emp}}(f_n) + \mathcal{R}_{\text{emp}}(f_{\mathcal{F}}) - R(f_{\mathcal{F}}) \\ &\leq 2 \sup_{f \in \mathcal{F}} |\mathcal{R}(f) - \mathcal{R}_{\text{emp}}(f)|. \end{aligned}$$

Dado que $\mathcal{R}_{\text{emp}}(f_n) - \mathcal{R}_{\text{emp}}(f_{\mathcal{F}}) \leq 0$ por definición de f_n, y luego acotando al tomar el supremo de todas las funciones en el espacio $\mathcal{F}$.

De esto podemos concluir:

$$P\left(|R(f_n) - R(f_{\mathcal{F}})| \geq \epsilon\right) \leq P\left(\sup_{f \in \mathcal{F}} |\mathcal{R}(f) - \mathcal{R}_{\text{emp}}(f)| \geq \epsilon/2\right).$$

Bajo la ley uniforme de los grandes números, el lado derecho tiende a 0, lo que lleva a la consistencia de la minimización del riesgo empírico con respecto a la clase de funciones subyacente $\mathcal{F}$. En otras palabras, la convergencia uniforme sobre $\mathcal{F}$ es una condición suficiente para la consistencia de la minimización del riesgo empírico sobre $\mathcal{F}$.

Parte de la elegancia de la teoría VC radica en que el recíproco también es cierto. Es decir, la convergencia uniforme no solo es una condición suficiente, sino también una condición necesaria para la consistencia de la minimización del riesgo empírico con respecto a $\mathcal{F}$. Esto se formaliza en el siguiente teorema:

Teorema 1.6.3.1 (Vapnik y Chervonenkis, 1971)**.** *La convergencia uniforme*

$$P\left(\sup_{f \in \mathcal{F}} |\mathcal{R}(f) - \mathcal{R}_{emp}(f)| > \epsilon\right) \to 0 \quad \textit{cuando } n \to \infty,$$

para todo $\epsilon > 0$, es una condición necesaria y suficiente para la consistencia de la minimización del riesgo empírico con respecto a $\mathcal{F}$.

En la sección 1.6.2 dimos un ejemplo donde consideramos el conjunto de todas las funciones posibles y mostramos que el aprendizaje era imposible. Ahora, la dependencia del aprendizaje en el conjunto de funciones subyacente ha regresado bajo una forma diferente: la condición de convergencia uniforme depende críticamente del conjunto de funciones para el cual debe cumplirse.

Intuitivamente, parece claro que cuanto mayor sea el espacio de funciones $\mathcal{F}$, mayor será:

$$\sup_{f \in \mathcal{F}} |\mathcal{R}(f) - \mathcal{R}_{\text{emp}}(f)|.$$

Por lo tanto, cuanto mayor sea $\mathcal{F}$, mayor será la medida de probabilidad:

$$P\left(\sup_{f \in \mathcal{F}} |\mathcal{R}(f) - \mathcal{R}_{\text{emp}}(f)| > \epsilon\right).$$

En consecuencia, cuanto mayor sea $\mathcal{F}$, más difícil será satisfacer la ley uniforme de los grandes números. Esto implica que para espacios de funciones más grandes, la consistencia es más difícil de lograr en comparación con espacios de funciones más pequeños.

Esta caracterización abstracta de la consistencia como una propiedad de convergencia uniforme es teóricamente fascinante, pero no siempre es útil en la práctica. La razón es que parece complicado determinar si la ley uniforme de los grandes números se cumple para una clase de funciones $\mathcal{F}$ particular. Por esta razón, se han desarrollado herramientas como la dimensión VC para analizar y limitar la capacidad de las clases de funciones, lo cual discutiremos en las siguientes secciones.

1.7. Cotas de generalización y medidas de capacidad

Para hacer afirmaciones sobre lo que ocurre después de observar solo un número finito de puntos de datos —que en la práctica será siempre el caso—, necesitamos examinar más de cerca la convergencia uniforme del riesgo empírico al verdadero.

Resultará que esto nos proporcionará cotas sobre el riesgo y también nos dará información sobre qué propiedades de las clases de funciones determinan si la convergencia uniforme puede tener lugar. Detengámonos en la probabilidad del Teorema anterior:

$$P\left(\sup_{f \in \mathcal{F}} |\mathcal{R}(f) - \mathcal{R}_{\text{emp}}(f)| > \epsilon\right). \tag{1.4}$$

Consideremos entonces un espacio de funciones finito, como el que nos encontramos en la práctica. Sea $\mathcal{F}_{emp} = \{f_1, f_2, \ldots, f_m\}$ un conjunto de m funciones. Cada una de las funciones $f_i \in \mathcal{F}_{emp}$ satisface la ley estándar de los grandes números en la forma de la cota de Chernoff:

$$P\left(|R(f_i) - \mathcal{R}_{\text{emp}}(f_i)| \geq \epsilon\right) \leq 2\exp(-2n\epsilon^2).$$

Ahora queremos transformar estas afirmaciones sobre funciones individuales f_i en una ley uniforme de los grandes números. Dada la subatividad de la probabilidad, tenemos:

$$P\left(\sup_{f\in\mathcal{F}}|\mathcal{R}(f)-\mathcal{R}_{\text{emp}}(f)|\geq\epsilon\right)\leq\sum_{i=1}^{m}P\left(|R(f_i)-\mathcal{R}_{\text{emp}}(f_i)|\geq\epsilon\right).$$

Y aplicando la cota de Chernoff a cada término en la suma, obtenemos:

$$P\left(\sup_{f\in\mathcal{F}}|\mathcal{R}(f)-\mathcal{R}_{\text{emp}}(f)|\geq\epsilon\right)\leq 2m\exp(-2n\epsilon^2). \tag{1.5}$$

Si el espacio de funciones $\mathcal{F}_{emp}$ es fijo, m puede considerarse una constante, y el término $2m\exp(-2n\epsilon^2)$ converge a 0 cuando $n\to\infty$. Por lo tanto, la minimización del riesgo empírico sobre un conjunto finito de funciones es consistente con respecto a $\mathcal{F}_{emp}$.

En la práctica, nuestros espacios de funciones estarán definidos por los paramétros que nos permitimos seleccionar para cada posible clasificador, como la ordenada al origen y la pendiente en el caso de la regresión lineal y serán usualmente , como en ese ejemplo, espacios infinitos. Luego veremos que podemos llegar a una desigualdad similar a esta última, reemplazando m por un valor que depende de cada tipo de espacio funcional y que, de ser finito, será necesariamente polinómico, lo que nos permitirá obtener cotas de generalización. Llamaremos a dicha cantidad, la *dimensión VC*.

1.7.1. Simetrización

La *simetrización* es un paso técnico importante para usar medidas de capacidad en espacios infinitos de funciones. Su propósito principal es reemplazar el evento

$$\sup_{f\in\mathcal{F}}|\mathcal{R}(f)-\mathcal{R}_{\text{emp}}(f)|$$

por un evento alternativo que pueda calcularse únicamente en una muestra dada. Supongamos que tenemos una muestra $(X_i,Y_i)_{i=1,\ldots,n}$. Ahora introducimos una nueva muestra llamada *muestra fantasma*. Esta es simplemente otra muestra $(X_i',Y_i')_{i=1,\ldots,n}$, que también se extrae de manera *iid* de la misma distribución subyacente y que es independiente de la primera muestra. No necesitamos generar esta muestra en la práctica; es solo una herramienta matemática para realizar el análisis.

Dfinimos el riesgo empírico con respecto a esta muestra como $\mathcal{R}'_{\text{emp}}(f)$. Con la ayuda de esta muestra fantasma, podemos demostrar el siguiente resultado:

Lema 1.7.1.1 (de simetrización). *Sea $m\epsilon^2 \geq 2$. Sea $\mathcal{F}$ el espacio de funciones definido por un algoritmo de aprendizaje. Sean $(X_i, Y_i)_{i=1,\ldots,n}$, $(X_i', Y_i')_{i=1,\ldots,n}$ muestras aleatorias iid. Dada $f \in \mathcal{F}$, sean $\mathcal{R}(f)$ su riesgo, $\mathcal{R}_{emp}(f)$ su riesgo empírico sobre la primera muestra y $\mathcal{R}'_{emp}(f)$ el riesgo sobre la segunda. Entonces, para cualquier $\epsilon > 0$*

$$P\left(\sup_{f\in\mathcal{F}} |\mathcal{R}(f) - \mathcal{R}_{emp}(f)| > \epsilon\right) \leq 2P\left(\sup_{f\in\mathcal{F}} |\mathcal{R}_{emp}(f) - \mathcal{R}'_{emp}(f)| > \frac{\epsilon}{2}\right).$$

Esta desigualdad, nos permite acotar el comportamiento de la diferencia $\mathcal{R}(f) - \mathcal{R}_{\text{emp}}(f)$ mediante la diferencia entre riesgos empíricos $\mathcal{R}_{\text{emp}}(f)$ y $\mathcal{R}'_{\text{emp}}(f)$ calculados sobre las muestras original y fantasma, respectivamente. Esto nos permite trabajar con eventos que depende únicamente de las muestras observadas.

Notemos que, incluso si $\mathcal{F}$ contiene un número infinito de funciones, las diferentes formas en que estas pueden clasificar un conjunto de entrenamiento de n puntos de muestra es finita. En efecto, para cualquier punto de entrenamiento dado en la muestra, una función puede tomar solo los valores -1 o $+1$. En una muestra de n puntos $\{X_1, \ldots, X_n\}$, una función puede actuar de, como máximo, 2^n formas diferentes: puede asignar a cada Y_i el valor -1 o $+1$. Esto tiene una consecuencia muy importante. Incluso si una clase de funciones $\mathcal{F}$ contiene infinitas funciones, hay como mucho 2^n formas diferentes en que esas funciones pueden clasificar los puntos de una muestra finita de n puntos. Tomando

$$\sup_{f\in\mathcal{F}} |R_{\text{emp}}(f) - R'_{\text{emp}}(f)|,$$

entonces el supremo efectivamente solo recorre una clase de funciones finita. Para entender esto, notemos que dos funciones $f, g \in \mathcal{F}$ que toman los mismos valores en la muestra dada tienen el mismo riesgo empírico, es decir, $R_{\text{emp}}(f) = R_{\text{emp}}(g)$. La afirmación análoga se cumple para la muestra fantasma y el riesgo empírico asociado R'_{emp}.

AGREGAR FIGURA DE EJEMPLO CON DOS FUNCIONES LINEALES SEPARANDO DE IGUAL MANERA LAS MUESTRAS

Por lo tanto, todas las funciones f, g que coinciden tanto en la muestra original como en la muestra fantasma producirán el mismo término $|R_{\text{emp}}(f) - R'_{\text{emp}}(f)|$. Así, las únicas funciones que necesitamos considerar para calcular el supremo son las 2^n funciones que podemos obtener en la muestra original y la muestra fantasma juntas. Por lo tanto, podemos reemplazar el supremo sobre $f \in \mathcal{F}$ por el supremo sobre una clase finita de funciones con como máximo 2^n funciones.

1.7.2. El coeficiente de fragmentación

Con la intención de analizar más a fondo la capacidad de una clase de funciones $\mathcal{F}$, intentaremos acotar la medida (1.4). Para esto, introducimos el concepto de coeficiente de fragmentación. Este coeficiente mide la cantidad de formas en que las funciones de $\mathcal{F}$ pueden separar los puntos de cualquier muestra tomada sobre X

Definición 1.7.2.1. *Sean $\mathcal{F}$ una clase de funciones y $Z_n := \{(X_1, Y_1), \ldots, (X_n, Y_n)\}$ un conjunto de n puntos etiquetados. Definimos la relación $\sim$ sobre $\mathcal{F}$:*

$$f_1 \sim f_2 \qquad si \qquad f_1(X_i) = f_2(X_i) \quad \forall i = 1, \ldots, n.$$

Observación 1.7.2.1. *La relación $\sim$ es una relación de equivalencia sobre $\mathcal{F}$. Luego, para una muestra fija, define una partición sobre X.*

Definición 1.7.2.2. *Definimos el **conjunto de fragmentación** $\mathcal{F}_{Z_n}$ a cualquier conjunto de representantes de $\mathcal{F}$ bajo la relación de equivalencia $\sim$.*

Definición 1.7.2.3. *Dado un espacio muestral X y $Z_n := \{(X_1, Y_1), \ldots, (X_n, Y_n)\}$ una muestra de tamaño n. Definimos el **coeficiente de fragmentación** del espacio funcional $\mathcal{F}$ como:*

$$\mathcal{N}(\mathcal{F}, n) = \max_{Z_n}\{|\mathcal{F}_{Z_n}|\} = \max_{\{X_1,\ldots,X_n\}}\{|\mathcal{F}_{\{X_1,\ldots,X_n\}}| \mid X_1, \ldots, X_n \in X\}.$$

El coeficiente $\mathcal{N}(\mathcal{F}, n)$ tiene una interpretación sencilla: mide cuántos conjuntos de etiquetas $\{Y_1, \ldots, Y_n\}$ pueden ser generados por las funciones de $\mathcal{F}$, considerando todas las posibles muestras de tamaño n. Es decir, es el número de formas en las que $\mathcal{F}$ puede separar en dos clases al espacio de entrada X.

En el caso de clasificación binaria, como el que estamos tratando, el mayor cardinal posible para un conjunto de fragmentación coincide con el cardinal del conjunto de partes de un conjunto de tamaño n (basta considerar cada subconjunto posible del mismo como aquel donde el clasificador etiqueta con $+1$ todos los valores en el subconjunto y con -1 los valores fuera de él). Entonces, si $\mathcal{N}(\mathcal{F}, n) = 2^n$, existe al menos una muestra de tamaño n que puede ser separada de todas las formas posibles por funciones en $\mathcal{F}$. En este caso, decimos que la clase de funciones $\mathcal{F}$ *fragmenta* al conjunto de n puntos.

En la definición, tomamos el máximo puesto que ciertas muestras pueden no ser fragmentables de la misma manera que otras por $\mathcal{F}$. Consideremos el espacio $\mathbb{R}^2$ y sea $\mathcal{F} = \{f : f(x) = \alpha x + \beta;\ \alpha, \beta \in \mathbb{R}\}$, el espacio de funciones lineales.

Es claro que podemos separar dos puntos de la manera que deseemos en este espacio, pero no podemos hacer lo mismo con tres puntos *cualesquiera*. En

efecto, sea $Z_3 = \{X_1, X_2, X_3\} \in \mathbb{R}^2$ y supongamos que los puntos en dicho conjuntos son colineales. Tomemos el caso en el que el primero y el tercero de ellos pertenecen a la misma categoría, mientras que el que queda en medio a la otra categoría, como en la figura 1.5 (a).

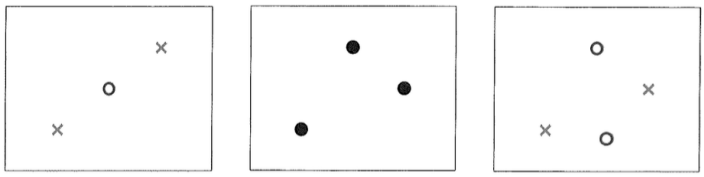

Figura 1.5: Fragmentación de muestras en $\mathbb{R}^2$ de tres y cuatro puntos.

Es claro que no es posible separar de todas las maneras posibles a esta muestra en particular, por lo que su conjunto de fragmentación tiene un cardinal menor a $2^3 = 8$. Pero sí es posible hacerlo con otras muestras de tres elementos, como por ejemplo cualquiera que tenga tres puntos no colineales, como en la figura 1.5 (b). De esto sigue que los clasificadores lineales de este tipo tienen un coeficiente de fragmentación máximo de $2^3 = 4$ para muestras de tres puntos en $\mathbb{R}^2$. Pero notemos que no es así cuando tomamos una muestra de cuatro puntos en $\mathbb{R}^2$. En este caso, es imposible separar de todas las maneras posibles a una muestra de cuatro puntos, sean colineales o no, como se aprecia en la parte (c) de la figura. Demostraremos estos hechos más adelante.

El coeficiente de fragmentación es una medida de la capacidad de una clase de funciones, es decir, mide el tamaño o la *complejidad* de $\mathcal{F}$ en un sentido particular. Si $\mathcal{F}$ contiene muchas funciones, entonces $\mathcal{N}(\mathcal{F}, n)$ tiende a ser grande. Sin embargo, el coeficiente de fragmentación también tiene en cuenta la relación entre $\mathcal{F}$ y el espacio de entrada de donde provienen los puntos de muestra. Más aún, nos permite asignar un número finito a cada espacio de clasificadores para cada tamaño de muestras, aún cuando los espacios de funciones sean infinitos. Este número puede ser en el peor de los casos exponencial en el tamaño de la muestra, pero nos ayudará a pasar del factor m, que depende de la cantidad de funciones en el espacio (en la ecuación 1.5), a un valor que es finito y, como veremos más adelante, potencialmente polinómico.

1.7.3. Cotas de convergencia uniforme

Mantenemos la atención sobre la ecuación (1.4). Dado un espacio de funciones arbitrario, posiblemente infinito, ahora queremos analizar la probabilidad de que el riesgo empírico difiera significativamente del riesgo verdadero. Con las

herramientas anteriores, podemos derivar una cota como sigue. Consideremos una muestra de $2n$ puntos, es decir, un conjunto Z_{2n}, donde interpretamos los primeros n puntos como la muestra original y los otros n puntos como la muestra fantasma.

La idea es reemplazar el supremo sobre $\mathcal{F}$ en términos de $R(f)$ y $R_{\text{emp}}(f)$ por un supremo sobre los riesgos empíricos calculados en ambas muestras, es decir, por un expresión que utilice el supremo sobre $\mathcal{F}_{Z_{2n}}$. En este espacio, existen a lo sumo $\mathcal{N}(\mathcal{F}, n) < 2^{2n}$ funciones distintas, lo que nos permite reemplazar, como indicabamos, el valor potencialmente infinito m por el coeficiente de fragmentación. Escribimos:

$$\begin{aligned} P\left(\sup_{f\in\mathcal{F}} |R(f) - R_{\text{emp}}(f)| > \epsilon\right) &\leq 2P\left(\sup_{f\in\mathcal{F}} |R_{\text{emp}}(f) - R'_{\text{emp}}(f)| > \frac{\epsilon}{2}\right). \\ &= 2P\left(\sup_{f\in\mathcal{F}_{Z_{2n}}} |R_{\text{emp}}(f) - R'_{\text{emp}}(f)| > \frac{\epsilon}{2}\right). \\ &\leq 2\mathcal{N}(\mathcal{F}, n) \cdot \exp\left(\frac{-n\epsilon^2}{4}\right). \end{aligned}$$

Primero, utilizando el argumento de simetrización, como antes. Luego, el supremo puede restringirse a un conjunto finito de funciones dado que estamos considerando solo los riesgos empíricos sobre dos muestras de tamaño n, que consideramos fijas, y de allí igualamos en el segundo paso. Por último, aplicamos la cota de Chernoff (1.2), sabiendo que el numero de funciones posibles m es a lo sumo el coeficiente de fragmentación.

Para formalizar esto, usamos el coeficiente de fragmentación $\mathcal{N}(\mathcal{F}, 2n)$, que proporciona el número máximo de particiones que $\mathcal{F}$ puede realizar sobre una muestra de $2n$ puntos. Aplicando el límite de la unión y la cota de Chernoff, obtenemos:

$$P\left(\sup_{f\in\mathcal{F}} |R(f) - R_{\text{emp}}(f)| > \epsilon\right) \leq 2\mathcal{N}(\mathcal{F}, 2n) \exp\left(-\frac{n\epsilon^2}{4}\right). \tag{1.6}$$

Esta es una cota de convergencia uniforme que conecta el coeficiente de fragmentación con la probabilidad de que el riesgo empírico se desvíe significativamente del riesgo verdadero. Para que la minimización del riesgo empírico sea consistente, es suficiente que $\mathcal{N}(\mathcal{F}, n)$ crezca a un ritmo razonablemente lento, por ejemplo, polinomialmente en n.

Tomemos $\mathcal{F}_{all}$, el espacio de todos los clasificadores posibles. Es claro que, en este caso, $\mathcal{N}(\mathcal{F}_{all}, n) = 2^{2n}$, pues cualquier muestra de $2n$ puntos puede ser separada de 2^{2n} formas posibles. Entonces, la cota de convergencia uniforme para este espacio es:

$$P\left(\sup_{f\in\mathcal{F}_{all}}|R(f)-R_{\text{emp}}(f)|>\epsilon\right)\leq 2{\cdot}2^{2n}\exp\left(-\frac{n\epsilon^2}{4}\right)=2\exp\left(n\left(2\log(2)-\frac{n\epsilon^2}{4}\right)\right).$$

Que es una expresión que no tiende a 0 cuando n tiende a infinito. Con lo que sabemos hasta ahora, esto no significa que podamos concluir que la minimización del riesgo empírico usando el espacio de todas las funciones clasificadoras posibles sea inconsistente, dado que la cota de convergencia uniforme anterior es solo una condicion suficiente. Pero si podemos dar el siguiente resultado para definir una condición necesaria y suficiente para la consistencia de la minimización del riesgo empírico.

Teorema 1.7.3.1. *La minimización del riesgo empírico es consistente con respecto a un espacio de funciones $\mathcal{F}$ si y solo si*

$$\frac{\log\left(\mathcal{N}(\mathcal{F},n)\right)}{n}\to 0 \qquad \textit{cuando } n\to\infty.$$

Demostración. cf Mendelson 2003. Para teoremas relacionados, ver Vapnik y Chernovenkis 1971,1981, section 12.4, Devroye et al 1996.

□

Corolario 1.7.3.1.1. *Si $\mathcal{N}(\mathcal{F},n)$ es polinómica, entonces la minimización del riesgo empírico es consistente con respecto a $\mathcal{F}$.*

Corolario 1.7.3.1.2. *Si $\mathcal{N}(\mathcal{F},n)=2^n$, entonces la minimización del riesgo empírico no es consistente con respecto a $\mathcal{F}$. En particular, no es consistente con respecto al espacio de todas los clasificadores posibles $\mathcal{F}_{all}$.*

1.7.4. Cotas de generalización

Es a veces útil reescribir la ecuación (1.6) de forma inversa. Es decir, en lugar de fijar ϵ y luego calcular la probabilidad de que el riesgo empírico se desvíe del riesgo verdadero en más de ϵ, podemos especificar la probabilidad con la que queremos que la cota se cumpla, y luego obtener una afirmación que nos indique qué tan cerca podemos esperar que esté el riesgo del riesgo empírico. Es decir, fijamos $\delta>0$ y resolviendo para ϵ. Como resultado, obtenemos que, con una probabilidad al menos $1-\delta$, cualquier función $f\in\mathcal{F}$ satisface:

$$R(f)\leq R_{\text{emp}}(f)+\sqrt{\frac{4}{n}\log(2\mathcal{N}(\mathcal{F},n))-\log(\delta)}. \tag{1.7}$$

De la misma manera que antes, podemos usar esta cota para derivar afirmaciones de consistencia. Por ejemplo, ahora es evidente que la minimización del riesgo empírico es consistente para la clase de funciones $\mathcal{F}$ si el término

$\log(2\mathcal{N}(\mathcal{F}, 2n))/n$ converge a 0 cuando $n \to \infty$. Una vez más, este es el caso si $\mathcal{N}(\mathcal{F}, 2n)$ crece de manera polinómica con n.

Nótese que la cota (1.7) se cumple para todas las funciones $f \in \mathcal{F}$. Por un lado, esto es una fortaleza de la cota, ya que se cumple en particular para la función f_n que minimiza el riesgo empírico, que es lo que buscamos. Además, muchos algoritmos de aprendizaje no minimizan exactamente el riesgo empírico, por lo que la cota también se cumple para ellos. Sin embargo, esto también puede interpretarse como una debilidad, ya que, al considerar un espacio más grande del necesario, obtenemos una cota menos ajustada de lo que podríamos si solo considerásemos funciones en el espacio del algoritmo de aprendizaje.

Intentemos obtener una intuición sobre esta cota. Nos dice que si tanto $R_{\text{emp}}(f)$ como el término de raíz cuadrada son pequeños simultáneamente, entonces podemos garantizar que, con alta probabilidad, el riesgo (es decir, el error en puntos futuros que aún no hemos visto) será pequeño. Esto puede parecer una afirmación sorprendente, sin embargo, no afirma nada imposible. Si usamos una clase de funciones con un $\mathcal{N}(\mathcal{F}, n)$ relativamente pequeño, es decir, una clase de funciones que no puede explicar muchas funciones posibles, y luego notamos que usando una función de esta clase podemos explicar los datos muestreados del problema en cuestión, entonces es probable que esto no sea una coincidencia y hayamos capturado algunos aspectos esenciales del problema.

Por otro lado, si el problema es demasiado difícil de aprender a partir de la cantidad de datos dada, entonces descubriremos que, para explicar los datos (es decir, para lograr un $R_{\text{emp}}(f)$ pequeño), necesitamos una clase de funciones que sea tan grande que pueda, en esencia, explicar cualquier cosa. En ese caso, el término de raíz cuadrada sería grande. Finalmente, nótese que la dificultad de aprender un problema está completamente determinada por si podemos encontrar una clase de funciones adecuada y, por lo tanto, por nuestro conocimiento previo sobre él. Incluso si la función óptima es subjetivamente muy compleja, si nuestra clase de funciones contiene esa función y pocas o ninguna otra, estamos en una excelente posición para aprender.

Existe una gran cantidad de cotas similares a (1.6) y su forma alternativa (1.7). Las diferencias ocurren en las constantes, tanto en el coeficiente frente a la exponencial como en su exponente. Las cotas también difieren en el exponente de ϵ y en la forma en que miden la capacidad.

1.7.5. La dimensión VC

Hasta ahora hemos definido cotas de generalización en términos del coeficiente de framentación. Aunque este es útil, a menudo es difícil calcularlo para clases de funciones arbitrarias. Existen otras medidas de capacidad que son más fáciles de calcular y que proporcionan cotas más ajustadas, con ventajas y des-

ventajas. Introducimos la más conocida de todas, la llamada **dimensión VC** (dimensión de Vapnik-Chervonenkis). En pocas palabras, esta medida intenta caratecterizar el crecimiento del coeficiente de fragmentación a medida que n crece, destilado en un solo valor.

Decimos que una muestra Z_n de tamaño n es fragmentada por la clase de funciones $\mathcal{F}$ si $\mathcal{N}(\mathcal{F}, n) = 2^n$. Es decir, $\mathcal{F}$ puede realizar cualquier separación posible de etiquetas sobre los puntos de la muestra. La dimensión VC de $\mathcal{F}$, denotada como $\mathrm{VC}(\mathcal{F})$, se define como el tamaño máximo de una muestra que puede ser fragmentada por $\mathcal{F}$. Formalmente:

Definición 1.7.5.1. *La dimensión VC de una clase de funciones $\mathcal{F}$ es el tamaño máximo de una muestra que puede ser fragmentada por $\mathcal{F}$, es decir:*

$$VC(\mathcal{F}) = \max\{n \in \mathbb{N} \mid \mathcal{N}(\mathcal{F}, n) = 2^n\}.$$

Si no existe tal máximo, definimos $VC(\mathcal{F}) = \infty$.

La dimensión VC tiene una interpretación combinatoria elegante y proporciona una forma práctica de analizar la capacidad de $\mathcal{F}$. Por ejemplo, si $\mathrm{VC}(\mathcal{F})$ es finita, podemos garantizar la consistencia de la minimización del riesgo empírico. Además, una clase de funciones con dimensión VC más baja tiende a tener una mejor capacidad de generalización.

Lema 1.7.5.1. *Sea $\mathcal{F}$ una clase de funciones con dimensión VC finita. Entonces*

$$\mathcal{N}(\mathcal{F}, n) \leq \sum_{i=0}^{VC(\mathcal{F})} \binom{n}{i}, \qquad \forall n \in \mathbb{N}.$$

En particular, para $n \geq VC(\mathcal{F})$ se tiene que

$$\mathcal{N}(\mathcal{F}, n) \leq \left(\frac{e \cdot n}{VC(\mathcal{F})} \right)^{VC(\mathcal{F})}.$$

La importancia de esta afirmación radica en el último hecho. Si $n \geq \mathrm{VC}(\mathcal{F})$, el coeficiente de fragmentación se comporta como una función polinómica del tamaño de la muestra n. Este es un resultado muy notable: una vez que sabemos que la dimensión VC de una clase de funciones $\mathcal{F}$ es finita, ya sabemos que los coeficientes de fragmentación crecen polinómicamente con n. Por los resultados de la sección anterior, esto implica la consistencia de la minimización del riesgo empírico.

Nótese que también tenemos una afirmación en la otra dirección. Si la dimensión VC es infinita, esto significa que para cada n existe alguna muestra que

puede ser fragmentada por $\mathcal{F}$, es decir, tal que $\mathcal{N}(\mathcal{F}, n) = 2^n$. Para este caso, ya vimos anteriormente que ERM no es consistente. Enunciamos todo esto en el siguiente Teorema.

Teorema 1.7.5.1. *La minimización del riesgo empírico es consistente con respecto a una clase de funciones $\mathcal{F}$ si y solo si $VC(\mathcal{F}) < \infty$.*

Una propiedad importante a notar tanto sobre el coeficiente de fragmentación como sobre la dimensión VC es que no dependen de la distribución subyacente P, sino únicamente de la clase de funciones $\mathcal{F}$. Por un lado, esto es una ventaja, ya que todas las cotas de generalización derivadas de estos conceptos se aplican a todas las distribuciones de probabilidad posibles. Por otro lado, esto también puede considerarse una desventaja, ya que los conceptos de capacidad no tienen en cuenta propiedades particulares de la distribución en cuestión. En este sentido, estos conceptos de capacidad a menudo conducen a cotas bastante amplias, pero es lo que desde un comienzo nos interesaba.

1.7.6. Complejidad de Rademacher

Un concepto diferente para medir la capacidad de un espacio funcional es la **complejidad de Rademacher**. En comparación con el coeficiente de fragmentación y la dimensión VC, esta sí depende de la distribución de probabilidad subyacente y, por lo general, conduce a cotas mucho más ajustadas que ambas.

La complejidad de Rademacher se define de la siguiente manera.

Definición 1.7.6.1. *Sean $\sigma_1, \sigma_2, \ldots$ variables aleatorias independientes que toman los valores $+1$ y -1 con probabilidad 0,5. Definimos la complejidad de Rademacher $\mathscr{R}(\mathcal{F})$ de un espacio funcional $\mathcal{F}$ como:*

$$\mathscr{R}(\mathcal{F}) := \mathbb{E} \sup_{f \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^{n} \sigma_i f(X_i).$$

Cada una de estas variables aleatorias se denominan *variables de Rademacher*. Por ejemplo, pueden interpretarse como los resultados de lanzar repetidamente una moneda no cargada.

Consideremos primero los valores σ_i como fijos e interpretémoslos como etiquetas asignadas a los puntos X_i. Como tanto σ_i como $f(X_i)$ solo toman los valores $+1$ o -1, el producto $\sigma_i f(X_i)$ toma el valor $+1$ si $\sigma_i = f(X_i)$, y -1 si $\sigma_i \neq f(X_i)$.

Como consecuencia, la suma en el lado derecho de la ecuación será grande si las etiquetas $f(X_i)$ coinciden con las etiquetas σ_i en muchos puntos de datos.

Esto significa que la función f se ajusta bien a las etiquetas σ_i: si las etiquetas σ_i fueran en general correctas, f tendría un error de entrenamiento (o riesgo empírico) pequeño. Luego

$$\sup_{f \in \mathcal{F}} \sum_{i=1}^{n} \sigma_i f(X_i)$$

es grande si existe una función en $\mathcal{F}$ que se ajusta bien a la secuencia dada de etiquetas σ_i.

Ahora recordemos que las etiquetas σ_i son variables aleatorias. Podemos considerarlas como etiquetas aleatorias asignadas a los puntos de datos X_i. Como tomamos la esperanza sobre ambos, los puntos de datos y las etiquetas aleatorias, la complejidad de Rademacher será alta si el espacio funcional $\mathcal{F}$ es capaz de ajustarse bien a etiquetas aleatorias. Esta intuición tiene sentido: un espacio funcional debe ser bastante grande para poder ajustarse a todo tipo de etiquetas aleatorias en todo tipo de conjuntos de datos. En este sentido, la complejidad de Rademacher mide qué tan complejo es el espacio funcional: cuanto mayor sea $\mathscr{R}(\mathcal{F})$, mayor será la complejidad de $\mathcal{F}$.

Desde un punto de vista matemático, la complejidad de Rademacher es conveniente para trabajar. Se pueden demostrar cotas de generalización de la siguiente forma:

Teorema 1.7.6.1. *Sea $\mathcal{F}$ un espacio funcional y Z_n una muestra de tamaño n. Entonces, para cualquier $\delta > 0$,*

$$\mathcal{R}(f) \leq \mathcal{R}_{emp}(f) + 2\mathscr{R}(\mathcal{F}) + \sqrt{\frac{\log(1/\delta)}{2n}}.$$

con probabilidad al menos $1 - \delta$.

La complejidad de Rademacher tienen algunas ventajas sobre los conceptos clásicos de capacidad, como la dimensión VC. En particular, las cotas obtenidas mediante complejidades de Rademacher tienden a ser mucho más precisas que las obtenidas mediante herramientas clásicas. Las técnicas de demostración son diferentes de las explicadas anteriormente, pero no entraremos en detalles aquí. Para literatura sobre cotas basadas en la complejidad de Rademacher, véase, por ejemplo, Mendelson (2003), Bousquet et al. (2003) o Boucheron et al. (2005) y las referencias allí citadas.

1.7.7. Cotas con grandes márgenes de separación

Finalmente, queremos introducir otro tipo de medida de capacidad de clases de funciones que es más especializada que las cantidades combinatorias generales presentadas anteriormente. Consideremos el caso particular en el que el espacio

de datos consiste en puntos en el espacio bidimensional $\mathbb{R}^2$ y donde queremos separar las clases mediante una línea recta.

Definición 1.7.7.1. *Dado un conjunto de puntos de entrenamiento y un clasificador f_n que puede separarlos perfectamente, definimos el* ***margen*** *del clasificador f_n como la menor distancia de cualquier punto de entrenamiento a la línea separadora definida por f_n.*

Esta definición está ilustrada en la Figura (1.6). De manera análoga, se puede definir un margen para clasificadores lineales en un espacio de dimensión arbitraria d.

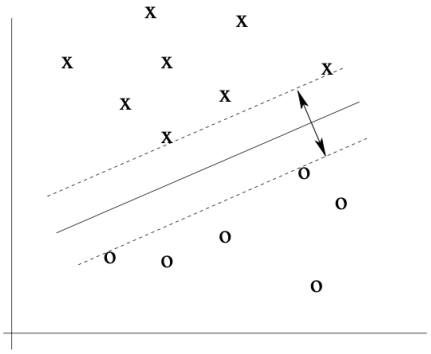

Figura 1.6: Margen de un clasificador lineal: las cruces representan puntos de entrenamiento con etiqueta $+1$, mientras que los círculos representan puntos de entrenamiento con etiqueta -1. La línea recta es el clasificador lineal f_n, y la línea discontinua muestra el margen. La anchura ρ del margen está representada por la flecha.

Se puede demostrar que la dimensión VC de la clase $\mathcal{F}_\rho$ de funciones lineales con margen al menos ρ puede ser acotada esencialmente por la razón entre el radio R de la esfera más pequeña que encierra los puntos de datos y el margen ρ, es decir,

$$\mathrm{VC}(\mathcal{F}_\rho) \leq \text{mín}\left\{d, \frac{4R^2}{\rho^2}\right\} + 1.$$

(cf. Teorema 5.1 en Vapnik, 1995). Es decir, cuanto mayor sea el margen ρ de las funciones en la clase $\mathcal{F}_\rho$, menor será su dimensión VC. De esta manera, se puede usar el margen de los clasificadores como un concepto de capacidad. Uno de los clasificadores más conocidos, la **máquina de soporte vectorial** (SVM, *Support Vector Machine*), se basa en este resultado. Para un tratamiento más completo, véase Schölkopf y Smola (2002).

Un ejemplo de cota de generalización que involucra el gran margen es el siguiente (para una declaración más precisa, véase el Teorema 7.3 en Schölkopf y Smola, 2002):

Teorema 1.7.7.1 (Cotas de grandes márgenes)**.** *Supongamos que el espacio de datos está contenido dentro de una esfera de radio R en $\mathbb{R}^d$. Consideremos el conjunto $\mathcal{F}_\rho$ de clasificadores lineales con margen al menos ρ. Supongamos que se nos dan n ejemplos de entrenamiento. Denotemos por $\nu(f)$ la fracción de ejemplos de entrenamiento con margen menor que ρ o que han sido clasificados erróneamente por algún clasificador $f \in \mathcal{F}_\rho$. Entonces, con probabilidad al menos $1-\delta$, el error verdadero de cualquier $f \in \mathcal{F}_\rho$ puede ser acotado por:*

$$\mathcal{R}(f) \leq \nu(f) + \sqrt{\frac{c}{n}\left(\frac{R^2}{\rho^2}\log(n)^2 + \log(1/\delta)\right)}.$$

1.7.8. Conclusiones acerca de las cotas de generalización

Hasta ahora, hemos introducido algunos conceptos de capacidad para clases de funciones: el coeficiente de fragmentación, la dimensión VC y la complejidad de Rademacher. En la literatura, existen muchos otros conceptos de capacidad, y presentarlos todos está fuera del alcance de esta visión general. Sin embargo, queremos destacar la forma general que suelen adoptar la mayoría de las cotas de generalización.

Usualmente, estas cotas están compuestas por tres términos diferentes y tienen la siguiente forma: con probabilidad al menos $1-\delta$,

$$\mathcal{R}(f) \leq R_{\text{emp}}(f) + \text{capacidad}(\mathcal{F}) + \text{confianza}(\delta).$$

Es decir, se puede acotar el riesgo verdadero de un clasificador por su riesgo empírico, un término de capacidad que, en el caso más simple, solo depende de la clase de funciones subyacente $\mathcal{F}$, y un término de confianza que depende de la probabilidad con la que se desea que la cota se cumpla.

Nótese que, por su naturaleza, todas las cotas de esta forma son cotas en el peor caso: dado que la cota se cumple para todas las funciones en la clase $\mathcal{F}$, el comportamiento de la cota está determinado por el peor clasificador dentro del espacio funcional. Este punto suele ser utilizado para criticar este enfoque en la teoría del aprendizaje estadístico, ya que los clasificadores naturales no tienden a seleccionar la peor función dentro de una clase.

Capítulo 2

La Teoría VC

En el capítulo previo, hemos intentado proveer una descripción liviana y en el orden de lo intuitivo de la teoría brindada por Vapnik y Chernovenkis. Ahora nos proponemos alcanzar la rigurosidad necesaria para demostrar los teoremas que hemos presentado previamente. Para ello, introduciremos una nueva notación y seguiremos [8], el célebre artículo de 1971 - escrito tres años antes en su idioma original - por los dos colosos rusos.

2.0.1. Definiciones preliminares

Sea X un conjunto de eventos sobre el cual se ha definido una medida de probabilidad P_X. Sea $\mathcal{S}$ una colección de eventos aleatorios, es decir, subconjuntos del espacio X que son medibles con respecto a la medida P_X.

Denotemos por $X^{(l)}$ el espacio de muestras de tamaño l en X. Sobre el espacio $X^{(l)}$, definimos una medida de probabilidad P tal que

$$P(Y_1 \times Y_2 \times \cdots \times Y_l) = P_X(Y_1)P_X(Y_2)\dots P_X(Y_l), \tag{2.1}$$

donde los Y_i son subconjuntos medibles de X.

Definición 2.0.1.1. *Cada muestra $\{X_1, \dots, X_l\}$ y cada evento $A \in \mathcal{S}$ determinan una* ***frecuencia relativa*** *para A, la cual es igual al cociente entre el número n_A de elementos de la muestra que pertenecen a A y el tamaño total de la muestra:*

$$\nu_A(\{X_1, \dots X_l\}) = \frac{n_A}{l}.$$

Observemos que la frecuencia relativa no es más que una probabilidad empírica, y en lo siguiente, representará al riesgo empírico. Un conocido resultado de

la relación entre esta probabilidad muestral y la probabilidad real en el espacio es el Teorema de Bernoulli, el cual enunciamos sin demostración:

Teorema 2.0.1.1 (de Bernoulli)**.** *Sea $A \in \mathcal{S}$ un evento aleatorio, $P(A)$ la probabilidad del evento A y $\nu_A(\{X_1, \dots X_l\})$ su frecuencia relativa para una muestra de tamaño l. Entonces, para cualquier $\varepsilon > 0$,*

$$\lim_{l\to\infty} P_X\left(|P(A) - \nu_A(\{X_1, \dots X_l\})| \geq \varepsilon\right) = 0.$$

Es decir, la frecuencia relativa (o probabilidad empírica) converge en probabilidad a la probabilidad real del evento A.

Nos interesará el supremo de esta diferencia sobre todos los posibles eventos en $\mathcal{S}$:

$$\pi^{(l)} = \sup_{A\in\mathcal{S}} |P(A) - \nu_A(\{X_1, \dots X_l\})|. \tag{2.2}$$

Notemos que analizando el comportamiento de la medida de dicho supremo al crecer el tamaño muestral, estaremos en condiciones de hacer afirmaciones acerca de la ecuación (1.4), que es una instancia particular de esta.

Observación 2.0.1.1. *La frecuencia relativa es una función puntual en el espacio $X^{(l)}$. Es decir*

$$\nu(A, X^{(l)}) : X^{(l)} \to \mathbb{R}$$

Además, supondremos que esta función es medible con respecto a la medida en $X^{(l)}$, es decir, que $\nu(A, X^{(l)})$ es una variable aleatoria.

Y si la variable $\nu(A, X^{(l)})$ converge en probabilidad a cero a medida que el tamaño de la muestra crece indefinidamente, diremos que la frecuencia relativa de los eventos $A \in \mathcal{S}$ tiende (en probabilidad) a la probabilidad de estos eventos de manera *uniforme* sobre la clase $\mathcal{S}$.

Los teoremas siguientes están dedicados a obtener estimaciones para la probabilidad de los eventos $\{\nu(A, X^{(l)}) > \varepsilon\}$ y a clarificar las condiciones bajo las cuales, para cualquier ε,

$$\lim_{l\to\infty} P\left(\pi^{(l)} > \varepsilon\right) = 0.$$

Definición 2.0.1.2. *Sea $Z_l = \{X_1, \dots, X_r\}$ una muestra finita de elementos en X. Cada conjunto A en $\mathcal{S}$ determina en esta muestra un subconjunto*

$$Z_l^A = Z_l \cap A = \{X_i \in Z_l \mid X_i \in A\},$$

*que consiste en aquellos elementos de la muestra Z_l que pertenecen a A. Diremos que el conjunto A **induce** el subconjunto Z_l^A en la muestra Z_l.*

Definición 2.0.1.3. *Denotamos el conjunto de todos los subconjuntos diferentes inducidos por los conjuntos de $\mathcal{S}$ en la muestra $Z_r = \{X_1, \ldots, X_r\}$ como $\mathcal{S}(X_1, \ldots, X_r)$ o $\mathcal{S}(Z_r)$. Es decir*

$$\mathcal{S}(Z_r) = \left\{ Z_r^A \mid A \in \mathcal{S} \right\}.$$

Definición 2.0.1.4. *El número de subconjuntos distintos de la muestra $Z_r = \{X_1, \ldots, X_r\}$ inducidos por los conjuntos en $\mathcal{S}$ será denominado el **índice del sistema** $\mathcal{S}$ con respecto a la muestra Z_r y será denotado por $\Delta^{\mathcal{S}}(Z_r)$. Es decir*

$$\Delta^{\mathcal{S}}(X_1, \ldots, X_r) = \#\left(\mathcal{S}(X_1, \ldots, X_r)\right).$$

Reparemos en que si consideramos los subconjuntos de $\mathcal{S}$ como los posibles conjuntos en los cuales los clasificadores binarios de un espacio funcional $\mathcal{F}$ pueden etiquetar a los elementos de X, entonces $\Delta^{\mathcal{S}}(Z_r)$ es el cardinal del conjunto de fragmentación $\mathcal{F}_{Z_r}$. La siguiente observación es también una generalización de lo que hemos dicho en tanto a la cardinalidad del conjunto de fragmentación.

Observación 2.0.1.2. *El índice del sistema puede ser a lo sumo 2 elevada a la cantidad de elementos en la muestra. Es decir*

$$\Delta^{\mathcal{S}}(X_1, \ldots, X_r) \leq 2^r$$

Por último, generalizamos la noción de coeficiente de fragmentación en la *función de crecimiento*, tomando el máximo sobre todos los índices posibles.

Definición 2.0.1.5. *Llamaremos **función de crecimiento** de orden r al índice máximo de un sistema sobre todas las muestras posibles de tamaño r sobre un espacio X.*

$$m^{\mathcal{S}}(r) = \max_{\{X_1, \ldots, X_r\}} \Delta^{\mathcal{S}}(X_1, \ldots, X_r)$$

Ejemplo 1: Sea X una línea recta y sea $\mathcal{S}$ el conjunto de todos los rayos de la forma $X \leq a$. En este caso, la función de crecimiento es

$$m^{\mathcal{S}}(r) = r + 1.$$

Ejemplo 2: Sea X el segmento $[0, 1]$. El conjunto $\mathcal{S}$ consiste en todos los conjuntos abiertos. En este caso,

$$m^{\mathcal{S}}(r) = 2.$$

Ejemplo 3: Sea $X = E_n$, el espacio euclidiano de dimensión n. El conjunto $\mathcal{S}$ de eventos consiste en todos los semiespacios de la forma

$$\langle X_0, \varphi \rangle \geq 1,$$

donde φ es un vector fijo y las llaves representan el producto interno en el espacio. Evaluemos la función de crecimiento $m^{\mathcal{S}}(r)$.

Consideremos, junto con el espacio E_n de vectores X_0, el espacio E_n de vectores φ. A cada vector $X_0 \in E_n$ le corresponde una partición del espacio E_n en dos subespacios: $\langle X_k, \varphi_2 \rangle \geq 1$ y $\langle X_k, \varphi_2 \rangle < 1$. Recíprocamente, cada vector φ determina un evento A en el sistema $\mathcal{S}$.

Tomemos r vectores $X_1, \ldots, X_r$. Estos generan una partición del espacio E_n en un número de componentes tales que los vectores φ dentro de cada componente determinan eventos $A \in \mathcal{S}$ que inducen el mismo subconjunto en la muestra $\{X_1, \ldots, X_r\}$.

Definición 2.0.1.6. *Sea $\Phi(n, r)$ el número máximo de componentes en los que es posible particionar el espacio n-dimensional mediante r hiperplanos. Llamaremos a esta función* ***función de crecimiento lineal*** *del sistema.*

Para el ejemplo 3, por definición:

$$m^{\mathcal{S}}(r) = \Phi(n, r).$$

Es decir, la función de crecimiento lineal es igual a la función de crecimiento cuando solo consideramos partición del espacio por medio de hiperplanos.

Observación 2.0.1.3. *Podemos definir a la función de crecimiento lineal en términos recursivos.*

$$\Phi(n, r) = \Phi(n, r-1) + \Phi(n-1, r-1), \quad \Phi(0, r) = 1, \quad \Phi(n, 0) = 1.$$

Esto a su vez nos permite dar la definición combinatoria de la funcion de crecimiento lineal:

$$\Phi(n, r) = \begin{cases} \sum_{k=0}^{n} \binom{r}{n}, & si\ r > n, \\ 2^r, & si\ r \leq n. \end{cases} \tag{2.3}$$

Por último, observamos que

Observación 2.0.1.4. *Para $n > 0$ y $r \geq 0$, se cumple que:*

$$\Phi(n, r) \leq r^n + 1.$$

Y tomaremos $\binom{n}{k} = 0$ en el caso que $n < k$.

2.0.2. La función de crecimiento

Una propiedad que nos interesa del coeficiente de fragmentación (y por extensión, de la dimensión VC) es que podamos asegurarnos de que, de ser posible, sean polinómicas. De esa manera la desigualdad (1.6) es absorvida por la exponencial y nos aseguramos la convergencia en probabilidad.

Ahora probaremos que la función de crecimiento $m^{\mathcal{S}}(r)$ es idénticamente igual a 2^r o está acotada superiormente por $r^n + 1$ (es decir, es a lo sumo polinómica), donde n es una constante que corresponde al valor de r para el cual la igualdad

$$m^{\mathcal{S}}(r) = 2^r$$

se viola por primera vez. Para demostrar este hecho, necesitamos el siguiente lema.

Lema 2.0.2.1. *Si para alguna muestra de tamaño $X_1, \ldots, X_i$ y algún número n, con $1 \leq n \leq i$, se cumple que*

$$\Delta^{\mathcal{S}}(X_1, \ldots, X_i) \geq \Phi(n, i),$$

entonces existe un subconjunto $X_1, \ldots, X_n$ de esta muestra tal que

$$\Delta^{\mathcal{S}}(X_1, \ldots, X_n) = 2^n.$$

Donde la función $\Phi(n, i)$ está definida de forma recursiva.

Una demostración de este lema puede hallarse en [8].

Ahora si, enunciamos y demostramos el resultado.

Teorema 2.0.2.1. *La función de crecimiento $m^{\mathcal{S}}(r)$ es idénticamente igual a 2^r o está acotada superiormente por la función potencia*

$$r^n + 1,$$

donde n es una constante positiva que corresponde al valor de r para el cual la igualdad

$$m^{\mathcal{S}}(r) = 2^r$$

se viola por primera vez.

Demostración. Ya hemos dicho que $m^{\mathcal{S}}(r) \leq 2^r$. Supongamos que $m^{\mathcal{S}}(r)$ no es igual a 2^r y que n es el primer valor de r para el cual $m^{\mathcal{S}}(r) \neq 2^r$. Entonces, para cualquier muestra de tamaño $r > n$,

$$\Delta^{\mathcal{S}}(X_1, \ldots, X_r) < \Phi(n, r).$$

De lo contrario, en base a la afirmación del lema, se podría encontrar una sub muestra $Z_n^A = \{X_i \in Z_n \mid X_i \in A\}$ tal que

$$\Delta^{\mathcal{S}}\left(Z_n^A\right) = 2^n.$$

Pero esto contradice la suposición inicial de que $m^{\mathcal{S}}(n) < 2^n$. Así, $m^{\mathcal{S}}(r)$ es o bien idénticamente igual a 2^r, o bien está acotada superiormente por $\Phi(n, r)$. A su vez, para $r > 0$, se cumple que $\Phi(n, r) \leq r^n + 1$ por la observación anterior. Sigue que

$$m^{\mathcal{S}}(r) \leq r^n + 1.$$

□

Detengámonos un momento en considerar que significan estos resultados en tanto hemos pensado los espacios funcionales de clasificadores en la sección anterior. Ya hemos dicho que el coeficiente de fragmentación de un espacio para un tamaño de muestra dada es una instancia de la función de crecimiento. Es decir

$$\mathcal{N}(\mathcal{F}, r) = m^{\mathcal{S}}(r)$$

Por otro lado, si n es el primer valor donde la igualdad $m^{\mathcal{S}}(r) = 2^r$ se viola, la dimensión VC de un espacio $\mathcal{F}$ por definición es $n - 1$, es decir el mayor tamaño de muestra que puede ser fragmentado por $\mathcal{F}$.

Considerando la definición combinatoria de la función de crecimiento lineal (2.3), hemos demostrado el Lema (1.7.5.1). Lo enunciamos de forma general en el siguiente corolario.

Corolario 2.0.2.1.1. *En las condiciones del teorema anterior, si $n \in \mathbb{N}$ es el mínimo valor del tamaño muestral r tal que*

$$m^{\mathcal{S}}(r) < 2^r$$

entonces

$$m^{\mathcal{S}}(r) \leq \sum_{k=0}^{n} \binom{r}{k}.$$

2.0.3. El Lema de simetrización

Enfoquémonos ahora en probar el lema de simetrización (1.7.1.1). Tomemos dos muestra de tamaño l y unifiquémolas en una sola muestra de tamaño $2l$, $Z_{2l} = \{X_1, \ldots, X_l, X_{l+1}, \ldots, X_{2l}\}$. Supongamos que las frecuencias relativas del evento $A \in \mathcal{S}$ han sido calculadas en la primera semimuestra $\{X_1, \ldots, X_l\}$ y en la segunda semimuestra $\{X_{l+1}, \ldots, X_{2l}\}$ por separado.

Para cualquier evento $A \in \mathcal{S}$, denotemos las respectivas frecuencias por ν_A^1 y ν_A^2, y consideremos la diferencia entre estas cantidades:

$$\rho_A^{(l)} = |\nu_A^1 - \nu_A^2|.$$

Nos interesa la máxima diferencia entre estas cantidades sobre todos los eventos en la clase $\mathcal{S}$,

$$\rho^{(l)} = \sup_{A \in \mathcal{S}} \rho_A(l).$$

Observemos que

$$\sup_{A \in \mathcal{S}} \rho_A^{(l)} = \max_{A \in \mathcal{S}} \rho_A^{(l)},$$

ya que, para un l fijo, $\rho_A^{(l)}$ solo toma un número finito de valores. Expliquemos esto. Al fijar dos muestras de tamaño l, tenemos a lo sumo $2l$ elementos distintos del espacio X. Tomemos un subconjunto $A \subset X$ tal que $A \in \mathcal{S}$. Para cada muestra Z_l, definimos la frecuencia muestral como

$$\nu(A, Z_l) = \frac{n(A, Z_l)}{l} = \frac{\#(\{X_i \in Z_l \mid X_i \in A\})}{l}.$$

Si bien los conjuntos A son potencialmente infinitos, podemos separarlos en un número finito de grupos. Basicamente, aquellos que contienen un elemento de la muestra Z_l, aquellos que contienen dos y así hasta aquellos que contienen los l elementos de la muestra. Es decir, hay l frecuencias posibles para cada muestra (que ya está fija), aún tomando infinitos $A \in \mathcal{S}$. Por lo tanto, sus diferencias $\rho_A^{(l)}$ también han de serlo.

Mostraremos que si $\rho^{(l)} \to 0$ cuando $l \to \infty$, entonces también $\pi^{(l)} \to 0$, y que las estimaciones de $\rho^{(l)}$ conducen a estimaciones para $\pi^{(l)}$. Recordemos que π es la diferencia entre la frecuencia muestral $\nu(A, Z_l)$ y la probabilidad del evento A definida en (2.2). De ahora en más, supondremos que $\rho^{(l)}$ es una función medible.

Es conveniente introducir la siguiente notación:

$$Q = \left\{\pi^{(l)} > \varepsilon\right\}, \quad C = \left\{\rho^{(l)} \geq \frac{1}{2}\varepsilon\right\}.$$

El siguiente es exactamente el Lema de simetrización que hemos presentado con anterioridad.

Lema 2.0.3.1 (de simetrización)**.** *Sea $l \geq 2/\varepsilon^2$, entonces*

$$P(C) \geq \frac{1}{2} P(Q).$$

Demostración. Por definición

$$P(C) = P\left(\rho^{(l)} - \frac{\varepsilon}{2} \geq 0\right) = \int_{Z^{(2l)}} \left(1 \cdot \chi_{\{\rho^{(l)} - \frac{\varepsilon}{2} \geq 0\}} + 0 \cdot \chi_{\{\rho^{(l)} - \frac{\varepsilon}{2} < 0\}}\right) dP$$

Donde integramos sobre el espacio muestral de tamaño $2l$ de tal manera que

$$Z^{(2l)} = Z_1^{(l)} \times Z_2^{(l)} = \{X_1, \ldots, X_l\} \times \{X_{l+1}, \ldots, X_{2l}\} \tag{2.4}$$

Definamos la función indicadora:

$$\theta(z) = \begin{cases} 1 \text{ si } z \geq 0 \\ 0 \text{ si } z < 0 \end{cases}$$

con la cual reescribimos la probabilidad de C como

$$P(C) = \int_{Z^{(2l)}} \left(\theta\left(\rho^{(l)} - \frac{\varepsilon}{2}\right)\right) dP \tag{2.5}$$

Descomponiendo como en (2.4) al espacio muestral, podemos escribir:

$$P(C) = \int_{Z_1^{(l)} \times Z_2^{(l)}} \left(\theta\left(\rho^{(l)} - \frac{\varepsilon}{2}\right)\right) dP$$

Y, tomando $P = P_1 \times P_2$, por el teorema de Fubini separamos los signos de integración

$$P(C) = \int_{Z_1^{(l)}} \left(\int_{Z_2^{(l)}} \theta\left(\rho^{(l)} - \frac{\varepsilon}{2}\right) dP_2 \right) dP_1$$

Ahora, consideremos el subevento Q sobre el espacio muestral $Z_1^{(l)}$, es decir aquellas muestras de l elementos en las cuales se verifica la definición de Q que hemos dado

$$Q = \left\{\pi^{(l)} > \varepsilon\right\} = \left\{\hat{Z}^{(l)} \in Z^{(l)} : \sup_{A \in \mathcal{S}} \left| P_A - \nu_A\left(\hat{Z}^{(l)}\right)\right| > \varepsilon\right\},$$

y restringimos la primer integral a este evento

$$P(C) \geq \int_Q \left(\int_{Z_2^{(l)}} \theta\left(\rho^{(l)} - \frac{\varepsilon}{2}\right) dP_2 \right) dP_1. \tag{2.6}$$

Por definición, para cada muestra $\hat{Z}^{(l)}$ en Q debe existir algún evento $A_0 \in \mathcal{S}$ tal que $|P_{A_0} - \nu_{A_0}\left(\hat{Z}^{(l)}\right)| > \varepsilon$. Para dicho evento, se cumple

$$\begin{aligned} \varepsilon < \left|P_{A_0} - \nu_{A_0}^{(1)}\right| &= \left|\nu_{A_0}^{(1)} - P_{A_0}\right| \\ &= \left|\nu_{A_0}^{(1)} - \nu_{A_0}^{(2)} + \nu_{A_0}^{(2)} - P_{A_0}\right| \\ &\leq \left|\nu_{A_0}^{(1)} - \nu_{A_0}^{(2)}\right| + \left|\nu_{A_0}^{(2)} - P_{A_0}\right|, \end{aligned}$$

y así

$$\varepsilon - \left|\nu_{A_0}^{(2)} - P_{A_0}\right| < \left|\nu_{A_0}^{(1)} - \nu_{A_0}^{(2)}\right|.$$

Por lo tanto, la desigualdad

$$\rho_{A_0}^{(l)} = |\nu_{A_0}^{(1)} - \nu_{A_0}^{(2)}| \geq \frac{\varepsilon}{2},$$

se verifica si

$$|\nu_{A_0}^{(2)} - P_{A_0}| \geq \frac{\varepsilon}{2}.$$

Retomando la desigualdad (2.6), reparemos un momento en la integral interior. Dada la monotonía de θ y quedándonos con tan solo un evento $A_0 \in \mathcal{S}$, en vez de tomar el supremo sobre todos los eventos, podemos escribir

$$\begin{aligned}\int_{Z_2^{(l)}} \theta\left(\rho^{(l)} - \frac{\varepsilon}{2}\right) dP_2 \;&\geq\; \int_{Z_2^{(l)}} \theta\left(\rho_{A_0}^{(l)} - \frac{\varepsilon}{2}\right) dP_2 \\ &=\; \int_{Z_2^{(l)}} \theta\left(\frac{\varepsilon}{2} - |\nu_{A_0}^{(2)} - P_{A_0}|\right) dP_2\end{aligned}$$

En donde en la última igualdad aplicamos lo recién explicado. Ahora notemos que

$$\int_{Z_2^{(l)}} \theta\left(\frac{\varepsilon}{2} - |\nu_{A_0}^{(2)} - P_{A_0}|\right) dP_2 \;=\; 1 - P\left(|\nu_{A_0}^{(2)} - P_{A_0}| > \frac{\varepsilon}{2}\right), \qquad (2.7)$$

dada la propia definición de θ, que vale 0 sólo cuando se da la desigualdad dentro de la probilidad de la derecha. Ahora procedemos a acotar dicha probabilidad.

Haciendo uso de la desigualdad de Chebyshev:

$$P\left\{|X - \mathbb{E}(X)| \geq k \cdot \sqrt{\mathrm{Var}(X)}\right\} \leq \frac{1}{k^2}$$

para la variable aleatoria $X = \nu_{A_0}^{(2)}$ y el valor

$$k = \frac{\varepsilon}{2 \cdot \sqrt{\mathrm{Var}\left(\nu_{A_0}^{(2)}\right)}},$$

escribimos

$$P\left(|\nu_{A_0}^{(2)} - \mathbb{E}\left(\nu_{A_0}^{(2)}\right)| \geq \frac{\varepsilon}{2}\right) \;\leq\; \frac{4}{\epsilon^2} \cdot \mathrm{Var}\left(\nu_{A_0}^{(2)}\right). \qquad (2.8)$$

Analicemos esta última desigualdad, factor a factor.

Podemos escribir la frecuencia relativa del evento A_0 sobre una muestra Z_l'' en el espacio muestral $Z_2^{(l)} = \{X_1, \ldots, X_l\}$ como la suma de variables aleatorias Bernoulli $\tilde{X}_i$ con $i = 1, 2, \ldots, l$ tales que

$$\tilde{X}_i = \begin{cases} 1 \text{ si } X \in A_0 \\ 0 \text{ si } X \notin A_0 \end{cases}$$

donde

$$\nu_{A_0}^{(2)} = \frac{1}{l} \sum_{i=1}^{l} \tilde{X}_i.$$

Es decir, la frecuencia relativa es una combinación lineal de variables Bernoulli, cada una con probabilidad P_{A_0} (que no es ni mas ni menos que la probabilidad de que cada muestra esté dentro del evento A_0).

Previamente, llamamos $n_{A_0}^{(2)}$ a esta suma, y utilizándola podemos escribir la esperanza de la frecuencia relativa como:

$$\begin{aligned} \mathbb{E}\left(\nu_{A_0}^{(2)}\right) = \mathbb{E}\left(\frac{n_{A_0}^{(2)}}{l}\right) &= \frac{1}{l} \cdot \mathbb{E}\left(\#\left\{X_i \in Z_l'' \mid X_i \in A_0\right\}\right) \\ &= \frac{1}{l} \cdot l \cdot P_{A_0} = P_{A_0} \end{aligned}$$

Que es lo esperable, dado que las muestras en promedio deberian tender a las probabilidades.

Ahora, centrémonos en la varianza de la desigualdad (2.8). Primero, recordemos que la varianza de toda variable aleatoria Bernoulli tiene forma

$$\text{Var}(X) = p(1-p), \qquad X \sim \text{Bernoulli } (p = P(X)).$$

Esto nos permite escribir la varianza de la frecuencia muestral ν_{A_0} en términos de la probabilidad del evento A_0:

$$\begin{aligned} \text{Var}\left(\nu_{A_0}^{(2)}\right) = \text{Var}\left(\frac{1}{l}\sum_{i=1}^{l} \tilde{X}_i\right) &= \frac{1}{l}\sum_{i=1}^{l} \text{Var}\left(\tilde{X}_i\right) \\ &= \frac{1}{l}\sum_{i=1}^{l} \left[P(\tilde{X}_i)(1 - P(\tilde{X}_i))\right] \\ &= \frac{1}{l} \cdot l \cdot P(\tilde{X}_i)(1 - P(\tilde{X}_i)) \end{aligned}$$

Ahora, como $P(\tilde{X}_i) = P_{A_0}$, tenemos que

$$\text{Var}\left(\nu_{A_0}^{(2)}\right) = P_{A_0}(1 - P_{A_0}).$$

Por lo tanto, podemos reescribir (2.8) como:

$$P\left(\left|\nu_{A_0}^{(2)} - P_{A_0}\right| > \frac{\varepsilon}{2}\right) \leq \frac{4(1-P_{A_0})P_{A_0}}{\varepsilon^2 l}.$$

Dado que P_{A_0} es una probabilidad, y que la función $f(x) = x(1-x)$ tiene un máximo en $x = \frac{1}{2}$, tenemos que $(1-P_{A_0})P_{A_0} \leq \frac{1}{4}$. Por lo tanto,

$$4(1-P_{A_0})P_{A_0} \leq 1,$$

luego

$$P\left(\left|\nu_{A_0}^{(2)} - P_{A_0}\right| > \frac{\varepsilon}{2}\right) \leq \frac{1}{\varepsilon^2 l}.$$

Retomando (2.7), tenemos

$$\int_{Z_2^{(l)}} \theta\left(\frac{\varepsilon}{2} - |\nu_{A_0}^{(2)} - P_{A_0}|\right) dP_2 \;\geq\; 1 - \frac{1}{\varepsilon^2 l}.$$

Por lo tanto, para $l \geq 2/\varepsilon^2$,

$$\int_{Z_2^{(l)}} \theta\left(\frac{\varepsilon}{2} - \left|\nu_{A_0}^{(2)} - P_{A_0}\right|\right) dP_2 \geq \frac{1}{2}.$$

De esto se sigue inmediatamente que, para $l \geq 2/\varepsilon^2$,

$$P(C) \geq \int_Q \left(\frac{1}{2}\right) dP_1 = \frac{1}{2}P(Q).$$

Y el lema queda demostrado.

□

Utilizamos el Lema de Simetrización para probar el siguiente teorema.

Teorema 2.0.3.1. *La probabilidad de que la frecuencia relativa de al menos un evento A en la clase $\mathcal{S}$ difiera de su probabilidad en una muestra de tamaño l por más de ε, para $l \geq 2/\varepsilon^2$, satisface la desigualdad*

$$P\left(\pi^{(l)} > \varepsilon\right) \leq 4\, m^{\mathcal{S}}(2l)\, e^{-\varepsilon^2 l/8}.$$

Demostración. En virtud del Lema de Simetrización, basta con estimar

$$P\left(\rho^{(l)} \geq \frac{\varepsilon}{2}\right) = \int_{Z^{(2l)}} \theta\left(\rho^{(l)} - \frac{\varepsilon}{2}\right) dP.$$

donde $\rho^{(l)}$ se considera como una función de la muestra

$$Z^{(2l)} = (X_1, \ldots, X_l, X_{l+1}, \ldots, X_{2l}).$$

Consideremos la aplicación del espacio $Z^{(2l)}$ sobre sí mismo, resultante de alguna permutación T_i de los elementos de la muestra $Z^{(2l)}$. Dada la simetría de la definición (2.1) de la medida P en $Z^{(2l)}$, se cumple la siguiente relación para cualquier función integrable f sobre cualquier muestra Z_{2l}:

$$\int_{Z^{(2l)}} f(Z_{2l})\, dP = \int_{Z^{(2l)}} f(T_i Z_{2l})\, dP.$$

Por lo tanto,

$$P\left(\rho^{(l)} \geq \frac{\varepsilon}{2}\right) = \int_{Z^{(2l)}} \frac{1}{(2l)!} \sum_{i=1}^{(2l)!} \theta\left(\rho^{(l)}(T_i Z_{2l}) - \frac{\varepsilon}{2}\right) dP. \tag{2.9}$$

donde simplemente promnediamos sobre las $(2l)!$ permutaciones posibles.

Observemos además que

$$\theta\left(\rho^{(l)} - \frac{\varepsilon}{2}\right) = \theta\left(\sup_{A\in\mathcal{S}} \left|\nu_A^{(1)} - \nu_A^{(2)}\right| - \frac{\varepsilon}{2}\right) = \sup_{A\in\mathcal{S}} \theta\left(\left|\nu_A^{(1)} - \nu_A^{(2)}\right| - \frac{\varepsilon}{2}\right).$$

Claramente, si dos conjuntos A_1 y A_2 inducen la misma submuestra en una muestra

$$Z_{2l} = (X_1, \ldots, X_l, X_{l+1}, \ldots, X_{2l}),$$

entonces

$$\nu_{A_1}^{(1)}(T_i Z_{2l}) = \nu_{A_2}^{(1)}(T_i Z_{2l}), \qquad \nu_{A_1}^{(2)}(T_i Z_{2l}) = \nu_{A_2}^{(2)}(T_i Z_{2l}).$$

Es decir, tienen idéntica frecuencia muestral relativa, para ambas muestras (1) y (2). Por lo tanto, las diferencias entre cada una de estas frecuencias también serán idénticas

$$\rho_{A_1}^{(l)}(T_i Z_{2l}) = \rho_{A_2}^{(l)}(T_i Z_{2l}),$$

para cualquier permutación T_i. Esto implica que si elegimos el subsistema $\mathcal{S}' \subseteq \mathcal{S}$ que consiste en todos los conjuntos A que inducen submuestras esencialmente diferentes en la muestra Z_{2l}, el valor del siguiente supremo no cambiará al tomarlo sobre la restricción S':

$$\sup_{A\in\mathcal{S}} \theta\left(\rho_A^{(l)}(T_i Z_{2l}) - \frac{\varepsilon}{2}\right) = \sup_{A\in\mathcal{S}'} \theta\left(\rho_A^{(l)}(T_i Z_{2l}) - \frac{\varepsilon}{2}\right)$$

A su vez, podemos acotar la expresión por la suma sobre un numero finito de eventos, los que nos permite escribir

$$\sup_{A\in\mathcal{S}} \theta\left(\rho_A^{(l)}(T_i Z_{2l}) - \frac{\varepsilon}{2}\right) \leq \sum_{A\in\mathcal{S}'} \theta\left(\rho_A^{(l)}(T_i Z_{2l}) - \frac{\varepsilon}{2}\right).$$

Notemos que este es exactamente el mismo argumento que utilizamos en el capítulo anterior para acotar la medida de probabilidad del supremo de la diferencia entre el riesgo y el riesgo empírico, tomado sobre todos los clasificadores en un espacio funcional, utilizando solo una cantidad finita de clasificadores en el espacio. Recordemos que llamamos indíce del sistema al cardinal del conjunto S', al que notamos como $\Delta^{\mathcal{S}'}(X_1, \ldots, X_{2l})$.

Ahora, podemos acotar la integral en (2.9) escribiendo

$$\frac{1}{(2l)!}\sum_{i=1}^{(2l)!} \theta\left(\rho^{(l)}(T_i Z^{2l}) - \frac{\varepsilon}{2}\right) = \frac{1}{(2l)!}\sum_{i=1}^{(2l)!} \sup_{A\in\mathcal{S}} \theta\left(\rho_A^{(l)}(T_i Z^{2l}) - \frac{\varepsilon}{2}\right)$$

$$\leq \sum_{A\in\mathcal{S}'}\left[\frac{1}{(2l)!}\sum_{i=1}^{(2l)!} \theta\left(\rho_A^{(l)}(T_i Z^{2l}) - \frac{\varepsilon}{2}\right)\right].$$

La expresión entre corchetes denota el cociente del número de permutaciones para una muestra fija para las cuales $|\nu_A^{(1)} - \nu_A^{(2)}| \leq \frac{1}{2}\varepsilon$ y el número total de posibles permutaciones, $(2l)!$. Esta expresión es exactamente igual a:

$$\Gamma = \sum_{k:|2\frac{k}{l} - \frac{n_A}{l}|\geq\frac{\varepsilon}{2}} \frac{\binom{n_A}{k}\binom{2l-n_A}{l-k}}{\binom{2l}{l}}.$$

donde n_A es el número de elementos en la muestra $\{X_1, \ldots, X_{2l}\}$ que pertenecen a A, como ya hemos dicho.

Esta expresión satisface la estimación

$$\Gamma \leq 2e^{-\varepsilon^2 l/8},$$

que se obtiene mediante un cálculo sencillo pero largo, por lo que omitimos la demostración, como los autores de [8] hacen.

Por lo tanto,

$$\begin{aligned}\frac{1}{(2l)!}\sum_{i=1}^{(2l)!} \theta\left(\rho^{(l)}(T_i Z^{2l}) - \frac{\varepsilon}{2}\right) &\leq \sum_{A\in\mathcal{S}'} 2e^{-\varepsilon^2 l/8} \\ &= 2\Delta^{\mathcal{S}}(X_1, \ldots, X_{2l})e^{-\varepsilon^2 l/8} \\ &\leq 2m^{S}(2l)e^{-\varepsilon^2 l/8}.\end{aligned}$$

Aplicando el Lema de Simetrización, tenemos

$$P\left(\pi^{(l)} > \varepsilon\right) \leq 2 \cdot 2m^S(2l)e^{-\varepsilon^2 l/8}.$$

Por último, notemos que

$$m^S(2l) < (2l)^n + 1$$

por la cota dada en el Teorema (2.0.2.1), lo que nos da

$$\lim_{l\to\infty} P\left(\pi^{(l)} > \varepsilon\right) \leq 4 \lim_{l\to\infty} ((2l)^n + 1)e^{-\varepsilon^2 l/8} = 0.$$

□

El siguiente corolario es inmediato.

Corolario 2.0.3.1.1. *Una condición suficiente para que las frecuencias relativas de los eventos en la clase $\mathcal{S}$ converjan uniformemente sobre $\mathcal{S}$ (en probabilidad) a sus correspondientes probabilidades es que exista un n finito tal que*

$$m^{\mathcal{S}}(l) \leq l^n + 1, \quad \textit{para todo } l.$$

Hemos dado condiciones suficientes para que las frecuencias relativas (nuestras probabilidades empíricas) converjan uniformemente a las probabilidades reales de los eventos en la clase $\mathcal{S}$. En la siguiente sección veremos que estas condiciones nos aseguran, además, convergencia uniforme de forma casi segura.

2.0.4. Condiciones de convergencia uniforme casi segura

Teorema 2.0.4.1. *Si $m^{\mathcal{S}}(l) \leq l^n + 1$, entonces*

$$P(\pi^{(l)} \to 0) = 1.$$

Demostración. Dado que

$$P(\pi^{(l)} > \varepsilon) \leq 4m^{\mathcal{S}}(2l)e^{-\varepsilon^2 l/8},$$

para $l \geq 2/\varepsilon^2 = L$, la serie

$$\sum_{l=1}^{\infty} P(\pi^{(l)} > \varepsilon) \leq \sum_{l=1}^{L} P(\pi^{(l)} > \varepsilon) + 4 \sum_{l=L+1}^{\infty} [(2l)^n + 1]e^{-\varepsilon^2 l/8}$$

es convergente para cualquier ε.
Por el lema, esto implica que

$$P(\pi^{(l)} \to 0) = 1.$$

□

Observemos primero que la definición del índice implica inmediatamente que

$$\Delta^{\mathcal{S}}(X_1,\ldots,X_k,X_{k+1},\ldots,X_l) \leq \Delta^{\mathcal{S}}(X_1,\ldots,X_k)\Delta^{\mathcal{S}}(X_{k+1},\ldots,X_l).$$

De aquí se sigue que

$$\log_2 \Delta^{\mathcal{S}}(X_1,\ldots,X_k,X_{k+1},\ldots,X_l) \leq \log_2 \Delta^{\mathcal{S}}(X_1,\ldots,X_k)+\log_2 \Delta^{\mathcal{S}}(X_{k+1},\ldots,X_l).$$

En lo que sigue, se asumirá que el índice $\Delta^{\mathcal{S}}(X_1,\ldots,X_k)$, visto como una función de $X_l = \{X_1,\ldots,X_l\}$, es medible con respecto a la medida P.

Definición 2.0.4.1. *Sea*

$$F^{\mathcal{S}}(z) = P\left(\log_2 \Delta^{\mathcal{S}}(X_1,\ldots,X_l) < z\right).$$

Definición 2.0.4.2. *La entropía del sistema de eventos $\mathcal{S}$ en muestras de tamaño l se define como*

$$H^{\mathcal{S}}(l) = \mathbb{E}\left[\log_2 \Delta^{\mathcal{S}}(X_1,\ldots,X_l)\right].$$

Dada la monotonía de la esperanza, la última desigualdad implica que

$$H^{\mathcal{S}}(l_1+l_2) \leq H^{\mathcal{S}}(l_1) + H^{\mathcal{S}}(l_2).$$

Los siguientes lemas nos ayudarán más adelante.

Lema 2.0.4.1. *La sucesión $\frac{H^{\mathcal{S}}(l)}{l}$ tiene un límite c, con $0 \leq c \leq 1$, cuando $l \to \infty$.*

Ahora mostraremos que, para valores grandes de l, la distribución de la variable aleatoria

$$\xi^{(l)} = \frac{1}{l}\log_2 \Delta^{\mathcal{S}}(X_1,\ldots,X_l)$$

se concentra cerca de c.

Lema 2.0.4.2.

$$\lim_{l\to\infty} P(|\xi^{(l)} - c| > \varepsilon) = 0, \quad \textit{para } \varepsilon > 0.$$

Teorema 2.0.4.2. *Una condición necesaria y suficiente para que las frecuencias relativas converjan (en probabilidad) a las probabilidades de manera uniforme sobre la clase de eventos $\mathcal{S}$ es que*

$$\lim_{l\to\infty} \frac{H^{\mathcal{S}}(l)}{l} = 0. \tag{2.10}$$

Observemos que, por el Lema 4, la condición (2,10) es equivalente a la siguiente afirmación:

$$\lim_{l\to\infty} P\left(\frac{1}{l}\log_2 \Delta^{\mathcal{S}}(X_1,\ldots,X_l) > \delta\right) = 0, \quad \text{para todo } \delta > 0. \tag{2.11}$$

Capítulo 3

Otros conceptos del Aprendizaje

3.1. Consistencia de Bayes y error de aproximación

Hasta ahora, hemos reparado tan solo en el enfoque estándar para acotar el error de estimación de un clasificador. Esto es suficiente para lograr consistencia con respecto a una clase de funciones dada $\mathcal{F}$. En esta sección, queremos analizar la pieza faltante para lograr la **consistencia de Bayes**: el error de aproximación.

Recordemos que el error de estimación se definía como $\mathcal{R}(f_n) - \mathcal{R}(f_{\mathcal{F}})$ y el error de aproximación como $\mathcal{R}(f_{\mathcal{F}}) - \mathcal{R}(f_{\text{Bayes}})$ (1.4.3.1). Para lograr consistencia de Bayes, ambos términos deben tender a cero cuando $n \to \infty$. Hemos visto que, para que el error de estimación converja a 0, debemos asegurarnos de que el espacio funcional $\mathcal{F}$ tenga una capacidad razonablemente baja.

Sin embargo, esto plantea un problema para el error de aproximación: si el espacio funcional $\mathcal{F}$ tiene una capacidad pequeña, esto significa, en particular, que el espacio $\mathcal{F}$ es considerablemente más pequeño que el espacio de todas las funciones posibles $\mathcal{F}_{\text{all}}$. Por lo tanto, si fijamos la clase de funciones $\mathcal{F}$ y f_{Bayes} no está contenido en $\mathcal{F}$, entonces el error de aproximación podría no ser cero.

Existen solo dos formas de resolver este problema. La primera es imponer condiciones sobre la forma funcional del clasificador de Bayes. Si $f_{\text{Bayes}} \in \mathcal{F}$ para algún espacio funcional conocido $\mathcal{F}$ con capacidad pequeña, entonces sabemos que el error de aproximación es 0. En este caso, la consistencia de Bayes se reduce a la consistencia con respecto a $\mathcal{F}$, lo cual puede lograrse mediante los métodos discutidos previamente. Sin embargo, si no queremos hacer suposi-

ciones sobre el clasificador de Bayes, entonces debemos optar por una enfoque diferente.

3.1.1. Espacios funcionales anidados

En esta construcción, no consideraremos una única clase de funciones $\mathcal{F}$, sino una secuencia de espacios funcionales $\mathcal{F}_1, \mathcal{F}_2, \ldots$. Al construir un clasificador sobre n puntos de datos, lo haremos a partir del espacio funcional $\mathcal{F}_n$. La clave de esta construcción es que el espacio funcional $\mathcal{F}_n$ debe volverse más complejo a medida que el tamaño de la muestra n aumenta.

El enfoque estándar es elegir los espacios $\mathcal{F}_n$ de modo que formen una secuencia creciente de espacios funcionales anidados, es decir,

$$\mathcal{F}_1 \subset \mathcal{F}_2 \subset \mathcal{F}_3 \subset \ldots$$

La intuición detrás de esto es que comenzamos con un espacio funcional simple y luego, gradualmente, agregamos funciones más complejas al espacio. Si ahora se nos da una muestra de n puntos de datos, seleccionamos nuestro clasificador del espacio $\mathcal{F}_n$. Para que este clasificador sea consistente con Bayes, debemos garantizar dos condiciones:

1. **El error de estimación debe converger a 0 cuando $n \to \infty$.** Para lograr esto, notemos que para cada n fijo, podemos acotar el error de estimación utilizando uno de los métodos conocidos. Esta cota disminuye conforme aumenta el tamaño de la muestra, pero crece cuando el término de complejidad aumenta. Debemos entonces asegurarnos de que el error de estimación total decrezca de forma neta, es decir, que el término de complejidad no domine al término del tamaño de la muestra. Para garantizar esto, debemos asegurarnos de que la complejidad de $\mathcal{F}_n$ no crezca demasiado rápido a medida que aumenta n.
2. **El error de aproximación debe converger a 0 cuando $n \to \infty$.** Intuimos que para lograrlo, debiéramos garantizar que, eventualmente, para algún n suficientemente grande, cada función en $\mathcal{F}_{\text{all}}$ esté contenida en $\mathcal{F}_n$, o que pueda ser aproximada por una función en $\mathcal{F}_n$. Discutiremos a continuación cómo puede lograrse.

Un ejemplo de cómo estas dos condiciones pueden formularse de manera rigurosa se presenta en el siguiente teorema, adaptado del Teorema 18.1 de Devroye et al. (1996):

Teorema 3.1.1.1. *Sea $\mathcal{F}_1, \mathcal{F}_2, \ldots$ una secuencia de espacios funcionales, y consideremos los clasificadores*

$$f_n = \arg \min_{f \in \mathcal{F}_n} R_{emp}(f).$$

Supongamos que, para cualquier distribución P, se satisfacen las siguientes dos condiciones:

1. *Las dimensiones VC de los espacios $\mathcal{F}_n$ satisfacen*

$$VC(\mathcal{F}_n) \cdot \log n/n \to 0$$

   cuando $n \to \infty$.

2. *El error de aproximación satisface $\mathcal{R}(f_{\mathcal{F}_n}) \to \mathcal{R}(f_{Bayes})$ cuando $n \to \infty$.*

Entonces, la secuencia de clasificadores f_n es consistente con Bayes.

Intentemos entender este teorema. Se nos da una secuencia creciente de espacios funcionales $\mathcal{F}_n$. Para cada muestra de tamaño n, elegimos la función en $\mathcal{F}_n$ que minimiza el riesgo empírico. Este es nuestro clasificador f_n. Si queremos que este clasificador sea consistente, deben cumplirse dos condiciones.

La primera condición establece que la complejidad de las clases de funciones, medida mediante la dimensión VC, debe crecer lentamente. Por ejemplo, si elegimos los espacios funcionales $\mathcal{F}_n$ de modo que $\mathrm{VC}(\mathcal{F}_n) \approx n^\alpha$ para algún $\alpha \in (0, 1)$, entonces la primera condición se satisface porque

$$\mathrm{VC}(\mathcal{F}_n) \cdot \frac{\log n}{n} \approx n^\alpha \cdot \frac{\log n}{n} = \frac{\log n}{n^{1-\alpha}} \to 0.$$

Sin embargo, si elegimos $\mathrm{VC}(\mathcal{F}_n) \approx n$ (es decir, $\alpha = 1$ en el cálculo anterior), entonces la condición ya no se cumple:

$$\mathrm{VC}(\mathcal{F}_n) \cdot \frac{\log n}{n} \approx \log n \to \infty.$$

La segunda condición del teorema simplemente establece que el error de aproximación debe converger a 0, pero el teorema no proporciona información sobre cómo lograr esto. Sin embargo, como discutimos anteriormente, es claro que esto solo se puede alcanzar mediante una secuencia creciente de clases de funciones.

3.1.2. Regularización

Una forma implícita de trabajar con espacios funcionales anidados es el **principio de regularización**. En lugar de minimizar el riesgo empírico $R_{\mathrm{emp}}(f)$ y luego expresar la capacidad de generalización del clasificador resultante f_n mediante alguna medida de capacidad de la clase funcional subyacente $\mathcal{F}$, se

puede seguir un enfoque más directo: se minimiza directamente el llamado **riesgo regularizado**:

$$\mathcal{R}_{\text{reg}}(f) = R_{\text{emp}}(f) + \lambda\Omega(f).$$

Aquí, $\Omega(f)$ es el llamado **regularizador**. Este regularizador está diseñado para penalizar funciones excesivamente complejas. Por ejemplo, a menudo se elige un regularizador que penaliza funciones con grandes fluctuaciones, es decir, se elige $\Omega(f)$ de modo que sea pequeño para funciones que varían lentamente y grande para funciones que fluctúan mucho. Como otro ejemplo, en clasificadores lineales se puede elegir $\Omega(f)$ como el inverso del margen de una función (definición 1.7.7.1).

El parámetro λ en la definición del riesgo regularizado es una constante de ajuste que equilibra la importancia de $R_{\text{emp}}(f)$ y $\Omega(f)$. Si λ es muy grande, tomamos la penalización inducida por $\Omega(f)$ muy seriamente y podríamos preferir funciones con $\Omega(f)$ pequeño, incluso si tienen un alto riesgo empírico. Por otro lado, si λ es pequeño, la influencia de la penalización disminuye y simplemente elegimos funciones basándonos en sus riesgos empíricos.

El principio de regularización consiste en elegir el clasificador f_n que minimiza el riesgo regularizado $\mathcal{R}_{\text{reg}}(f)$. Muchos de los clasificadores ampliamente utilizados pueden formularse dentro del marco de la regularización, por ejemplo, la máquina de soporte vectorial (conocidas como SVM o Support Vector Machines) (ver Schölkopf y Smola, 2002, para más detalles).

Para demostrar la **consistencia de Bayes** de clasificadores regularizados, se procede esencialmente como se describió en la subsección anterior: para alguna secuencia lentamente creciente $\omega_1, \omega_2, \ldots$, consideramos espacios funcionales anidados $\mathcal{F}_{\omega_1}, \mathcal{F}_{\omega_2}, \ldots$, donde cada $\mathcal{F}_{\omega_i}$ contiene todas las funciones f que satisfacen $\Omega(f) \leq \omega_i$. Eventualmente, si i es lo suficientemente grande, el espacio $\mathcal{F}_{\omega_i}$ aproximará el espacio $\mathcal{F}_{\text{all}}$ de todas las funciones.

Para garantizar la consistencia, se debe hacer que la constante λ tienda a 0 cuando $n \to \infty$. Esto asegura que, para n grande, efectivamente se nos permita elegir funciones de un espacio cercano a $\mathcal{F}_{\text{all}}$. Por otro lado, la constante λ no debe converger a 0 demasiado rápido, ya que, de lo contrario, comenzaríamos a sobreajustar para valores pequeños de n (si λ es muy pequeño, esencialmente se ignora el regularizador y, en consecuencia, se realiza algo similar a la minimización de riesgo empírico pero sobre un conjunto muy grande de funciones).

Nótese que existe una diferencia conceptual importante entre la minimización del riesgo empírico y la regularización. En la regularización, se introduce una función Ω que mide la **complejidad** de una función individual f. En ERM, por otro lado, nunca observamos la complejidad de funciones individuales, sino solo la complejidad de una clase funcional completa. Esta última, sin embargo,

es más bien una medida de **capacidad**, es decir, una medida de la cantidad de funciones en $\mathcal{F}$, y solo indirectamente una medida de qué tan complejas son las funciones individuales en la clase. Desde un punto de vista intuitivo, el primer enfoque es a menudo más fácil de comprender, ya que la complejidad de una función individual es un concepto más tangible que la capacidad de una clase funcional completa.

El teorema anterior muestra el principio general de cómo podemos lograr la **consistencia de Bayes**. Sin embargo, el teorema simplemente establece como su segunda condición que el error de aproximación debe converger a 0. ¿Cómo se puede lograr esto en la práctica? Resulta que hay muchas situaciones en las que esto no es tan difícil. Esencialmente, debemos asegurarnos de que cada función en $\mathcal{F}_{\text{all}}$ esté contenida en $\mathcal{F}_n$ para algún n grande, o que pueda ser aproximada arbitrariamente bien por una función de $\mathcal{F}_n$.

El área de las matemáticas que estudia este tipo de problemas se llama **teoría de la aproximación**, pero para los propósitos de la teoría del aprendizaje, resultados de aproximación simples suelen ser suficientes. El único problema técnico que debemos resolver es obtener una afirmación de la siguiente forma: si dos funciones son cercanas entre sí, entonces sus valores de riesgo también son cercanos.

Afirmaciones de este tipo suelen ser bastante fáciles de obtener. Por ejemplo, es directo ver que si f es una función de clasificación binaria (es decir, con $f(x) \in \{\pm 1\}$) y g es cualquier otra función medible arbitraria, y la distancia L_1 entre f y g es menor que δ, entonces su diferencia en el riesgo 0-1 también es menor que δ, es decir,

$$P(f(x) \neq \operatorname{sgn}(g(x))) < \delta.$$

Esto significa que, para demostrar que el error de aproximación de un espacio funcional $\mathcal{F}$ es menor que δ, solo necesitamos saber que cada función en $\mathcal{F}_{\text{all}}$ puede ser aproximada hasta δ en la norma L_1 por funciones de $\mathcal{F}$.

Resultados de este tipo son abundantes en la literatura matemática. Por ejemplo, si X es un subconjunto acotado de los números reales, es bien sabido que se puede aproximar cualquier función medible sobre este conjunto de manera arbitrariamente precisa mediante un polinomio. Por lo tanto, podríamos elegir los espacios $\mathcal{F}_n$ como los espacios de polinomios de grado a lo sumo d_n, donde d_n crece lentamente con n. Esto es suficiente para garantizar la convergencia del error de aproximación.

3.2. Los teoremas de la chancha y los veinte

Hemos presentado, además de parte de la historia de la disciplina, algunos resultados positivos en la teoría del aprendizaje estadístico. Formalizamos el pro-

blema de aprendizaje, hemos definido el objetivo del aprendizaje -minimizar el riesgo-, hemos especificado qué propiedades debe tener un clasificador -hablamos de la consistencia- y también hemos desarrollado un marco para pensar estas propiedades de manera fundamental. Además, vimos que existen diferentes formas de lograr clasificadores consistentes, de las cuales hemos mencionado algunas (como los k-vecinos más cercanos y la minimización del riesgo empírico).

Ahora surge de manera natural la siguiente pregunta: ¿cuál de todos los clasificadores consistentes es *el mejor* clasificador? ¿Existe un clasificador universalmente bueno?

Intentemos reformular esta pregunta de una manera más formal. Manteniendo la restricción de muestreo iid a partir de alguna distribución de probabilidad subyacente pero desconocida, ¿existe un clasificador que, en promedio sobre todas las distribuciones de probabilidad, obtenga mejores resultados que cualquier otro clasificador? ¿Podemos comparar dos clasificadores f y g, en promedio sobre todas las distribuciones?

La razón por la cual consideramos afirmaciones "en promedio sobre todas las distribuciones" radica en que no queremos hacer ninguna suposición sobre la distribución subyacente. Por lo tanto, parece natural estudiar el comportamiento de los clasificadores en cualquier posible distribución. Veamos entonces que, no solamente no existe un clasificador universalmente bueno, sino que, en promedio sobre todas las distribuciones posibles, todos los clasificadores son equivalentes.

Para tener alguna intuición al respecto, supongamos que nuestro espacio de entrada X solo consiste en un conjunto finito de puntos, es decir, $X = \{x_1, \ldots, x_m\}$ para algún número grande m. Ahora consideremos todas las formas posibles de asignar etiquetas a esos puntos de datos, es decir, consideremos todas las distribuciones de probabilidad posibles sobre X.

Dado un pequeño subconjunto de puntos de entrenamiento $(X_i, Y_i)_{i=1,\ldots,n}$, con $n < m$, utilizamos alguna regla de clasificación para construir un clasificador a partir de esos puntos. Es decir, fijamos un clasificador f construido con n puntos etiquetados del espacio finito X.

Ahora, consideremos todos los puntos de X que no han sido usados como puntos de entrenamiento y sus etiquetas. Solemos llamar conjunto de prueba (*test set*) a este conjunto. Por un lado, existe una asignación de etiquetas posible P_1 sobre el conjunto de prueba (los datos que el clasificador no conoce) en la que el clasificador no comete ningún error, específicamente aquella que coincide exactamente con la asignación de etiquetas producida por el clasificador mismo. Sin embargo, también existe una asignación de etiquetas P_2 en la que el clasificador comete el mayor error posible: basta tomar la distribución que asigna a los datos de prueba etiquetas de forma inversa a la del clasificador.

De la misma manera, podemos ver que, esencialmente, para cualquier error R dado, podemos construir una distribución de probabilidad sobre X tal que el error del clasificador f_n en el conjunto de prueba sea exactamente R. El mismo razonamiento se aplica a cualquier otro clasificador. Por lo tanto, si promediamos sobre todas las distribuciones de probabilidad posibles en X, todos los clasificadores f_n alcanzarán el mismo error de prueba: cada vez que existe una distribución en la que el clasificador obtiene un buen rendimiento, existe una distribución (digamos) opuesta sobre el conjunto de prueba en la que el clasificador obtiene un mal rendimiento. En particular, en promedio sobre todas las distribuciones de probabilidad, ningún clasificador puede ser mejor que un clasificador aleatorio en el conjunto de prueba.

Si bien este resultado pareciera atacar las mismas bases de la teoría del aprendizaje estadítico que nos hemos esmerado en construir, en realidad es algo bastante esperable cuando reparamos en el hecho de que estamos planteando medir la capacidad de un clasificador cualquiera sobre *todas* las distribuciones posibles. Como a priori no imponemos condiciones de ningún tipo sobre la distribución de probilidad, dado un clasificador cualquiera construido en base a n puntos, siempre podríamos hallar una distribución -al menos teórica- sobre la que dicho clasificador no tenga manera de generalizar los aprendido. Por ejemplo, se puede construir una distribución de probabilidad sobre las etiquetas simplemente lanzando una moneda y, para cada punto de datos, decidir su etiqueta verdadera según el resultado del lanzamiento aleatorio.

Parece plausible que, en un escenario así, no sirva de nada conocer las etiquetas de los puntos de entrenamiento. La única manera de aprender es excluir tales casos artificiales. Necesitamos asegurarnos de que existe algún mecanismo inherente por el cual podamos usar las etiquetas de entrenamiento para generalizar exitosamente a las etiquetas de prueba. Formalmente, esto significa que debemos restringir el espacio de distribuciones de probabilidad bajo consideración. Es decir, no podemos partir sin asumir nada acerca de la distribución.

En inglés, se utiliza la expresión *no free lunch* para hacer referencia a situaciones donde es necesario hacer compromisos y ceder, y de ahí deriva el nombre de los varios teoremas que formalizan la problemática de esta sección. En un intento de hacernos propio el tema, los llamaremos telúricamente, *los teoremas de la chancha y los veinte*. A continuación, enunciamos y probamos uno de los teoremas mencionados, resultado de Devroye en 1982, que encuadra la situación desde la perspectiva de la convergencia del riesgo.

Teorema 3.2.0.1 (de la chancha y los veinte)**.** *Sea $\varepsilon > 0$. Para cualquier entero n y cualquier regla de clasificación f_n, existe una distribución $P(X, Y)$ con riesgo bayesiano $\mathcal{R}(f_{Bayes}) = 0$ tal que:*

$$\mathbb{E}[\ell(X, Y, f_n)] \geq \frac{1}{2} - \varepsilon.$$

Demostración. Construiremos una familia de distribuciones conjuntas para el par de variables aleatorias (X, Y) y mostraremos que para al menos un elemento de esta familia, cualquier clasificador f_n -construido en base a n puntos en X-, presenta un rendimiento deficiente.

Consideremos un espacio de entrada discreto X que se distribuye uniformemente sobre el conjunto finito $\{1, 2, \ldots, K\}$, según la función distribución de probabilidad discreta:

$$P(X = i) = \begin{cases} \frac{1}{K}, & \text{si } i \in \{1, 2, \ldots, K\}, \\ 0, & \text{en otro caso.} \end{cases}$$

Construida la distribución de X, definamos la distribución conjunta $P(X, Y)$. Como X es finita, es posible parametrizar la distribución conjunta con un solo valor $b \in [0, 1)$. Luego, cada distribución posible de (X, Y) puede es representada por un único valor b. Tomando la expansión binaria de b, tenemos:

$$b = 0.b_0 b_1 b_2 \ldots.$$

Definimos entonces la etiqueta Y como función de X:

$$Y = b_X,$$

donde b_X coincide con el X-ésimo dígito binario de b. Utilizar la expansión binaria de b facilita la interpretación de la parametrización, dado que cada dígito i-ésimo es exactamente la etiqueta correspondiente a la muestra i-ésima. Por ejemplo, si $K = 3$ y estamos en el caso donde $f_{Bayes}(X_1) = 0$, $f_{Bayes}(X_2) = 1$ y $f_{Bayes}(X_3) = 1$, la distribucíon conjunta estaría representada por $b = 0, 011$.

Dado que Y es una función de X, existe una regla de decisión perfecta y el clasificador de Bayes alcanza una pérdida cero:

$$\ell(X, Y, f_{Bayes}) = 0.$$

Consideramos a f_n un clasificador sobre X construido a partir de las primeras n muestras. Escribimos entonces

$$f_n(X) = f_n(X, Z_n),$$

para explicitar que f_n es función también de la muestra de entrenamiento Z_n. Luego, definimos el riesgo sobre b como el riesgo de f_n para esa distribución asociado a la pérdida-0-1:

$$\begin{aligned}\mathcal{R}_n(b) &= \mathbb{E}[\ell((X,Y)(b), f_n(X))] \\ &= \sum_{i=1}^{K} 1 \cdot P(f_n(X_i, Z_n) \neq b_i) + \sum_{i=1}^{K} 0 \cdot P(f_n(X_i, Z_n) = b_i) \\ &= \sum_{i=1}^{K} P(f_n(X_i, Z_n) \neq b_i) \\ &= P(f_n(X, Z_n) \neq Y).\end{aligned}$$

Ahora introducimos una variable aleatoria B, distribuida uniformemente en $[0, 1)$ e independiente de X y de los datos de entrenamiento Z_n. Su expansión binaria es: $B = 0, B_1 B_2 B_3 \ldots$, que es una sucesión de variables aleatorias independientes en $\{0, 1\}$ con probabilidad $\frac{1}{2}$. Evidentemente, el rol de B es generar una distribución conjunta sobre X e Y de manera aleatoria. También podemos generar de manera aleatoria una distribución para Y, que podemos llamar $Y = B_X$, usando la misma convención que antes donde, de nuevo, B_X es el dígito que corresponde al punto X en la expansión binaria de B. Notemos que X puede no pertenecer al conjunto de entrenamiento Z_n.
Podemos expresar el riesgo de este nuevo parámetro como:

$$\mathcal{R}_n(B) = P(f_n(X, Z_n) \neq B_X).$$

y dado que B representa a cualquiera de las distribuciones parametrizadas por b, es cierto que

$$\sup_{b \in [0,1)} \mathcal{R}_n(b) \geq \mathcal{R}_n(B). \tag{3.1}$$

Notemos que B_X es una variable de Bernoulli(1/2) independiente de $f_n(X, Z_n)$ cuando X no pertenece al conjunto de entrenamiento. De aquí se obtiene:

$$P(f_n(X, Z_n) \neq B_X \mid X, Z_n) = \frac{1}{2}, \quad \text{si } X \notin \{X_1, \ldots, X_n\}.$$

Esto se debe a que el clasificador no tiene información sobre la etiqueta de X si X no está en el conjunto de entrenamiento, como es el caso y, dado que B_X se genera aleatoriamente, la probabilidad de que f_n acierte es de 1/2.

Esto nos sirve para dar la siguiente expresión del riesgo de B:

$$\begin{aligned}\mathcal{R}_n(B) &= P(f_n(X, Z_n) \neq B_X) \\ &= P(f_n(X, Z_n) \neq B_X \mid X \notin \{X_1, \ldots, X_n\}) \cdot P(X \notin \{X_1, \ldots, X_n\}) \\ &\quad + P(f_n(X, Z_n) \neq B_X \mid X \in \{X_1, \ldots, X_n\}) \cdot P(X \in \{X_1, \ldots, X_n\}) \\ &\geq P(f_n(X, Z_n) \neq B_X \mid X \notin \{X_1, \ldots, X_n\}) \cdot P(X \notin \{X_1, \ldots, X_n\}) \\ &= \frac{1}{2} \cdot P(X \notin \{X_1, \ldots, X_n\})\end{aligned}$$

Donde simplemente separamos los eventos en donde el punto X sobre el que analizamos al clasificador está o no está en el conjunto de entrenamiento, y desechamos el caso afirmativo. Luego

$$\mathcal{R}_n(B) = P(f_n(X, Z_n) \neq B_X) \geq \frac{1}{2} \cdot P(X \notin \{X_1, \ldots, X_n\})$$

Por otro lado, como las muestras son tomadas de forma iid:

$$\begin{aligned} P(X \notin \{X_1, \ldots, X_n\}) &= P(X \neq X_1, X \neq X_2, \ldots, X \neq X_n) \\ &= P(X \neq X_1)^n \\ &= \left(1 - \frac{1}{K}\right)^n \end{aligned}$$

Luego

$$\mathcal{R}_n(B) \geq \frac{1}{2} \cdot \left(1 - \frac{1}{K}\right)^n$$

Considerando la cota 3.1, sabemos que existe una distribución conjunta $(X, Y)(b)$ tal que

$$\mathcal{R}_n(b) \;=\; \mathbb{E}[\ell((X, Y)(b), f_n(X))] \;\geq\; \frac{1}{2} \cdot \left(1 - \frac{1}{K}\right)^n \xrightarrow[K \to \infty]{} \frac{1}{2}$$

□

Este teorema establece que, aunque existen reglas de clasificación que son universalmente consistentes, es decir, que asintóticamente proporcionan el rendimiento óptimo para cualquier distribución, su desempeño en muestras finitas siempre será muy malo -cercano al de clasificar utilizando el lanzamiento de una moneda- para algunas distribuciones.

Esto significa que ningún clasificador puede garantizar que logremos un error aceptable para todas las distribuciones, aún teniendo una cantidad de muestras enorme (pero finita). Sin embargo, dado que la distribución adversa depende de n, el Teorema no nos permite concluir que existe una única distribución para la cual la probabilidad de error sea mayor que, por ejemplo, $\ell(f_{Bayes}) + 1/4$ para todo n. Tal afirmación iría en contradicción la propia existencia de reglas universalmente consistentes.

Las restricciones sobre la distribución de probabilidad pueden adoptar diversas formas. Podríamos suponer que la distribución subyacente tiene una función de densidad que *se comporta bien*. O que existe una función de distancia en el espacio de entrada y que las etiquetas dependen de alguna manera continua de la distancia, es decir, que los puntos cercanos tienden a tener etiquetas similares. Si hacemos tales suposiciones, es posible construir clasificadores que las exploten (por ejemplo, el clasificador k-NN para aprovechar la estructura de distancia). Desde ya que el teorema sigue siendo válido, lo que significa que habrá conjuntos

de datos para los cuales este clasificador fallará estrepitosamente. Sin embargo, estos serán conjuntos de datos provenientes de distribuciones donde las suposiciones han sido gravemente violadas. Y en tales casos, tiene sentido que un clasificador que se basa en esas suposiciones no tenga ninguna posibilidad de éxito.

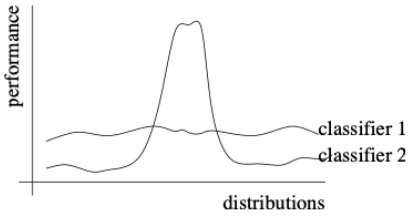

Figura 3.1: Tomando en consideración todas las distribuciones posibles -eje X-, el clasificador 1 es en general mejor que el clasificador 2, y este último es muchísimo mejor para un subconjunto de distribuciones. Podemos pensar que 1 es un clasificador general y 2 está construído para un tipo de problemas específico. El teorema de la chancha y los veinte nos dice que el área debajo de ambas curvas debe ser igual, es decir que en promedio ambos clasificadores tienen el mismo error.

Un problema que suele surgir en el contexto del teorema de la chancha y los veinte es cómo se relaciona con los teoremas de consistencia demostrados anteriormente. Por ejemplo, vimos que el clasificador de los k-vecinos más cercanos es universalmente consistente, es decir, es consistente para cualquier distribución subyacente P. La solución a esta aparente paradoja radica en el hecho de que las afirmaciones de consistencia solo tratan el caso límite cuando $n \to \infty$. En el ejemplo del espacio de datos finito mencionado anteriormente, notemos que, tan pronto como el tamaño de la muestra es lo suficientemente grande como para haber muestreado esencialmente cada zona del espacio al menos una vez, un clasificador que memoriza los datos de entrenamiento ya no cometerá más errores. Afirmaciones similares (aunque algo más complicadas) también son válidas en casos de espacios de datos infinitos. Así, los teoremas de la chanca y los veinte hacen afirmaciones sobre un tamaño de muestra finito n, mientras que la consistencia considera el límite cuando $n \to \infty$.

Otra manera de expresar esto es la siguiente: si las etiquetas de los puntos se asignan de manera completamente aleatoria, entonces las etiquetas de los puntos de prueba serán completamente independientes de las etiquetas de los puntos de entrenamiento. En tal escenario, el aprendizaje es imposible.

Bibliografía

[1] *Statistical Learning Theory: Models, Concepts and Results* - von Luxburg, B. Schölkopf (2008).

[2] *The Nature of Statistical Learning Theory* - Vladimir Vapnik (1995).

[3] *Statistical Learning Theory* - Vladimir Vapnik (1998).

[4] *A Probabilistic Theory of Pattern Recognition* - Luc Devroye, László Györfi, Gábor Lugosi (1996).

[5] *Learning with Kernels* - B. Schölkopf, A. Smola (1996).

[6] *Statistical Inference* - Casella, Berger .

[7] *Pattern Recognition and Machine Learning* - C. Bishop (2006).

[8] *On the uniform convergence of relative frequencies of events to their probabilities* - V. N. VAPNIK and A. YA. CHERVONENKIS (1971)