1

Un comienzo tranquilo

1.1 El marco de aprendizaje estadístico

- **Conjunto de dominio:** Los puntos del dominio son *instancias* y a $\mathcal{X}$ cómo *espacios de instancias*.
- **Conjunto de etiquetas:** Sea $\mathcal{Y}$ el conjunto de etiquetas, donde $\{0, 1\}$.
- **Datos de entrenamiento:** Sea $\mathcal{X} \times \mathcal{Y}$, una secuencia finita de pares, se llama un conjunto de entrenamiento a
$$S = ((x_1, y_1), \ldots, (x_m, y_m)).$$
- **Salida del aprendizaje:** Sea $h : \mathcal{X} \to \mathcal{Y}$ un predictor, hipótesis o clasificador. Se utiliza para predecir nuevos puntos del dominio. $A(S)$ se denota cómo el predictor de que un algoritmo de aprendizaje, A, regresa al recibir la secuencia de entrenamiento S.
- **Modelo simple de generación de datos:** Denotamos a $\mathcal{D}$ la probabilidad sobre $\mathcal{X}$. Luego, tenemos
$$f : \mathcal{X} \to \mathcal{Y}$$
una función de etiquetas *correctas*.
En resumen, cada par en los datos de entrenamiento S se genera muestreando primero un punto x_i de acuerdo con $\mathcal{D}$ y luego etiquetándolo con f.
- **Medidas de éxito:** Definimos el error de un clasificador o predictor cómo la probabilidad de que no prediga la etiqueta correcta en un punto. Es decir, el error de h es la probabilidad de extraer una instancia aleatoria x, según la distribución $\mathcal{D}$, tal que $h(x)$ no sea igual a $f(x)$.
formalmente, dado un subconjunto del dominio $A \subset \mathcal{X}$, la distribución de probabilidad, $\mathcal{D}$, asigna un número, $\mathcal{D}(A)$, que determina la probabilidad de observar un punto $x \in A$. A nos referimos a un evento y lo expresamos con la función $\pi : \mathcal{X} \to \{0, 1\}$. Es decir,
$$A = \{x \in \mathcal{X} : \pi(x) = 1\}.$$
También podemos utilizar la notación:
$$\mathbb{P}_{x \sim \mathcal{D}}[\pi(x)],$$
para expresar $\mathcal{D}(A)$.

  Definimos el error de una regla de predicción, $h : \mathcal{X} \to \mathcal{Y}$, como
$$L_{\mathcal{D},f}(h) = \mathbb{P}_{x \sim \mathcal{D}}[h(x) \neq f(x)] = \mathcal{D}(\{x : h(x) \neq f(x)\}). \tag{1.1}$$
  - $(\mathcal{D}, f)$ indica que el error se mide con respecto a la distribución de probabilidad $\mathcal{D}$ y la función de etiquetado correcta f.
  - $L_{(\mathcal{D},f)}(h)$ es el error de generalización, riesgo o error verdadero de h. L es el error que consideramos cómo la pérdida del aprendizaje.

1.2 Minimización empírica del riesgo

Una noción útil de error que el aprendizaje puede calcular es el error de entrenamiento, dado por:

$$L_S(h) = \frac{|\{i \in [m] : h(x_i) \neq y_i\}|}{m} \tag{1.2}$$

donde $[m] = \{1, \ldots, m\}$. Este paradigma de aprendizaje (proponer un predictor h que minimice $L_S(h)$) se denomina *Minimización Empírica del Riesgo o ERM*, para abreviar.

1.2.1 Sobreajuste

El sobre ajuste se produce cuando encontramos un predictor cuyo desempeño en el conjunto de entrenamiento es excelente, pero en el mundo real es muy pobre.

Se tiene un cuadrado C con área 2 que contiene varias instancias. Dentro de C existe otro cuadrado con área 1. Consideremos el siguiente predictor:

$$h_S(x) = \begin{cases} y_i & \text{si} \quad \exists i \in [m]\ x_i = x \\ \\ 0 & \text{En otro caso} \end{cases} \tag{1.3}$$

Si, utilizamos un algoritmo ERM, se tiene que ningún clasificador puede tener un error menor a $L_S(h_S) = 0$. Por otro lado, el error verdadero de cualquier clasificador que predice la etiqueta 1, sólo en un número finito de instancias es, en este caso $1/2$, de donde $L_{\mathcal{D}}(h_S) = 1/2$.

El sobreajuste del emparejamiento polinomial: El objetivo de este ejercicio es mostrar que puede describirse como un polinomio con umbral. Es decir, demostrar que dado un conjunto de entrenamiento $S = \{(x_i, f(x_i))\}_{i=1}^m \subseteq (\mathbb{R}^d \times \{0,1\})^m$, existe un polinomio p_S tal que $h_S(x) = 1$ si y sólo si $p_S(x) \geq 0$, donde h_S es como se define en la Ecuación (2.3). De ello se deduce que aprender la clase de todos los polinomios con umbral utilizando la regla ERM puede conducir a un sobreajuste.

1.3 Minimización empírica del riesgo con sesgo inductivo

Acabamos de ver que ERM podría conducir a un sobreajuste. En lugar de renunciar al paradigma de ERM, buscaremos condiciones bajo las cuales exista una garantía de no sobreajuste. Es decir, condiciones bajo las cuales el predictor de ERM tiene un buen desempeño con respecto a los datos de entrenamiento.

Formalmente se debe elegir a priori un conjunto de predictores. Este conjunto se llama clase de hipótesis y se denota por $\mathcal{H}$. Cada $h \in \mathcal{H}$ es una función que se asigna de $\mathcal{X}$ a $\mathcal{Y}$. Para una clase $\mathcal{H}$ dada y una muestra de entrenamiento, S, el $\text{ERM}_{\mathcal{H}}$, usa la regla ERM para elegir un predictor, $h \in \mathcal{H}$. Con el menor error posible sobre S. Formalmente:

$$\text{ERM}_{\mathcal{H}}(S) \in \text{argmin}_{h \in \mathcal{H}} L_S(h),$$

donde armin representa el conjunto de hipótesis en H que alcanzan el valor mínimo de $L_S(h)$ sobre $\mathcal{H}$. Al restringir a elegir un predictor de $\mathcal{H}$, lo sesgamos hacia un conjunto particular de predictores. Esta restricción suelen denominarse *sesgo inductivo*, dado que la elección de dicha restricción se determina antes de que el alumno vea los datos de entrenamiento, idealmente debería basarse en algún conocimiento previo sobre el problema que se va a aprender.

Elegir una clase de hipótesis más restringida nos protege mejor contra el sobreajuste, pero al mismo tiempo podría causarnos un sesgo inductivo más fuerte.

1.3.1 Clases de hipótesis finitas

El tipo de restricción más simple a una clase es imponer un límite superior a su Tamaño al número de predictores h en $\mathcal{H}$. Demostraremos que si $\mathcal{H}$ es una clase finita, entonces $\text{ERM}_{\mathcal{H}}$ no se sobreajustará, siempre que se base en una muestra de entrenamiento suficientemente grande, que dependerá del Tamaño de $\mathcal{H}$.

Analicemos el desempeño de la regla de aprendizaje $\text{ERM}_{\mathcal{H}}$ suponiendo que $\mathcal{H}$ es una clase finita. Para una muestra de entrenamiento, S, etiquetada de acuerdo con alguna $f : \mathcal{X} \to \mathcal{Y}$, sea h_S el resultado de aplicar $\text{ERM}_{\mathcal{H}}$ a S. Es decir,

$$h_S \in \text{argmin}_{h \in \mathcal{H}} L_S(h). \tag{1.4}$$

Definición 1.1 **El supuesto de realizabilidad.** Existe $h^* \in \mathcal{H}$ tal que $L_{\mathcal{D},f}(h^*) = 0$. Este supuesto implica que con probabilidad 1 sobre muestras aleatorias, S, donde las instancias de S se muestrean de acuerdo a $\mathcal{D}$ y están etiquetadas por f, tenemos $L_S(h^*) = 0$. En otras palabras, significa que hay una función perfecta que puede predecir sin errores todos los ejemplos en la distribución de los datos.

La suposición más común en ML estadístico es que la muestra de entrenamiento S se genera mediante puntos de muestreo de la distribución $\mathcal{D}$ independientemente unos de otros. Formalmente:

El *i.i.d. Supuesto:* Cada x_i en S se muestrea recientemente de acuerdo con $\mathcal{D}$ y luego se etiqueta de acuerdo con la función de etiquetado, f. Denotamos esta suposición como $S \sim \mathcal{D}^m$ donde m es el tamaño de S, y $\mathcal{D}^m$ denota la probabilidad sobre m-tuplas inducida al aplicar $\mathcal{D}$ para seleccionar cada elemento de la tupla independientemente de los otros miembros de la tupla. Intuitivamente, el conjunto de entrenamiento S es una ventana a través de la cual se obtiene información parcial sobre la distribución $\mathcal{D}$ en el mundo y la función de etiquetado, f. Cuanto más grande sea la muestra, más probable será que refleje con mayor precisión la distribución y el etiquetado utilizados para generarla.

En el marco de aprendizaje estadístico, la selección aleatoria del conjunto de entrenamiento S introduce una variabilidad en la elección del predictor h_S y, por ende, en el riesgo empírico $L_{\mathcal{D},f}(h_S)$. Este riesgo empírico es una medida de qué tan bien el predictor seleccionado se ajusta a los datos de entrenamiento. Sin embargo, debido a la aleatoriedad en la selección de S, existe la posibilidad de que este conjunto no sea representativo de la distribución de datos subyacente $\mathcal{D}$. Para cuantificar esta incertidumbre, introducimos la probabilidad δ de obtener una muestra no representativa, y su complemento $(1 - \delta)$, que representa el nivel de confianza en nuestra predicción.

El *parámetro de precisión* ϵ es crucial para evaluar la calidad del aprendizaje. Si el riesgo verdadero $L_{(\mathcal{D},f)}(h_S)$ excede ϵ, interpretamos que el aprendizaje ha fallado, ya que el predictor no es suficientemente preciso. En contraste, si $L_{(\mathcal{D},f)}(h_S) \leq \epsilon$, consideramos que el predictor es aproximadamente correcto.
El objetivo es limitar la probabilidad de seleccionar muestras de entrenamiento que conduzcan a un predictor con un riesgo excesivo. Formalmente, buscamos un límite superior para la probabilidad:

$$\mathcal{D}^m \left(\left\{ S|_x : L_{(\mathcal{D},f)}(h_S) > \epsilon \right\} \right).$$

El conjunto $\mathcal{H}_B$ contiene las hipótesis 'malas', aquellas cuyo riesgo supera ϵ. Estas son las hipótesis que deseamos evitar, ya que no cumplen con nuestro criterio de precisión.
El conjunto M se define como las muestras engañosas, donde para cada $S|_x \in M$, existe al menos una hipótesis mala en $\mathcal{H}_B = \left\{ h \in \mathcal{H} : L_{(\mathcal{D},f)}(h) > \epsilon \right\}$ que parece buena en $S|_x$, es decir, tiene un riesgo empírico de cero:

$$M = \{ S|_x : \exists h \in \mathcal{H}_B, L_S(h) = 0 \}.$$

Bajo el supuesto de realizabilidad, que asume que $L_S(h_S) = 0$ para la hipótesis seleccionada por el algoritmo, el evento $L_{(\mathcal{D},f)}(h_S) > \epsilon$ solo puede ocurrir si $S|_x$ pertenece a M. Esto significa que:

$$\left\{S|_x : L_{(\mathcal{D},f)}(h_S) > \epsilon\right\} \subseteq M.$$

Podemos expresar M como la unión de los conjuntos de muestras de entrenamiento que hacen que cualquier hipótesis mala parezca buena:

$$M = \bigcup_{h \in \mathcal{H}_B} \{S|_x : L_S(h) = 0\}. \tag{1.5}$$

Por lo tanto, la probabilidad de que el riesgo del predictor seleccionado sea mayor que ϵ, está acotada por la probabilidad de que la muestra de entrenamiento pertenezca a M:

$$\mathcal{D}^m\left(\left\{S|_x : L_{(\mathcal{D},f)}(h_S) > \epsilon\right\}\right) \leq \mathcal{D}^m(M) = \mathcal{D}^m\left(\bigcup_{h \in \mathcal{H}_B} \{S|_x : L_S(h) = 0\}\right). \tag{1.6}$$

Esta relación es fundamental para comprender cómo la calidad de los datos de entrenamiento afecta la selección de un predictor y, en última instancia, la generalización del modelo. Si los datos de entrenamiento son de mala calidad, es decir, no representativos o engañosos, aumenta la probabilidad de seleccionar un predictor que no generalice bien, lo que se traduce en un riesgo verdadero que supera el umbral de precisión ϵ. Por lo tanto, es esencial asegurarse de que los datos de entrenamiento sean de alta calidad y representativos de la distribución real de los datos para desarrollar modelos robustos y confiables.
Ahora, acotamos superiormente el lado derecho de la Ecuación anterior usando el límite de unión, una propiedad básica de probabilidades.

Lema 1.1 **Límite de unión.** Para cualesquera dos conjuntos A, B y una distribución $\mathcal{D}$ tenemos

$$\mathcal{D}(A \cup B) \leq \mathcal{D}(A) + \mathcal{D}(B).$$

■

En el contexto del aprendizaje automático, el análisis de la probabilidad de que un algoritmo seleccione un predictor inadecuado es fundamental. Este predictor, denotado h_S, es evaluado por su riesgo verdadero $L_{(\mathcal{D},f)}(h_S)$, que mide cuán bien las predicciones coinciden con los resultados reales. Un riesgo que supera el umbral ϵ es indicativo de un fallo en el aprendizaje, señalando que el predictor no generaliza adecuadamente a datos no vistos.
Para establecer una cota superior en la probabilidad de seleccionar tal predictor, aplicamos el lema del límite de unión, que nos dice que la probabilidad de la unión de dos eventos no puede ser mayor que la suma de sus probabilidades individuales. En nuestro análisis, consideramos la unión de eventos donde cada hipótesis mala en $\mathcal{H}_B$—aquellas cuyo riesgo verdadero excede ϵ—parece ajustarse perfectamente al conjunto de entrenamiento. Matemáticamente, esto se expresa como:

$$\mathcal{D}^m\left(\left\{S|_x : L_{(\mathcal{D},f)}(h_S) > \epsilon\right\}\right) \leq \sum_{h \in \mathcal{H}_B} \mathcal{D}^m\left(\{S|_x : L_S(h) = 0\}\right). \tag{1.7}$$

Para cada hipótesis mala h en $\mathcal{H}_B$, la probabilidad de que h clasifique correctamente todos los ejemplos en el conjunto de entrenamiento S es el producto de las probabilidades individuales de que h clasifique correctamente cada ejemplo. Dado que los ejemplos se asumen independientes e idénticamente distribuidos (i.i.d.), esta probabilidad se calcula como:

$$\mathcal{D}^m\left(\{S|_x : L_S(h) = 0\}\right) = \mathcal{D}^m\left(\{S|_x : \forall i, h(x_i) = f(x_i)\}\right) = \prod_{i=1}^{m} \mathcal{D}\left(\{x : h(x) = f(x)\}\right). \tag{1.8}$$

La probabilidad de que h clasifique correctamente un único ejemplo es $1 - L_{(\mathcal{D},f)}(h)$, que es menor o igual a $1 - \epsilon$ debido a que h es una hipótesis mala. Utilizando la desigualdad exponencial $1 - \epsilon \leq e^{-\epsilon}$, obtenemos una cota más estricta para la probabilidad de que h se ajuste perfectamente a S:

$$\mathcal{D}^m\left(\{S|_x : L_S(h) = 0\}\right) \leq (1-\epsilon)^m \leq e^{-m\epsilon}. \tag{1.9}$$

Al combinar esta cota con la cantidad de hipótesis malas, concluimos que la probabilidad de que el riesgo del predictor seleccionado supere ϵ está acotada por $|\mathcal{H}_B|e^{-m\epsilon}$, y dado que $|\mathcal{H}_B|$ es a lo sumo igual al tamaño total de la clase de hipótesis $|\mathcal{H}|$, la cota final es:

$$\mathcal{D}^m\left(\left\{S|_x : L_{(\mathcal{D},f)}(h_S) > \epsilon\right\}\right) \leq |\mathcal{H}_B|e^{-\epsilon m} \leq |\mathcal{H}|e^{-\epsilon m}.$$

Este resultado es significativo porque relaciona la complejidad de la clase de hipótesis y el tamaño del conjunto de entrenamiento con la confianza en la selección de un predictor adecuado. Cuanto mayor sea el conjunto de entrenamiento y más pequeña la clase de hipótesis, menor será la probabilidad de seleccionar un predictor con un riesgo alto, lo que favorece la generalización del modelo a nuevos datos. Cada punto en el círculo grande representa una posible m-tupla de instancias. Cada óvalo coloreado representa el conjunto de m-tuplas "engañosas" de instancias para algún predictor "malo" $h \in \mathcal{H}_B$. El ERM puede potencialmente sobreajustarse siempre que obtenga un conjunto de entrenamiento S engañoso. Es decir, para algunos $h \in \mathcal{H}_B$ tenemos $L_S(h) = 0$. La ecuación (2.9) garantiza que para cada mala hipótesis individual, $h \in \mathcal{H}_B$, como máximo $(1-\epsilon)^m$ la fracción m de los conjuntos de entrenamiento sería engañosa. En particular, **cuanto mayor es m, más pequeño se vuelve cada uno de estos óvalos coloreados**. La unión formaliza el hecho de que el área que representa los conjuntos de entrenamiento que son engañosos con respecto a algunos $h \in \mathcal{H}_B$ (es decir, los conjuntos de entrenamiento en M) es como máximo la suma de las áreas de los óvalos coloreados. Por lo tanto, está acotado por $|\mathcal{H}_B|$ veces el tamaño máximo de un óvalo de color. Cualquier muestra S fuera de los óvalos coloreados no puede hacer que la regla ERM se ajuste demasiado.

Corolario 1.1 Sea $\mathcal{H}$ una clase de hipótesis finita y sean $\delta \in (0,1)$ y $\epsilon > 0$, y sea m un número entero que satisfaga

$$m \geq \frac{\log(|\mathcal{H}|/\delta)}{\epsilon}.$$

Entonces, para cualquier función de etiquetado, f, y para cualquier distribución, $\mathcal{D}$, para la cual se cumple el supuesto de realizabilidad (es decir, para alguna $h \in \mathcal{H}, L_{(\mathcal{D},f)}(h) = 0$), con probabilidad de al menos $1-\delta$ sobre la elección de un i.i.d. muestra S de tamaño m, tenemos que para cada hipótesis de ERM, h_S, se cumple que

$$L_{(\mathcal{D},f)}(h_S) \leq \epsilon.$$

■

El corolario nos dice que para una m suficientemente grande, la regla $\text{ERM}_{\mathcal{H}}$ sobre una clase de hipótesis finita será probablemente (con confianza $1-\delta$) aproximadamente (hasta un error de ϵ) correcta. En el siguiente capítulo definimos formalmente el modelo de aprendizaje probablemente aproximadamente correcto (PAC).

2

Un modelo de aprendizaje formal

2.1 Aprendizaje PAC

En el capítulo anterior hemos demostrado que para una clase de hipótesis finita, si la regla ERM con respecto a esa clase se aplica en una muestra de entrenamiento suficientemente grande (cuyo tamaño es independiente de la distribución subyacente o la función de etiquetado), entonces la hipótesis de salida será probablemente aproximadamente correcto. De manera más general, definamos aprendizaje probablemente aproximadamente correcto (PAC).

Definición 2.1 **Capacidad de aprendizaje PAC.** Una clase de hipótesis $\mathcal{H}$ es PAC aprendizable si existe una función $m_{\mathcal{H}} : (0,1)^2 \to \mathbb{N}$ y un algoritmo de aprendizaje con la siguiente propiedad: Para cada , $\delta \in (0,1)$, para cada distribución $\mathcal{D}$ sobre $\mathcal{X}$, y para cada función de etiquetado $f : \mathcal{X} \to \{0,1\}$, si el supuesto realizable se cumple con respecto a $\mathcal{H}, \mathcal{D}, f$, entonces cuando se ejecuta el algoritmo de aprendizaje en $m \geq m_{\mathcal{H}}(\epsilon, \delta)$ i.i.d. En los ejemplos generados por $\mathcal{D}$ y etiquetados por f, el algoritmo devuelve una hipótesis h tal que, con una probabilidad de al menos $1 - \delta$ (sobre la elección de los ejemplos), $L_{(\mathcal{D},f)}(h) \leq \epsilon$.

- Clase de Hipótesis ($\mathcal{H}$): Es un conjunto de modelos.
- Función ($m_{\mathcal{H}}$): Determinaría cuantos ejemplos necesita revisar el algoritmo para poder identificar con confianza un modelo que prediga
- Distribución ($\mathcal{D}$) sobre ($\mathcal{X}$): Representaría la distribución de probabilidad de las instancias.
- Función de Etiquetado (f): Es la "verdad de fondo" que etiqueta correctamente. (Las etiquetas en el conjunto de datos de entrenamiento son ejemplos de la función).
- Suposición Realizable: Esta suposición implicaría que existe al menos un modelo en ($\mathcal{H}$) que puede modelar con precisión perfecta, según (f).
- Hipótesis (h): Después de entrenar con suficientes datos, el algoritmo seleccionaría un modelo específico (h) que predice con precisión.
- $L_{(\mathcal{D},f)}(h)$: La pérdida empírica indica qué tan bien el modelo seleccionado (h) está prediciendo con la función de etiquetado verdadera (f).
- ϵ (Epsilon): Es el margen de error que estamos dispuestos a tolerar en las predicciones de nuestro algoritmo de aprendizaje. En otras palabras, es una medida de cuánto error es aceptable en la hipótesis que el algoritmo produce. Un ϵ pequeño significa que queremos que nuestra hipótesis sea muy precisa, mientras que un ϵ más grande indica que estamos dispuestos a aceptar más errores.

Nos permitirá perdonar al clasificador por cometer errores menores. Nos dice que tan lejos puede estar el clasificador de salida del óptimo.

- δ (Delta): Es la probabilidad de que nuestro algoritmo de aprendizaje falle en encontrar una hipótesis con un error menor o igual a ϵ. Es una medida de confianza en el rendimiento del algoritmo. Un δ pequeño significa que queremos estar muy seguros de que nuestro algoritmo encontrará una hipótesis buena, mientras que un δ más grande significa que estamos dispuestos a aceptar un mayor riesgo de fallar.
  Indica la probabilidad de que el clasificador cumpla con ese requisito de precisión ϵ.

  En la práctica, cuando establecemos valores para ϵ y δ, estamos definiendo nuestras expectativas y limitaciones para el algoritmo de aprendizaje. Por ejemplo, si establecemos $\epsilon = 0.05$ y $\delta = 0.01$, estamos diciendo que queremos que el algoritmo encuentre una hipótesis que tenga un 5% de error o menos, y queremos estar 99% seguros de que lo logrará.

2.1.1 Complejidad de muestra

$m_{\mathcal{H}}$ es una función que determina cuántos ejemplos necesitará el algoritmo para encontrar una hipótesis con un error menor o igual a ϵ y con una probabilidad de al menos $1 - \delta$. Definiremos la complejidad muestral del aprendizaje $\mathcal{H}$ como la *función mínima* en el sentido de que para cualquier, δ, $m_{\mathcal{H}}(\epsilon, \delta)$ es el entero mínimo que satisface los requisitos del aprendizaje PAC, con precisión ϵ y confianza δ.

Corolario 2.1 Cada clase de hipótesis finita es PAC aprendizable con complejidad muestral

$$m_{\mathcal{H}}(\epsilon, \delta) \leq \left\lceil \frac{\log(|\mathcal{H}|/\delta)}{\epsilon} \right\rceil .$$

■

2.2 Un modelo de aprendizaje más general: Liberación del supuesto de realizabilidad: aprendizaje PAC agnóstico

El modelo que acabamos de describir puede generalizarse fácilmente, de modo que pueda resultar relevante para un ámbito más amplio de tareas de aprendizaje. Consideramos generalizaciones en dos aspectos:

- Eliminación del supuesto de realizabilidad.
- Problemas de aprendizaje más allá de la clasificación binaria.

Un modelo más realista para la distribución generadora de datos

Recordemos que el supuesto de realizabilidad requiera que:

$$\mathbb{P}_{x \sim \mathcal{D}} \left[h^*(x) = f(x) \right] = 1.$$

A continuación suavizaremos el supuesto de realizabilidad reemplazando la función de etiquetado objetivo con una noción más flexible, una distribución generadora de etiquetas de datos.

Formalmente, sea $\mathcal{D}$ una distribución de probabilidad sobre $\mathcal{X} \times \mathcal{Y}$, donde, cómo antes, $\mathcal{X}$ es nuestro conjunto de instancias e $\mathcal{Y}$ es un conjunto de etiquetas ($\{0, 1\}$). Es decir, $\mathcal{D}$ es una distribución conjunta entre puntos y etiquetas de instancias. Se puede ver dicha distribución como compuesta de dos partes:

- Una distribución $\mathcal{D}_x$a sobre putos de instancias no etiquetas (distribución marginal) y

- una probabilidad condicional sobre etiquetas para cada punto del dominio, $\mathcal{D}((x,y)|x)$.

En el ejemplo de la papaya, $\mathcal{D}_x$ determina la probabilidad de encontrar una papaya cuyo color y dureza se encuentren en algún dominio de valores de color-dureza, y la probabilidad condicional es la probabilidad de que una papaya con color y dureza representados por x sea sabrosa. De hecho, tal modelado permite que dos papayas que comparten el mismo color y dureza pertenezcan a diferentes categorías de sabor.

El error empírico y el verdadero revisado

Para una distribución de probabilidad, $\mathcal{D}$, sobre $\mathcal{X} \times \mathcal{Y}$, se puede medir la probabilidad de que h cometa un error cuando los puntos etiquetados se extraen aleatoriamente de acuerdo con D. Redefinimos el error (o riesgo) verdadero de una regla de predicción h como:

$$L_{\mathcal{D}}(h) = \mathbb{P}_{(x,y)\sim\mathcal{D}}\left[h(x) \neq y\right] = \mathcal{D}\left(\{(x,y) : h(x) \neq y\}\right). \tag{2.1}$$

Nos gustaría minimizar este error, pero no conocemos los datos que generan $\mathcal{D}$. Pero si tenemos conocimiento de los datos de entrenamiento, S. De donde la definición de riesgo empírico sigue siendo:

$$L_S(h) = \frac{|\{i \in [m] : h(x_i) \neq y_i\}|}{m}.$$

Dado S, podemos calcular $L_S(h)$, para cualquier función $h : \mathcal{X} \to \{0,1\}$. Tenga en cuenta que $L_S(h) = L_{\mathcal{D}(uniforme\ sobre\ S)}(h)$.

El objetivo

El objetivo es encontrar alguna hipótesis, $h : \mathcal{X} \to \mathcal{Y}$ que probablemente minimice el riesgo real, $L_{\mathcal{D}}(h)$.

El predictor óptimo de Bayes

Dada cualquier distribución de probabilidad $\mathcal{D}$ sobre $\mathcal{X} \times \{0,1\}$, la mejor función de predicción de etiquetas de $\mathcal{X}$ a $\{0,1\}$ será:

$$f_{\mathcal{D}(x)} = \begin{cases} 1 & si \quad \mathbb{P}[y=1|x] \geq 1/2 \\ 0 & \text{en otro caso} \end{cases}$$

Teorema 2.1 **El predictor optimo de Bayes.** Demostrar que para cada distribución de probabilidad $\mathcal{D}$, el predictor óptimo de Bayes $f_{\mathcal{D}}$ es óptimo, en el sentido de que para cada clasificador g de $\mathcal{X}$ a $\{0,1\}$, $L_{\mathcal{D}}(f_{\mathcal{D}}) \leq L_{\mathcal{D}}(g)$.
Demostración.- Sean,

$$f_{\mathcal{D}(x)} = \begin{cases} 1 & si \quad \mathbb{P}[y=1|x] \geq 1/2 \\ 0 & \text{en otro caso} \end{cases}$$

la función de predicción de etiquetas y

$$L_S(h) = \frac{|\{i \in [m] : h(x_i) \neq y_i\}|}{m}.$$

El riesgo verdadero de $f_{\mathcal{D}}$.

Consideramos dos casos para cada punto x en el dominio de $\mathcal{X}$:

1. Si $\mathbb{P}[y=1|x] \geq 1/2$, entonces $f_{\mathcal{D}}(x) = 1$. Si $g(x) \neq f_{\mathcal{D}}(x)$, entonces $g(x) = 0$. Por lo tanto, g comete un error siempre que $y = 1$, lo cual ocurre con probabilidad de al menos $1/2$.
2. Si $\mathbb{P}[y=1|x] < 1/2$, entonces $f_{\mathcal{D}}(x) = 0$. Si $g(x) \neq f_{\mathcal{D}}(x)$, entonces $g(x) = 1$. Por lo tanto, g comete un error siempre que $y = 0$, lo cual ocurre con probabilidad mayor a $1/2$.

En ambos casos, el clasificador g tiene un riesgo mayor o igual al de $f_{\mathcal{D}}$. Dado que esto es cierto para cada punto x, se sigue que el riesgo total de g es al menos tan grande cómo el de $f_{\mathcal{D}}$, lo que muestra que $f_{\mathcal{D}} \leq L_{\mathcal{D}}(g)$. ■

Lo que nos dice el teorema es que ningún clasificador, g tal que $\mathcal{X} \to \{0,1\}$, tiene un error menor. Desafortunadamente cómo no conocemos $\mathcal{D}$, no podemos calcular $f_{\mathcal{D}}$.

Ahora podemos presentar la definición formal de capacidad de aprendizaje PAC agnóstico (más realista).

Más adelante requeriremos que el algoritmo de aprendizaje encuentre un predictor cuyo error nos sea mucho mayor que el mejor error posible de un predictor en alguna clase de hipótesis de referencia determinada, por su puesto la fuerza de tal requisito depende de la elección de esa clase de hipótesis.

Definición 2.2 **Capacidad de aprendizaje PAC agnóstico.** Una clase de hipótesis $\mathcal{H}$ es PAC agnóstica, si existe una función $m_{\mathcal{H}} : (0,1)^2 \to \mathbb{N}$ y un algoritmo de aprendizaje con la siguiente propiedad: Para cada, $\delta \in (0,1)$ y para cada distribución $\mathcal{D}$ sobre $\mathcal{X} \times \mathcal{Y}$, cuando se ejecuta el algoritmo de aprendizaje en $m \geq m_{\mathcal{H}}(\epsilon, \delta)$ i.i.d. ejemplos generados por $\mathcal{D}$, el algoritmo devuelve una hipótesis h tal que, con una probabilidad de al menos $1 - \delta$ (sobre la elección de los m ejemplos de entrenamiento),

$$Ł_{\mathcal{D}}(h) \leq \min_{h' \in \mathcal{H}} L_{\mathcal{D}}(h') + \epsilon.$$

El aprendizaje PAC agnóstico generaliza la definición de aprendizaje PAC, independiente de que si el supuesto de realizable se cumple o no. Según esta generalización podemos declarar éxito si su error no es mucho mayor que el mejor error alcanzable de un predictor de la clase $\mathcal{H}$.

2.2.1 El alcance de los problemas de aprendizaje modelados

A continuación ampliamos nuestro modelo para que pueda aplicarse a una amplia variedad de tareas de aprendizaje. Consideremos algunos ejemplos de diferentes tareas de aprendizaje.

- Clasificación multiclase.
- Regresión.- La medida de acierto en este caso es diferente. Podemos evaluar la calidad de una función de hipótesis, $h : \mathcal{X} \to \mathcal{Y}$, mediante *la diferencia cuadrática esperada* entre las etiquetas verdaderas y sus valores predichos. Es decir,

$$L_{\mathcal{D}}(h) = \mathbb{E}_{(x,y)\sim\mathcal{D}} \left[(h(x) - y)^2 \right]. \tag{2.2}$$

  Ahora, generalizamos nuestro formalismo de la medida de éxisto o acierto de la siguiente manera:

Funciones de pérdida generalizada

Dado cualquier conjunto $\mathcal{H}$ (que desempeña el papel de nuestras hipótesis o modelos) y algún dominio Z. Sea ℓ cualquier función desde $\mathcal{H} \times Z$ hasta el conjunto de números reales no negativo,

$$\ell : \mathcal{H} \times Z \to \mathbb{R}^+.$$

A estas funciones las llamamos funciones de pérdida. Z se generaliza para problemas de predicción, dominio de instancias o de etiquetas, etc.

Ahora, *definamos la función de riesgo cómo la pérdida esperada de un clasificador*, $h \in \mathcal{H}$, con respecto a una distribución de probabilidad $\mathcal{D}$ sobre Z. Es decir,

$$L_{\mathcal{D}}(h) = \mathbb{E}_{z\sim\mathcal{D}} \left[\ell(h, z) \right]. \tag{2.3}$$

Esto es, consideramos la expectativa de pérdida de h sobre los objetos de z elegidos aleatoriamente según $\mathcal{D}$. De manera similar, definimos *el riesgo empírico cómo la pérdida esperada sobre una muestra dada* $S = (z_1, \ldots, z_m) \in Z^m$. Es decir,

$$L_S(h) = \frac{1}{m} \sum_{i=1}^{m} \ell(h, z_i). \tag{2.4}$$

Las funciones de pérdida utilizadas en los ejemplos anteriores de tareas de clasificación y regresión son las siguientes:

- **Pérdida 0-1**: Aquí, nuestra variable aleatoria z abarca el conjunto de pares $\mathcal{X} \times \mathcal{Y}$ y la función de pérdida es:

$$l_{0-1}(h,(x,y)) = \begin{cases} 0 & si \quad h(x) = y \\ 1 & si \quad h(x) \neq y \end{cases}$$

  Esta función de pérdida se utiliza en problemas de clasificación binaria o en multiclase. Cabe señalar que, para una variable aleatoria, α tomando los valores $\{0,1\}$. $\mathbb{E}_{\alpha\sim\mathcal{D}}[\alpha] = \mathbb{P}_{\alpha\sim\mathcal{D}}[\alpha = 1]$. En consecuencia, para la función de pérdida, las definiciones de $L_{\mathcal{D}}(h)$ dadas en la ecuación (3.3) y (3.1) coincidan.

- **Pérdida cuadrática**: Aquí, nuestra variable aleatoria z abarca el conjunto de pares $\mathcal{X} \times \mathcal{Y}$ y la función de pérdida es:

$$\ell(h,(x,y)) = (h(x) - y)^2.$$

  Esta función de pérdida se utiliza en problemas de regresión.

Definimos la capacidad de aprendizaje PAC agnóstico para funciones de pérdida generalizadas.

Definición 2.3 **Capacidad de aprendizaje de PAC agnóstico para funciones de pérdida general.** Una clase de hipótesis $\mathcal{H}$ es aprendible de PAC agnóstico con respecto a un conjunto Z y una función de pérdida $\ell : \mathcal{H} \times Z \to \mathbb{R}^+$, si existe una función $m_{\mathcal{H}} : (0,1)^2 \to \mathbb{N}$ y un algoritmo de aprendizaje con la siguiente propiedad: Para cada , $\delta \in (0,1)$ y para cada distribución $\mathcal{D}$ sobre Z, cuando se ejecuta el algoritmo de aprendizaje en $m \geq m_{\mathcal{H}}(\epsilon,\delta)$ ejemplos i.i.d. generados por $\mathcal{D}$, el algoritmo devuelve $h \in \mathcal{H}$ tal que, con una probabilidad de al menos $1-\delta$ (sobre la elección de los m ejemplos de entrenamiento),

$$L_{\mathcal{D}}(h) \leq \min_{h' \in \mathcal{H}} L_{\mathcal{D}}(h') + \epsilon,$$

donde $L_{\mathcal{D}}(h) = \mathbb{E}_{z\sim\mathcal{D}}[\ell(h,z)]$.

Nota sobre Medibilidad: En la definición mencionada, para cada $h \in \mathcal{H}$, consideramos la función $\ell(h,\cdot) : Z \to \mathbb{R}^+$ como una variable aleatoria y definimos $L_{\mathcal{D}(h)}$ para ser el valor esperado de esta variable aleatoria. Para eso, necesitamos requerir que la función $\ell(h,\cdot)$ sea medible. Formalmente, asumimos que hay una σ-álgebra de subconjuntos de Z, sobre la cual se define la probabilidad $\mathcal{D}$, y que la preimagen de cada segmento inicial en $\mathbb{R}^+$ está en esta σ-álgebra. En el caso específico de la clasificación binaria con la pérdida $0-1$, la σ-álgebra está sobre $\mathcal{X} \times \{0,1\}$ y nuestra suposición sobre ℓ es equivalente a la suposición de que para cada h, el conjunto $\{(x,h(x)) : x \in \mathbb{X}\}$ está en la σ-álgebra.

La medibilidad garantiza que todas las operaciones que realizamos con variables aleatorias son consistentes con la estructura de probabilidad del espacio subyacente, permitiendo así que se definan y calculen las probabilidades y expectativas de manera adecuada. Esto es crucial en el aprendizaje automático y la clasificación binaria, donde queremos asegurarnos de que las predicciones y evaluaciones de los modelos sean compatibles con la teoría de la probabilidad.

NOTA: Explicación detallada de la definición de capacidad de aprendizaje PAC agnóstico para funciones de pérdida generalizadas. Explicaremos la definición de aprendizaje PAC agnóstico con funciones de pérdida generales, detallando cada elemento que aparece en ella.
La Definición 3.4 establece que una clase de hipótesis $\mathcal{H}$ es agnósticamente PAC aprendible con respecto a un conjunto Z y una función de pérdida $\ell : \mathcal{H} \times Z \to \mathbb{R}_+$ si existen ciertos componentes que cumplen con propiedades específicas:

1. **Conjunto Z**: Es el espacio de todas las posibles instancias o ejemplos sobre los que queremos hacer predicciones. Cada elemento de Z es un dato que puede ser evaluado por las hipótesis en $\mathcal{H}$.

2. **Función de pérdida ℓ**: Es una función que toma una hipótesis h y un ejemplo z, y devuelve un número real no negativo que representa el "costo" de predecir z usando h. La función de pérdida mide qué tan bien la hipótesis se desempeña en el ejemplo dado.

3. **Clase de hipótesis $\mathcal{H}$**: Es el conjunto de todas las hipótesis que el algoritmo de aprendizaje puede seleccionar. Una hipótesis es una función específica que intenta predecir o clasificar los ejemplos en Z.

4. **Función m_H**: Es una función que determina el tamaño mínimo de muestra necesario para aprender de manera efectiva. Depende de dos parámetros, ϵ y δ, y asigna a estos parámetros un número natural $\mathbb{N}$. La función $m_H(\epsilon, \delta)$ nos dice cuántos ejemplos i.i.d. necesitamos para que el algoritmo de aprendizaje funcione con una confianza y precisión deseadas.

5. **Parámetros ϵ (epsilon) y δ (delta)**: Son umbrales que definimos para el aprendizaje. ϵ es el margen de error que estamos dispuestos a tolerar en la predicción, y δ es la probabilidad máxima de que el algoritmo de aprendizaje exceda este margen de error.

6. **Distribución D sobre Z**: Es la distribución de probabilidad según la cual se generan los ejemplos. Representa cómo se espera que los datos se distribuyan en el mundo real.

7. **Riesgo esperado $L_D(h)$**: Es el riesgo promedio o esperado de una hipótesis h cuando se evalúa con la función de pérdida ℓ sobre la distribución D. Se calcula como la expectativa de la función de pérdida $\ell(h, z)$ para un ejemplo z tomado de la distribución D.

Imagina que tienes un conjunto de herramientas, que son tus hipótesis en $\mathcal{H}$. Entre todas ellas, hay una que es la mejor, es decir, la que comete menos errores al predecir o clasificar los datos. Ahora bien, cuando usas un algoritmo de aprendizaje automático, estás tratando de encontrar una herramienta (hipótesis) que sea casi tan buena como la mejor que tienes, pero sin necesidad de probarlas todas.
La definición dice que si eliges suficientes ejemplos (datos) de acuerdo con la función $m_{\mathcal{H}}(\epsilon, \delta)$, entonces el algoritmo de aprendizaje te dará una hipótesis h que no será mucho peor que la mejor de tu conjunto. En términos más precisos, el "peor" se cuantifica con ϵ, que es una pequeña cantidad de error adicional que estás dispuesto a aceptar. Y la "confianza" se refiere a la probabilidad de que esto sea cierto, que queremos que sea muy alta, al menos $1 - \delta$.
Por lo tanto, la definición asegura que, con una alta probabilidad, el algoritmo encontrará una hipótesis cuyo rendimiento (en términos de error) no exceda el de la mejor hipótesis por más de un pequeño margen ϵ, siempre y cuando haya suficientes datos para aprender de ellos. Esto es lo que significa ser agnósticamente PAC aprendible con funciones de pérdida generales. Es una forma de decir que podemos aprender bien, incluso si no podemos ser perfectos.

3

Aprendizaje a través de la convergencia uniforme

En este capítulo desarrollaremos una herramienta general, la convergencia uniforme, y la aplicaremos para demostrar que cualquier clase finita se puede aprender en el modelo PAC agnóstico con funciones de pérdida generales, siempre que la función de pérdida de rango esté acotada.

3.1 La convergencia uniforme es suficiente para la capacidad de aprendizaje

La idea detrás de la condición de aprendizaje analizada en este capítulo es muy simple. Recuerde que, dada una clase de hipótesis, $\mathcal{H}$, el paradigma de aprendizaje ERM funciona de la siguiente manera: al recibir una muestra de entrenamiento, S, el alumno evalúa el riesgo (o error) de cada h en $\mathcal{H}$ en la muestra dada y genera un miembro de $\mathcal{H}$ que minimiza este riesgo empírico. La esperanza es que una h que minimiza el riesgo empírico con respecto a S sea un minimizador de riesgo (o tenga un riesgo cercano al mínimo) con respecto a la distribución de probabilidad de los datos verdaderos también. Para ello, basta con asegurar que los riesgos empíricos de todos los miembros de $\mathcal{H}$ sean buenas aproximaciones de su riesgo real. Dicho de otra manera, necesitamos que, de manera uniforme, en todas las hipótesis de la clase de hipótesis, el riesgo empírico sea cercano al riesgo verdadero, como se formaliza a continuación.

Definición 3.1 **ϵ-muestras representativas.** Un conjunto de entrenamiento S se llama ϵ-representativo (con respecto a el dominio Z, la clase de hipótesis $\mathcal{H}$, la función de pérdida ℓ y la distribución $\mathcal{D}$) si

$$\forall h \in \mathcal{H}, |L_S(h) - L_D(h)| \leq \epsilon.$$

El siguiente lema simple establece que siempre que la muestra sea $(\epsilon/2)$-representativa, se garantiza que la regla de aprendizaje de ERM devolverá una buena hipótesis.

Lema 3.1 Supongamos que un conjunto de entrenamiento S es $\frac{\epsilon}{2}$-representativo (con respecto a el dominio Z, clase de hipótesis $\mathcal{H}$, función de pérdida ℓ y distribución $\mathcal{D}$). Entonces, cualquier salida de $\text{ERM}_{\mathcal{H}}(S)$. Es decir, cualquier $h_S \in \text{argmin}_{h \in \mathcal{H}} L_S(h)$, satisface

$$L_D(h_S) \leq \min_{h \in \mathcal{H}} L_D(h) + \epsilon.$$

Demostración.- Para cada $h \in \mathcal{H}$,

$$L_D(h_S) \leq L_S(h_S) + \frac{\epsilon}{2} \leq L_S(h) + \frac{\epsilon}{2} \leq L_D(h) + \frac{\epsilon}{2} + \frac{\epsilon}{2} = L_D(h) + \epsilon.$$

Donde la primera y tercera desigualdad se deben al supuesto de que S es $\frac{\epsilon}{2}$-representativo (Definición 4.1) y la segunda desigualdad se cumple ya que h_S es un predictor del ERM. ■

El lema anterior implica que para garantizar que la regla ERM sea un aprendizaje PAC agnóstico, basta con demostrar que con una probabilidad de al menos $1-\delta$ sobre la elección aleatoria de un conjunto de entrenamiento, será un conjunto de entrenamiento representativo. La condición de convergencia uniforme formaliza este requisito.

Definición 3.2 **Convergencia uniforme.** Decimos que una clase de hipótesis $\mathcal{H}$ tiene la propiedad de convergencia uniforme (con respecto a el dominio de Z, y una función de pérdida ℓ) si existe una función $m^{UC}_{\mathcal{H}} : (0,1)^2 \to \mathbb{N}$ tal que para cada $\delta \in (0,1)$ y para cada distribución de probabilidad $\mathcal{D}$ sobre Z, si S es una muestra $m \geq m^{UC}_{\mathcal{H}}(\epsilon, \delta)$ ejemplos i.i.d. según $\mathcal{D}$. Entonces, con una probabilidad de al menos $1-\delta$, S es representativo.

La función $m^{UC}_{\mathcal{H}}$ modela la complejidad mínima de la muestra para obtener la propiedad de convergencia uniforme. Es decir, cuántos ejemplos necesitamos para asegurarnos de que con una probabilidad de al menos $1-\delta$ la muestra es ϵ-representativa. El término uniforme aquí se refiere a tener un tamaño de muestra fijo que funcione para todos los miembros de $\mathcal{H}$ y para todas las distribuciones de probabilidad posibles en el dominio. El siguiente corolario se deriva directamente del Lema 4.2 y de la definición de convergencia uniforme.

Corolario 3.1 Si una clase $\mathcal{H}$ tiene la propiedad de convergencia uniforme con un función $m^{UC}_{\mathcal{H}}$. Entonces, la clase es PAC agnóstico aprendible con la complejidad de muestra $m^{UC}_{\mathcal{H}}(\epsilon/2, \delta)$. Además en ese caso, el paradigma $\text{ERM}_{\mathcal{H}}$ es un aprendizaje exitoso de PAC agnóstico para $\mathcal{H}$. ■

3.2 Las clases finitas son agnósticas y se pueden aprender en PAC

En vista del corolario 4.1, la afirmación de que cada clase de hipótesis finita es agnóstica y se puede aprender en PAC se seguirá una vez que establezcamos que la convergencia uniforme es válida para una clase de hipótesis finita.

Para demostrar que se cumple la convergencia uniforme, seguimos un argumento de dos pasos, similar a la derivación del Capítulo 2. El primer paso aplica la cota de unión mientras que el segundo emplea una medida de desigualdad de concentración. Ahora explicamos estos dos pasos en detalle. Arreglar algunos, δ. Necesitamos encontrar un tamaño de muestra m que garantice que para cualquier $\mathcal{D}$, con una probabilidad de al menos $1-\delta$ de la elección de $S = (z_1, \ldots, z_m)$ i.d.d. muestreados de $\mathcal{D}$, tenemos que para todo $h \in \mathcal{H}$, $|L_S(h) - L_D(h)| \leq \epsilon$. Esto es,

$$D^m\left(\{S : \forall h \in \mathcal{H}, |L_S(h) - L_D(h)| \leq \epsilon\}\right) \geq 1-\delta.$$

Equivalentemente, necesitamos mostrar que

$$D^m\left(\{S : \exists h \in \mathcal{H}, |L_S(h) - L_D(h)| > \epsilon\}\right) < \delta.$$

Esto es,

$$\{S : \exists h \in \mathcal{H}, |L_S(h) - L_D(h)| > \epsilon\} = \bigcup_{h \in \mathcal{H}} \{S : |L_S(h) - L_D(h)| > \epsilon\}.$$

Y aplicando la cota de unión (Lema 2.1), obtenemos

$$D^m\left(S : \exists h \in \mathcal{H}, |L_S(h) - L_D(h)| > \epsilon\right) \leq \sum_{h \in \mathcal{H}} D^m\left(S : |L_S(h) - L_D(h)| > \epsilon\right). \tag{3.1}$$

Nuestro segundo paso será argumentar que cada suma del lado derecho de esta desigualdad es lo suficientemente pequeña (para una m suficientemente grande). Es decir, mostraremos que para cualquier hipótesis fija, h, (que se elige de antemano antes del muestreo del conjunto de entrenamiento), la brecha entre los riesgos verdaderos y empíricos, $|L_S(h) - L_D(h)|$, es probable que sea pequeño.
Recuerde que $L_D(h) = \mathbb{E}_{z\sim D}[\ell(h,z)]$ y que $L_S(h) = \frac{1}{m}\sum_{i=1}^{m}\ell(h,z_i)$. Dado que cada z_i se muestrea i.i.d. de D, el valor esperado de la variable aleatoria '(h,z_i) es $L_D(h)$. Por la linealidad de la expectativa, se deduce que $L_D(h)$ es también el valor esperado de $L_S(h)$. Por tanto, la cantidad $|L_D(h) - L_S(h)|$ es la desviación de la variable aleatoria $L_S(h)$ de su expectativa. Por lo tanto, necesitamos demostrar que la medida de $L_S(h)$ se concentra alrededor de su valor esperado. Un hecho estadístico básico, la ley de los grandes números, establece que cuando m llega al infinito, los promedios empíricos convergen a su verdadera expectativa. Esto es cierto para $L_S(h)$, ya que es el promedio empírico de variables aleatorias m i.i.d. Sin embargo, dado que la ley de los grandes números es sólo un resultado asintótico, no proporciona información sobre la brecha entre el error estimado empíricamente y su valor verdadero para cualquier tamaño de muestra finito dado. En su lugar, utilizaremos una medida de desigualdad de concentración debida a Hoeffding, que cuantifica la brecha entre los promedios empíricos y su valor esperado.

Nota: Explicación de clase finita agnósticamente PAC aprendible Para comprender la afirmación de que cada clase de hipótesis finita es agnósticamente PAC aprendible, necesitamos establecer que la convergencia uniforme se mantiene para una clase de hipótesis finita. Esto se hace siguiendo un argumento de dos pasos que involucra la aplicación de la cota de unión y el uso de una desigualdad de concentración de medida.

Comenzamos fijando dos valores, ϵ y δ. Estos valores representan, respectivamente, el margen de error que estamos dispuestos a tolerar y la probabilidad máxima de que este margen sea superado. Nuestro objetivo es encontrar un tamaño de muestra m que garantice que, para cualquier distribución D, con una probabilidad de al menos $1-\delta$, el conjunto de muestras $S = (z_1, ..., z_m)$ tomadas de manera i.i.d. de D cumpla que para todas las hipótesis h en nuestra clase de hipótesis $\mathcal{H}$, la diferencia entre el riesgo empírico $L_S(h)$ y el riesgo verdadero $L_D(h)$ sea menor o igual a ϵ.
Matemáticamente, queremos que:

$$D^m(\{S : \forall h \in H, |L_S(h) - L_D(h)| \leq \epsilon\}) \geq 1 - \delta.$$

De manera equivalente, necesitamos demostrar que la probabilidad de que la diferencia entre el riesgo empírico y el verdadero sea mayor que ϵ para alguna hipótesis h es menor que δ:

$$D^m(\{S : \exists h \in H, |L_S(h) - L_D(h)| > \epsilon\}) < \delta.$$

Para abordar esta probabilidad sobre todas las hipótesis posibles, aplicamos la cota de unión, que nos permite sumar las probabilidades individuales de eventos para cada hipótesis. Esto se representa como la unión de todos los conjuntos de muestras para los cuales la diferencia de riesgo es mayor que ϵ para alguna hipótesis:

$$\{S : \exists h \in H, |L_S(h) - L_D(h)| > \epsilon\} = \bigcup_{h\in H}\{S : |L_S(h) - L_D(h)| > \epsilon\},$$

y aplicando la cota de unión obtenemos:

$$D^m(\{S : \exists h \in H, |L_S(h) - L_D(h)| > \epsilon\}) \leq \sum_{h\in H} D^m(\{S : |L_S(h) - L_D(h)| > \epsilon\}).$$

El segundo paso es argumentar que cada sumando del lado derecho de esta desigualdad es lo suficientemente pequeño para un m suficientemente grande. Esto significa que para cualquier hipótesis fija h, la diferencia entre el riesgo verdadero y el empírico, $|L_S(h) - L_D(h)|$, probablemente será pequeña.

Recordamos que el riesgo verdadero $L_D(h)$ es el valor esperado de la pérdida sobre la distribución D, y el riesgo empírico $L_S(h)$ es el promedio de la pérdida sobre las muestras. Dado que cada muestra z_i

se toma i.i.d. de D, el valor esperado de la pérdida para una muestra individual es igual al riesgo verdadero. Por lo tanto, la cantidad $|L_D(h) - L_S(h)|$ es la desviación de la variable aleatoria $L_S(h)$ de su expectativa.

Para demostrar que la medida de $L_S(h)$ está concentrada alrededor de su valor esperado, utilizamos la desigualdad de Hoeffding, que cuantifica la brecha entre los promedios empíricos y su valor esperado. Aunque la ley de los grandes números nos dice que los promedios empíricos convergen a su verdadera expectativa cuando el número de muestras tiende a infinito, la desigualdad de Hoeffding nos proporciona una garantía cuantitativa para tamaños de muestra finitos.

En resumen, estos dos pasos demuestran que para clases de hipótesis finitas, podemos garantizar que con una muestra suficientemente grande, la diferencia entre el riesgo empírico y el verdadero será pequeña con alta probabilidad, lo que satisface la condición de convergencia uniforme y, por lo tanto, la clase de hipótesis es agnósticamente PAC aprendible. Esto es fundamental para la teoría del aprendizaje automático, ya que proporciona una base para comprender y confiar en los algoritmos de aprendizaje automático que se utilizan en la práctica.

Para entender la desigualdad de Hoeffding, comencemos con una desigualdad que se llama la desigualdad de Markov. Sea Z una variable aleatoria no negativa. La esperanza de Z se puede escribir de la siguiente manera:

$$E[Z] = \int_0^\infty P[Z \geq x]\, dx \quad (B.1)$$

Dado que

$$P[Z \geq x]$$

es monótonamente no creciente, obtenemos

$$\forall a \geq 0,\ E[Z] \geq \int_0^a P[Z \geq x]\, dx \geq \int_0^a P[Z \geq a]\, dx = aP[Z \geq a]. \quad (B.2)$$

Reorganizando la desigualdad obtenemos la desigualdad de Markov:

$$\forall a \geq 0,\ P[Z \geq a] \leq \frac{E[Z]}{a} \quad (B.3)$$

Para variables aleatorias que toman valores en $[0, 1]$, podemos derivar de la desigualdad de Markov lo siguiente.

Teorema 3.1 Sea Z una variable aleatoria que toma valores en $[0, 1]$. Supongamos que $E[Z] = \mu$. Entonces, para cualquier $a \in (0, 1)$,

$$P[Z > 1 - a] \geq \frac{\mu - (1 - a)}{a}$$

Esto también implica que para cada $a \in (0, 1)$,

$$P[Z > a] \geq \frac{\mu - a}{1 - a} \geq \mu - a.$$

Demostración.- Sea

$$Y = 1 - Z.$$

Entonces, Y es una variable aleatoria no negativa con

$$E[Y] = 1 - E[Z] = 1 - \mu.$$

Aplicando la desigualdad de Markov en Y obtenemos

$$P[Z \leq 1 - a] = P[1 - Z \geq a] = P[Y \geq a] \leq \frac{E[Y]}{a} = \frac{1 - \mu}{a}$$

Por lo tanto,

$$P[Z > 1 - a] \geq 1 - \frac{1-\mu}{a} = \frac{a+\mu-1}{a}$$

■

Teorema 3.2 **Teorema de Hoeffding.** Sea X una variable aleatoria que toma valores en el intervalo $[a, b]$ y tal que $E[X] = 0$. Entonces, para todo $\lambda > 0$,

$$E[e^{\lambda X}] \leq e^{\frac{\lambda^2 (b-a)^2}{8}}.$$

Demostración.- Dado que $f(x) = e^{\lambda x}$ es una función convexa, tenemos que para todo $\alpha \in (0,1)$, y $x \in [a, b]$,

$$f(x) \leq \alpha f(a) + (1-\alpha) f(b).$$

Estableciendo $\alpha = \frac{b-x}{b-a} \in [0,1]$ obtenemos

$$e^{\lambda x} \leq \frac{b-x}{b-a} e^{\lambda a} + \frac{x-a}{b-a} e^{\lambda b}.$$

Tomando la esperanza, obtenemos que

$$E[e^{\lambda X}] \leq \frac{b - E[X]}{b-a} e^{\lambda a} + \frac{E[x]-a}{b-a} e^{\lambda b} = \frac{b}{b-a} e^{\lambda a} - \frac{a}{b-a} e^{\lambda b},$$

donde usamos el hecho de que $E[X] = 0$. Denotemos $h = \lambda(b-a)$, $p = \frac{-a}{b-a}$, y

$$L(h) = -hp + \log(1 - p + pe^h).$$

Entonces, la expresión en el lado derecho de lo anterior se puede reescribir como $e^{L(h)}$. Por lo tanto, para concluir nuestra prueba basta con mostrar que $L(h) \leq \frac{h^2}{8}$. Esto se sigue del teorema de Taylor utilizando los hechos:

$$L(0) = L'(0) = 0 \quad \text{y} \quad L''(h) \leq \frac{1}{4} \quad \text{para todo} \quad h.$$

■

Lema 3.2 **Desigualdad de Hoeffding.** Sea $\theta_1, \ldots, \theta_m$ una secuencia de variables eleatorias i.i.d. y suponiendo que para todo $i, \mathbb{E}[\theta_i] = \mu$ y $\mathbb{P}[a \leq \theta_i \leq b] = 1$. Entonces, para cualquier $\epsilon > 0$,

$$\mathbb{P}\left[\left|\frac{1}{m}\sum_{i=1}^{m} \theta_i - \mu\right| > \epsilon\right] \leq 2e^{-\frac{2m\epsilon^2}{(b-a)^2}}.$$

Demostración.- Denotemos a $X_i = Z_i - \mathbb{E}[Z_i]$, y $\overline{X} = \frac{1}{n}\sum_i X_i$. Usando la monotocidad de la función exponente y la desigualdad de Markov, se tiene que para cada $\lambda > 0$ y $\epsilon > 0$,

$$\mathbb{P}\left[\overline{X} \geq \epsilon\right] = \mathbb{P}\left[e^{\lambda \overline{X}} \geq e^{\lambda \epsilon}\right] \leq e^{-\lambda\epsilon}\, \mathbb{E}[e^{\lambda \overline{X}}].$$

Usando el supuesto de independencia también tenemos

$$\mathbb{E}[e^{\lambda \overline{X}}] = \mathbb{E}\left[\prod_i e^{\lambda X_i/m}\right] = \prod_i \mathbb{E}[e^{\lambda X_i/m}].$$

Según el teorema de Hoeffding , por cada i tenemos

$$\mathbb{E}[e^{\lambda X_i/m}] \leq e^{\frac{\lambda^2 (b-a)^2}{8m^2}}.$$

Por lo tanto,

$$\mathbb{P}\left[\overline{X} \geq \epsilon\right] \leq e^{-\lambda\epsilon} \prod_i e^{\frac{\lambda^2(b-a)^2}{8m^2}} = e^{-\lambda\epsilon + \frac{\lambda^2(b-a)^2}{8m^2}}.$$

Haciendo $\lambda = 4m\epsilon/(b-a)^2$ se obtiene

$$\mathbb{P}\left[\overline{X} \geq \epsilon\right] \leq e^{\frac{-2m\epsilon^2}{(b-a)^2}}.$$

Aplicando los mismos argumentos a la variable $-\overline{X}$ obtenemos que $\mathbb{P}[\overline{X} \leq -\epsilon] \leq e^{-\frac{2ms^2}{(b-a)^2}}$. El teorema se sigue aplicando la unión ligada en los dos casos. ■

Volviendo a nuestro problema, sea θ_i la variable aleatoria (h, z_i). Dado que h es fijo y $z_1, \ldots, z_m$ son muestreados de manera i.i.d., se sigue que $\theta_1, \ldots, \theta_m$ también son variables aleatorias i.i.d. Además, $LS(h) = \frac{1}{m}\sum_{i=1}^m \theta_i$ y $LD(h) = \mu$. Supongamos además que el rango de ℓ es $[0,1]$ y por lo tanto $\theta_i \in [0,1]$. Por lo tanto, obtenemos que

$$\mathbb{D}_m(\{S : |LS(h) - LD(h)| > \epsilon\}) = \mathbb{P}\left(\left|\frac{1}{m}\sum_{i=1}^m \theta_i - \mu\right| > \epsilon\right) \leq 2\exp\left(-2m\epsilon^2\right). \tag{3.2}$$

Combinando esto con la Ecuación (4.1) se obtiene

$$\mathbb{D}_m(\{S : \exists h \in H, |LS(h) - LD(h)| > \epsilon\}) \leq \sum_{h\in H} 2\exp\left(-2m\epsilon^2\right) = 2|H|\exp\left(-2m\epsilon^2\right).$$

Finalmente, si elegimos

$$m \geq \frac{\log(2|H|/\delta)}{2\epsilon^2}$$

entonces

$$\mathbb{D}_m(\{S : \exists h \in H, |LS(h) - LD(h)| > \epsilon\}) \leq \delta.$$

Corolario 3.2 Sea $\mathcal{H}$ una clase de hipótesis finita, sea Z un dominio, y sea $\ell : \mathcal{H} \times Z \to [0,1]$ una función de pérdida. Entonces, $\mathcal{H}$ disfruta de la propiedad de convergencia uniforme con complejidad de muestra

$$m^H_{UC}(\epsilon, \delta) \leq \left\lceil \frac{\log(2|H|/\delta)}{2\epsilon^2} \right\rceil.$$

Además, la clase es agnósticamente PAC aprendible usando el algoritmo ERM con complejidad de muestra

$$m^H(\epsilon, \delta) \leq m^H_{UC}(\epsilon/2, \delta) \leq \left\lceil \frac{2\log(2|H|/\delta)}{\epsilon^2} \right\rceil.$$

■

NOTA: El "Truco de Discretización" Mientras que el corolario precedente solo se aplica a clases de hipótesis finitas, hay un truco simple que nos permite obtener una muy buena estimación de la complejidad de muestra práctica de clases de hipótesis infinitas. Considere una clase de hipótesis que está parametrizada por d parámetros. Por ejemplo, sea $X = \mathbb{R}$, $Y = \{\pm 1\}$, y la clase de hipótesis, $\mathcal{H}$, sea todas las funciones de la forma

$$h_\theta(x) = \text{sign}(x - \theta)$$

. Es decir, cada hipótesis está parametrizada por un parámetro, $\theta \in \mathbb{R}$, y la hipótesis da como resultado 1 para todas las instancias mayores que θ y da como resultado -1 para instancias menores que θ. Esta es una clase de hipótesis de tamaño infinito. Sin embargo, si vamos a aprender esta clase de hipótesis en la práctica, usando una computadora, probablemente mantendremos números reales usando la representación de punto flotante, digamos, de 64 bits. Se sigue que en la práctica, nuestra clase de hipótesis

está parametrizada por el conjunto de escalares que pueden ser representados usando un número de punto flotante de 64 bits. Hay a lo sumo 2^{64} tales números; por lo tanto, el tamaño real de nuestra clase de hipótesis es a lo sumo 2^{64}. Más generalmente, si nuestra clase de hipótesis está parametrizada por d números, en la práctica aprendemos una clase de hipótesis de tamaño a lo sumo 2^{64d}. Aplicando el Corolario 4.2 obtenemos que la complejidad de muestra de s clases está acotada por

$$\frac{128d + 2\log(2/\delta)}{\epsilon^2}.$$

Este límite superior en la complejidad de muestra tiene la deficiencia de depender de la representación específica de números reales utilizada por nuestra máquina. En el Capítulo 6 introduciremos una manera rigurosa de analizar la complejidad de muestra de clases de hipótesis de tamaño infinito. No obstante, el truco de discretización puede ser utilizado para obtener una estimación aproximada de la complejidad de muestra en muchas situaciones prácticas.

4

El intercambio entre sesgo y complejidad

En el Capítulo 2 vimos que a menos que se tenga cuidado, los datos de entrenamiento pueden engañar al aprendiz y resultar en **sobreajuste**. Para superar este problema, restringimos el espacio de búsqueda a alguna clase de hipótesis $\mathcal{H}$. Tal clase de hipótesis puede verse como reflejo de algún conocimiento previo que el aprendiz tiene sobre la tarea: una creencia de que uno de los miembros de la clase $\mathcal{H}$ es un modelo de bajo error para la tarea. Por ejemplo, en nuestro problema del sabor de las papayas, basándonos en nuestra experiencia previa con otras frutas, podríamos asumir que algún rectángulo en el plano color-dureza predice (al menos aproximadamente) la deliciosidad de la papaya.

¿Es realmente necesario tal conocimiento previo para el éxito del aprendizaje? Quizás exista algún tipo de aprendiz universal, es decir, un aprendiz que no tiene conocimiento previo sobre una tarea determinada y está listo para ser desafiado por cualquier tarea. Profundicemos en este punto. Una tarea de aprendizaje específica está definida por una distribución desconocida D sobre $\mathcal{X} \times \mathcal{Y}$, donde el objetivo del aprendiz es encontrar un predictor $h : \mathcal{X} \to \mathcal{Y}$, cuyo riesgo, $L_D(h)$, sea lo suficientemente pequeño. La pregunta es, por lo tanto, si existe un algoritmo de aprendizaje A y un tamaño de conjunto de entrenamiento m, tal que para cada distribución D, si A recibe m ejemplos i.i.d. de D, hay una alta probabilidad de que produzca un predictor h que tenga un bajo riesgo.

La primera parte de este capítulo aborda esta pregunta formalmente. El teorema de No-Free-Lunch afirma que no existe tal aprendiz universal. Para ser más precisos, *el teorema establece que para tareas de predicción de clasificación binaria, para cada aprendiz existe una distribución en la que falla*. Decimos que el aprendiz falla si, al recibir ejemplos i.i.d. de esa distribución, su hipótesis de salida probablemente tenga un gran riesgo, digamos, ≥ 0.3, mientras que para la misma distribución, existe otro aprendiz que producirá una hipótesis con un pequeño riesgo. En otras palabras, **el teorema afirma que ningún aprendiz puede tener éxito en todas las tareas aprendibles: cada aprendiz tiene tareas en las que falla mientras otros aprendices tienen éxito.**

Por lo tanto, **al abordar un problema de aprendizaje particular, definido por alguna distribución D, deberíamos tener algún conocimiento previo sobre D. Un tipo de tal conocimiento previo es que D proviene de alguna familia paramétrica específica de distribuciones**. Estudiaremos el aprendizaje bajo tales suposiciones más adelante en el Capítulo 24. **Otro tipo de conocimiento previo sobre D, que asumimos al definir el modelo de aprendizaje PAC, es que existe h en alguna clase de hipótesis predefinida $\mathcal{H}$, tal que $L_D(h) = 0$. Un tipo más suave de conocimiento previo sobre D es asumir que $\min_{h \in \mathcal{H}} L_D(h)$ es pequeño. En cierto sentido, esta suposición más débil sobre D es un prerrequisito para usar el modelo PAC agnóstico, en el cual requerimos que el riesgo de la hipótesis de salida no sea mucho mayor que $\min_{h \in \mathcal{H}} L_D(h)$.**

En la segunda parte de este capítulo estudiamos los beneficios y las trampas de usar una clase de hipótesis como medio para formalizar el conocimiento previo. Descomponemos el error de un algoritmo ERM sobre una clase $\mathcal{H}$ en dos componentes. El primer componente refleja la calidad de nuestro conocimiento previo, medido por el riesgo mínimo de una hipótesis en nuestra clase de hipótesis, $\min_{h \in \mathcal{H}} L_D(h)$. Este componente también se llama el **error de aproximación, o el sesgo del algoritmo hacia la elección de**

una hipótesis de $\mathcal{H}$. El segundo componente es el **error debido al sobreajuste, que depende del tamaño o la complejidad de la clase $\mathcal{H}$ y se llama el error de estimación**. Estos dos términos **implican un equilibrio entre elegir una $\mathcal{H}$ más compleja (que puede disminuir el sesgo pero aumenta el riesgo de sobreajuste) o una $\mathcal{H}$ menos compleja (que podría aumentar el sesgo pero disminuye el potencial de sobreajuste).**

4.1 Teorema de No-Free-Lunch

En esta parte demostramos que no existe un aprendiz universal. Lo hacemos mostrando que ningún aprendiz puede tener éxito en todas las tareas de aprendizaje, como se formaliza en el siguiente teorema:

Teorema 4.1 **No-Free-Lunch.** Sea A cualquier algoritmo de aprendizaje para la tarea de clasificación binaria con respecto a la pérdida $0-1$ sobre un dominio $\mathcal{X}$. Sea m cualquier número menor que $|\mathcal{X}|/2$, representando un tamaño de conjunto de entrenamiento. Entonces, existe una distribución D sobre $\mathcal{X} \times \{0,1\}$ tal que:

i) Existe una función $f : \mathcal{X} \to \{0,1\}$ con $L_D(f) = 0$.

ii) Con probabilidad de al menos $1/7$ sobre la elección de $S \sim D^m$ tenemos que $L_D(A(S)) \geq 1/8$.

Este teorema afirma que para cada aprendiz, existe una tarea en la que falla, aunque esa tarea pueda ser aprendida con éxito por otro aprendiz. De hecho, un aprendiz trivialmente exitoso en este caso sería un aprendiz ERM con la clase de hipótesis $\mathcal{H} = \{f\}$, o más generalmente, ERM con respecto a cualquier clase de hipótesis finita que contenga f y cuyo tamaño satisfaga la ecuación $m \geq 8\log(7|\mathcal{H}|/6)$ (ver Corolario 2.3).

Prueba Sea C un subconjunto de X de tamaño $2m$. La intuición de la prueba es que cualquier algoritmo de aprendizaje que solo observe la mitad de las instancias en C no tiene información sobre cuáles deberían ser las etiquetas del resto de las instancias en C. Por lo tanto, existe una "realidad", es decir, alguna función objetivo f, que contradiría las etiquetas que $A(S)$ predice sobre las instancias no observadas en C.
Note que hay $T = 2^{2m}$ posibles funciones de C a $\{0,1\}$. Denotemos estas funciones por $f_1, \ldots, f_T$. Para cada una de estas funciones, sea D_i una distribución sobre $C \times \{0,1\}$ definida por

$$D_i(\{(x,y)\}) = \begin{cases} \frac{1}{|C|} & \text{si } y = f_i(x) \\ 0 & \text{de lo contrario.} \end{cases}$$

Es decir, la probabilidad de elegir un par (x,y) es $\frac{1}{|C|}$ si la etiqueta y es efectivamente la etiqueta verdadera según f_i, y la probabilidad es 0 si $y \neq f_i(x)$. Claramente, $L_{D_i}(f_i) = 0$.
Mostraremos que para cada algoritmo A, que recibe un conjunto de entrenamiento de m ejemplos de $C \times \{0,1\}$ y devuelve una función $A(S) : C \to \{0,1\}$, se cumple que

$$\max_{i \in [T]} \mathop{\mathbb{E}}_{S \sim D_i^m} [L_{D_i}(A(S))] \geq \frac{1}{4}. \tag{4.1}$$

Claramente, esto significa que para cada algoritmo A', que recibe un conjunto de entrenamiento de m ejemplos de $X \times \{0,1\}$ existe una función $f : X \to \{0,1\}$ y una distribución D sobre $X \times \{0,1\}$, tal que $L_D(f) = 0$ y

$$\mathop{\mathbb{E}}_{S \sim D^m} [L_D(A'(S))] \geq \frac{1}{4}. \tag{4.2}$$

Es fácil verificar que lo anterior es suficiente para mostrar que $\mathbb{P}[L_D(A'(S)) \geq \frac{1}{8}] \geq \frac{1}{7}$, que es lo que necesitamos probar (ver Ejercicio 1).

Ahora pasamos a demostrar que la Ecuación (5.1) se cumple. Hay $k = (2m)^m$ secuencias posibles de m ejemplos de C. Denotemos estas secuencias por $S_1, \ldots, S_k$. Además, si $S_j = (x_1, \ldots, x_m)$ denotamos por S_i^j la secuencia que contiene las instancias en S_j etiquetadas por la función f_i, es decir, $S_i^j = ((x_1, f_i(x_1)), \ldots, (x_m, f_i(x_m)))$. Si la distribución es D_i entonces los posibles conjuntos de entrenamiento que A puede recibir son $S_i^1, \ldots, S_i^k$, y todos estos conjuntos de entrenamiento tienen la misma probabilidad de ser muestreados. Por lo tanto,

$$\mathop{\mathbb{E}}_{S \sim D_i^m}[L_{D_i}(A(S))] = \frac{1}{k} \sum_{j=1}^{k} L_{D_i}(A(S_i^j)). \tag{4.3}$$

Usando los hechos de que "máximo" es mayor que "promedio" y que "promedio" es mayor que "mínimo", tenemos

$$\begin{aligned} \max_{i \in [T]} \frac{1}{k} \sum_{j=1}^{k} L_{D_i}(A(S_i^j)) &\geq \frac{1}{T} \sum_{i=1}^{T} \frac{1}{k} \sum_{j=1}^{k} L_{D_i}(A(S_i^j)) \\ &= \frac{1}{k} \sum_{j=1}^{k} \frac{1}{T} \sum_{i=1}^{T} L_{D_i}(A(S_i^j)) \\ &\geq \min_{j \in [k]} \frac{1}{T} \sum_{i=1}^{T} L_{D_i}(A(S_i^j)). \end{aligned} \tag{4.4}$$

A continuación, fijamos algún $j \in [k]$. Denotemos $S_j = (x_1, \ldots, x_m)$ y dejemos que $v_1, \ldots, v_p$ sean los ejemplos en C que no aparecen en S_j. Claramente, $p \geq m$. Por lo tanto, para cada función $h : C \to \{0, 1\}$ y cada i tenemos

$$\begin{aligned} L_{D_i}(h) &= \frac{1}{2m} \sum_{x \in C} 1[h(x) \neq f_i(x)] \\ &\geq \tfrac{1}{2m} \textstyle\sum_{r=1}^{p} 1[h(v_r) \neq f_i(v_r)] \\ &\geq \tfrac{1}{2p} \textstyle\sum_{r=1}^{p} 1[h(v_r) \neq f_i(v_r)]. \end{aligned} \tag{4.5}$$

Por lo tanto,

$$\begin{aligned} \tfrac{1}{T} \textstyle\sum_{i=1}^{T} L_{D_i}(A(S_i^j)) &\geq \frac{1}{T} \sum_{i=1}^{T} \frac{1}{2p} \sum_{r=1}^{p} 1[A(S_i^j)(v_r) \neq f_i(v_r)] \\ &= \frac{1}{2p} \sum_{r=1}^{p} \frac{1}{T} \sum_{i=1}^{T} 1[A(S_i^j)(v_r) \neq f_i(v_r)] \\ &\geq \frac{1}{2} \cdot \min_{r \in [p]} \frac{1}{T} \sum_{i=1}^{T} 1[A(S_i^j)(v_r) \neq f_i(v_r)]. \end{aligned} \tag{4.6}$$

A continuación, fijamos algún $r \in [p]$. Podemos particionar todas las funciones en $f_1, \ldots, f_T$ en $\frac{T}{2}$ pares disjuntos, donde para un par $(f_i, f_{i'})$ tenemos que para cada $c \in C$, $f_i(c) \neq f_{i'}(c)$ si y solo si $c = v_r$. Dado que para tal par debemos tener $S_i^j = S_{i'}^j$, se sigue que

$$1[A(S_i^j)(v_r) \neq f_i(v_r)] + 1[A(S_{i'}^j)(v_r) \neq f_{i'}(v_r)] = 1,$$

lo que produce

$$\frac{1}{T} \sum_{i=1}^{T} 1[A(S_i^j)(v_r) \neq f_i(v_r)] = \frac{1}{2}.$$

Combinando esto con la Ecuación (5.6), la Ecuación (5.4) y la Ecuación (5.3), obtenemos que la Ecuación (5.1) se cumple, lo que concluye nuestra prueba. ■

4.1.1 No-Free-Lunch y Conocimiento Previo

¿Cómo se relaciona el resultado de No-Free-Lunch con la necesidad de conocimiento previo? Consideremos un predictor ERM sobre la clase de hipótesis $\mathcal{H}$ de todas las funciones f de $\mathcal{X}$ a $\{0,1\}$. Esta clase representa la falta de conocimiento previo: cada función posible del dominio al conjunto de etiquetas es considerada un buen candidato. Según el teorema de No-Free-Lunch, cualquier algoritmo que elija su salida de hipótesis en $\mathcal{H}$, y en particular el predictor ERM, fallará en alguna tarea de aprendizaje. Por lo tanto, esta clase no es PAC aprendible, como se formaliza en el siguiente corolario:

Corolario 4.1 Sea $\mathcal{X}$ un conjunto de dominio infinito y sea $\mathcal{H}$ el conjunto de todas las funciones de $\mathcal{X}$ a $\{0,1\}$. Entonces, $\mathcal{H}$ no es PAC aprendible.

Prueba Supongamos, por contradicción, que la clase es aprendible. Elija algún $\epsilon < 1/8$ y $\delta < 1/7$. Por la definición de PAC aprendibilidad, debe existir algún algoritmo de aprendizaje A y un entero $m = m(\epsilon, \delta)$, tal que para cualquier distribución generadora de datos sobre $\mathcal{X} \times \{0,1\}$, si para alguna función $f : \mathcal{X} \to \{0,1\}$, $LD(f) = 0$, entonces con probabilidad mayor que $1 - \delta$ cuando A se aplica a muestras S de tamaño m, generadas i.i.d. por D, $LD(A(S)) \leq \epsilon$. Sin embargo, aplicando el teorema de No-Hay-Almuerzo-Gratis, ya que $|\mathcal{X}| > 2^m$, para cada algoritmo de aprendizaje (y en particular para el algoritmo A), existe una distribución D tal que con probabilidad mayor que $1/7 > \delta$, $LD(A(S)) > 1/8 > \epsilon$, lo que lleva a la contradicción deseada. ■

¿Cómo podemos prevenir tales fallos? Podemos escapar de los peligros previstos por el teorema de No-Free-Lunch utilizando nuestro conocimiento previo sobre una tarea de aprendizaje específica, para evitar las distribuciones que nos causarán fallar al aprender esa tarea. Dicho conocimiento previo puede ser expresado restringiendo nuestra clase de hipótesis.

Pero, ¿cómo deberíamos elegir una buena clase de hipótesis? Por un lado, queremos creer que esta clase incluye la hipótesis que no tiene error en absoluto (en el entorno PAC), o al menos que el error más pequeño alcanzable por una hipótesis de esta clase es de hecho bastante pequeño (en el entorno agnóstico). Por otro lado, acabamos de ver que no podemos simplemente elegir la clase más rica - la clase de todas las funciones sobre el dominio dado. Este compromiso se discute en la siguiente sección.

4.2 Descomposición del Error

Para responder a esta pregunta descomponemos el error de un predictor ERM $_{\mathcal{H}}$ en dos componentes de la siguiente manera. Sea h_S una hipótesis ERM $_{\mathcal{H}}$. Entonces, podemos escribir

$$L_{\mathcal{D}}(h_S) = \epsilon_{\text{app}} + \epsilon_{\text{est}}$$

donde:

$$\epsilon_{\text{app}} = \min_{h \in \mathcal{H}} L_{\mathcal{D}}(h),$$

$$\epsilon_{\text{est}} = L_{\mathcal{D}}(h_S) - \epsilon_{\text{app}}.$$

- **El Error de Aproximación** – el riesgo mínimo alcanzable por un predictor en la clase de hipótesis. Este término mide cuánto riesgo tenemos porque nos restringimos a una clase específica, es decir, cuánto sesgo inductivo tenemos. **El error de aproximación no depende del tamaño de la muestra y está determinado por la clase de hipótesis elegida. Ampliar la clase de hipótesis puede disminuir el error de aproximación.**

  Bajo la suposición de realizabilidad, el error de aproximación es cero. En el caso agnóstico, sin embargo, el error de aproximación puede ser grande. De hecho, siempre incluye el error del predictor

óptimo de Bayes (ver Capítulo 3), el error mínimo pero inevitable, debido al posible no determinismo del mundo en este modelo. A veces en la literatura el término error de aproximación se refiere no a $\min_{h \in \mathcal{H}} LD(h)$, sino más bien al exceso de error sobre el del predictor óptimo de Bayes, es decir, $\min_{h \in \mathcal{H}} LD(h) - \epsilon_{\text{Bayes}}$.

- **El Error de Estimación** – la diferencia entre el error de aproximación y el error alcanzado por el predictor ERM. El error de estimación resulta porque el riesgo empírico (es decir, error de entrenamiento) es solo una estimación del riesgo verdadero, y por lo tanto el predictor que minimiza el riesgo empírico es solo una estimación del predictor que minimiza el riesgo verdadero. La calidad de esta estimación depende del tamaño del conjunto de entrenamiento y del tamaño, o complejidad, de la clase de hipótesis. **Como hemos mostrado, para una clase de hipótesis finita, ϵ_{est} aumenta (logarítmicamente) con $|\mathcal{H}|$ y disminuye con m.** Podemos pensar en el tamaño de $\mathcal{H}$ como una medida de su complejidad. En capítulos futuros definiremos otras medidas de complejidad de clases de hipótesis.

Dado que nuestro objetivo es minimizar el riesgo total, enfrentamos un compromiso, llamado el compromiso sesgo-complejidad. Por un lado, elegir $\mathcal{H}$ para ser una clase muy rica disminuye el error de aproximación pero al mismo tiempo podría aumentar el error de estimación, ya que un $\mathcal{H}$ rico podría llevar a un sobreajuste. Por otro lado, elegir $\mathcal{H}$ para ser un conjunto muy pequeño reduce el error de estimación pero podría aumentar el error de aproximación o, en otras palabras, podría llevar a un **subajuste**. Por supuesto, una gran elección para $\mathcal{H}$ es la clase que contiene solo un clasificador – el clasificador óptimo de Bayes. Pero el clasificador óptimo de Bayes depende de la distribución subyacente D, que no conocemos (de hecho, el aprendizaje habría sido innecesario si hubiéramos conocido D).

La teoría del aprendizaje estudia cuán rica podemos hacer $\mathcal{H}$ mientras todavía mantenemos un error de estimación razonable. En muchos casos, la investigación empírica se centra en diseñar buenas clases de hipótesis para un cierto dominio. Aquí, "bueno" significa clases para las cuales el error de aproximación no sería excesivamente alto. La idea es que aunque no somos expertos y no sabemos cómo construir el clasificador óptimo, todavía tenemos algún conocimiento previo del problema específico en cuestión, lo que nos permite diseñar clases de hipótesis para las cuales tanto el error de aproximación como el error de estimación no son demasiado grandes. Volviendo a nuestro ejemplo de las papayas, no sabemos exactamente cómo el color y la dureza de una papaya predicen su sabor, pero sabemos que la papaya es una fruta y basándonos en experiencias previas con otras frutas conjeturamos que un rectángulo en el espacio de color-dureza puede ser un buen predictor.

Resumen

El teorema de No-Free-Lunch afirma que no hay un aprendiz universal. **Cada aprendiz tiene que ser especificado para alguna tarea, y usar algún conocimiento previo sobre esa tarea, para tener éxito**. Hasta ahora hemos modelado nuestro conocimiento previo restringiendo nuestra salida de hipótesis a ser miembro de una clase de hipótesis elegida. Al elegir esta clase de hipótesis, enfrentamos un compromiso, entre una clase más grande, o más compleja, que es más probable que tenga un pequeño error de aproximación, y una clase más restringida que garantizaría que el error de estimación será pequeña. En el próximo capítulo estudiaremos con más detalle el comportamiento del error de estimación. En el Capítulo 7 discutiremos formas alternativas de expresar conocimientos previos.

5

La dimensión VC

En el capítulo anterior, descompusimos el error de la regla $\text{ERM}_{\mathcal{H}}$ en error de aproximación y error de estimación. El error de aproximación depende del ajuste de nuestro conocimiento previo (reflejado por la elección de la clase de hipótesis $\mathcal{H}$) a la distribución desconocida subyacente. En contraste, la definición de aprendizaje PAC requiere que el error de estimación esté acotado uniformemente sobre todas las distribuciones.

Nuestro objetivo actual es averiguar qué clases H son aprendibles PAC, y caracterizar exactamente la complejidad de muestra para aprender una clase de hipótesis dada. *Hasta ahora hemos visto que las clases finitas son aprendibles, pero que la clase de todas las funciones (sobre un dominio de tamaño infinito) no lo es*. ¿Qué hace que una clase sea aprendible y la otra no? ¿Pueden las clases de tamaño infinito ser aprendibles y, si es así, qué determina su complejidad de muestra?

Comenzamos el capítulo mostrando que las clases infinitas pueden ser efectivamente aprendibles, y por lo tanto, la finitud de la clase de hipótesis no es una condición necesaria para el aprendizaje. Luego presentamos una caracterización notablemente nítida de la familia de clases aprendibles en la configuración de clasificación binaria con la pérdida de cero-uno. Esta caracterización fue descubierta por primera vez por Vladimir Vapnik y Alexey Chervonenkis en 1970 y se basa en una noción combinatoria llamada dimensión Vapnik-Chervonenkis (dimensión VC). Definimos formalmente la dimensión VC, proporcionamos varios ejemplos y luego enunciamos el teorema fundamental de la teoría del aprendizaje estadístico, que integra los conceptos de aprendizaje, dimensión VC, la regla ERM y convergencia uniforme.

5.1 Las clases de tamaño infinito pueden ser aprendibles

En el Capítulo 4 vimos que las clases finitas son aprendibles, y de hecho la complejidad de muestra de una clase de hipótesis está acotada superiormente por el logaritmo de su tamaño. Para demostrar que el tamaño de la clase de hipótesis no es la caracterización correcta de su complejidad de muestra, primero presentamos un ejemplo simple de una clase de hipótesis de tamaño infinito que es aprendible.

Ejemplo 5.1 Sea $\mathcal{H}$ el conjunto de funciones de umbral sobre la línea real, es decir, $\mathcal{H} = \{h_a : a \in \mathbb{R}\}$, donde $h_a : \mathbb{R} \to \{0, 1\}$ es una función tal que $h_a(x) = \mathbb{1}_{[x<a]}$. Para recordar al lector, $\mathbb{1}[x < a]$ es 1 si $x < a$ y 0 en caso contrario. Claramente, $\mathcal{H}$ es de tamaño infinito. Sin embargo, el siguiente lema muestra que $\mathcal{H}$ es aprendible en el modelo PAC utilizando el algoritmo ERM. ■

Lema 5.1 Sea H la clase de umbrales como se definió anteriormente. Entonces, H es PAC aprendible, utilizando la regla ERM, con complejidad de muestra de $m_H(\varepsilon, \delta) \leq \lceil \log(2/\delta)/\varepsilon \rceil$.

Demostración.- Sea a^* un umbral tal que la hipótesis $h^*(x) = 1[x < a^*]$ logra $L_{\mathcal{D}}(h^*) = 0$. Sea $\mathcal{D}_x$ la distribución marginal sobre el dominio X y sea $a_0 < a^* < a_1$ tal que

$$\mathbb{P}_{x \sim \mathcal{D}_x}[x \in (a_0, a^*)] = \mathbb{P}_{x \sim \mathcal{D}_x}[x \in (a^*, a_1)] = \varepsilon.$$

(Si $\mathcal{D}_x(-\infty, a^*) \leq \varepsilon$ establecemos $a_0 = -\infty$ y de manera similar para a_1). Dado un conjunto de entrenamiento S, sea $b_0 = \max\{x : (x, 1) \in S\}$ y $b_1 = \min\{x : (x, 0) \in S\}$ (si no hay ejemplo en S es positivo establecemos $b_0 = -\infty$ y si no hay ejemplo en S es negativo establecemos $b_1 = \infty$). Sea b_S un umbral correspondiente a una hipótesis ERM, h_S, lo que implica que $b_S \in (b_0, b_1)$. Por lo tanto, una condición suficiente para $L_D(h_S) \leq \varepsilon$ es que tanto $b_0 \geq a_0$ como $b_1 \leq a_1$. En otras palabras,

$$\mathbb{P}_{S \sim \mathcal{D}_m}[L_D(h_S) > \varepsilon] \leq \mathbb{P}_{S \sim D_m}[b_0 < a_0 \vee b_1 > a_1],$$

y usando la cota de la unión podemos acotar lo anterior por

$$\mathbb{P}_{S \sim D_m}[L_D(h_S) > \varepsilon] \leq \mathbb{P}_{S \sim D_m}[b_0 < a_0] + \mathbb{P}_{S \sim D_m}[b_1 > a_1]. \tag{5.1}$$

El evento $b_0 < a_0$ ocurre si y solo si todos los ejemplos en S no están en el intervalo (a_0, a^*), cuya masa de probabilidad se define como ε, es decir,

$$\mathbb{P}_{S \sim D_m}[b_0 < a_0] = \mathbb{P}_{S \sim D_m}[\forall (x, y) \in S, x \notin (a_0, a^*)] = (1 - \varepsilon)^m \leq e^{-\varepsilon m}.$$

Dado que asumimos $m > \log(2/\delta)/\varepsilon$ se sigue que la ecuación es como máximo $\delta/2$. De la misma manera, es fácil ver que $P_{S \sim D_m}[b_1 > a_1] \leq \delta/2$. Combinando con la Ecuación (6.1) concluimos nuestra prueba. ■