Fundamentos de aprendizaje automático

Universidad del Valle de Guatemala

Profesor Pedro Aguilar
Notas de clase de Rodrigo Leonardo

Contents

1 Probabilidad

Definición 1.1. El resultado de lanzar una moneda al aire es una variable aleatoria X. Al espacio de valores que toma la variable aleatoria le llamamos el espacio muestral. En este caso, $X = \{0, 1\}$. Sobre S podemos definir una distribución de probabilidad de forma que $p(0) = p(1) = 1/2$.

Para $\mathbb{Z}^+$, podemos definir $p(i) = \frac{6}{\pi^2} \cdot \frac{1}{i^2},\ i \in S$. En otros casos no podemos definir una distribución de probabilidad. Por ejemplo si $S = \mathbb{R}$, entonces se define una distribución de probabilidad, i.e.:

$$\mathbb{P}(a < x < b) = \int_a^b p(x)dx \tag{1}$$

Definición 1.2. La distribución de probabilidad cumulativa es:

$$F(a) = \int_{-\infty}^a p(x)dx = \mathbb{P}(x < a) \tag{2}$$

1.1 Espacios muestrales, eventos e independencias

Definición 1.3. Supongamos que lanzamos n veces una moneda, entonces tenemos variables aleatorieas $x_1, \ldots, x_n$. El espacio muestral está dado por $X = \{0, 1\}^n$. Definimos un evento como un subconjunto del espacio muestral.

Definición 1.4. Sean A y B dos eventos. La ocurrencia de A y B se denota por $A \cap B$. Definimos la probabilidad condicional de que ocurra A después de que ocurra el evento B. Denotado por:

$$\mathbb{P}(A|B) = \frac{\mathbb{P}(A \cap B)}{\mathbb{P}(B)} \tag{3}$$

Definición 1.5. Los eventos A y B son independientes si: $\mathbb{P}(A|B) = \mathbb{P}(A)$, entonces $\mathbb{P}(A)\mathbb{P}(B) = \mathbb{P}(A \cap B)$.

Definición 1.6. Dos variables aleatorias X, Y son independientes si para cualesquiera eventos A y B de valores de X y Y, respectivamente, A y B son independientes. Para una colección de variables aleatorieas $x_1, \ldots, x_n$ independientes:

$$\mathbb{P}(x_1 \in A_1, \ldots, x_n \in A_n) = \prod_{i=1}^n \mathbb{P}(x_i \in A_i) \tag{4}$$

1.2 Valor esperado

Definición 1.7. Definimos $\mathbb{E}(X) = \sum xp(x)$ o $\mathbb{E}(X) = \int_{-\infty}^{\infty} xp(x)dx$ como el valor esperado para una variable aleatoria X. Nótese que:

$$\begin{aligned}\mathbb{E}(X+Y) &= \int_S (x+y)p(x,y)d\Sigma \\ &= \int_S xp(x,y)d\Sigma + \int_S yp(x,y)d\Sigma \\ &= \mathbb{E}(X) + \mathbb{E}(Y)\end{aligned}$$

1.3 Cota de la unión

Nota 1.1. Consideremos los eventos $A_1, \ldots, A_n$, entonces:

$$\begin{aligned}\mathbb{P}(\bigcup_{i=1}^{n} A_i) &= \sum_{i=1}^{n} \mathbb{P}(A_i) - \sum_{i,j=1}^{n} \mathbb{P}(A_i \cap A_j) \pm \cdots \pm \mathbb{P}(\bigcap\{A_i\}) \\ &\leq \sum_{i=1}^{n} \mathbb{P}(A_i)\end{aligned}$$

1.4 Función característica

Definición 1.8. Consideremos S y $A \subseteq S$, el espacio muestral y un evento. Definimos a la función característica:

$$\chi_A(x) = \begin{cases} 1 & x \in A \\ 0 & x \notin A \end{cases} \tag{5}$$

Supongamos que $S = \{1, \ldots, n\}$. Por ejemplo, consideremos $\Sigma = \{1, \ldots, n\}$ y A es el subconjunto de puntos fijos bajo $\sigma \in A(\Sigma) = S_n$. El tamaño de A es $\Sigma_{i=1}^{n} \chi_A(i)$, entonces:

$$\begin{aligned}\mathbb{E}(\sum_{i=1}^{n} \chi_A(i)) &= \sum_{i=1}^{n} \mathbb{E}(\chi_A(i)) \\ &= n\mathbb{E}(\chi_A(1)) = \frac{n(n-1)!}{n!} = 1\end{aligned}$$

1.5 Varianza

Definición 1.9. La varianza de una variable X se denota por σ^2 o $\mathbb{V}(X)$, y se define como:

$$\sigma^2 = \mathbb{E}((x - \mathbb{E}(x))^2) \tag{6}$$

Definición 1.10. Definimos la desviación estándar $\sigma = \sqrt{\mathbb{V}(X)}$.

Propiedad 1.1. Podemos probar que:

$$\begin{aligned}\sigma^2 &= \mathbb{E}(X - \mathbb{E}(X))^2 = \mathbb{E}(X^2) - 2\mathbb{E}(X\mathbb{E}(X)) + (\mathbb{E}(X))^2 \\ &= \mathbb{E}(X^2) - 2\mathbb{E}^2(X) + \mathbb{E}^2(X) = \mathbb{E}(X^2) - \mathbb{E}^2(X)\end{aligned}$$

Propiedad 1.2. La varianza de la suma de variables aleatorias:

$$\begin{aligned}\mathbb{V}(X+Y) &= \mathbb{E}(X+Y)^2 - (\mathbb{E}(X+Y))^2 \\ &= \mathbb{E}(X^2) + 2\mathbb{E}(XY) + \mathbb{E}(Y^2) - \mathbb{E}^2(X) - 2\mathbb{E}(X)\mathbb{E}(Y) - \mathbb{E}^2(Y) \\ &= \mathbb{V}(X) + \mathbb{V}(Y) + 2\mathbb{E}(XY) - 2\mathbb{E}(X)\mathbb{E}(Y)\end{aligned}$$

Propiedad 1.3. Si X y Y son independientes, $\mathbb{V}(X+Y) = \mathbb{V}(X) + \mathbb{V}(Y)$. Si $X_1, \ldots, X_n$ son v.a. independientes:

$$\mathbb{V}(\sum_{i=1}^{n} X_i) = \sum_{i=1}^{n} \mathbb{V}(X_i) \tag{7}$$

1.6 Teorema del límite central

Ejemplo 1.1. Supongamos que $X_1, \ldots, X_n$ son v.a. idéntica e independientemente distribuidas (iid). Sean $\mathbb{E}(X_i) = \frac{1}{2}$, $\mathbb{V}(X_i) = \frac{1}{2}(0 - \frac{1}{2})^2 + \frac{1}{2}(1 - \frac{1}{2})^2 = \frac{1}{4}$, donde:

$$X_i = \begin{cases} 0 & \text{con probabilidad } \frac{1}{2} \\ 1 & \text{con probabilidad } \frac{1}{2} \end{cases} \tag{8}$$

Sea $S = X_1 + \cdots + X_n \implies \mathbb{E}(S) = \frac{n}{2}$, $\mathbb{V} = \frac{n}{4}$ y $\sigma = \frac{\sqrt{n}}{2}$. Nótese que $\frac{\sigma}{\mathbb{E}(S)} \to 0$ cuando $n \to \infty$. Si las X_i están iid con desviación estándar σ, entonces $S = \sum_{i=1}^{n} \chi_i$ tiene desviación estándar $\sqrt{n}\sigma$. Por lo tanto, $\frac{1}{\sqrt{n}} \sum_{i=1}^{n} \chi_i$ tiene distribución σ.

Teorema 1.1 (Límite central)**.** Sean $X_1, \ldots, X_n$ una colección de v.a. iid con $\mathbb{E}(X) = \mu$ y $\mathbb{V}(X_i) = \sigma^2$, $\forall i = 1, \ldots, n$. La distribución de:

$$\frac{1}{\sqrt{n}}(\sum_{i=1}^{n} X_i - n\mu)$$

converge a la distribución de una gaussiana con media 0 y varianza σ^2, $N(0, \sigma^2)$.

1.7 Distribuciones de probabilidad

Definición 1.11. La distribución normal/gaussiana $N(\mu, \sigma)$ se define como:

$$p(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{\frac{-(x-\mu)^2}{2\sigma^2}} \tag{9}$$

Nota 1.2. Note que:

$$\int_{\mathbb{R}} e^{-x^2} dx = \sqrt{\pi} \tag{10}$$

Definición 1.12. Un experimento A, B con probabilidades $p,\ 1-p$ respectivamente. Repetimos el experimento n veces y queremos la probabilidad de obtener A k veces:

$$B(n,p) = \binom{n}{k} p^k (1-p)^{n-k} \tag{11}$$

La media es np y la varianza $np(1-p)$.

Propiedad 1.4. El valor esperado de la distribución binomial es np. En efecto:

$$\begin{aligned}
\mathbb{E}(K) &= \sum_{k=0}^{n} K B_k(n,p) = \sum_{k=0}^{n} K \frac{n!}{k!(n-k)!} p^k (1-p)^{n-k} \\
&= np \sum_{k=0}^{n} \frac{(n-1)!}{(k-1)!((n-1)-(k-1))!} p^{k-1} (1-p)^{(n-1)-(k-1)} \\
&= np \sum_{k=1}^{n} \frac{(n-1)!}{(k-1)!((n-1)-(k-1))!} p^{k-1} (1-p)^{(n-1)-(k-1)} \\
&= np \sum_{k=0}^{n-1} \frac{(n-1)!}{k!((n-1)!-k!)} p^k (1-p)^{n-1-k} = np
\end{aligned}$$

La varianza:

$$\begin{aligned}
\mathbb{V}(X) &= p(1-p)^2 + (1-p)(0-p)^2 = p(1-p)^2 + (1-p)p^2 \\
&= (1-p)(p - p^2 + p^2) = p(1-p)
\end{aligned}$$

Entonces, $\mathbb{V}(K) = np(1-p)$.

Nota 1.3. La distribución normal cumple con:

$$\mathbb{P}(\mu \pm \sigma) = \frac{e^{-1/2}}{\sqrt{2\pi}\sigma} = \frac{1}{\sqrt{2\pi e}\sigma} \tag{12}$$

Definición 1.13. Si en la distribución binomial considera a $d = np$, entonces si $n >> k$:

$$\frac{n^k}{k!}(\frac{d}{n})^2(1-\frac{d}{n})^n \approx \frac{d^k}{k!} e^{-d}$$

Entonces la distribución de Poisson es:

$$\text{Poisson}(\lambda) = \frac{\lambda^k e^{-\lambda}}{k!} \tag{13}$$

1.8 Regla de Bayes

Definición 1.14. La regla de Bayes relaciona la probabilidad condicional $\mathbb{P}(A|B)$ con $\mathbb{P}(B|A)$:

$$\mathbb{P}(A|B) = \frac{\mathbb{P}(A \cap B)}{\mathbb{P}(B)} = \frac{\mathbb{P}(B \cap A)}{\mathbb{P}(A)} \tag{14}$$

Se le conoce a $\mathbb{P}(A)$ como la probabilidad a priori y a $\mathbb{P}(A|B)$ como la probabilidad a posteriori.

Ejemplo 1.2. Sea A el evento en el que un producto es defectuoso y B el evento en el que una prueba dice que el producto es defectuoso. Supóngase que $\mathbb{P}(A) = 0.001$, $\mathbb{P}(B|A) = 0.99$ y $\mathbb{P}(B|A^c) = 0.02$. Entonces:

$$\begin{aligned}\mathbb{P}(B) &= \mathbb{P}(B|A)\mathbb{P}(A) + \mathbb{P}(B|A^c)\mathbb{P}(A^c) \\ &= \mathbb{P}(B \cap A) + \mathbb{P}(B \cap A^c) \\ &= 0.99 \cdot 0.001 + 0.02 \cdot (1 - 0.001) = 0.02087\end{aligned}$$

Finalmente:

$$\mathbb{P}(A|B) = \frac{\mathbb{P}(B|A)\mathbb{P}(A)}{\mathbb{P}(B)} = 0.0471$$

2 Modelo Formal de Aprendizaje

Nota 2.1. ¿Cuándo se necesita machine learning? Para replicar actividades que realizan humanos, tomando en cuenta la adaptabilidad del sistema y la capacidad sobrehumana en cuanto a memoria.

Nota 2.2. Tipos de aprendizaje:

1. Supervisado vs No supervisado: datos etiquetados para que el aprendiz entrene con valores de verdad se conoce como aprendizaje supervisado. Si el aprendiz recibe datos con el objetivo de clasificarlos de acuerdo a características innatas, se le conoce como aprendizaje no supervisado.

2. Aprendiz activo vs pasivo: un aprendiz activo interactúa con los datos en el entrenamiento

3. Ayuda del maestro: al aprendiz se le evalúa en un peor escenario para que, al momento de encontrar un maestro cualquiera, pueda adaptarse a mejores condiciones.

4. Online vs protocolo de aprendizaje por lotes: velocidad de respuesta y necesidad de la misma. Algunos aprendices rinden mejor luego de aprender de grandes cantidades de información sin siquiera realizar una predicción.

En machine learning se hace énfasis en asumir lo menos posible acerca de los datos con los que se trabajará.

Nota 2.3. Lo que conoce el aprendiz:

1. Dominio: un conjunto arbitrario X. Si $x \in X$ es un ejemplo, X es el espacio de ejemplos.
2. Espacio de etiquetas: un conjunto finito Y.
3. Datos de entrenamiento: son ejemplos etiquedatos,

$$S = \{(x_1, y_1), \ldots, (x_n, y_n)\} \subset X \times Y, \quad nin\mathbb{Z}^+$$

Lo que esperamos del aprendiz:

1. Predicción: $h : X \to Y$, le llaman predictor, hipótesis o clasificador.

Un modelo de generación de datos:

1. Los datos en S se generan de la siguiente manera: sea D una distribución de probabilidad definida en X que extrae datos aleatoriamente.
2. Se etiqueta según una función f (ambas desconocidas por el aprendiz). Se genera (x_i, y_i) por $X \xrightarrow{D} x_i \xrightarrow{f} y_i$.

Definición 2.1. La medida de éxito se define como el error del aprendiz, el cual es la probabilidad de que si se extrae algún $x_D \sim X$, $h(x) \neq f(x)$. La medida de error, conocido como una función de pérdida, es una probabilidad

$$L_{D,f}(h) = D(\{x | h(x) \neq f(x)\}) \tag{15}$$

es el tamaño del evento de los ejemplos que no se clasifican correctamente, dado la distribución.

Nota 2.4. Sea $A \subset X$ y $\pi(x)$ la función característica. Entonces $A = \{x \in \chi \ni \pi(x) = 1\}$.

Nota 2.5. A $L_{D,f}(h)$ se le conoce como el error de generalización, riesgo o error verdadero de h.

2.1 Minimización de riesgo empírico (ERM)

Nota 2.6. Dado que el aprendiz no tiene acceso a D ni a f, si se obtienen ejemplos $S = \{(x_1, y_1), \ldots, (x_m, y_m)\}$, $m \in \mathbb{Z}^+$, con $\chi \xrightarrow{D,f} S$.

Definición 2.2. El error de entrenamiento se define como:

$$L_S(h) = \big|\{i \in \{1, \ldots, m\} : h(x_i) \neq f(y_i)\}\big| \tag{16}$$

Nota 2.7. A la definición anterior se le conoce como minimización de riesgo empírico (ERM).

Ejemplo 2.1. Sea:

$$h(x) = \begin{cases} y_i & \text{si } \exists i \ni x_i = x \\ 0 & \text{en otro caso} \end{cases}$$

Entonces $L_s(h) = 0$. Si la distribución es uniforme, $L_{D,f}(h) = 1/2$. Esto se llama *overfitting* o sobreajuste.

2.2 ERM con sesgo

Definición 2.3. Del conjunto de predictores $\{f : X \to Y\}$ definimos a H como la clase de hipótesis, $h \in H$ un predictor tal que $h : X \to Y$. Realizando ERM sobre la clase H:

$$\text{ERM}_H \in \arg\min_{h \in H} L_S(h)$$

2.3 Clase de hipótesis finitas

Nota 2.8. Vamos a demostrar que con estas clases se evita el sobreajuste y podemos acercarnos lo suficiente al clasificador real. Dada una clase de hipótesis H escogemos el predictor que minimiza el error de entrenamiento $h_s \in \arg\min_{h \in H} L_S(h)$ y consideramos las siguientes hipótesis.

1. Hipótesis de realizabilidad: $\exists h^* \in H \ni L_{D,f}(h^*) = 0$. Esto implica que $L_S(h^*) = 0$ y $h_S \in \text{ERM}_H,\ L_S(h_S) = 0$.

2. Suposición de iid: los ejemplos en el conjunto de entrenamiento S están independiente e idénticamente distribuidos. Se denota por $S \sim D^m$, donde D es la distribución dada.

Notemos que hay un riesgo en escoger elementos de S que no sean significativos del dominio.

Definición 2.4. Definimos δ como la probabilidad de extraer una muestra no representativa. Definimos el parámetro de confianza como $1 - \delta$.

Definición 2.5. Si $L_{D,f}(h_S) > \epsilon$, el aprendiz falló. Si $L_{D,f}(h_S) \leq \epsilon$, el aprendiz tuvo éxito y se ha generado un predictor aproximadamente correcto.

Nota 2.9. Ahora proponemos una cota de la probabilidad de que el aprendiz falle. Extraemos m elementos del dominio: $S|_X = (x_1, \ldots, x_m)$. Queremos acotar $D^m(\{S|_X : L_{(D,f)}(h_S) > \epsilon\})$. Sea H_B el conjunto de malas hipótesis, i.e.

$$H_B = \{h \in H : L_{(D,f)}(h) > \epsilon\} \tag{17}$$

Sea $M = \{S|_X : \exists h \in H_B,\ L_S(h) = 0\}$, el conjunto de datos de prueba engañosos. Supongamos que tenemos una muestra $S|_X \ni L_{D,f}(h_S) > \epsilon,\ L_S(h_S) = 0 \implies \{S|_X : L_{(D,f)}(h_S) > \epsilon\} \subseteq M$. Notemos que $M = \bigcup_{h \in H_B} \{S|_X : L_S(h) = 0\}$, entonces:

$$\begin{aligned} D^m(\{S|_X : L_{(D,f)}(h_S) > \epsilon\}) \leq D^m(M) &= D^m(\bigcup_{h \in H_B} \{S|_X : L(h) = 0\}) \\ &\leq \sum_{h \in H_B} D^m(\{S|_X : L(h) = 0\}) \end{aligned}$$

Pero:

$$\begin{aligned} D^m(\{S|_X : L_s(h) = 0\}) &= \prod_{i=1}^{m} D(\{x_i | f(x_i) = h(x_i)\}) \\ &= \prod_{i=1}^{m} (1 - L_{D,f}(h)) = \sum_m \prod_{i=1}^{m} (1 - \epsilon) = (1 - \epsilon)^m \leq e^{-\epsilon m} \end{aligned}$$

Finalmente,

$$\begin{aligned} D^m(\{S|_X : L_{(D,f)}(h_S) > \epsilon\}) &\leq \sum_{h \in H_B} e^{-\epsilon m} \\ &= |H_B|\, e^{-\epsilon m} \end{aligned}$$

Corolario 2.1. Sea H una clase de hipótesis finita. Sea $\delta \in (0,1)$, $\epsilon > 0$ y $m \in \mathbb{Z}^+ \ni m \geq \frac{\ln(|H|/\delta)}{\epsilon}$. Entonces para una función de clasificación f, una distribución D, con la condición de realizabilidad y con probabilidad de al menos $1 - \delta$ en la elección de la muestra iid de tamaño m, tenemos que para cada hipótesis ERM_H, h_S

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

Se dice que el proceso de ERM es probablemente aproximado correcto (PAC).

3 Aprendizaje PAC

Definición 3.1. Una clase de hipótesis H es PAC aprendible si existe una función $m_H : (0,1)^2 \to \mathbb{N}$ y un algoritmo de aprendizaje con la siguiente propiedad. Para todo $\epsilon, \delta \in (0,1)$, para cada distribución D y función de clasificación $f : X \to \{0,1\}$. Si se satisface realizabilidad, entonces si para algún $m \geq m_H$ datos de entrenamiento, el algoritmo produce la hipótesis h tal que $L_{(D,f)}(h) < \epsilon$, con probabilidad $1 - \delta$ (sobre la elección de S). Se hacen dos aproximaciones:

1. ϵ: parámetro de precisión, controla el error de la hipótesis.
2. δ: parámetro de confianza, probabilidad de que la hipótesis alcance dicho error.

3.1 Complejidad de la muestra

Definición 3.2. La complejidad de la muestra se define como $m_H : (0,1)^2 \to \mathbb{N}$.

Nota 3.1. m_H es la función mínima dado H, ϵ y δ. Además, m_H es el mínimo entero para el cual H es PAC aprendible.

Nota 3.2. Cada clase de hipótesis finita es PAC aprendible con complejidad de la muestra:

$$m_H(\epsilon, \delta) \leq \frac{\ln\big(|H|/\delta\big)}{\epsilon}$$

Entonces se pueden generalizar los modelos de aprendizaje:

1. Relajando la hipótesis de realizabilidad.
2. Clasificación binaria.

3.2 Aprendizaje PAC agnóstico

Suponemos que D es una distribución sobre $X \times Y$ sobre el dominio y las etiquetas. Podemos pensarlo como una distribución D_X que extrae datos del dominio y $D((x,y)|x)$ es una distribución para las etiquedas para cada punto x.

3.3 Error empírico y real

Nota 3.3. Extraemos datos de entrenamiento según una distribución D sobre $X \times Y$. Entonces podemos medir la probabilidad de que una hipótesis falle como:

$$L_D(h) = D(\{(x,y) : h(x) \neq y\}) = \mathbb{P}_{(x,y)\sim D}(h(x) \neq y) \tag{18}$$

Definición 3.3. El error empírico lo definimos como:

$$L_S(h) = \frac{\big|\{i \in \{1,\dots,m\} : h(x_i) \neq y_i\big|}{m} \tag{19}$$

Con $S = \{(x_1, y_1), \dots, (x_m, y_m)\}$.

Nota 3.4. Nosotros queremos una hipótesis $h \in H$ que minimice el error verdadero probablemente aproximadamente.

3.4 Predictor óptimo de Bayes

Definición 3.4. Dado una distribución D sobre $\chi \times \{0,1\}$, definimos el predictor:

$$f_D(x) = \begin{cases} 1 & \text{si } \mathbb{P}(y=1|x) \geq \frac{1}{2} \\ 0 & \text{si } \mathbb{P}(y=0|x) < \frac{1}{2} \end{cases} \tag{20}$$

f_D es óptimo en el sentido de que $L_D(f_D) \leq L_D(g),\ \forall g : X \to Y$.

Definición 3.5 (Aprendizaje PAC agnóstico)**.** Una clase de hipótesis H es PAC agóstica si $\exists m_H : (0,1)^2 \to \mathbb{Z}^+$ y un algoritmo de aprendizaje que satisface que $\forall \epsilon, \delta \in (0,1)$ y cada distribución D sobre $X \times Y$ si tomamos $m > m_H(\epsilon, \delta)$ ejemplos de entrenamiento iid, entonces el resultado h del algoritmo de entrenamiento es tal que:

$$L_D(h) < \inf_{h' \in H} \{L_p(h')\} + \epsilon \tag{21}$$

con probabilidad $1 - \delta$.

Nota 3.5.
1. La clasificación multiclase tiene a X como dominio y a Y como conjunto discreto.
2. La regresión tiene a X como dominio y a Y como conjunto continuo de forma que
$$L_D(h) = \mathbb{E}_{(x,y)\sim D}((h(x) - y)^2) \tag{22}$$
content...

3.5 Funciones de pérdida generalizadas

Definición 3.6. Sea H nuestra clase de hipótesis y Z un dominio. El espacio donde está definida la distribución en el caso anterior, $Z = X \times Y$. Se dice que ℓ es una función de pérdida $\ell : H \times Z \to \mathbb{R}^+$.

Definición 3.7. Definimos la función de riesgo como el valor esperado para un clasificador $h \in H$ sobre los elementos de Z con la distribución D,

$$L_D(h) = \mathbb{E}_{Z\sim D}(\ell(h, z)) \tag{23}$$

Definición 3.8. Definimos el riesgo empírico para una muestra $S = (z_1, \ldots, z_n)$ como

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

Ejemplo 3.1.
1. Binario, la variable $z \in X \times Y$,
$$\ell(h, (x, y)) = \begin{cases} 0 & \text{si } h(x) = y \\ 1 & \text{si } h(x) \neq y \end{cases}$$
2. Pérdida cuadrática, la variable z está en $X \times Y$
$$\ell(h, (x, y)) = (h(x) - y)^2$$

4 Aprendizaje bajo convergencia uniforme

Definición 4.1. Fijamos $Z = X \times Y,\ D,\ \ell$ y H. Decimos que una muestra S es ϵ-representativa si $\forall h \in H$

$$|L_D(h) - L_S(h)| < \epsilon$$

Lema 4.1. Supongamos que S es $\frac{\epsilon}{2}$-representativa. Entonces el resultado de h_S de ERM_H satisface

$$L_D(h_S) < \min_{h \in H} L_D(h) + \epsilon$$

Proof. Sea $h \in H$,

$$\begin{aligned} L_D(h_S) &< L_S(h_S) + \epsilon \\ &\leq L_S(h) + \epsilon \\ &< L_D(h) + \epsilon, \quad \forall h \in H \end{aligned}$$

□

Definición 4.2. Una clase de hipótesis H tiene convergencia uniforme (una vez fijado el dominio y el error) si existe una función $m_H^{uc} : (0,1)^2 \to \mathbb{N} \ni \forall \epsilon,\ \delta \in (0,1)$ y cada D, se satisfaga que S es una muestra con m elementos, $m \geq m_H^{uc}$, entonces S es ϵ-representativo con probabilidad $1 - \delta$.

Corolario 4.1. Si H es una clase de hipótesis con convergencia uniforme, entonces H es PAC agnóstica con complejidad $m_H(\epsilon, \delta) \leq m_H^{uc}(\epsilon/2, \delta)$, usando ERM_H.

4.1 Clases de hipótesis finitas son PAC aprendibles

Teorema 4.1 (Desigualdad de Houffding)**.** Sean $\theta_1, \ldots, \theta_n$ un conjunto de variables aleatorias iid con $\mathbb{E}(\theta_i) = \mu$ y $\mathbb{P}(a \leq \theta_i \leq L) = 1,\ \forall i = 1, \ldots, n$. Entonces $\forall \epsilon > 0$

$$\mathbb{P}(|\frac{1}{m}\sum_{i=1}^{m} \theta_i - \mu| > \epsilon) \leq 2\exp\{-\frac{2m\epsilon^2}{(b-a)^2}\}$$

Nota 4.1. Queremos mostrar que si H es una clase finita, entonces H tiene convergencia uniforme. Sean $\epsilon, \delta \in (0,1)$. Queremos un tamaño de muestra tal que, con probabilidad $1 - \delta$,

$$|L_S(h) - L_D(h)| < \epsilon, \quad \forall h \in H$$

Como S se extrae iid,

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

de manera equivalente,

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

Nótese que

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

$$\begin{aligned} D^m(\{S : \exists h \in H, \quad |L_S(h) - L_D(h)| > \epsilon\}) &= D^m(\bigcup_{h \in H} \{S : |L_S(h) - L_D(h)\} > \epsilon) \\ &\leq \sum_{h \in H} D^m(\{S : |L_S(h) - L_D(h)| > \epsilon\}) \end{aligned}$$

Por otro lado,

$$L_D(h) = \mathbb{E}_{Z \sim D}[\ell(h, z)], \quad L_S(h) = \frac{1}{m} \sum_{i=1}^{m} \ell(h, z_i)$$

Si z es aleatoria, h fijo, entonces $\ell(h, z)$ es aleatoria y $\mathbb{E}(L_S(h)) = L_D(h) \implies$ $|L_S(h) - L_D(h)|$ es una medida de la desviación.

Si $\ell(h, z_i)$ es nuestra variable aleatoria, $\mathbb{E}(\ell(h, z)) = L_D(h)$. Suponemos que $\ell(h, z) \in [0, 1]$. Entonces

$$D^m(\{S : |L_S(h) - L_D(h)| > \epsilon\}) = \mathbb{P}(|\frac{1}{m} \sum_{i=1}^{m} \ell(h, z_i) - L_D(h)| > \epsilon) \leq 2e^{-2m\epsilon^2}, \quad \text{entonces}$$

$$D^m(\{S : h \in H, \quad |L_S(h) - L_D(h)| > \epsilon\}) \leq 2|H|e^{-2m\epsilon^2} \leq \epsilon$$

Por lo tanto

$$m = \frac{\ln\big(2|H|\ \delta\big)}{2\epsilon^2}$$

Problema 4.1. Recordemos los PAC no agnósticos, entonces $\exists p(x) \ni \text{sign}(p(x)) = h(x)$.

$$h(x) = \begin{cases} y_i & \text{si } x = x_i \\ 1 & \text{en otro caso} \end{cases}$$

Consideremos

$$p(x) = \prod_{x_i,\ y_i = 0} (x - x_i)^2$$

Problema 4.2. Demuestre $\mathbb{E}(L_S(h)) = L_D(h), \quad S = \{(x_1, y_1), \ldots, (x_m, y_m)\}$.

Proof. Note que $L_S(h) = \frac{1}{m} \sum_{i=1}^{m} \ell(h, x_i) \implies$

$$\mathbb{E}(L_S(h)) = \mathbb{E}(\frac{1}{m} \sum_{i=1}^{m} \ell(h, x_i)) = \frac{1}{m} \sum_{i=1}^{m} \mathbb{E}(\ell(h, x_i))$$

Si S está formado por datos idénticamente distribuidos, entonces

$$L_S(h) = \frac{1}{m}\sum_{i=1}^{m} L_D(h) = L_D(h)$$

□

Problema 4.3. Recordemos la definición de PAC aprendible. Si fijamos δ y tomamos $0 < \epsilon_1 < \epsilon_2 < 1$, entonces $m_H(\epsilon_2, \delta) \leq m_H(\epsilon_1, \delta)$. Si S_m tiene m datos, $m \geq m_H(\epsilon_H, \delta)$, entonces $L_D(h) < \epsilon_1 < \epsilon_2 \implies m \geq m_H(\epsilon_2, \delta)$. Por lo tanto $m_H(\epsilon_1, \delta) \geq m_H(\epsilon_2, \delta)$. Entonces m_H es mutuamente creciente respecto a ϵ. Similarmente, podemos demostrar que es monótonamente respecto a δ.

Propiedad 4.1. H es una clase agnóstica PAC aprendible $\implies$ H es PAC aprendible.

Ejemplo 4.1. Sea X un conjuto finito y sea $H = \{h_z\} \cup \{h_-\}$ donde

$$h_z(h) = \begin{cases} 1 & x = z \\ 0 & x \neq z \end{cases}, \quad h_-(x) = 0$$

Estamos asumiendo realizabilidad. Sea S una muestra con m elementos. $S = \{(x_1, y_1), \ldots, (x_m, y_m)\}$. Sea x_0 el único punto tal que $f(x_0) = 1$. Si $x_0 \notin \{x_1, \ldots, x_m\}$, ERM $\implies h_-$. Si $x_0 \in \{x_1, \ldots, x_m\}$, ERM $\implies h_{x_0}$. Dado ϵ, δ, queremos encontrar $m_H(\epsilon, \delta)$,

$$D^m(\{S_m : L_D(h) < \epsilon\}) > 1 - \delta \iff D^m(\{S_m : L_D(h) > \epsilon\}) < \delta$$

Supongamos que queremos aprender h_-, formamos $S_m \implies$ (ERM) h_- con $L_D(h_-) = 0$. Si queremos aprender h_{x_0}, formamos S_m (que contiene x_0) $\implies$ (ERM) h_{x_0}. Si queremos aprender h_{x_0}, formamos S_m y $x_0 \notin S_m$. ERM $\to h_-$, $\epsilon < L_D(h) = \mathbb{P}(x : h_-(x) \neq h_{x_0}(x)) = \mathbb{P}(\{x_0\})$. Entonces $\mathbb{P}(x : x \neq x_0) < 1 - \epsilon$.

$$D^m(\{S : L_p(h) > \epsilon\}) < (1 - \epsilon)^m < \delta \implies m = \frac{\ln(\delta)}{\ln(1 - \epsilon)}$$

Siguiendo hipótesis de realizabilidad, $X = \mathbb{R}^2$, $y = \{0, 1\}$, $H = \{h_r\}$

$$h_r(x) = \begin{cases} 1 & |x| < r \\ 0 & \text{othw} \end{cases}$$

S_m, entonces ERM $\implies$ círculo con menor radio que contiene a todos los puntos de clase 1. Fijamos ϵ, δ, queremos encontra $m \ni D^m(\{S : L_D(h) > \epsilon\})$. Sea h^* el disco que queremos aprender. Sea h_ϵ el disco tal que $\mathbb{P}(x : x \in h^* - h_\epsilon) = \epsilon$. Sea S_m una muestra tal que $L_D(h) > \epsilon$. Esto significa que ningún punto en $\{x_1, \ldots, x_m\}$ está en

$h^* - h_\epsilon$. La probabilidad de que $x \in h^* - h_\epsilon = \epsilon \implies \mathbb{P}(x : x \notin h^* - h_\epsilon) = 1 - \epsilon \implies$ la probabilidad de formar S^m donde ningún punto está en $h^* - h_\epsilon$ es $(1-\epsilon)^m < \delta \implies$

$$m < \frac{\ln \delta}{\ln(1-\epsilon)}$$

Pero $(1-\epsilon)^m < e^{-m\epsilon} < \delta \implies m > \frac{\ln(1/\delta)}{\epsilon}$.

5 Predictores Lineales

Nota 5.1. Algunas clases de hipótesis son espacios mitad, regresión lineal y regresión logística. Los algoritmos de clasificación son programación lineal, regresión lineal y regresión logística.

Definición 5.1. Definimos a la clase de funciones afines en $\mathbb{R}^d$ como

$$L_d = \{h_{w,b} : w \in \mathbb{R}^d,\ b \in \mathbb{R}\}$$

tal que $h_{w,b} : \mathbb{R}^d \to \mathbb{R}$, donde

$$h_{w,b}(x) = \langle w, x \rangle + b = \sum_{i=1}^{d} w_i x_i + b$$

Entonces, $L_d = \{x \mapsto \langle w, x \rangle + b,\ x, w \in \mathbb{R}^d,\ b \in \mathbb{R}\}$. Podemos absorber a b en el vector w. Definimos

$$w' = (b, w_1, \ldots, w_d), \quad x' = (1, x_1, \ldots, x_d), \quad h_{w,b}(x) = \langle w', x' \rangle$$

5.1 Espacios mitad

Definición 5.2. Dada la labor de clasificación binaria, $X = \mathbb{R}^d \to Y = \{\pm 1\}$. Definimos la clase de hipótesis HS_d (half-spaces) como

$$HS_d = \{h : \mathbb{R}^d \to \{\pm 1\} \ni h(x) = sign(h_{w,b}(x)),\ h_{w,b} \in L_d\}$$

Sean $S = \{(x_1, y_1), \ldots, (x_m, y_m)\},\ x_i \in \mathbb{R}^d,\ y_i \in \{\pm 1\}$. Como estamos suponiendo realizabilidad, ERM $\to h_S$ y $L_S(h_S) = 0$. Por otro lado, $h_w(x) - \langle w, x \rangle \implies \mathrm{sign}(\langle w, x_i \rangle) = y_i,\ \forall i = 1, \ldots, m$. Nótese que, excluyendo puntos frontera, $y_i \langle w, x_i \rangle > 0$.

Definición 5.3. Definimos $\gamma = \min_i \langle w, x_i \rangle$ y $\tilde{w} = \frac{w^*}{\gamma}$, donde $w^* = \mathrm{EMR}$. Entonces,

$$y_i \langle \tilde{w}, x_i \rangle = \tfrac{1}{\gamma} \langle w^*, x_i \rangle \geq 1,\ \forall i = 1, \ldots, m \implies A = y_i x_{ij} \text{ y } v = \begin{bmatrix} 1 \\ \vdots \\ 1 \end{bmatrix} \implies A\tilde{w} \geq v.$$

5.2 Perceptrón

Definición 5.4. Si tenemos $S = \{(x_1, y_1), \ldots, (x_m, y_m)\}$. Sea $w^{(1)} = (0, \ldots, 0)$, si $\exists (x_i, y_i)$ mal clasificado, i.e. $y_i \langle x_i, w \rangle < 0$, entonces $w^{(1)} = w^{(0)} + y_i x_i$. En general,

$$w^{(t+1)} = w^{(t)} + x_i y_i \tag{24}$$

Podemos probar que

$$\begin{aligned} y_i \langle x_i, x^{(t+1)} \rangle &= y_i \langle x_i, w^{(t)} + y_i x_i \rangle \\ &= y_i \langle x_i, w^{(t)} \rangle + \|x_i\|^2 \end{aligned}$$

Sea $R = \max_i |x_i|$ y

$$B = |\inf\{w : y_i \langle w, x_i \rangle > 1, \ \forall i\}|$$

El número de pasos máximo del algoritmo de perceptrón es RB^2.

5.3 Regresión lineal

Definición 5.5. Sean $X \subseteq \mathbb{R}^d$, $Y \subseteq \mathbb{R}$. La clase de hipótesis es

$$H_{REG} = \{x \mapsto \langle w, x \rangle + b, \ w \in \mathbb{R}^d, \ b \in \mathbb{R}\}$$

Definición 5.6. La función de pérdida continua es $\ell(h, (x, y)) = (h(x) - y)^2$.

Definición 5.7. La función de riesgo es $L_S(h) = \sum_{i=1}^m (h(x_i) - y_i)^2$.

Definición 5.8. La función de mínimos cuadrados es

$$\arg\min_w L_S(h_w) = \arg\min_w \frac{1}{m} \sum_{i=1}^m (\langle w, x_i \rangle - y_i)^2 \tag{25}$$

Nota 5.2. En la definición anterior,

$$\frac{\mathrm{d}}{\mathrm{d}w} L_S(h_w) = \frac{2}{m} \sum_{i=1}^m (\langle w, x_i \rangle - y_i) x_i$$

Definimos $(x_{ai}) = (\vec{x}_1, \vec{x}_2, \ldots, \vec{x}_m)$. Entonces

$$\begin{aligned} 0 = \sum_{i=1}^m (\langle w, x_i \rangle - y_i) x_{bi} &= \sum_{i=1}^m \sum_{a=1}^d (w_a x_{ai} - y_i) x_{bi} \\ &= \sum_{i=1}^m \sum_{a=1}^d x_{bi} x_{ia}^T w_a - x_{bi} y_i = (X X^T w)_b - (Xy)_b \end{aligned}$$

Lo cual implica que $\underbrace{XX^T}_{A} w = \underbrace{Xy}_{b} \implies Aw = b$. Como $A = XX^T$, entonces A es simétrica. Entonces, $A = VDV^T \implies A^{-1} = VD^{-1}V^T$, pero si D no tiene inversa, definimos $D^+ \ni D^+_{ii} = 0$ ssi $D_{ii} = 0$ y $D^+_{ii} = \frac{1}{D_{ii}}$ ssi $D_{ii} \neq 0$. Entonces, $A^+ = VD^+V^T$. Definimos $\hat{w} = A^+b$, entonces

$$\begin{aligned} A\hat{w} &= (VDV^T)A^+b = (VDV^T)VD^+V^Tb \\ &= VDD^+V^Tb = \sum_{i:D_{ii}\neq 0} V_iV_i^Tb = b \end{aligned}$$

5.4 Regresión logística

Definición 5.9. Sea $h : \mathbb{R}^d \to [0,1],\ X = \mathbb{R}^d,\ y = \{\pm 1\}$. Y $h(x_i)$ es la probabildiad de que y_i sea 1. Sea

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

y

$$H_{sig} = \sigma \circ L_d = \{x \mapsto \sigma(\langle w, x\rangle + b) : w \in \mathbb{R}^d,\ b \in \mathbb{R}\}$$

Y definimos su función de pérdida

$$\ell(h,(x,y)) = \ln\big(1 + \exp\{-y\langle w, x\rangle\}\big) \tag{26}$$

con

$$L_S(h) = \frac{1}{m}\sum_{i=1}^{m} \ln\big(1 + \exp\{-y_i\langle w, x_i\rangle\}\big)$$

6 VC dimension

En términos de conjuntos, para todo $B \subset A,\ \exists X_n \cap A = B$.

Ejemplo 6.1. $X = \mathbb{R}^2,\ y = \{0,1\},\ H$ rectángulos alineados con ejes $\mathrm{VCdim}(H) = 4$. No puede ser 5 dado que, si considera los puntos con coordenadas min/max en x, y, el quinto punto se encuentra necesariamente dentro del cuadrado.

Ejemplo 6.2. $X = \mathbb{R}^n$ y clase de hipótesis anterior. Entonces $\mathrm{VCdim}(H) = 2n$ con

$$x_i = (0, \dots, 0, \pm 1, 0, \dots, 0)$$

Propiedad 6.1. Si $H \subset H'$, entonces $\mathrm{VCdim}(H) \leq \mathrm{VCdim}(H')$.

Propiedad 6.2. $2^n = |H| \geq 2^{\mathrm{VC\,dim}(H)} \implies n \geq \mathrm{VC\,dim}(H)$.

Ejemplo 6.3. Sea $X = \{0,1\}^n$. Sea $I \subseteq \{1,2,\ldots,n\}$. Definimos una función de paridad asociada a I. Sea $(x_1,\ldots,x_n),\ h_I(x) = \sum_{i\in I} x_i \mod 2$. Entonces

$$H_{n\text{-parity}} = \{h_I : I \subseteq \{1,\ldots,n\}\}$$

Entonces para $h_I = \emptyset$ funciona en $\mathbb{R}^3$. Ahora,

$$h_I(e_k) = \sum_{i\in I} x_i \mod 2 = 1$$

y es 0 para el resto. Sea B un subconjunto de $\{e_i,\ i = 1,\ldots,n\},\ I = \{i : e_i \in B\}$

$$h_I(e_i) = \begin{cases} 1 & i \in I \\ 0 & i \notin I \end{cases}$$

6.1 Teorema Fundamental del aprendizaje PAC

Teorema 6.1. Sea H una clase de hipótesis, $H = \{h : X \to \{0,1\}\}$. Sea la función de pérdida la función 0–1. Entonces son equivalentes

1. H tiene convergencia uniforme.
2. Cualquier algoritmo ERM es un aprendiz PAC agnóstico exitoso para H.
3. H es PAC agnóstico aprendible.
4. H es PAC aprendible.
5. Cualquier algoritmo ERM es un aprendiz PAC exitoso para H.
6. H tiene dimensión VC finita.

6.2 Función de crecimiento

Definición 6.1. Sea H una clase de hipótesis. La función de crecimiento de H, $\tau_H : \mathbb{N} \to \mathbb{N}$ se define como

$$\tau_H(m) = \max_{C\subset X : |C| = m} |H_c|$$

Ejemplo 6.4. Si $\text{VCdim}(H) = d$, para $m \leq d,\ \tau_H(m) = 2^m$.

Lema 6.1. Sea H una clase de hipótesis con dimensión $\text{VCdim}(H) \leq d < \infty$. Entonces, para todo m

$$\tau_H(m) \leq \sum_{i=0}^{d} \binom{m}{i}$$

En otras palabras, para $m > d+1$,

$$\tau_H(m) \leq \left(\frac{em}{d}\right)^d$$

Teorema 6.2. Sea H una clase de hipótesis y sea τ_H su función de crecimiento. Entonces, para cada D y cada $\delta \in (0,1)$, tenemos

$$|L_D(h) - L_S(h)| \leq \frac{4 + \sqrt{\ln \tau_H(2m)}}{\delta\sqrt{2m}}$$

con probabilidad $1 - \delta$.

Nota 6.1. H dimensión VC$= d$. Para $m > d,\ \tau_H(m) \leq (em/d)^d$. Entonces

$$|L_D(h) - L_S(h)| \leq \frac{4 + \sqrt{d \ln\big(em/d\big)}}{\delta\sqrt{2m}}$$

7 Aprendizaje no uniforme

Definición 7.1. Una clase de hipótesis es no-uniformemente aprendible si existe un algoritmo A y una función $m_H^{NU} : (0,1) \times (0,1) \times H \to \mathbb{N}$ tal que si $\epsilon, \delta \in (0,1)$ y $hinH$, entonces para toda distribución D, se tiene que

$$L_D(A(S)) \leq L_D(h) + \epsilon$$

Nota 7.1. Comparación con esquema PAC agnóstico.

Ejemplo 7.1. Sea $X = \mathbb{R},\ y = \{\pm 1\}$. $H = \{h_p\},\ h_p(x) = sgn(p(x))$. VCdim$(H) \to \infty$. Si $H_2 = \{h_p\},\ h_{p(x)} = sgn(a + bx + cx^2)$, VCdim$(H_2) = n + 1$.

Teorema 7.1. Sea H una clase de hipótesis que satisfaga

1. $H = \bigcup_{n\in\mathbb{N}} H_n$
2. Cada H_n tiene la propiedad de convergencia.

entonces H es no-uniformemente aprendible.

Teorema 7.2. Una clase de hipótesis es no-uniformemente aprendible ssi H es una unión contable de clases de hipótesis PAC agnósticas.

Proof. Si $H = \bigcup_{h\in H} H_n$, donde H_n es PAC-aprendible, entonces por el TFAS, H_n tiene la propiedad de convergencia uniforme y H es NU. □

7.1 Minimización estructural de riesgo

Definición 7.2. Sea H una clase no uniformemente aprendible. Entonces, $H = \bigcup_n H_n$ donde H_n tiene la propiedad de convergencia uniforme. Definimos $\epsilon_n : \mathbb{N} \times (0,1) \to (0,1)$

$$\epsilon_n(m, \delta) = \min\{\epsilon \in (0,1) : m_{H_n}^{UC}(\epsilon, \delta) \leq m\}$$

Y ϵ_n es el mínimo error que se puede alcanzar aprendiendo H_n con ciertos ejemplos y cierta confianza. Entonces

$$|L_D(h) - L_s(h)| < \epsilon_n(m, \delta), \quad \forall h \in H_n$$

con probabilidad $1 - \delta$. Sea $w : \mathbb{N} \to [0, 1]$ una función tal que $\sum_n w(n) \leq 1$. Esta es una función que le asigna un peso a $H_1, \ldots$.

Teorema 7.3. Sea w una función de peso. Sea H una clase de hipótesis tal que $H = \bigcup_n H_n$ donde cada H_n tiene convergencia uniforme. Entonces, para cada $\delta \in (0, 1)$ y distribución con probabilidad $1-\delta$ sobre $S \sim D^m$, se satisface para cada $n \in \mathbb{N}$ y $h \in H_n$

$$|L_D(h) - L_s(H)| \leq \epsilon_n(m, w(n)\delta)$$

Entonces, para cada $\delta \in (0, 1)$ y distribución D con probabilidad al menos $1-\delta$, $\forall h \in H$

$$L_D(h) \leq L_S(h) + \min_{n:h\in H_n} \epsilon(m, w(n)\delta)$$

Proof. Para cada n, tomamos $\delta_n = w(n)\delta$. Entonces, como H_n tiene convergencia uniforme,

$$\forall h \in H_n, \quad |L_D(h) - L_S(h) \leq \epsilon_n(m, \delta_n)$$

Entonces, existe $h \in H_n$,

$$|L_D(h) - L_S(h)| > \epsilon_n(m, \delta_n)$$

con probabilidad δ_n. Por lo tanto

$$\begin{aligned}\mathbb{P}_{S\sim D^m}[\exists h \in H : |L_D(h) - L_S(h)| > \epsilon_n \forall n] &\leq \sum_n \mathbb{P}[\exists h \in H_n : |L_D(h) - L_S(h)| > \epsilon_n] \\ &\leq \sum_n \delta_n = \sum_n \delta w(n) \leq \delta\end{aligned}$$

Entonces, con probabilidad $1 - \delta$, sobre $S \sim D^m$, $\forall h \in H$

$$L_D(h) \leq L_S(h) + \min_{n:h\in H_n} \epsilon(m, w(n)\delta) = L_S(h) + \epsilon_{n(h)}(m, w(n)\delta)$$

si definimos $n(h) = \min\{n : h \in H_n\}$. □

Nota 7.2. Lo que sé es $H = \bigcup_n$ con H_n convergencia uniforme, m_H^{UC}, w función de peso, $\sum_n w(n) \leq 1$. Definimos ϵ_n, $n(h)$. La entrada es $S \sim D^m, \delta$. Y la salida es

$$h \in \arg\min_{h\in H} L_S(h) + \epsilon_{n(h)}(m, w(n(h))\delta)$$

Teorema 7.4. Sea H una clase de hipótesis tal que $H = \bigcup_n H_n$ con cada H_n con convergencia uniforme. Entonces, usando SRM, H es no-uniformemente aprendible, con complejidad

$$m_H^{NUL}(\epsilon, \delta, h) \leq m_{H_{n(h)}}^{UC}(\epsilon/2, w(n)\delta)$$

donde w es la función de peso.

Proof. Usando SRM como nuestro algoritmo A, respecto a una función de peso w. Sean, $\epsilon, \delta \in (0,1)$ y $h \in H$. Sea $m > m_{H_{n(h)}}^{UC}(\epsilon, w(n(h))\delta)$. Entonces, con probabilidad $1-\delta$, usando el teorema anterior

$$L_D(h) \leq L_S(h) + \epsilon_{n(h)}(m, w(n(h))\delta)$$

En particular para el resultado $A(S)$ de SRM,

$$\begin{aligned} L_D(A(S)) &\leq \min_{h' \in H}[L_S(h') + \epsilon_{n(h')}(m, w(n(h'))\delta)] \\ &\leq L_S(h) + \epsilon_{n(h)}(m, w(n(h))\delta) \end{aligned}$$

Si $m \geq m_{H_{n(h)}}^{UC}(\epsilon/2, w(n(h))\delta)$, entonces $\epsilon_{n(h)}(m, w(n(h))\delta) \leq \epsilon/2$. Entonces, como $H_{n(h)}$ tiene convergencia uniforme, $L_S(h) \leq L_D(h) + \epsilon/2$ con probabilidad $1-\delta$, entonces

$$L_D(A(S)) \leq L_D(h) + \epsilon/2 + \epsilon/2, \forall h \in H$$

□

7.2 Longitud de mínimo descripción

H es una clase de hipótesis numerable

$$H = \bigcup_{n \in \mathbb{N}} H_n = \bigcup_{n \in \mathbb{N}} \{h_n\}$$

Entonces

$$\theta = \frac{1}{m} \sum_{i=1}^{m} \ell(h, z_i) \tag{27}$$

tal que

$$\mathbb{E}_{z \sim D}[\ell(h, z)] = L_D(h)$$

y

$$\mathbb{P}[|\frac{1}{m} \sum_{i=1}^{m} \ell(h, z_i) - L_D(h)| > \epsilon] \leq 2e^{-2m\epsilon^2} < \delta \tag{28}$$

Cada clase $\{h_m\}$ tiene complejidad de muestra

$$m_{H_n}^{UC}(\epsilon, \delta) = \frac{\ln(2/\delta)}{2\epsilon^2}$$

En este caso, podemos ver a $w : H \to [0,1]$ y SRM

$$h \in \arg\min_{h' \in H} L_S(h') + \sqrt{\frac{-ln(w(h)) + \ln\big(2/\delta\big)}{2m}}$$

Vamos a asignar pesos w en relación con la representación de las hipótesis. Sea H una clase de hipótesis y sea Σ un conjunto finito de caracteres $\Sigma = \{0,1\}$. Una palabra es una secuencia finita de símbolos en Σ y Σ^* es el conjunto de símbolos de longitud finita en Σ. Por ejemplo, $\sigma = \{10010\} \in \Sigma^*$, con $|\sigma| = 5$. Entonces, Σ^* es libre de prefijos si, dados $\sigma, \sigma' \in \Sigma^*$, entonces σ no es un prefijo de σ'. Definimos un lenguaje para representar a H como un mapeo $d : H \to \Sigma^*$, dada h tenemos su representación $d(h)$.

Lema 7.1 (Desigualdad de Kraft)**.** Sea $S \subset \{0,1\}^*$ libre de prefijos, entonces

$$\Sigma_{\sigma \in S} \frac{1}{2^{|\sigma|}} \leq 1 \tag{29}$$

Dada la palabra σ, $P(\sigma)$ es la probabilidad de formar la palabra σ, entonces

$$P(\sigma) = \frac{1}{2^{|\sigma|}}$$

Como las probabilidades suman hasta 1

$$\sum_{\sigma \in S} \frac{1}{2^{|\sigma|}} \leq 1$$

y

$$h \in \arg\min_{h \in H} L_S(h) + \sqrt{\frac{|d(h) + \ln\big(2/\delta\big)|}{2m}}, \quad w(h) = \frac{1}{2^{|d(h)|}}$$

Teorema 7.5. Sea H una clase de hipótesis numerable y $d : H \to \{0,1\}^*$ un lenguaje para representar a H libre de prefijos. Entonces, para cada m, δ, D, con probabilidad no menor que $1 - \delta$, tenemos que

$$\forall h \in H, \quad L_D(h) \leq L_S(h) + \sqrt{\frac{|d(h)| + \ln\big(2/\delta\big)}{2m}}$$

Teorema 7.6 (Desigualdad de Hoeffing)**.** Si tenemos variables aleatorias $X_1, \ldots, X_m$ tales que $\mathbb{E}[x_i] = \mu$ y definimos

$$\theta = \frac{1}{m} \sum_{i=1}^{m} X_i$$

entonces

$$\mathbb{P}[|\frac{1}{m} \sum_{i=1}^{m} X_i - \mu| > \epsilon] \leq 2$$

7.3 Longitud de mínima descripción (MDL)

Lo que sabemos, H clase numerable, $d : H \to \{0,1\}$, $|d(h)| = |h|$. La entrada es $S \sim D^m$, δ. y salida

$$h \in \arg\min_{h \in H} L_S(h) + \sqrt{\frac{|d(h) + \ln\big(2/\delta\big)|}{2m}}$$

8 Complejidad computacional de aprendizaje

Un algoritmo A. Tenemos un dominio $Z = X \times Y$, una clase de hipótesis H, una función de pérdida $\ell : H \times Z \to \mathbb{R}$, $S \sim D^m$ extraída de Z según una distribución D de manera iid. Queremos que A genere $h \in H$ tal que, dados ϵ, δ,

$$L_D(h) \leq \min_{h' \in H} L_D(h) + \epsilon$$

con probabilidad $1 - \delta$. Complejidad de muestra

$$m_H^{ag}(\epsilon, \delta) = \frac{2ln(2|H|/\delta)}{\epsilon^2}$$

Definición 8.1 (Complejidad computacional de aprendizaje estadístico)**.** Dada una función $f : (0,1)^2 \to \mathbb{N}$, un problema de aprendizaje $(X \times Y, H, \ell)$ y un algoritmo A, decimos que A resuelve el problema de aprendizaje en un tiepo $\mathcal{O}(f)$ si existe un número c tal que, dados $\epsilon, \delta \in (0,1)$ y para toda D,

1. A termina en un tiempo no mayor a $cf(\epsilon, \delta)$.
2. El resultado de A puede ser usado para predecir la etiqueta de un ejemplo nuevo.
3. El resultado de A es PAC.

Dada una secuencia $(X_n \times Y_n, H_n, \ell_n)_{n=1}^{\infty}$ de problemas de apredizaje y un algoritmo A aplicable a esta secuencia, una función $g : \mathbb{N} \times (0,1)^2 \to \mathbb{N}$, decimos que la complejidad computacional es $\mathcal{O}(g)$ su para todo n el problema se resuelve en un tiepo $\mathcal{O}(f_n)$, donde $f_n(\epsilon, \delta) = g(n, \epsilon, \delta)$. Por otro lado,

$$\begin{aligned} g(d, \epsilon, \delta) &= dm_H(\epsilon, \delta) \\ &= \frac{d \ln\big(|H|/\delta\big)}{\epsilon} \end{aligned}$$

Entonces decimos que A es eficiente si su complejidad es $\mathcal{O}(p(n, \frac{1}{\epsilon}, \frac{1}{\delta}))$.

9 Selección y validación de modelo

Regresión de $L : \mathbb{R} \to \mathbb{R}$

Usando SRM, $H_1, H_2, \ldots$

$$m_{H_d}^{UC}(\epsilon, \delta) \leq \frac{g(d) \ln\big(1/\delta\big)}{\epsilon^2}$$

Con $g : \mathbb{N} \to \mathbb{N}$ monótona creciente. Usando SRM, tenemos una hipótesis H que minimiza

$$L_D(h) \leq L_S(h) + \sqrt{\frac{g(d)(\ln\big(1/\delta\big) + 2\ln d + \ln\big(\pi^2/b\big)}{m}}$$

con probabilidad $1 - \delta$ sobre $S \sim D^m$.

9.1 Validación

Sea $V = \{(x_1, y_1), \ldots, (x_{m_v}, y_{m_v})\}$ conjunto de validación y $S = \{(x_1, y_1), \ldots, (x_m, y_m)\}$.

Teorema 9.1. Sea h un predictor cualquiera y $/ell : H \times Z \to [0, 1]$ una función de pérdida. Entonces, si $\delta \in (0, 1)$, tenemos que

$$|L_D(h) - L_v(h)| \leq \sqrt{\frac{\ln\big(2/\delta\big)}{m_v}} \tag{30}$$

con probabilidad $1 - \delta$ sobre $V\mathrm{D}^m$.

Proof. Sea h una hipótesis. Entonces $\ell(h, z)$ es una variable aleatoria. Usando el lema de Hoeffding,

$$\mathbb{P}_{V \sim D^{m_v}}[|\frac{1}{m_v} \sum_{i=1}^{m_v} \ell(h, z_i) - L_D(h)| > \sqrt{\frac{\ln\big(2/\delta\big)}{m_v}}]$$

□

Podemos usar validación para seleccionar un modelo m. Si tenemos $H_1, \ldots, H_r$, definimos $H = \{h_1, \ldots, h_r\}$ como los resultados de nuestro algoritmo de aprendizaje sobre cada una de las clases y usamos los datos en V para aprender sobre H.

Teorema 9.2. Sea $H = \{h_1, \ldots, h_r\}$, ℓ una función de pérdida con valores en $[0, 1]$. Usamos V con m ejemplos. Entonces tenemos

$$\forall h \in,\ |L_D(h) - L_V(h)| \leq \sqrt{\frac{\ln\big(2|H|/\delta\big)}{2m_V}}$$

con probabilidad al menos $1 - \delta$.

Proof. $|L_D(h) - L_V(h)| \leq \sqrt{\frac{\ln\big(2|H|/\delta\big)}{2m_V}}$ con probabilidad $1 - \frac{\delta}{|H|}$ □

9.2 Cross validation

entrada

S datos de entrenamiento

parámetros

algoritmo A

entero k

Se hace una partición de S en $S_1, \ldots, S_k$. Para cada parámetro $\theta \in \Theta$

para cada $i \in [1, \ldots, k]$

$h_{i,\theta} = A(S/S_{i,\theta})$

$L(\theta) = \sum_{i=1}^{k} L_{S_i}(h_{i,\theta})$

salida

$\theta^* \in \arg\min(L(\theta))$

$h_{\theta^*} = A(S, \theta^*)$

Nota 9.1. Note que

$$L_D(h) = \underbrace{L_D(h^*)}_{\text{error de aprox.}} + \underbrace{L_D(h) - L_D(h^*)}_{\text{error de est.}}$$

El error de aproximación depende de D, H y no depende de S. Se busca aumentar H, que sea más complejo. El error de estimación mejora al aumentar S, reducir H.

Nota 9.2. Ahora,

$$L_D(h_S) = (L_D(h_S) - L_V(h_S)) + \underbrace{L_V(h_S) - L_S(h_S)}_{\text{sobreajuste}} + \underbrace{L_S(h_S)}_{\text{subajuste}}$$

Ejemplo 9.1.

$$\sum_{i=0}^{\infty} x^2 - y - \alpha + \aleph_0$$

dim VC aprendizaje no uniforme