Universidad Nacional Autónoma de México.
IIMAS

Aprendizaje de Maquina

Semestre 2026-1.

D.C.C. Carlos Ignacio Hernández Castellanos

José Alberto Alonso González

Tarea 1

Integrantes:

- Villalón Pineda Luis Enrique .

1. (**10** puntos) Monotonía de la Complejidad de Muestra: Sea $\mathcal{H}$ una clase de hipótesis para una tarea de clasificación binaria. Supón que $\mathcal{H}$ es PAC-aprendible y que su complejidad de muestra está dada por $m_{\mathcal{H}}(\cdot,\cdot)$. Demuestre que $m_{\mathcal{H}}$ es monótonamente no creciente en cada uno de sus parámetros. Es decir:

- Si $0 < \epsilon_1 \leq \epsilon_2 < 1$, entonces $m_{\mathcal{H}}(\epsilon_1, \delta) \geq m_{\mathcal{H}}(\epsilon_2, \delta)$.
- Si $0 < \delta_1 \leq \delta_2 < 1$, entonces $m_{\mathcal{H}}(\epsilon, \delta_1) \geq m_{\mathcal{H}}(\epsilon, \delta_2)$.

Demostración:

Demostremos que la complejidad de la muestra disminuye de forma monótona en el parámetro de precisión ϵ. La prueba de que la complejidad de la muestra disminuye de forma monótona en el parámetro de confianza δ es análoga. Denotemos por $\mathcal{D}$ una distribución desconocida sobre $\mathcal{X}$, y sea $f \in \mathcal{H}$ la hipótesis objetivo. Denotemos por A un algoritmo que aprende $\mathcal{H}$ con complejidad de muestra $m_{\mathcal{H}}(\cdot,\cdot)$. Fijemos algún $\delta \in (0,1)$. Supongamos que $0 < \epsilon_1 \leq \epsilon_2 \leq 1$. Necesitamos demostrar que $m_1 \overset{\text{def}}{=} m_{\mathcal{H}}(\epsilon_1, \delta) \geq m_{\mathcal{H}}(\epsilon_2, \delta) \overset{\text{def}}{=} m_2$. Dada una secuencia de entrenamiento i.i.d. de tamaño $m \geq m_1$, tenemos que, con una probabilidad de al menos $1-\delta$, A devuelve una hipótesis h tal que

$$L_{\mathcal{D},f}(h) \leq \epsilon_1 \leq \epsilon_2$$

Por la minimalidad de m_2, concluimos que $m_2 \leq m_1$.

2. (10 puntos) Valor Esperado del Riesgo Empírico: Sea $\mathcal{H}$ una clase de clasificadores binarios sobre un dominio $\mathcal{X}$. Sea $\mathcal{D}$ una distribución desconocida sobre $\mathcal{X}$ y f la hipótesis verdadera. Fijado $h \in \mathcal{H}$, muestra que el valor esperado del error empírico $L_S(h)$ sobre la elección de S es igual al riesgo verdadero $L_{(\mathcal{D},f)}(h)$:

$$\mathbb{E}_{S|x\sim\mathcal{D}^m}\left[L_S(h)\right] = L_{(\mathcal{D},f)}(h).$$

Demostración:

Por la linealidad del valor esperado:

$$\begin{aligned}
\mathbb{E}_{S|x\sim\mathcal{D}^m}[L_S(h)] &= \mathbb{E}_{S|x\sim\mathcal{D}^m}\left[\frac{1}{m}\sum_{i=1}^{m} \mathbb{1}_{\{h(x_i)\neq f(x_i)\}}\right] \\
&= \frac{1}{m}\sum_{i=1}^{m} \mathbb{E}_{x_i\sim\mathcal{D}}[\mathbb{1}_{\{h(x_i)\neq f(x_i)\}}] \\
&= \frac{1}{m}\sum_{i=1}^{m} \mathbb{P}_{x_i\sim\mathcal{D}}[h(x_i) \neq f(x_i)] \\
&= \frac{1}{m}\cdot m \cdot L_{(\mathcal{D},f)}(h) \\
&= L_{(\mathcal{D},f)}(h)
\end{aligned}$$

3. (5 puntos) Círculos Concéntricos: Sea $\mathcal{X} = \mathbb{R}^2, \mathcal{Y} = \{0,1\}$ y la clase de hipótesis $\mathcal{H}$ definida como:

$$\mathcal{H} = \left\{ h_r : r \in \mathbb{R}_+, h_r(x) = \mathbf{1}_{\|x\| \leq r} \right\}.$$

Demuestre que $\mathcal{H}$ es PAC-aprendible (bajo el supuesto de realizabilidad) y su complejidad de muestra por:

$$m_{\mathcal{H}}(\epsilon, \delta) \leq \left\lceil \frac{\log(1/\delta)}{\epsilon} \right\rceil.$$

Demostración:

Consideremos el algoritmo ERM A que, dada una secuencia de entrenamiento $S = ((\mathbf{x}i, y_i))\, i = 1^m$, devuelve la hipótesis $\hat{h}$ correspondiente al círculo «más ajustado» que contiene todas las instancias positivas. Denotemos el radio de esta hipótesis por $\hat{r}$. Supongamos que es realizable y sea $h^\star$ un círculo con error de generalización cero. Denotemos su radio por $r^\star$. Sea $\epsilon, \delta \in (0,1)$. Sea $\bar{r} \leq r$ un escalar tal que $\mathcal{D}_{\mathcal{X}}(\{x : \bar{r} \leq \|\mathbf{x}\| \leq r\}) = \epsilon$. Definamos $E = \{\mathbf{x} \in \mathbb{R}^2 : \bar{r} \leq \|\mathbf{x}\| \leq r^\star\}$. La probabilidad (sobre el muestreo S) de que $L_{\mathcal{D}}(h_S) \geq \epsilon$ está limitada por la probabilidad de que ningún punto en S pertenezca a E. La probabilidad de este evento está limitada por

$$(1-\epsilon)^m \leq e^{-\epsilon m}$$

El límite deseado en la complejidad de la muestra se obtiene al requerir que $e^{-\epsilon m} \leq \delta$.

4. (5 puntos) Conjunciones Booleanas: Sea $\mathcal{X} = \{0,1\}^d$ y $\mathcal{Y} = \{0,1\}$. Sea $\mathcal{H}$ la clase de todas las conjunciones booleanas (positivas y negativas) sobre d variables. Asume realizabilidad. Demuestra que esta clase es PAC-aprendible y acota su complejidad de muestra. Propón un algoritmo que implemente la regla ERM y cuya complejidad sea polinomial en d y m.

   Demostracion :

   En primer lugar, observamos que $\mathcal{H}$ es finito. Calculemos su tamaño con precisión. Cada hipótesis, además de la hipótesis totalmente negativa, se determina decidiendo para cada variable x_i si $x_i, \bar{x}_i$ o ninguna de ellas aparece en la conjunción correspondiente. Por lo tanto, $|\mathcal{H}| = 3^d + 1$. Concluimos que $\mathcal{H}$ es PAC aprendible y su complejidad de muestra puede limitarse por

   $$m_{\mathcal{H}}(\epsilon, \delta) \leq \left\lceil \frac{d \log 3 + \log(1/\delta)}{\epsilon} \right\rceil$$

   Dado que el algoritmo tarda un tiempo lineal (en términos de la dimensión d) en procesar cada ejemplo, el tiempo de ejecución está limitado por $O(m \cdot d)$.

   Ahora veamos el algoritmo. Definimos $h_0 = x_1 \cap \bar{x}_1 \cap \ldots \cap x_d \cap \bar{x}_d$. Obsérvese que h_0 es la hipótesis siempre negativa. Sea $((\mathbf{a}^1, y^1), \ldots, (\mathbf{a}^m, y^m))$ una secuencia de entrenamiento i.i.d. de tamaño m. Dado que no podemos obtener ninguna información de los ejemplos negativos, nuestro algoritmo los ignora. Para cada ejemplo positivo a, eliminamos de h_i todos los literales que faltan en a. Es decir, si $a_i = 1$, eliminamos $\bar{x}_i$ de h y si $a_i = 0$, eliminamos x_i de h_i. Finalmente, nuestro algoritmo devuelve h_m. Por construcción y

realizabilidad, h_i etiqueta positivamente todos los ejemplos positivos entre $\mathbf{a}^1, \ldots, \mathbf{a}^i$. Por las mismas razones, el conjunto de literales en h_i contiene el conjunto de literales en la hipótesis objetivo. Por lo tanto, h_i clasifica correctamente los elementos negativos entre $\mathbf{a}^1, \ldots, \mathbf{a}^i$. Esto implica que h_m es un ERM.

5. (10 puntos) PAC Agnóstico: Sea $\mathcal{H}$ una clase de clasificadores binarios. Demuestra que si $\mathcal{H}$ es agnósticamente PAC-aprendible, entonces también es PAC-aprendible. Además, si un algoritmo A es un aprendiz agnóstico exitoso, también lo es para el caso PAC bajo realizabilidad.

   Demostracion:

   Supongamos que $\mathcal{H}$ es agnóstico PAC aprendible, y sea A un algoritmo de aprendizaje que aprende $\mathcal{H}$ con complejidad de muestra $m_{\mathcal{H}}(\cdot,\cdot)$. Demostramos que $\mathcal{H}$ es PAC aprendible utilizando A.

   Sea $\mathcal{D}, f$ una distribución (desconocida) sobre $\mathcal{X}$ y la función objetivo, respectivamente. Podemos suponer, sin pérdida de generalidad, que $\mathcal{D}$ es una distribución conjunta sobre $\mathcal{X} \times \{0,1\}$, donde la probabilidad condicional de y dada x se determina de forma determinista por f. Dado que asumimos la realizabilidad, tenemos ínf $h \in \mathcal{H}L\mathcal{D}(h) = 0$. Sea $\epsilon, \delta \in (0,1)$. Entonces, para cada entero positivo $m \geq m_{\mathcal{H}}(\epsilon,\delta)$, si equipamos A con un conjunto de entrenamiento S que consiste en m instancias i.i.d. etiquetadas por f, entonces con una probabilidad de al menos $1-\delta$ (sobre la elección de $S|_x$), devuelve una hipótesis h con

$$\begin{aligned} L_{\mathcal{D}}(h) &\leq \text{ínf}\, h' \in \mathcal{H}L\mathcal{D}\left(h'\right) + \epsilon \\ &= 0 + \epsilon \\ &= \epsilon \end{aligned}$$

6. (5 puntos) Predictor Bayesiano Óptimo: Demuestra que para toda distribución $\mathcal{D}$, el predictor bayesiano $f_{\mathcal{D}}$ minimiza el riesgo verdadero:

$$L_{\mathcal{D}}\left(f_{\mathcal{D}}\right) \leq L_{\mathcal{D}}(g), \qquad \text{para todo } g : \mathcal{X} \to \{0,1\}$$

   Demostración:

   Sea $x \in \mathcal{X}$. Sea α_x la probabilidad condicional de una etiqueta positiva dada x. Tenemos que

$$\begin{aligned} \mathbb{P}\left[f_{\mathcal{D}}(X)\ neqy \mid X = x\right] &= \mathbb{1}[\alpha_x \geq 1/2] \cdot \mathbb{P}[Y = 0 \mid X = x] + \mathbb{1}[\alpha_x < 1/2] \cdot \mathbb{P}[Y = 1 \mid X = x] \\ &= \mathbb{1}[\alpha_x \geq 1/2] \cdot (1 - \alpha_x) + \mathbb{1}[\alpha_x < 1/2] \cdot \alpha_x \\ &= \text{mín}\left\{\alpha_x, 1 - \alpha_x\right\}. \end{aligned}$$

Sea g un clasificador [1] de $\mathcal{X}$ a $\{0,1\}$. Tenemos que

$$\begin{aligned}
\mathbb{P}[g(X) \neq Y \mid X = x] &= \mathbb{P}[g(X) = 0 \mid X = x] \cdot \mathbb{P}[Y = 1 \mid X = x] \\
&\quad + \mathbb{P}[g(X) = 1 \mid X = x] \cdot \mathbb{P}[Y = 0 \mid X = x] \\
&= \mathbb{P}[g(X) = 0 \mid X = x] \cdot \alpha_x + \mathbb{P}[g(X) = 1 \mid X = x] \cdot (1 - \alpha_x) \\
&\geq \mathbb{P}[g(X) = 0 \mid X = x] \cdot \text{mín}\{\alpha_x, 1 - \alpha_x\} \\
&\quad + \mathbb{P}[g(X) = 1 \mid x] \cdot \text{mín}\{\alpha_x, 1 - \alpha_x\} \\
&= \text{mín}\{\alpha_x, 1 - \alpha_x\},
\end{aligned}$$

La afirmación se deduce ahora del hecho de que lo anterior es cierto para cada $x \in \mathcal{X}$. Más formalmente, por la ley de la esperanza total,

$$\begin{aligned}
L_{\mathcal{D}}(f_{\mathcal{D}}) &= \mathbb{E}(x,y) \sim \mathcal{D}[\mathbb{1}[f_{\mathcal{D}}(x) \neq y]] \\
&= \mathbb{E}x \sim \mathcal{D}_X[\mathbb{E}y \sim \mathcal{D}Y \mid x[\mathbb{1}[f_{\mathcal{D}}(x) \neq y] \mid X = x]] \\
&= \mathbb{E}_{x \sim \mathcal{D}_X}[\alpha_x] \\
&\leq \mathbb{E}x \sim \mathcal{D}_X[\mathbb{E}y \sim \mathcal{D}Y \mid x[\mathbb{1}[g(x) \neq y] \mid X = x]] \\
&= L_{\mathcal{D}}(g).
\end{aligned}$$

7. (5 puntos) Comparación de Algoritmos de Aprendizaje

a) Demuestre que para toda distribución generadora de datos $\mathcal{D}$ sobre $\mathcal{X} \times \{0,1\}$, el predictor bayesiano minimiza el riesgo con respecto a la pérdida $|h(x) - y|$ entre todos los predictores probabilísticos.

b) Demuestre que para toda distribución $\mathcal{D}$, existe un algoritmo $A_{\mathcal{D}}$ que es mejor que cualquier otro algoritmo de aprendizaje en términos del riesgo.

c) Demuestre que para cada algoritmo de aprendizaje A, existe una distribución $\mathcal{D}$ y un algoritmo B tal que A no es mejor que B respecto a $\mathcal{D}$.

Demostraciones:
(a) Esto se demostró en el ejercicio anterior.
(b) En el ejercicio anterior demostramos que, para cada distribución $\mathcal{D}$, el predictor óptimo bayesiano $f_{\mathcal{D}}$ es óptimo con respecto a $\mathcal{D}$.
(c) Elija cualquier distribución $\mathcal{D}$. Entonces, A no es mejor que $f_{\mathcal{D}}$ con respecto a D.