Universidad Nacional Autónoma de México.
IIMAS

Aprendizaje de Máquina

Semestre 2026-1.

D.C.C. Carlos Ignacio Hernández Castellanos

José Alberto Alonso González

Tarea 1

Integrantes:

- Villalón Pineda Luis Enrique .

Ejercicios y demostraciones

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:

   Sea A un algoritmo PAC que aprende $\mathcal{H}$ con complejidad de muestra $m_{\mathcal{H}}(\epsilon, \delta)$. Fijemos $\delta \in (0,1)$ y supongamos $0 < \epsilon_1 \leq \epsilon_2 < 1$. Por definición de $m_{\mathcal{H}}(\epsilon_1, \delta)$, para cualquier $m \geq m_{\mathcal{H}}(\epsilon_1, \delta)$, si S es una muestra i.i.d. de tamaño m etiquetada por la hipótesis objetivo realizable, entonces con probabilidad al menos $1-\delta$ (sobre la elección de $S|_x$) el algoritmo A devuelve una hipótesis h tal que

   $$L_{\mathcal{D}}(h) \leq \epsilon_1.$$

   Pero entonces también se cumple $L_{\mathcal{D}}(h) \leq \epsilon_2$ (porque $\epsilon_1 \leq \epsilon_2$). Por la definición mínima de $m_{\mathcal{H}}(\epsilon_2, \delta)$ (es el mínimo m que asegura error $\leq \epsilon_2$ con prob. $1-\delta$), necesariamente

   $$m_{\mathcal{H}}(\epsilon_2, \delta) \leq m_{\mathcal{H}}(\epsilon_1, \delta),$$

   lo que demuestra la primera desigualdad.

   La prueba para el parámetro de confianza δ es análoga: si se requiere una mayor confianza (es decir, $\delta_1 \leq \delta_2$), el tamaño muestral mínimo para garantizarla no puede disminuir; formalmente, para $m \geq m_{\mathcal{H}}(\epsilon, \delta_1)$ la probabilidad de éxito es al menos $1-\delta_1 \geq 1-\delta_2$, así que
   textbfDemostración:

   Por definición, $m_{\mathcal{H}}(\epsilon, \delta)$ denota el *mínimo* entero m tal que, para toda distribución $\mathcal{D}$ y toda muestra $S \sim \mathcal{D}^m$, se cumple que con probabilidad al menos $1-\delta$, el algoritmo A devuelve una hipótesis h con error de generalización

   $$L_{\mathcal{D}}(h) \leq \epsilon.$$

   Sea A un algoritmo PAC que aprende $\mathcal{H}$ con complejidad de muestra $m_{\mathcal{H}}(\epsilon, \delta)$. Fijemos $\delta \in (0,1)$ y supongamos $0 < \epsilon_1 \leq \epsilon_2 < 1$. Por definición de $m_{\mathcal{H}}(\epsilon_1, \delta)$, para cualquier $m \geq m_{\mathcal{H}}(\epsilon_1, \delta)$, si S es una muestra i.i.d. de tamaño m etiquetada por la hipótesis objetivo realizable, entonces con probabilidad al menos $1-\delta$ (sobre la elección de $S|_x$) el algoritmo A devuelve una hipótesis h tal que

   $$L_{\mathcal{D}}(h) \leq \epsilon_1.$$

   Pero entonces también se cumple $L_{\mathcal{D}}(h) \leq \epsilon_2$ (porque $\epsilon_1 \leq \epsilon_2$). Por la definición mínima de $m_{\mathcal{H}}(\epsilon_2, \delta)$ (es el mínimo m que asegura error $\leq \epsilon_2$ con prob. $1-\delta$), necesariamente

   $$m_{\mathcal{H}}(\epsilon_2, \delta) \leq m_{\mathcal{H}}(\epsilon_1, \delta),$$

lo que demuestra la primera desigualdad.

La prueba para el parámetro de confianza δ es análoga: si se requiere una mayor confianza (es decir, $\delta_1 \leq \delta_2$), el tamaño muestral mínimo para garantizarla no puede disminuir; formalmente, para $m \geq m_{\mathcal{H}}(\epsilon, \delta_1)$ la probabilidad de éxito es al menos $1 - \delta_1 \geq 1 - \delta_2$, así que

$$m_{\mathcal{H}}(\epsilon, \delta_2) \leq m_{\mathcal{H}}(\epsilon, \delta_1).$$

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 $L_{(\mathcal{D},f)}(h)$:

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

Demostración: Sea $S = \{(x_i, y_i)\}_{i=1}^m$ con $x_i \overset{i.i.d.}{\sim} \mathcal{D}$ y $y_i = f(x_i)$. Con pérdida 0–1, el riesgo empírico es

$$L_S(h) = \frac{1}{m} \sum_{i=1}^{m} \mathbb{1}\{h(x_i) \neq y_i\}.$$

Por linealidad de la esperanza e idéntica distribución de los sumandos,

$$\mathbb{E}_S[L_S(h)] = \frac{1}{m} \sum_{i=1}^{m} \mathbb{E}_{(x_i,y_i)}\big[\mathbb{1}\{h(x_i) \neq y_i\}\big] = \mathbb{E}_{(X,Y)}\big[\mathbb{1}\{h(X) \neq Y\}\big].$$

Bajo realizabilidad $Y = f(X)$, luego

$$\mathbb{E}_S[L_S(h)] = \Pr_{X \sim \mathcal{D}}\big[h(X) \neq f(X)\big] = L_{(\mathcal{D},f)}(h).$$

3. (5 puntos) **Círculos concéntricos (aprendibilidad PAC).** Sea $\mathcal{X} = \mathbb{R}^2, \mathcal{Y} = \{0, 1\}$ y la clase de hipótesis

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

Demuestre que $\mathcal{H}$ es PAC-aprendible (bajo realizabilidad) y que

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

Demostración:

Sea $\mathcal{D}$ una distribución sobre $\mathbb{R}^2$ y supongamos realizabilidad: existe $r^\star$ tal que la hipótesis objetivo es $h_{r^\star}$ (es decir, las etiquetas son 1 exactamente para $\|x\| \leq r^\star$ y 0 en otro caso). Denotemos por $\mathcal{D}_X$ la marginal sobre $\mathbb{R}^2$ y definamos la función de distribución radial

$$F(r) := \Pr_{X \sim \mathcal{D}_X}\big(\|X\| \leq r\big).$$

Fijados $\epsilon, \delta \in (0,1)$, definamos

$$r_\epsilon := \sup\{r \leq r^\star : F(r) \leq 1 - \epsilon\}.$$

Observemos que por la definición de r_ϵ se cumple

$$F(r_\epsilon) \leq 1 - \epsilon,$$

y, por la monotonía de F, la masa en el anillo $(r_\epsilon, r^\star]$ satisface

$$\Pr\left(r_\epsilon < \|X\| \leq r^\star\right) = F(r^\star) - \lim_{r \uparrow r_\epsilon} F(r) \leq 1 - F(r_\epsilon) \leq \epsilon.$$

(En particular la masa de cualquier conjunto que contenga $(r_\epsilon, r^\star]$ es a lo sumo ϵ.)

Consideramos el algoritmo ERM que, dada una muestra $S = \{(x_i, y_i)\}_{i=1}^m$, devuelve el menor radio $\hat{r}$ que contiene todos los ejemplos con etiqueta positiva (por ejemplo, como $\hat{r} = \max\{\|x_i\| : y_i = 1\}$; si no hay positivos se puede devolver $\hat{r} = 0$). Si en la muestra existe al menos un punto positivo con radio en el intervalo $(r_\epsilon, r^\star]$, entonces $\hat{r} \geq r_\epsilon$ y, por tanto, el error de generalización de $h_{\hat{r}}$ está acotado por la masa del anillo:

$$L_{\mathcal{D}}(h_{\hat{r}}) = \Pr\left(\hat{r} < \|X\| \leq r^\star\right) \leq \Pr\left(r_\epsilon < \|X\| \leq r^\star\right) \leq \epsilon.$$

Lo contrario es que *ninguno* de los m puntos de la muestra caiga en el anillo $(r_\epsilon, r^\star]$. Pero ya que cada punto cae fuera del anillo con probabilidad al menos $1 - \epsilon$, la probabilidad de que los m puntos estén todos fuera es

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

Por lo tanto, si elegimos

$$m \geq \frac{\log(1/\delta)}{\epsilon},$$

entonces $e^{-\epsilon m} \leq \delta$, y concluimos que con probabilidad al menos $1 - \delta$ sobre la muestra el algoritmo ERM devuelve una hipótesis con error generalización $\leq \epsilon$.

Finalmente, tomando el techo obtenemos la cota anunciada:

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

Esto demuestra que $\mathcal{H}$ es PAC-aprendible bajo realizabilidad.

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 ERM eficiente.

   Demostración y descripción del algoritmo:

   Tamaño de $\mathcal{H}$. Para cada variable x_i hay tres opciones en una conjunción: incluir x_i, incluir $\bar{x}_i$, o no incluir ninguna de las dos. Por tanto hay a lo sumo 3^d conjunciones de

este tipo. Adicionalmente podemos incluir la hipótesis que es siempre negativa (que puede asociarse, por ejemplo, con la presencia simultánea de x_i y $\bar{x}_i$ para alguna i), con lo que una cota válida es

$$|\mathcal{H}| \leq 3^d + 1.$$

Aplicando la cota estándar para clases finitas (lema de la unión) obtenemos

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

Algoritmo ERM eficiente ("eliminar literales"). Representaremos la hipótesis como un vector de d entradas, donde para cada i el estado puede ser uno de $\{+$ (incluir x_i), $-$ (incluir $\bar{x}_i$), 0 (ninguno)$\}$. Inicializamos la hipótesis en el estado más conservador conforme a positivos: para cada i ponemos el estado ambivalente que contiene *ambos* literales. En la práctica basta representar esto como permitir la eliminación de cada literal cuando se vea un ejemplo positivo que lo contradiga.

El procedimiento es:

a) Inicializa la hipótesis h con, para cada i, la posibilidad de incluir x_i y $\bar{x}_i$.

b) Para cada ejemplo de entrenamiento (a, y):

- Si $y = 1$ (ejemplo positivo): para cada coordenada i
  - si $a_i = 1$ entonces elimina $\bar{x}_i$ de la conjunción (si estaba presente);
  - si $a_i = 0$ entonces elimina x_i de la conjunción (si estaba presente).
- Si $y = 0$ (ejemplo negativo): no se hace nada (los negativos sólo descartan hipótesis que ya serían inconsistentes, pero bajo realizabilidad no hay que usarles para eliminar literales).

c) Devuelve la conjunción resultante (si alguna variable quedó con ambos literales eliminados, se omite. Si quedó con ambos literales presentes, existe una contradicción y puede interpretarse como la hipótesis siempre negativa).

Corrección bajo realizabilidad. Sea $h^\star$ la conjunción objetivo (asumimos que existente por realizabilidad). Cualquier ejemplo positivo es consistente con $h^\star$, por lo que cuando procesamos un positivo no eliminamos ningún literal que pertenezca a $h^\star$. Por lo que la hipótesis construida contiene todos los literales de $h^\star$ y por tanto clasifica correctamente todos los positivos. Bajo realizabilidad, los negativos serán también correctamente clasificados. DE eeste modo, el algoritmo produce una hipótesis consistente con la muestra; por el principio ERM y la cota para clases finitas, con el número de muestras dado por la cota anterior logra PAC-aprendibilidad.

Complejidad. Cada actualización por ejemplo recorre las d coordenadas, así que la complejidad temporal es $O(m \cdot d)$. La cota muestral se dio arriba:

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

5. (10 puntos) **PAC agnóstico ⇒ PAC (realizable).** 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.

   Demostración:

   Recordemos la definición: $\mathcal{H}$ es agnósticamente PAC-aprendible si existe un algoritmo A y una función $m(\epsilon, \delta)$ tal que para cualquier distribución $\mathcal{D}$ sobre $\mathcal{X} \times \{0, 1\}$, con probabilidad al menos $1 - \delta$ sobre muestras de tamaño $m \geq m(\epsilon, \delta)$, A devuelve una hipótesis h satisfaciendo

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

   Si además asumimos realizabilidad, existe $h^\star \in \mathcal{H}$ con $L_{\mathcal{D}}(h^\star) = 0$, por lo que

   $$\inf_{h' \in \mathcal{H}} L_{\mathcal{D}}(h') = 0.$$

   Entonces la garantía agnóstica se reduce a

   $$L_{\mathcal{D}}(h) \leq 0 + \epsilon = \epsilon,$$

   con probabilidad al menos $1 - \delta$. Eso es exactamente la definición de PAC-aprendibilidad bajo realizabilidad. Además, el mismo algoritmo A con la misma función $m(\epsilon, \delta)$ sirve en el caso realizable. Por tanto, agnóstico PAC ⇒ PAC (realizable), y el aprendiz agnóstico también es aprendiz PAC en el caso realizable.

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

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

   Demostración. Fijemos $x \in \mathcal{X}$ y consideremos la probabilidad condicional de error de un clasificador cualquiera g dado $X = x$:

   $$\Pr\big(g(x) \neq Y \mid X = x\big) = \begin{cases} \Pr(Y = 1 \mid X = x) = \eta(x), & \text{si } g(x) = 0, \\ \Pr(Y = 0 \mid X = x) = 1 - \eta(x), & \text{si } g(x) = 1. \end{cases}$$

   Por tanto, para un punto x dado, el error condicional de g es

   $$\Pr\big(g(x) \neq Y \mid X = x\big) = \begin{cases} \eta(x), & g(x) = 0, \\ 1 - \eta(x), & g(x) = 1. \end{cases}$$

   El predictor bayesiano $f_{\mathcal{D}}$ elige para cada x la etiqueta que minimiza esta cantidad punto a punto; es decir,

   $$\Pr\big(f_{\mathcal{D}}(x) \neq Y \mid X = x\big) = \min\{\eta(x),\, 1 - \eta(x)\}.$$

Para cualquier otro clasificador g se tiene, por la definición de mínimo,

$$\Pr\big(g(x) \neq Y \mid X = x\big) \geq \min\{\eta(x),\, 1-\eta(x)\} = \Pr\big(f_{\mathcal{D}}(x) \neq Y \mid X = x\big).$$

Integrando (esperanza total) respecto a la marginal de X se obtiene la desigualdad global de riesgos:

$$L_{\mathcal{D}}(g) = \mathbb{E}_X\big[\Pr(g(X) \neq Y \mid X)\big] \geq \mathbb{E}_X\big[\Pr(f_{\mathcal{D}}(X) \neq Y \mid X)\big] = L_{\mathcal{D}}\big(f_{\mathcal{D}}\big).$$

Esto prueba que $f_{\mathcal{D}}$ minimiza el riesgo verdadero entre todos los clasificadores.

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.

   Respuesta: Esto es una reformulación del inciso anterior: entre todos los clasificadores deterministas o probabilísticos la elección que minimiza el riesgo punto a punto es la que selecciona la etiqueta con mayor probabilidad posterior $\Pr[Y = 1 \mid X = x]$ (si empatan, cualquier desempate mínimo sirve). Por tanto el predictor bayesiano minimiza el riesgo esperado respecto a la pérdida absoluta.

   (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.

   Respuesta (construcción): Para cada distribución $\mathcal{D}$ podemos definir el algoritmo $A_{\mathcal{D}}$ que, ignorando la muestra, devuelve el predictor bayesiano $f_{\mathcal{D}}$ (que depende de $\mathcal{D}$). Por construcción $f_{\mathcal{D}}$ minimiza el riesgo sobre $\mathcal{D}$, por lo tanto $A_{\mathcal{D}}$ es óptimo frente a $\mathcal{D}$. Obsérvese que $A_{\mathcal{D}}$ no es computable en general (porque desconoce $\mathcal{D}$), pero la afirmación pide existencia teórica, no computabilidad.

   (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 $\mathcal{D}$.

   Respuesta: Fijemos un algoritmo A. Escogemos la distribución $\mathcal{D}$ tal que el predictor bayesiano $f_{\mathcal{D}}$ tiene riesgo menor que el riesgo medio que A puede asegurar (por ejemplo, construir $\mathcal{D}$ concentrando masa en puntos donde A falla sistemáticamente). Definimos B como el algoritmo que devuelve $f_{\mathcal{D}}$. Entonces B supera a A en $\mathcal{D}$. Este argumento formaliza la no-existencia de un algoritmo universal que sea estrictamente mejor que todos los demás en todas las distribuciones.