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) Demuestre que para toda dos clases de hipótesis, si $\mathcal{H}' \subseteq \mathcal{H}$, entonces

$$\mathrm{VCdim}(\mathcal{H}') \leq \mathrm{VCdim}(\mathcal{H}).$$

Demostración. a) Dado un conjunto finito $C \subseteq \mathcal{X}$, donde esta definido por $\{0,1\}^C$ al conjunto de todas las *etiquetas* posibles de C, es decir, $y : C \to \{0,1\}$.

b) Para una clase de hipótesis $\mathcal{G} \subseteq \{0,1\}^{\mathcal{X}}$ definimos su proyección a C como

$$\mathcal{G}\restriction_C := \{\, h\restriction_C : h \in \mathcal{G} \,\} \subseteq \{0,1\}^C,$$

donde $h\restriction_C$ es la función $x \mapsto h(x)$ restringida al dominio C.

c) Ahora bien decimos que $\mathcal{G}$ *destroza* a C si

$$\mathcal{G}\restriction_C = \{0,1\}^C,$$

básicamente es que si para toda etiqueta $y \in \{0,1\}^C$ existe $h \in \mathcal{G}$ tal que $h\restriction_C = y$.

d) La dimensión VC de $\mathcal{G}$ es

$$\mathrm{VCdim}(\mathcal{G}) = \sup\{\, |C| : C \subseteq \mathcal{X} \text{ finito y } \mathcal{G} \text{ destroza a } C \,\},$$

con el hecho de que el supremo puede ser $+\infty$.

Ahora bien sea $C \subseteq \mathcal{X}$ un conjunto finito cualquiera. En el supuesto que $\mathcal{H}' \subseteq \mathcal{H}$ se deduce, aplicando la definición de restricción, la monotonía de las proyecciones:

$$\mathcal{H}'\restriction_C \subseteq \mathcal{H}\restriction_C .$$

Se cumple que si $g \in \mathcal{H}'\restriction_C$, entonces existe $h \in \mathcal{H}'$ con $g = h\restriction_C$; pero $h \in \mathcal{H}' \subseteq \mathcal{H}$, de modo que $g = h\restriction_C \in \mathcal{H}\restriction_C$.

Bien, ahora, supongamos que C es destrozado por $\mathcal{H}'$. Por definición,

$$\mathcal{H}'\restriction_C = \{0,1\}^C.$$

Usando lo anterior, obtenemos que

$$\{0,1\}^C = \mathcal{H}'\restriction_C \subseteq \mathcal{H}\restriction_C \subseteq \{0,1\}^C.$$

Por lo tanto, $\mathcal{H}\restriction_C = \{0,1\}^C$ y C también es destrozado por $\mathcal{H}$.

Ya probamos que *todo* conjunto finito que es destrozado por $\mathcal{H}'$ también lo es por $\mathcal{H}$. Por lo tanto,

$$\mathrm{VCdim}(\mathcal{H}') \leq \mathrm{VCdim}(\mathcal{H}).$$

Finalmente, si $\mathrm{VCdim}(\mathcal{H}') = +\infty$, entonces para cada $m \in \mathbb{N}$ existe C con $|C| = m$ destrozado por $\mathcal{H}'$, y por el argumento anterior también por $\mathcal{H}$. Por lo tanto $\mathrm{VCdim}(\mathcal{H}) = +\infty$. $\square$

2. (10 puntos) Sea $\mathcal{X}$ un conjunto de dominio finito y sea $k \leq |\mathcal{X}|$. Calcule la dimensión VC de cada una de las siguientes clases y justifique sus respuestas:

$$\mathcal{H}_{=k}^{\mathcal{X}} = \left\{h \in \{0,1\}^{\mathcal{X}} : |\{x : h(x) = 1\}| = k\right\},$$
$$\mathcal{H}_{\text{at-most-}k}^{\mathcal{X}} = \left\{h \in \{0,1\}^{\mathcal{X}} : |\{x : h(x) = 1\}| \leq k \text{ o } |\{x : h(x) = 0\}| \leq k\right\}.$$

Nota: Vamos a definir lo que es una rotulacion , ya que esto se va a ocupar en este ejercicio y en otro mas adelante

Definición .1 (Rotulación / Etiquetado / Labeling)**.** *Sea* $C = \{x_1, \ldots, x_m\} \subseteq \mathcal{X}$ *finito. Una* rotulación *de* C *es un vector* $(y_1, \ldots, y_m) \in \{0,1\}^m$*, donde* y_i *es la etiqueta asignada a* x_i.

Observación .1 (Destrozar)**.** *Una clase* $\mathcal{H} \subseteq \{0,1\}^{\mathcal{X}}$ destroza *a* C *si para toda rotulación* $(y_1, \ldots, y_m) \in \{0,1\}^m$ *existe* $h \in \mathcal{H}$ *con* $h(x_i) = y_i$ *para todo* i*. Equivalente:* $\mathcal{H} \restriction_C = \{0,1\}^C$.

Teorema .1. *Con la notación anterior,*

$$\text{VCdim}(\mathcal{H}_{=k}^{\mathcal{X}}) = \min\{k,\ n-k\}, \qquad \text{VCdim}(\mathcal{H}_{\textit{at-most-k}}^{\mathcal{X}}) = \min\{2k+1,\ n\}.$$

Demostración. Sea $C \subseteq \mathcal{X}$ con $|C| = m$ y una rotulación arbitraria de C con s unos (y $m - s$ ceros).

(1) Clase $\mathcal{H}_{=k}^{\mathcal{X}}$**.** Primero, si $k = 0$ o $k = n$ hay una sola función (constante), así que la VC-dimensión es 0. Bien, ahora supongamos $1 \leq k \leq n-1$. Una rotulación sobre C que puede extenderse a una hipótesis $h \in \mathcal{H}_{=k}^{\mathcal{X}}$ si y solo si

$$s \leq k \quad \text{y} \quad k - s \leq n - m,$$

pues necesitamos usar exactamente s unos dentro de C y completar con $k - s$ unos fuera de C.

Para que $\mathcal{H}_{=k}^{\mathcal{X}}$ destroce C, las dos desigualdades anteriores deben valer para *todo* $s \in \{0, \ldots, m\}$. En particular, tomando $s = m$ y $s = 0$ se obtiene la condición necesaria

$$m \leq k \quad \text{y} \quad m \leq n - k.$$

De la misma manera, si $m \leq \min\{k, n-k\}$, entonces para cualquier rotulación: (i) $s \leq m \leq k$ y (ii) $k - s \leq k \leq n - m$ (porque $m \leq n - k \Rightarrow k \leq n - m$), luego se puede completar a exactamente k unos. Por tanto, C es destrozado.

Si $m = \min\{k, n-k\} + 1$, alguna rotulación ya no es realizable:

- si $m \geq k + 1$, la "todos unos" tiene $s = m > k$;

- si $m \geq n - k + 1$, la "todos ceros" exige $k - 0 = k > n - m$.

Por lo que $\text{VCdim}(\mathcal{H}_{=k}^{\mathcal{X}}) = \min\{k, n-k\}$.

(2) Clase $\mathcal{H}_{\textbf{at-most-}k}^{\mathcal{X}}$**.**

Por definición, $h \in \mathcal{H}^{\mathcal{X}}_{\text{at-most-}k}$ si satisface $|\{x : h(x) = 1\}| \leq k$ *o* $|\{x : h(x) = 0\}| \leq k$. Para una rotulación con s unos en C:

- Si $s \leq k$, podemos fijar $h = 0$ en $\mathcal{X} \setminus C$; el total de unos queda $s \leq k$.

- Si $m - s \leq k$, podemos fijar $h = 1$ en $\mathcal{X} \setminus C$; el total de ceros queda $m - s \leq k$.

Por lo tanto, C es destrozado si para todo $s \in \{0, \ldots, m\}$ se cumple

$$s \leq k \quad \text{o} \quad s \geq m - k,$$

es decir, si los intervalos $[0, k]$ y $[m-k, m]$ cubren $\{0, \ldots, m\}$, lo cual ocurre exactamente cuando $m \leq 2k+1$. La necesidad es inmediata: si $m = 2k+2$, la rotulación con $s = k+1$ no satisface ninguna de las dos desigualdades ($k + 1 \nleq k$ y $m - s = k + 1 \nleq k$).

Por lo que, siempre $m \leq n$ por ser $C \subseteq \mathcal{X}$, de modo que $\text{VCdim}(\mathcal{H}^{\mathcal{X}}_{\text{at-most-}k}) = \text{mín}\{2k + 1, n\}$. □

Observación .2 (Nota)**.** *Para $\mathcal{H}^{\mathcal{X}}_{=k}$: $k = 0$ o $k = n$ dan VC-dim $= 0$.*

Para $\mathcal{H}^{\mathcal{X}}_{\text{at-most-}k}$: $k = 0$ produce solo las dos funciones constantes, por lo que la VC-dim es $1 = \text{mín}\{1, n\}$.

3. (10 puntos) Sea $\mathcal{H}^d_{\text{rec}}$ la clase de rectángulos alineados a los ejes en $\mathbb{R}^d$. Ya se ha visto que
$$\text{VCdim}(\mathcal{H}^2_{\text{rec}}) = 4.$$
Pruebe que, en general,
$$\text{VCdim}(\mathcal{H}^d_{\text{rec}}) = 2d.$$

Demostración. **Cota inferior:** $\text{VCdim} \geq 2d$**.**

Tomemos $S = \{e_1, \ldots, e_d, -e_1, \ldots, -e_d\} \subset \mathbb{R}^d$ (los vectores canónicos y sus opuestos). Dada una rotulación $y \in \{0, 1\}^{2d}$, para cada $i \in [d]$ elija

$$a_i = \begin{cases} -1, & \text{si } y_{d+i} = 1 \text{ (incluir } -e_i), \\ 0, & \text{si } y_{d+i} = 0 \text{ (excluir } -e_i), \end{cases} \qquad b_i = \begin{cases} 1, & \text{si } y_i = 1 \text{ (incluir } e_i), \\ 0, & \text{si } y_i = 0 \text{ (excluir } e_i). \end{cases}$$

Para $j \neq i$, siempre $0 \in [a_j, b_j]$ (pues $[a_j, b_j] \in \{[-1, 1], [0, 1], [-1, 0], [0, 0]\}$). Así,

$$e_i \in \prod_k [a_k, b_k] \iff b_i = 1 \iff y_i = 1, \qquad -e_i \in \prod_k [a_k, b_k] \iff a_i = -1 \iff y_{d+i} = 1.$$

Por tanto, existe $h \in \mathcal{H}^d_{\text{rec}}$ que realiza cualquier rotulación en S, y S es destrozado(Como tambien vimos en el inciso anteior por eso se cumple). Entonces $\text{VCdim}(\mathcal{H}^d_{\text{rec}}) \geq 2d$.

Cota superior: $\text{VCdim} \leq 2d$**.**

Ahora bien , sea $C \subset \mathbb{R}^d$ con $|C| = m \geq 2d + 1$. Para cada coordenada i, tome puntos

$$x^{\text{mín},i} \in \arg\min_{x \in C} x_i, \qquad x^{\text{máx},i} \in \arg\max_{x \in C} x_i.$$

Por lo que hay a lo más $2d$ puntos "extremos" distintos. Como $m \geq 2d+1$, existe $x^\star \in C$ que no es extremo en ninguna coordenada. Tomemos la rotulación $y(x) = 1$ para $x \in C \setminus \{x^\star\}$ y $y(x^\star) = 0$.

Sea $R = \prod_{i=1}^{d}[a_i, b_i]$ una caja que contiene a todos los puntos de $C \setminus \{x^\star\}$. Entonces,

$$a_i \leq \min_{x \in C \setminus \{x^\star\}} x_i \leq x_i^\star \leq \max_{x \in C \setminus \{x^\star\}} x_i \leq b_i, \quad \forall i,$$

pues quitar un punto no extremo no cambia mínimos ni máximos por coordenada. De aquí $x^\star \in R$, contradicción con $y(x^\star) = 0$. Por lo que ningún conjunto de tamaño $2d+1$ puede ser destrozado y, por tanto, $\mathrm{VCdim}(\mathcal{H}^d_{\mathrm{rec}}) \leq 2d$.

Entonces juntando ambas cotas nos da que , $\mathrm{VCdim}(\mathcal{H}^d_{\mathrm{rec}}) = 2d$. □

4. (5 puntos) Sea $\mathcal{H}$ la clase de intervalos sobre la recta (formalmente equivalente a rectángulos alineados a los ejes en dimensión $n = 1$). Proponga una implementación de la regla de aprendizaje ERM$\mathcal{H}$ (en el caso agnóstico) que, dado un conjunto de entrenamiento de tamaño m, ejecute en tiempo $O(m^2)$.
   Hint: use programación dinámica.

   Sea $S = \{(x_i, y_i)\}_{i=1}^{m}$ con $x_1 < \cdots < x_m$ y $y_i \in \{0,1\}$. Para un intervalo $[a,b]$, la hipótesis $h_{[a,b]}(x) = \mathbf{1}[\, a \leq x \leq b\,]$ nos da la pérdida empírica

   $$\ell([a,b]) \;=\; \#\{i:\; x_i \in [a,b],\, y_i = 0\} \;+\; \#\{i:\; x_i \notin [a,b],\, y_i = 1\}.$$

   En un conjunto finito, existe un ERM cuyos extremos coinciden con muestras (o bien un intervalo que no cubra ninguna muestra). Con eso basta considerar intervalos $[x_i, x_j]$ con $1 \leq i \leq j \leq m$ y, además, un *intervalo "vacío" dentro de la clase*, por ejemplo:

   $$[a_\varnothing, b_\varnothing] = \begin{cases} \left[\frac{x_1+x_2}{2},\, \frac{x_1+x_2}{2}\right], & \text{si } m \geq 2, \\ [x_1+1,\, x_1+1], & \text{si } m = 1, \end{cases}$$

   que no contiene a ningún x_i; sobre S predice 0 en todos los puntos.

   Precómputo. Sea $P[k] = \sum_{r=1}^{k} y_r$ con $P[0] = 0$. Para $[i,j] \equiv [x_i, x_j]$,

   $$\ell_{i,j} \;=\; \underbrace{\big(P[i-1] + (P[m] - P[j])\big)}_{\text{positivos fuera}} \;+\; \underbrace{\big((j-i+1) - (P[j] - P[i-1])\big)}_{\text{negativos dentro}}.$$

   Recurrencias locales (DP). Para $1 \leq i \leq j \leq m$:

   $$\boxed{\ell_{i+1,j} \;=\; \ell_{i,j} \;+\; (2y_i - 1)} \qquad \text{y} \qquad \boxed{\ell_{i,j+1} \;=\; \ell_{i,j} \;+\; (1 - 2y_{j+1})}.$$

   Inicialización.

   $$\ell_{1,1} \;=\; (P[m] - P[1]) \;+\; (1 - y_1) \;=\; P[m] + 1 - 2y_1, \quad \ell_\varnothing \;=\; \sum_{i=1}^{m} y_i \;=\; P[m].$$

   La primera fila: $\ell_{1,j+1} = \ell_{1,j} + 1 - 2y_{j+1}$ para $j = 1, \ldots, m-1$.

   Rellenado fila por fila. Para $i = 1, \ldots, m-1$:

a) $\ell_{i+1,i+1} \leftarrow \ell_{i,i+1} + (2y_i - 1)$.

b) Para $j = i+1, \ldots, m-1$: $\ell_{i+1,j+1} \leftarrow \ell_{i+1,j} + (1 - 2y_{j+1})$.

Selección del mejor intervalo.

Mantenemos a $(\hat{i}, \hat{j}) \in \arg\min_{1 \le i \le j \le m} \ell_{i,j}$.

Para después comparar con $\ell_\varnothing = P[m]$.

Así devolver:

$$\widehat{h}(x) = \begin{cases} \mathbf{1}\big[x \in [x_{\hat{i}}, x_{\hat{j}}]\big], & \text{si } \ell_{\hat{i},\hat{j}} \le \ell_\varnothing, \\ \mathbf{1}\big[x \in [a_\varnothing, b_\varnothing]\big], & \text{si } \ell_\varnothing < \ell_{\hat{i},\hat{j}}], \end{cases}$$

donde $[a_\varnothing, b_\varnothing]$ es el intervalo "vacío" que definimos (pertenece a $\mathcal{H}$ y no cubre ninguna muestra).

Correctitud y complejidad.

(i) Entre los ERM hay uno con extremos en muestras o bien un intervalo que no cubre ninguna;

(ii) Las recurrencias se pueden calcular instantáneamente usando la expresión con prefijos;

(iii) Exploramos todas las parejas (i, j) y el caso "vacío" en $\Theta(m^2)$ celdas, cada una en $O(1)$.

Tiempo total $O(m^2)$ (más el ordenamiento $O(m \log m)$ si es que hace falta).

5. (5 puntos) Demuestre cómo expresar el problema ERM de regresión lineal respecto a la función de pérdida de valor absoluto,

$$\ell(h, (x, y)) = |h(x) - y|,$$

como un programa lineal; es decir, cómo escribir el problema

$$\min_w \sum_{i=1}^{m} |\langle w, x_i \rangle - y_i|$$

en forma de programa lineal.
Hint: comience probando que para cualquier $c \in \mathbb{R}$:

$$|c| = \min_{a \ge 0} a \quad \text{s.a. } c \le a \text{ y } c \ge -a.$$

Demostración:

Dado $S = \{(x_i, y_i)\}_{i=1}^m$ con $x_i \in \mathbb{R}^d$, queremos

$$\min_{w \in \mathbb{R}^d} \sum_{i=1}^{m} \big|\langle w, x_i \rangle - y_i\big|.$$

Usando que $|c| = \text{mín}_{a\geq 0}\{a: \ c \leq a, \ -c \leq a\}$, sean las variables $s_i \geq 0$ y resolvemos en

$$\min_{w,s} \sum_{i=1}^{m} s_i \quad \text{s.a.} \quad \begin{cases} \langle w, x_i\rangle - y_i \leq s_i, \\ -(\langle w, x_i\rangle - y_i) \leq s_i, & \qquad i = 1, \dots, m. \\ s_i \geq 0, \end{cases}$$

Forma matricial. Sea $X \in \mathbb{R}^{m\times d}$ con filas $X_{i\to} = x_i^\top$, y definimos a la variable

$$v = \begin{bmatrix} w \\ s \end{bmatrix} \in \mathbb{R}^{d+m}, \qquad c = \begin{bmatrix} 0_d \\ 1_m \end{bmatrix}.$$

Construimos

$$A = \begin{bmatrix} X & -I_m \\ -X & -I_m \\ 0_{m\times d} & -I_m \end{bmatrix} \in \mathbb{R}^{(2m+m)\times(d+m)}, \qquad b = \begin{bmatrix} y \\ -y \\ 0_m \end{bmatrix}.$$

Entonces el ERM-ℓ_1 se escribe como el programa lineal

$$\begin{aligned} &\text{mín} \quad c^\top v \\ &\text{s.a.} \quad A\,v \ \leq \ b. \end{aligned}$$

Observación .3. *Las restricciones $s_i \geq 0$ son opcionales (pues al minimizar con $s_i \geq \langle w, x_i\rangle - y_i$ y $s_i \geq y_i - \langle w, x_i\rangle$ quedan dentro de), aun que podemos incluir el bloque $[\,0 \quad -I_m\,]\,v \leq 0$ para la no-negatividad. Ahora bien si se deseamos un intercepto, basta con aumentar cada x_i con una coordenada 1.*

6. (10 puntos) Construya un ejemplo que demuestre que la función de pérdida 0-1 puede sufrir de mínimos locales; es decir, construya un conjunto de entrenamiento

$$S \in (X \times \{\pm 1\})^m$$

(por ejemplo, para $X = \mathbb{R}^2$), para el cual existan un vector w y algún $\epsilon > 0$ tales que:

a) Para todo w' tal que $\|w - w'\| \leq \epsilon$, se cumple que

$$L_S(w) \leq L_S(w'),$$

donde la pérdida es la pérdida 0-1. Esto significa que w es un mínimo local de L_S.

b) Existe algún w^* tal que

$$L_S(w^*) < L_S(w).$$

Esto significa que w no es un mínimo global de L_S.

Consideremos la clase de semiespacios *homogéneos* en $\mathbb{R}^d$, $h_w(x) = \text{sign}(\langle w, x \rangle)$, y el conjunto de entrenamiento

$$S = \{(x, y)\} \quad \text{con} \quad x = \mathbf{e}_1,\ y = +1.$$

Sea $w = -\mathbf{e}_1$. Entonces $\langle w, x \rangle = -1$, por lo que $h_w(x) = -1 \neq y$,por lo tanto,

$$L_S(w) = \mathbf{1}[h_w(x) \neq y] = 1.$$

(1) w es mínimo local. Tomemos $\varepsilon \in (0, 1)$ y cualquier w' con $\|w' - w\|_2 \leq \varepsilon$. Por Cauchy–Schwarz y $\|x\|_2 = \|\mathbf{e}_1\|_2 = 1$,

$$\langle w', x \rangle = \langle w, x \rangle + \langle w' - w, x \rangle \leq -1 + \|w' - w\|_2 \|x\|_2 < -1 + 1 = 0.$$

Entonces $\text{sign}(\langle w', x \rangle) = -1 \neq y$, y $L_S(w') = 1 = L_S(w)$.

Por lo que se cumple $L_S(w) \leq L_S(w')$ para todo w' en la bola $\|w' - w\| \leq \varepsilon$: w es un mínimo local.

(2) No es mínimo global.

Tomemos $w^* = \mathbf{e}_1$. Entonces $\langle w^*, x \rangle = 1 > 0$, $h_{w^*}(x) = +1 = y$ y

$$L_S(w^*) = 0 < L_S(w) = 1.$$

Por lo que, w no es un mínimo global.

Observación .4. *El problema surge porque la función de pérdida 0-1 forma "escalones": cerca de cualquier punto w, hay una zona donde las predicciones no cambian, creando superficies planas que parecen mínimos locales pero no son el mejor resultado posible. Esto ocurre para cualquier punto normalizado eligiendo w = -yx.*