Clase 04: El Modelo de Aprendizaje

Responsable: Jordi Legorreta Montaño

EST-25134, Primavera 2021
Dr. Alfredo Garbuno Iñigo
Enero 26, 2021

- Seguiremos considerando el modelo PAC (modelo Probable y Aproximadamente Correcto).
- Más adelante veremos otras nociones de aprendizaje.
- Pioneros en escribir este modelo formal de aprendizaje: Vapnik y Chervonenkis.

1. PAC Aprendible o Aprendizaje PAC

$$|\mathcal{H}| < \infty \quad \oplus \quad ERM \quad \oplus \quad m \geq M_0$$

Definición 1.1. Decimos que una clase $\mathcal{H}$ es aprendible, en el sentido PAC, si existe una función $m_{\mathcal{H}} : (0,1)^2 \to \mathbb{N}$, y tenemos un algoritmo de aprendizaje con la siguiente propiedad: Al considerar cualquier $\varepsilon, \delta \in (0,1)$, cualquier distribución $\mathcal{D}$ sobre $\mathcal{X}$, y cualquier función de etiquetado $f : \mathcal{X} \to \mathcal{Y} = \{0,1\}$.

Si se satisface la hipótesis de realizabilidad con respecto a $\mathcal{H}$, $\mathcal{D}$ y f, entonces si utilizamos el algoritmo de aprendizaje con una cantidad suficiente de datos, $m \geq m_{\mathcal{H}}(\varepsilon, \delta)$, el algoritmo regresa una hipótesis h_S tal que, con una alta probabilidad —es decir una probabilidad mayor a $(1-\delta)$— se encuentra una error de generalización suficientemente preciso, $L_{D,f}(h_s) \leq \varepsilon$.

En este sentido lso parámetros tienen la siguiente interpretación:

- ε : precisión (qué tan lejos va a estar el clasificador con respecto a el óptimo).
- δ : parámetro de confianza (qué tan probable será que el clasificador que encontremos alcance el requerimiento de precisión).

1.1. ¿Hay forma de evitar estas aproximaciones (en términos de ε y δ)?

No, no se pueden evitar, ya que no tenemos acceso a $\mathcal{D}$, sino sólo a una muestra S, y no podemos garantizar que una muestra sea representativa.

$m \geq m_{\mathcal{H}}(\varepsilon, \delta)$: complejidad muestral, es decir, el número mínimo de observaciones independientes de la distribución que necesitamos para poder establecer las garantías anteriores.

Observación : si $\mathcal{H}$ es PAC aprendible, existen muchas posibilidades para $m_{\mathcal{H}}$, por lo que acotamos $m_{\mathcal{H}}$ para que sea la mínima función en el sentido de que nos dará el mínimo entero necesario, considerando ε y δ.

Corolario : toda clase finita de hipótesis ($\mathcal{H}$) es PAC aprendible si:

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

1.2. ¿En qué sentido podemos generalizar PAC?

- ¿Sólo clasificación?, es decir, ¿sólo 0 ó 1?
- La hipótesis de realizabilidad podría no ser la correcta, en el sentido de que posiblemente no podríamos garantizar que un algoritmo de aprendizaje fuese exitoso para una distribución fija y una función, en cualquier otro contexto de datos.

1.2.1. Aprendizaje PAC agnóstico

Antes: $\exists h^* \in \mathcal{H} \quad | \quad \mathbb{P}_{\mathcal{X}\sim D}\Big(h^*(x) = f(x)\Big) = 1$

Ahora:

D como una medida de probabilidad en el espacio $\mathcal{X} \times \mathcal{Y}$
D_x distribución marginal en $\mathcal{X}$
$D_{y|x}$ distribución condicional para $\mathcal{Y}$

Es decir, estamos ahora admitiendo la posibilidad de que dos elementos que tengan características muy similares (si no es que iguales) puedan tener una etiqueta distinta.

Por lo tanto, rompiendo con la hipótesis de realizabilidad, podemos definir una función de pérdida:

$$L_D(h) \overset{\text{def}}{=} \mathbb{P}_{(x,y)\sim D}\Big(h(x) \neq y\Big) \overset{\text{def}}{=} D\Big(\{(x,y) : h(x) \neq y\}\Big)$$

$$L_S(h) = \frac{|\{i : h(x^{(i)}) \neq y^{(i)}|}{m}, \qquad (x^{(i)}, y^{(i)}) \overset{iid}{\sim} D$$

Antes: la aleatoriedad provenía sólo de $\mathcal{X}$.

Ahora: extendemos D para que sea una medida de probabilidad en la pareja (x, y).

Se puede ver como: $D_{y|x}$ sustituye a $f(x) = y$, ya que la relción entre $\mathcal{X}$ e $\mathcal{Y}$ no es determinista, no es una función, sino que ahora es un modelo probabilístico.

Definición 1.2 (Clasificador Bayesiano Óptimo)**.** Dada la distribución D sobre $\mathcal{X} \times \{0,1\}$, el mejor predictor de etiquetas es:

$$f_D(x) = \begin{cases} 1 & \text{si} \quad \mathbb{P}(y=1 \mid x) \geq 1/2 \\ 0 & \text{si} \quad \mathbb{P}(y=1 \mid x) < 1/2 \end{cases} .$$

es óptimo porque cualquier otro clasificador tiene un error (riesgo) mayor que f_D.

$$g : \mathcal{X} \to \{0,1\}, \quad \text{entonces} \quad L_D(f_D) \leq L_D(g) .$$

Definición 1.3 (Capacidad de aprendizaje PAC agnóstico)**.** Una clase de hipótesis es PAC aprendible en un sentido agnóstico si existe una función de complejidad muestral $m_{\mathcal{H}}(0,1)^2 \to \mathbb{N}$ y un algoritmo de aprendizaje con la siguiente propiedad:

$$\forall(\varepsilon, \delta) \in (0,1) \quad \text{y} \quad \forall D \text{ sobre } \mathcal{X} \times \{0,1\}$$

si el algoritmo de aprendizaje utiliza un número suficiente de observaciones $m \geq m_{\mathcal{H}}(\varepsilon, \delta)$ iid de D, entonces encontraremos una hipótesis h con probabilidad $(1-\delta)$ que satisface:

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

- Si no se satisface la hipótesis de realizabilidad, entonces no podemos garantizar que el error de generalización ($L_D(h)$) sea arbitrariamente chico.
- PAC agnóstico puede ser exitoso aún cuando no alcancemos el mejor error posible sobre la clase de hipótesis con la que estamos trabajando.
- La precisión de nuestra generalización se vuelve relativa a la clase con la que estamos trabajando, precisamente por la constante $\min_{h' \in \mathcal{H}} L_D(h')$

1.2.2. Más allá de clasificación

Regresión :

$y \in \mathbb{R} \quad \therefore \quad \mathcal{Y} = \mathbb{R}$ (recta real).
$l(x,y) = (h(x) - y)^2$, donde $l(x,y)$ es la función de pérdida.
$L_D(h) = \mathbb{E}_{(x,y)\sim D}[l(x,y)]$, donde $L_D(h)$ es la función de riesgo.

Antes: conjunto de etiquetas $\mathcal{Y}$

Ahora: conjunto objetivo $\mathcal{Y}$

Funciones de pérdida :

$\mathcal{H}$: clase de hipótesis, y el dominio $Z = \mathcal{X} \times \mathcal{Y}$
$l : \mathcal{H} \times Z \to \mathbb{R}^+$
$L_D(h) = \mathbb{E}_{Z\sim D}[l(h,z)]$
$L_S(h) = \frac{1}{m} \sum_{i=1}^{m} l(h, z^{(i)})$
$Z^{(i)} \overset{iid}{\sim} D$
$L_S(h)$ es un estimador insesgado