TEMA 3 - FUNDAMENTOS DE AA

(de forma exacta)

+ Es imposible aprender "f" viendo solo una muestra (por inducción)
+ Sin embargo, podemos aprender propiedades que la estimen, asumiendo % error.

* Si los elementos de una muestra D son → independientes, idénticamente distribuidos (no cambia la distrib.)
  → Hay dependencia probabilística entre → variable aleatoria (población) → muestra

* Teorema No-Free-Lunch
  → Todos los algoritmos de aprendizaje son igual de buenos de media si consideramos todos los problemas de aprendizaje.
  * Por tanto, cada aprendiz $(A, \mathcal{H})$ debe aplicarse a distribuciones P que pueda aprender correctamente → Explotar con. específico.
  → $\forall$ algoritmo A, $\exists$ distribución P con el que falla.

INFIRIENDO DE LA MUESTRA: SOLUCIÓN PAC

MODELO BIN

+ Tenemos una bolsa de canicas { - verdes - rojas
  → ¿% canicas verdes en la bolsa?
  * Hay demasiadas "canicas" → Sacamos muestra y vemos %.
+ ¿Podemos garantizar que $\%\text{verdes}_{\text{muestra}} \approx \%\text{verdes}_{\text{bolsa}}$?
  * No al 100%, pero es probable.
* Desigualdad de Hoeffding:
  $$\mathbb{P}(\mathcal{D} : |\%\text{muestra} - \%\text{población}| > \epsilon) \leq 2e^{-2\epsilon^2 N} \quad \forall \epsilon > 0.$$
  * Dada una muestra i.i.d. de tamaño N, la probabilidad de que la diferencia entre la medición en la muestra y la real sea mayor que ϵ está acotada.
    ϵ: la diferencia entre "dentro" y "fuera". (Tolerancia)
* La cota depende solo de → N: el tamaño de la muestra.
  → ↑↑↑ N → ↓↓ cota (mayor precisión).

RELACIÓN BIN - APRENDIZAJE

+ Ahora:
  → Cada canica es un ejemplo de la población $(\mathcal{X})$, con sus características.
  → Esta será "verde" o "roja" dependiendo de si una hipótesis h clasifica correctamente el ejemplo o no (clasificación binaria).
  → % muestra ⇒ $E_{in}(h)$; % población ⇒ $E_{out}^{(h)} = \mathbb{P}_x[h(x) \neq f(x)]$.
  Aquí influye la distribución de la población.

* Desigualdad de Hoeffding para una única hipótesis.
  $$\mathbb{P}(\mathcal{D} : |E_{in}(h) - E_{out}(h)| > \epsilon) \leq 2e^{-2N\epsilon^2} \quad \text{para cualquier } \epsilon > 0.$$
  * Importante: esta probabilidad se cumple solo cuando $|\mathcal{H}| = 1$ (solo tomamos una hipótesis), y h es elegida sin ver los datos.
* Este es el "resultado PAC" (Probably Approximately Correct), que cuantifica N tal que un algoritmo de aprendizaje produzca resultado PAC
→ E_{in} es "aleatorio", pero conocido.
→ Si $E_{in}(h) \approx 0 \xrightarrow{N \gg (\text{grande})} E_{out}(h) \approx 0$ con alta probabilidad → $f \approx h$ en $\mathcal{X}$.
→ Si $E_{in} \gg 0 \longrightarrow E_{out}(h)$ también.
  * Peor $E_{in}^{(h)} = 0'5$ (equivale a elegir aleatoriamente).
  * Si $E_{in}^{(h)} > 0'5 \rightarrow \bar{h}$ (función complementaria de h): $E_{in}(\bar{h}) < 0'5$.

+ Desarrollando la desigualdad:
  * $\mathbb{P}(\mathcal{D} : |E_{out}(h) - E_{in}(h)| \leq \epsilon) > 1 - \delta \rightarrow E_{out}(h) \leq E_{in}(h) + \epsilon$ → (cota)
  → $E_{out}(h) \leq E_{in}(h) + \underbrace{\sqrt{\frac{1}{2N} \log \frac{2}{\delta}}}_{\epsilon(N,\delta)}$ (con prob. en $\mathcal{D}$ de $1-\delta$, por lo menos).
  * ↑↑ N → ↓↓ intervalo ↓↓ δ → ↑↑ intervalo
    (reducir ϵ → aumentar precisión)

* Desigualdad Hoeffding con varias hipótesis.
  * Esta vez, fijamos la "hipótesis g" a partir de los datos → "Hemos mirado más de una hipótesis" → ↑↑ % de que $E_{in} = 0$, pero no que $E_{out} \approx E_{in}$.
+ Si $|\mathcal{H}|$ es finito, estamos fijando ese nº de modelos:
  $$\mathbb{P}(\mathcal{D} : |E_{in}(g) - E_{out}(g)| > \epsilon) = \mathbb{P} \bigcup_{h_i \in \mathcal{H}} (\mathcal{D} : |E_{in}(h_i) - E_{out}(h_i)| > \epsilon).$$
  Como $P(\bigcup_{|\mathcal{H}|} B_i) \leq \sum_{i=1}^{|\mathcal{H}|} P(B_i) \rightarrow P(\mathcal{D} : |E_{in}(g) - E_{out}(g)| > \epsilon) < 2|\mathcal{H}|e^{-2\epsilon^2 N}, \quad \epsilon > 0.$

RELACIÓN BIN - APRENDIZAJE (II)

$O\left(\sqrt{\frac{\ln |\mathcal{H}|}{N}}\right)$

* Cota de inecuación de Hoeffding.

+ Con probabilidad $1-\delta$, $E_{out}(g) \le E_{in}(g) + \sqrt{\frac{1}{2N} \ln \frac{2|\mathcal{H}|}{\delta}}$.

$= \epsilon(N, \delta, |\mathcal{H}|)$

→ Si $|\mathcal{H}|$ es pequeño y $N >> \ln |\mathcal{H}|$, entonces $E_{out}(g) \approx E_{in}(g)$.

* Solo requiere $P(x)$ para generar muestras i.i.d.

HIPÓTESIS DE REALIZABILIDAD

+ Consiste en suponer que $\exists h^* \in \mathcal{H}$ tal que $E_{out\,P,f}(h^*) = 0$.
  * Entonces, $\mathcal{H}$ incluye al menos una función que $\forall P, f \rightarrow E_{in} = 0$.
+ Bajo esta hipótesis → ERM en $\mathcal{H}$ alcanza $h_s : E_{in}(h_s) = 0$.
  * Por tanto → $E_{out}(h_s) \le \frac{1}{N} \log \frac{|\mathcal{H}|}{\delta}$ (con prob. $\ge 1-\delta$).

DEFINICIÓN FORMAL DEL APRENDIZAJE PAC.

+ Una clase de funciones $\mathcal{H}$ es PAC-aprendible si existen:
  1. Una función $m_{\mathcal{H}}(\epsilon, \delta) \rightarrow \mathbb{N}$ (nº muestras para tener $< \frac{\epsilon}{\delta}$).
  2. Un algoritmo de aprendizaje A.

  tales que $\forall \epsilon, \delta \in (0,1)$, $\forall P(X)$, $\forall f$: al ejecutar A con $N \ge m_{\mathcal{H}}(\epsilon, \delta)$ ejemplos i.i.d generados por P y etiq. por f; A devuelve $h \in \mathcal{H} : \mathbb{P}(ERM_{P,f}(h) \le \epsilon) > 1-\delta$.

+ ¿$m_H(\epsilon, \delta)$? → Depende de $\begin{cases} \epsilon \\ \delta \\ \mathcal{H} \\ \text{rango de ERM (o otra } f \text{ pérdida).} \end{cases}$

+ Complejidad de muestra para cierto $\mathcal{H}$: mínimo nº de ejemplos $m_H(\epsilon, \delta)$ que satisface los requerimientos de PAC.

+ Si la f pérdida tiene rango $[0,1]$ → toda $\mathcal{H}$ finita realizable es PAC-aprendible, con complejidad de muestra:

$$m_H(\epsilon, \delta) \le \left| \frac{1}{\epsilon} \log \frac{|\mathcal{H}|}{\delta} \right|$$

(~~DEFINICIÓN~~) APRENDIZAJE PAC AGNÓSTICO

+ Una clase de funciones $\mathcal{H}$ es PAC-aprendible agnóstica si $\exists\, m_{\mathcal{H}}(\epsilon, \delta) \rightarrow \mathbb{N}$, $\exists A$ algoritmo de aprendizaje; tal que:
  * $\forall \epsilon, \delta \in (0,1]$, $\forall P(X)$. ($\cancel{\forall f}$)
  al ejecutar A sobre $N \ge m_H(\delta, \epsilon)$ ejemplos → devuelve h:

$$\mathbb{P}\left[E_{in\,P}(h) \le \min_{h' \in \mathcal{H}} E_{in\,P}(h') + \epsilon\right] > 1-\delta.$$

  * A devuelve h cercana a la mejor posible dentro de $\mathcal{H}$.
+ Regla ERM es un ejemplo de PAC-agnóstico.
+ En este caso → $m_H(\epsilon, \delta) \le \left\lceil \frac{2}{\epsilon^2} \log \frac{2|\mathcal{H}|}{\delta} \right\rceil$ → $O\left(\epsilon^{-2} \log \frac{|\mathcal{H}|}{\delta}\right)$
  * ¡Menor que en PAC!

FACTIBILIDAD DEL APRENDIZAJE

+ El aprendizaje es posible probabilísticamente $\begin{cases} \text{con muestras i.i.d.} \\ \text{con } P(X) \text{ estable.} \end{cases}$
+ Para tener éxito, debemos encontrar $h : E_{out}(h) \approx 0$.
  * Por ahora, solo garantizamos la inecuación de Hoeffding (que está acotado). ($< 2|\mathcal{H}| e^{-\epsilon^2 N}$)

Juntas → $E_{out} = 0$
→ ¿Podemos asegurar que $E_{out} \approx E_{in}$ lo suficiente? → Inecuación Hoeffding.
→ ¿Podemos reducir E_{in} lo suficiente? → Algoritmo de aprendizaje.

+ Dilema: ¿qué $\mathcal{H}$ elegir?
  * Si $\mathcal{H}$ es compleja → $|\mathcal{H}| \uparrow\uparrow$ $\begin{cases} E_{in} \downarrow\downarrow \\ \uparrow\uparrow |E_{in} - E_{out}| \end{cases}$ → Necesitamos $\uparrow\uparrow N$.
  * Si $\mathcal{H}$ es simple → $|\mathcal{H}| \downarrow\downarrow$ $\begin{cases} E_{in} \uparrow\uparrow \text{ (no se ajusta tanto)} \\ E_{in} \approx E_{out} \text{ con menos ejemplos.} \end{cases}$

→ NOTA: $|\mathcal{H}|$ sube cuantos más parámetros tenga que ajustar.

TEOREMA FUNDAMENTAL DEL APRENDIZAJE PAC (son equivalentes)

+ $\mathcal{H}$ es PAC-aprendible ↔ $\mathcal{H}$ es PAC-aprendible agnóstica ↔ Cualquier ERM es un aprendiz eficaz en PAC / PAC-agnóstico ↔ $\mathcal{H}$ tiene VC-dimensión finita. (T.4)