AA - TEMA 4: TEORÍA DE LA GENERALIZACIÓN

TRUCO DE LA DISCRETIZACIÓN

+ Si $|\mathcal{H}| = \infty$, la desigualdad de Hoeffding no nos sirve.
+ Sin embargo, $\mathcal{H}$ al final ajusta valores discretos (con precisión de 64 bits, por ejemplo) $\rightarrow |\mathcal{H}| = 2^{64 \cdot d}$ (d: Nº Parámetros)

$$\rightarrow m_{\mathcal{H}}(\epsilon, \delta) \leq \left\lceil \frac{2}{\epsilon^2} \log \frac{2|\mathcal{H}|}{\delta} \right\rceil = \frac{128 d + 2 \log\left(\frac{2}{\delta}\right)}{\epsilon^2}$$

+ Es una cota horrible de la complejidad de la muestra.

TRAMPA DE LA DESIGUALDAD UNIFORME

+ Para llegar a la desigualdad de Hoeffding para $|\mathcal{H}| > 1$, consideramos: que $P\left(\bigcup_{i=1:|\mathcal{H}|} B_i\right) \leq \sum_{i=1}^{|\mathcal{H}|} P(B_i)$.
  * Esto es, que no nos preocupamos por el valor real, sino que tomamos que todas las probabilidades son disjuntas ($B_i \cap B_j = \emptyset$), lo cual no es verdad.
+ Por tanto, $P(D: |E_{in}(g) - E_{out}(g)| > \epsilon) < 2|\mathcal{H}| e^{-2N\epsilon^2}$ es muy mejorable.

FUNCIÓN DE CRECIMIENTO *Clasificación binaria

+ Medida del nº de "funciones efectivas" en una clase, dados N puntos.
  * "Función efectiva": función que crea una clasificación concreta. Si dos f crean la misma clasificación con N puntos $\rightarrow$ Son 1 f. efectiva.
+ Función de crecimiento $(m_{\mathcal{H}}(N)) = \max_{x_1 \dots x_N} |\mathcal{H}(x_1, \dots x_N)| \leq 2^N$.
  ↳ Nº elementos del "conjunto de las dicotomías generables por $\mathcal{H}$.
  * Si $m_{\mathcal{H}}(N) = 2^N \rightarrow \mathcal{H}$ "shatters" the set.
  * Es independiente de $P \rightarrow$ Peor caso.
+ A través de $m_{\mathcal{H}}(N)$, podemos deducir (no trivialmente).
  * $P(D: |E_{in}(h) - E_{out}(h)| > \epsilon) \leq 4\, m_{\mathcal{H}}(2N) e^{-\frac{1}{8}N\epsilon^2}$.
  * Despejando ϵ: $E_{out}(h) \leq E_{in}(h) + \sqrt{\frac{8}{N} \log \frac{4 m_{\mathcal{H}}(2N)}{\delta}}$
  * Problema: tenemos que calcular $m_{\mathcal{H}}(N) \rightarrow$ Sabemos que $\leq 2^N$.

PUNTO DE RUPTURA

+ Si para algún valor k $m_{\mathcal{H}}(k) < 2^k \rightarrow k$ es pto. ruptura de $\mathcal{H}$
  * $\mathcal{H}$ no puede separar una muestra de tamaño k.
+ Resultado Sauer - Shelah - Perles: si k es punto de ruptura de $\mathcal{H}$: $\forall N,\ m_{\mathcal{H}}(N) \leq \sum_{i=0}^{k-1} \binom{N}{i}$, que es un polinomio de grado $k-1$.
  * Si $\mathcal{H}$ no tiene break point $\rightarrow m_{\mathcal{H}}(N) = 2^N$.
+ Sustituyendo:
  * Si $\mathcal{H}$ tiene break point $\rightarrow \delta = 4 \cdot O(N^k) \cdot e^{-\frac{1}{8}N\epsilon^2} \xrightarrow[N\to\infty]{} 0$.
  * Si $\mathcal{H}$ no tiene break point $\rightarrow \delta = 4 \cdot 2^N \cdot e^{-\frac{1}{8}N\epsilon^2} \not\rightarrow 0$.
    ↳ Aumentando N, no nos aseguramos de disminuir E_{out}.

DIMENSIÓN VC - VAPNIK & CHERVONENKIS

+ La dimensión VC de $\mathcal{H}$ ($d_{vc}(\mathcal{H})$) es el mayor valor de N tal que N no es un punto de ruptura.
  * Si $\mathcal{H}$ no tiene b.p $\rightarrow d_{vc}(\mathcal{H}) = \infty$.
+ Puede probarse que $m_{\mathcal{H}}(N) \leq \sum_{i=0}^{d_{vc}} \binom{N}{i} \leq \begin{cases} N^{d_{vc}} + 1 \\ \left(\frac{eN}{d_{vc}}\right)^{d_{vc}} \end{cases}$
+ La dimensión VC denota el nº de parámetros "efectivos" de $\mathcal{H}$.
  * P.e: $d_{vc} = d+1$ para modelos lineales de "d" dimensiones.
+ Sustituyendo:

$$E_{out}(h) \leq E_{in}(h) + \sqrt{\frac{8}{N} \log \frac{4((2N)^{d_{vc}}+1)}{\delta}} = E_{in}(h) + O\left(\sqrt{d_{vc} \frac{\log N}{N}}\right)$$

  * Si d_{vc} es finito y N muy grande $\rightarrow$ Garantizamos generalización
  * Si $d_{vc} = \infty$, $\mathcal{H}$ no sirve para aplicar la regla ERM.

COMPLEJIDAD DE LA MUESTRA

+ ¿N → conseguir ε (dif. entre dentro y fuera de la muestra). ?
  δ (prob. de que se cumpla la premisa, dado ε).

$$\sqrt{\frac{8}{N}\ln\frac{4m_H(2N)}{\delta}} \le \varepsilon \to N \ge \frac{8}{\varepsilon^2}\ln\left(\frac{4m_H(2N)}{\delta}\right) \to$$

$$\to N \ge \frac{8}{\varepsilon^2}\ln\left(\frac{4((2N)^{d_{vc}}+1)}{\delta}\right)$$

Ecuación implícita → Resolver iterativamente.

* Ejemplo: Perceptrón ($d_{vc} = 3$). (Init.)

$$N \ge \frac{8}{0'1^2}\ln\left(\frac{4(2N)^3+4}{0'1}\right) \xrightarrow{N=1000} N \ge 21.100 \longrightarrow N \ge 30.000 \text{ aprox.}$$

ε = Dif. Ein/out — δ = Prob. de encontrar "ejemplo malo" — punto fijo de N.

+ Según esta fórmula, las cotas crecen proporcionalmente a d_{vc}.
+ Regla práctica: $N > 10 \cdot d_{vc}$ (¡Qué poco!)

VC COMO RESTRICCIÓN DE LA COMPLEJIDAD DEL MODELO.

+ Normalmente → N es fijo. ¿Qué podemos esperar alcanzar con N?
  * $E_{out}(g) \le E_{in}(g) + \sqrt{\frac{8}{N}\ln\frac{4m_H(2N)}{\delta}} \le E_{in}(g) + \sqrt{\frac{8}{N}\ln\frac{4(2N)^{d_{vc}}+4}{\delta}}$
+ Penalización por complejidad del modelo:
  * $\Omega(N,\mathcal{H},\delta) = \sqrt{\frac{8}{N}\ln\frac{4(2N)^{d_{vc}}+4}{\delta}} = O\left(\sqrt{\frac{d_{vc}\cdot \ln N - \ln\delta}{N}}\right)$
+ Intercambio $\begin{cases} - \uparrow\uparrow d_{vc} \to \uparrow\uparrow \text{ probabilidad de que } f\ (E_{in} \approx 0). \\ - \downarrow\downarrow d_{vc} \to \uparrow\uparrow \text{ probabilidad de generalizar mejor } (E_{in} \approx E_{out}). \end{cases}$

$$E_{out} \le E_{in} + \Omega(d_{vc}).$$

* Análisis VC sólo depende de $\mathcal{H}$ → Independiente de $\begin{cases} - f \\ - P(x) \\ - A \end{cases}$
* Principalmente aplicable a clasificación y regresión.
* Cota superior muy amplia.

CONJUNTOS DE TEST

+ Idea: usar parte de la muestra para estimar E_{out} → Test set.
  * Estos ejemplos no se usan para entrenar.
  * Deben ser i.i.d de la misma $P(X)$ del conjunto de entrenam.
+ Error: $E_{test} \approx E_{out}$
+ ¿Por qué funciona? → Desigualdad de Hoeffding.
  * $P(|E_{test}(g) - E_{out}(g)| > \epsilon) \le 2e^{-2N\epsilon^2}$
    ↳ Para $N = 1000$: $E_{test} = \pm 5\%\ E_{out}$ con probabilidad $\ge 98\%$.
+ Contras: no usamos los ejemplos para entrenar → $\uparrow\uparrow E_{in}$.

DISCUSIÓN NLT

+ ¿Influyen las transformaciones no lineales en la cota VC?
  * Si fijamos $\mathcal{H}$ a ciegas → $d_{vc}(\mathcal{H}_{\text{rectas}}) = d_{vc}(\mathcal{H}_{\text{círculos}})$ con probabilidad $1-\delta$.
- ¿Qué pasa si probamos una $\mathcal{H}'$, fallamos, y probamos otra $\mathcal{H}'$?
  * Equivale a usar Φ < manteniendo características recta / añadiendo términos al cuadrado.
  * $\uparrow\uparrow d_{vc}$.
- ¿Qué pasa si miramos los datos sin fijar antes una hipótesis?
  * Debemos incluir las hipótesis que hemos pensado → dimensión de transformación.
  * Hemos decidido que el problema es la muestra, y no la población.
+ Si necesitamos ajustar bien → NLT $\begin{cases} - \uparrow\uparrow d_{vc} \to \uparrow N \text{ para PAC.} \\ - \uparrow\uparrow X \text{ (espacio de características).} \end{cases}$

APRENDIZAJE NO UNIFORME

+ ¿Qué pasa si $d_{vc} = +\infty$? → No nos sirve la regla ERM.
+ Consideramos $\mathcal{H} = \bigcup_n \mathcal{H}_n$, $d_{vc}(\mathcal{H}_n) < \infty \ \forall n$.
  * $\mathcal{H}$ es la unión de infinitas clases con d_{vc} finita.
+ SRM (Structural Risk Minimization) $\begin{cases} 1. \text{ Seleccionar secuencia de } \mathcal{H}\text{'s anidadas.} \\ 2. \text{ Estimar } g \text{ para cada set } \mathcal{H}. \end{cases}$
  * Dos tipos $\begin{cases} - \text{Fija complejidad} \to \downarrow \text{error empírico.} - g^* = \arg\min(E_{in}(g) + \Omega(\mathcal{H}_i)) \\ - \text{Fija error emp.} \to \downarrow \text{VC - dimensión (si } \uparrow\uparrow \text{dim} \to \uparrow\uparrow \text{variabilidad).} \end{cases}$
  * Útil cuando $\frac{N}{d_{vc}} < 20$ → ERM no tiene garantías.

AA - TEMA 4 : TEORÍA DE LA GENERALIZACIÓN (II)

INTERCAMBIO BIAS - VARIANZA

+ $E_{out}(g^{(D)}) = \mathbb{E}_x[(g^{(D)}(x) - f(x))^2]$ → Error cuadrático medio de $g^{(D)}$

 ↳ Esperanza (media) en población $P(x)$.

 ↓ Hipótesis g obtenida con A a partir de muestra D.

* Toma en cuenta $\mathcal{H}$ y A.

+ $g^{(D)}(x)$ será la clasificación que la hipótesis g, creada a partir de D, dará a X. Esta variará según D de forma aleatoria.

+ $\bar{g}(x) = \mathbb{E}_D[g^D(x)] \approx \frac{1}{k}(g^1(x) + \dots + g^k(x))$

 * Es la predicción media de x.

+ $var(x) = \mathbb{E}_D[(g^D(x) - \bar{g}(x))^2] = \mathbb{E}_D[g^D(x)^2] - \bar{g}(x)^2$.

 * Es lo que varía, de media, el error cuadrático de una predicción con su valor medio. Es decir, la varianza.

+ $E^D_{out}(x) = (g^D(x) - f(x))^2$ * Error cuadrático de g^D.

+ $E_{out}(x) = \mathbb{E}_D[E^D_{out}(x)]$ * Media de los errores de g^D respecto a f.

DESCOMPOSICIÓN BIAS - VARIANZA

$E_{out} = \mathbb{E}_D[E_{out}(g^D)] = \mathbb{E}_D[\mathbb{E}_x[(g^D(x) - f(x))^2]]$ - - - →

 En este caso, podemos intercambiar orden.

--→ $\mathbb{E}_D[(g^D(x) - f(x))^2] = \mathbb{E}_D(g^D(x)^2) - 2\,\mathbb{E}_D(g^D(x) \cdot f(x)) + f(x)^2$ --→

* Sustituimos $\mathbb{E}_D(g^D(x)) = \bar{g}(x)$:

--→ $\mathbb{E}_D(g^D(x)^2) - 2\,\cancel{\mathbb{E}_D(g(x))}^{\,\bar{g}(x)} \cdot f(x) + f(x)^2 =$

$= \underbrace{\mathbb{E}_D(g^D(x)^2) - \bar{g}(x)^2}_{var(x)} + \underbrace{\bar{g}(x)^2 - 2\bar{g}(x)f(x) + f(x)^2}_{bias(x)}$ // Añadimos y restamos $\bar{g}(x)^2$

$$\boxed{E_{out} = \mathbb{E}_x[var(x) + bias(x)] = bias + variance}$$

CONSECUENCIAS

+ $\mathbb{E}_D[E_{out}(g^D(x))] = \sigma^2 + bias + varianza$ (con señal ruidosa).

 * σ^2 es la varianza del ruido.

 * σ^2 es inevitable → nos centramos en el resto.

+ Hay dos posibilidades:

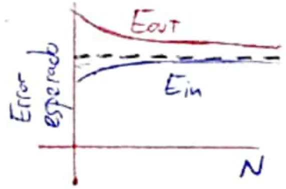
 → ↓↓↓ varianza con ↑ bias : Regularización.

 → ↓↓↓ bias con ↑ varianza ; Conocimiento previo. Para minimizar bias, tenemos que "conocer" f.

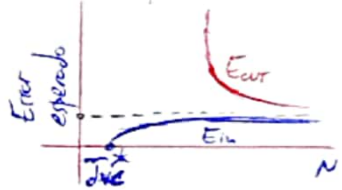
+ El poder del aprendizaje está vinculado a N (tamaño de la muestra).

CURVA DE APRENDIZAJE

+ Resume el comportamiento de E_{in} y E_{out} (en nuestro caso, el error medio de todas las muestras D posibles), cuando variamos N.



Función simple (polinomio grado 2)



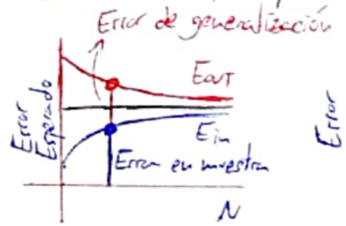
Modelo complejo (polinomio grado 10)

* Si $N \leq d_{vc}$, entonces la función es capaz de ajustarse totalmente

 ↓

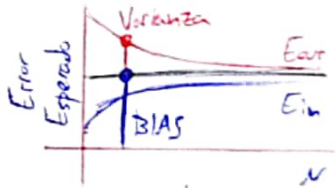
 $E_{in} = 0$

+ Según el enfoque, podemos interpretarlas de diferente manera:



Análisis VC

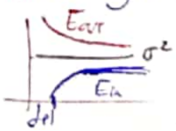
+ H ⇉ que generalice / encaje con datos



Análisis bias - varianza

+ H, A ⇉ Aproxime f / No tenga comportamiento con ↑↑ varianza.

+ En regresión lineal ⇉ E_{in} aumenta cuando el modelo absorbe info con $d+1$ param. / E_{out} ~~aumenta~~ disminuye hasta alcanzar el ruido residual.



Si N ↑↑ (sube)