

TEMA 6: VALIDACIÓN Y SELECCIÓN DE MODELOS

VALIDACIÓN

- + Idea: usar parte de $D \rightarrow$ estimar E_{out} directamente.
- * Si tenemos k elementos de D y entrenamos con el resto:

$$\mathbb{E}(E_{VAL}(h)) = E_{out}(h)$$

\hookrightarrow Media de los errores en los k elementos.

$$\text{var}(E_{VAL}(h)) = \frac{1}{k^2} \sum_{i,j} \text{cov}(e_i, e_j) = \frac{1}{k^2} \sum_i \text{var}(e_i) = \frac{\sigma^2}{k}$$

(k elem. son independientes $\rightarrow \text{cov}(k_1, k_2) = 0 \parallel \text{cov}(k_1, k_1) = \text{var}(k)$)

* Entonces, $E_{VAL}(h) = E_{out}(h) \pm O\left(\frac{1}{\sqrt{k}}\right)$ (desv. típica σ).

+ Proceso general:

1. Tomamos k elementos de D para validación.
2. Aprendemos con $D_{train} \rightarrow g^-$ (hipótesis de D_{train}).
3. Usamos D_{VAL} para test $\rightarrow E_{VAL}$.
4. Usamos E_{VAL} para estimar E_{out} .

* Una vez tenemos la estimación \rightarrow entrenamos con $D \rightarrow g$.
Como usamos más ejemplos que con g^- , podemos esperar $E_{out} \downarrow$.

$$E_{out}(g) \leq E_{out}(g^-) \leq E_{VAL}(g^-) + O\left(\frac{1}{\sqrt{k}}\right)$$

(N ejemplos) ($N-k$ ejemplos)

ELECCIÓN DE k .

- + Si k es pequeño $\rightarrow \uparrow O\left(\frac{1}{\sqrt{k}}\right) \rightarrow$ Nuestra estimación "Eval" será peor, pues tenemos mucha varianza con pocos puntos.
- + Si k es grande $\rightarrow \downarrow$ n° de ejemplos con los que entrenamos $\rightarrow g^-$ será peor $\rightarrow E_{VAL}$ no estimará bien $E_{out}(g)$, ya que entrenamos con muchos más ejemplos que con g^- .
- * Regla práctica: usar $\frac{1}{5}$ de los datos para D_{VAL} .

SELECCIÓN DE MODELOS

- + Uno de los principales usos de $D_{VAL} \rightarrow$ Elegir modelo \mathcal{H} .
- * Para varios \mathcal{H} y λ 's (si usa regularización), entrenamos con D_{train} y evaluamos con D_{test} .
- * Elegimos (\mathcal{H}, λ) tq tenga menor E_{VAL} : g_{min}^* $\rightarrow N \cdot \mathcal{H}$'s entrenadas.

⚠ Estaríamos tomando una decisión en base a $D_{VAL} \rightarrow$ Estamos "seleccionando"/contaminando la muestra.

* ¿Qué significa eso? $\rightarrow E_{VAL}(g_{min}^*)$ deja de ser un estimador de E_{out} insesgado, ya que elegimos deliberadamente "el mejor" en la muestra. Si utilizamos muchos modelos ($\rightarrow \infty$), entonces es posible que el mejor de ellos sea "por casualidad".

\rightarrow Si elegimos entre M modelos:

$$\text{Hoeffding: } E_{out}(g_{min}^*) = E_{VAL}(g_{min}^*) + O\left(\sqrt{\frac{\log M}{k}}\right)$$

(versión unión).

$$\text{Hoeffding: } E_{out}(g_{min}^*) = E_{VAL}(g_{min}^*) + O\left(\sqrt{dvc \frac{\log k}{k}}\right)$$

(versión VC)

\hookrightarrow Por tanto, elegir λ realmente es $dvc = 1$, y no $M \rightarrow \infty$.

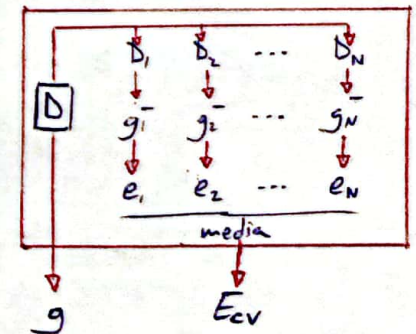
* Regla práctica: 10 ejemplos / dimensión.

LEAVE - ONE - OUT

- + Gestiona el problema de elección de k .
- + Para cada punto presente en D , generamos un modelo con D sin dicho punto (será D_{VAL}).

$$* E_{cv} \leq \frac{1}{N} \sum_{i=1}^N E_{VAL}(g_i^-)$$

* A pesar de que hay covarianza, resulta que $E_{cv} \approx E_{out}(N-1)$.



VALIDACIÓN CRUZADA

- + Generalización de leave-one-out.
- + Surge para aligerar el proceso de aprendizaje (es menos costosa).
- + Se divide en V -partes D ("V-Fold Cross Validation").
 - Usas $V-1$ para entrenar y 1 para validar.
 - Repites el proceso para cada "fold" (partición).
 - $E_{cv} = \text{media}(E_{VAL}(D_v))$.

APÉNDICE: TRES PRINCIPIOS DE APRENDIZAJE


NAVAJA DE OCCAM

- + Si tenemos que elegir entre varias opciones, a priori iguales, elige la más sencilla → 44% de ser cierta.
- + ¿Por qué?
 - * Matemáticamente: \leq Hoeffding... (dimensión VC)
 - * Intuitivamente: H s simples tienen menos diversidad.
 - Si encuentras una h que se ajusta a los datos, es más significativo si es simple, pues había 66% de que pasase.
- + Axioma de no falsificabilidad: si un experimento no tiene la oportunidad de contradecir la hipótesis → No significa nada.
 - * Si H puede dividir los datos de cualquier manera, ¿de qué sirve?
 - * $TP[\text{falsificación}] \geq 1 - \frac{m_H(N)}{2^N}$ → Función de crecimiento (si $d_{VC} = \infty \rightarrow m_H(N) = 2^N$)
Tasa de funciones/divisiones no realizables por H . $TP[\text{falsificación}] = 1$
- + Incluso, a veces podemos optar por un modelo más simple de lo realmente necesario → El precio a pagar por H compleja puede ser mayor que el beneficio del ajuste que da.

MUESTRA SESGADA

- + Si los datos de aprendizaje están sesgados, el aprendizaje produce un algoritmo igualmente sesgado.
- + Podemos intentar arreglarlo si somos conscientes de ello
 - Adaptamos D para que la distribución de los datos en él se aproxime a la población.
 - Esto no funciona si no tenemos NINGÚN dato así en D , pero después lo hay en D_{test} .

DATA SNOOPING

- + Si la muestra influye en algo el proceso de aprendizaje, no podemos asegurarnos de producir resultados buenos.
- * Mirar los datos 
 - Eres tú el algoritmo de aprendizaje. ¿Qué duc tienes?
 - * Sin embargo, puedes (y debes) recopilar información sobre f , las entradas (rangos, relaciones, varianzas) → 44% elegir mejor H .
- * Usar los datos de test para elegir H (para eso está D_{VAL}) o para calcular lo que sea que influya en el aprendizaje (p.e. normalizar).
- * "Si torturas los datos lo suficiente, confesarán".
 - Probamos un montón de modelos, y al final uno funciona.
 - ¿duc(H)? = unión de todos los modelos.
 - Si lees lo que otros hicieron con el mismo D , también afecta.
- + Solución 1: evitar todo lo anterior (¡suerte!).
- + Solución 2: vivir con ello, y tenerlo en cuenta.
 - Calcula la contaminación de los datos que realizas, y considérala.