Cuestionario 2 - AA

Óscar David López Arcos.

---

1. Identificar de forma precisa dos condiciones imprescindibles para que un problema de predicción puede ser aproximado por inducción desde una muestra de datos. Justificar la respuesta usando los resultados teóricos estudiados.

Las muestras han de ser independientes e idénticamente distribuidas.

Independientes: La ocurrencia de cada valor no debe afectar la probabilidad de que ocurra otro.

Idénticamente distribuidas: Todas las muestras tienen la misma distribución de probabilidad, es decir, tomando una muestra arbitraria un valor es más probable que otro solamente si éste tiene mayor presencia dentro del conjunto.

De este modo, un número alto de muestras aleatorias representaría fielmente al conjunto, ya que la diferencia entre la muestra y el conjunto se aproxima a 0 según la inecuación de Hoeffding:

$$\mathbb{P}(D : |\mu - \nu| > \epsilon) \leqslant 2e^{-2\epsilon^2 N}$$

2. El jefe de investigación de una empresa con mucha experiencia en problemas de predicción de datos tras analizar los resultados de los muchos algoritmos de aprendizaje usados sobre todos los problemas en los que la empresa ha trabajado a lo largo de su muy dilatada existencia, decide que para facilitar el mantenimiento del código de la empresa van a seleccionar un único algoritmo y una única clase de funciones con la que aproximar todas las soluciones a sus problemas presentes y futuros. ¿Considera que dicha decisión es correcta y beneficiará a la empresa? Argumentar la respuesta usando los resultados teóricos estudiados.

La decisión no es correcta, para solucionar cada problema se debe buscar el algoritmo y tipo de funciones que mejor se adapten a éste.

Esto se ve reflejado (para el caso del algoritmo) en el Teorema de No Barra Libre (NFL Theorem), que explica que no existe un único algoritmo que resuelva todos los problemas ya que todos tienen una probabilidad de fallar.

Del mismo modo, según la distribución de las muestras algunas clases de funciones clasificarán mejor que otras. Esto lo pudimos comprobar en prácticas, cuando clasificar puntos con funciones más complejas no necesariamente generaba mejores resultados que las lineales.

3. Supongamos un conjunto de datos D de 25 ejemplos extraídos de una función desconocida $f : X \to Y$, donde $X = \mathbb{R}$ e $Y = \{-1, +1\}$. Para aprender f usamos un conjunto simple de hipótesis $H = \{h_1, h_2\}$ donde h_1 es la función constante igual a h_1 y h_2 la función constante igual a -1. Consideramos dos algoritmos de aprendizaje, S(smart) y C(crazy). S elige la hipótesis que mejor ajusta los datos y C elige deliberativamente la otra hipótesis.

Fuera de la muestra no es posible garantizar nada. Es cierto que si los datos cumplen las condiciones i.i.d. la probabilidad de que S genere un mejor comportamiento que C es alta y aumenta conforme el tamaño de la muestra. Sin embargo, nada garantiza que el comportamiento de C fuera de la muestra sea mejor, puesto que la población es desconocida.

4. Con el mismo enunciado de la pregunta 3:

a) Asumir desde ahora que todos los ejemplos en D tienen $y_n = +1$. ¿Es posible que la hipótesis que produce C sea mejor que la hipótesis que produce S?. Justificar la respuesta.

Al ser una muestra no tiene por qué ser representativa de la totalidad del problema. Puede darse el caso que los 25 datos extraídos en la muestra sean los únicos del conjunto con etiqueta +1.

Por tanto, sería posible que la hipótesis de C fuese mejor que la producida por S. Como dice en las transparencias: "El mal caso es posible, aunque poco probable".

5. Considere la cota para la probabilidad del conjunto de muestras de error D de la hipótesis solución g de un problema de aprendizaje, a partir de la desigualdad de Hoeffding, sobre una clase finita de hipótesis,

$$\mathbb{P}[\,|E_{out}(g) - E_{in}(g)| > \epsilon)] < \delta(\epsilon, N, |H|)$$

a) Dar una expresión explícita para $\delta(\epsilon, N, |H|)$

$$2|H|e^{-2\epsilon^2 N}$$

b) Si fijamos $\epsilon = 0{,}05$ y queremos que el valor de δ sea como máximo $0{,}03$, ¿cual será el valor más pequeño de N que verifique estas condiciones cuando $H = 1$?

$$2|H|e^{-2\epsilon^2 N} = \delta(\epsilon, N, |H|) \quad \Rightarrow \quad 2e^{-2\epsilon^2 N} \leq 0{,}03 \quad \Rightarrow \quad e^{-2\epsilon^2 N} \leq 0{,}03/2 \quad \Rightarrow$$

$$\Rightarrow \log(e^{-2\epsilon^2 N}) \leq \log(0{,}03/2) \Rightarrow -2\epsilon^2 N \leq \log(0{,}03/2) \Rightarrow N \geq \log(0{,}03/2)/-2\epsilon^2 N \Rightarrow$$

$$\Rightarrow \quad N \geq \log(0{,}03/2)/-2\epsilon^2 N \quad \Rightarrow \quad N \geq 839{,}9 \Rightarrow N = 840$$

c) Repetir para $H = 10$ y para $H = 100$

$$H = 10 \Rightarrow \quad N \geq \log(0{,}03/20)/-2\epsilon^2 N \quad \Rightarrow \quad N \geq 1300{,}4 \Rightarrow N = 1301$$

$$H = 100 \Rightarrow \quad N \geq \log(0{,}03/200)/-2\epsilon^2 N \quad \Rightarrow \quad N \geq 1760{,}98 \Rightarrow N = 1761$$

6. Considere la cota para la probabilidad del conjunto de muestras de error D de la hipótesis solución g de un problema de aprendizaje, a partir de la desigualdad de Hoeffding, sobre una clase finita de hipótesis,

$$\mathbb{P}[\,|E_{out}(g) - E_{in}(g)| > \epsilon)] < \delta$$

a) ¿Cuál es el algoritmo de aprendizaje que se usa para elegir g?

La función que mejor ajuste la muestra de datos, es decir, aquella de menor $\left[f(x) \neq g(x)\right]$

b) Si elegimos g de forma aleatoria, ¿seguiría verificando la desigualdad?

Sí, la desigualdad de Hoeffding se verifica para toda g, independientemente de como esta se elija. Como δ depende del valor de ϵ, siempre se puede garantizar que $E_{out} \leq E_{in} + \epsilon$ con una probabilidad de $1 - \delta$

c) ¿Depende g del algoritmo usado?

Sí. Dependiendo del algoritmo, su criterio de parada o el método de ajuste, la función g podría ser diferente.

d) ¿Es una cota ajustada o una cota laxa?

Es una cota ajustada, ya que viene dada por la inecuación para las clases finitas $m_H(\epsilon, \delta) \leq \left\lceil \frac{1}{\epsilon^2} log \frac{2|H|}{\delta} \right\rceil$ que da un ajuste muy aproximado para la complejidad requerida.

7. ¿Por qué la desigualdad de Hoeffding no es aplicable de forma directa cuando el número de hipótesis de H es mayor de 1? Justificar la respuesta.

La desigualdad de Hoeffding supone una clase $H = h$, de una única función. Esta h está fijada antes de conocer la muestra, es por eso que podemos calcular de forma directa $E_{in}(h)$ y $E_{out}(h)$.

Sin embargo, para un $|H| > 1$, las probabilidades de error de todos los $E_{in}(h_i)$, $E_{out}(h_i)$ con $i \in \{1,2,...,|H|\}$ y $h_i \in H$ no son calculables directamente debido a que todos estos sucesos son compatibles. Para calcularlos, habría que tener en cuenta la probabilidad de intersección de estos, tal y como dictan las leyes de probabilidad de unión de sucesos compatibles.

Es por eso que para $|H| > 1$ la hipótesis g es identificada según los datos y no fijada previamente, haciendo imposible aplicar la desigualdad de forma directa.

8. Si queremos mostrar que k^* es un punto de ruptura para una clase de funciones H cuales de las siguientes afirmaciones nos servirían para ello:

a) Mostrar que existe un conjunto de k^* puntos $x_1, \ldots, x_{k^*}$ que H puede separar ("shatter").

Falso. Si existe dicho conjunto, k^* no sería punto de ruptura.

b) Mostrar que H puede separar cualquier conjunto de k^* puntos.

Falso. En este caso k^* tampoco sería punto de ruptura

c) Mostrar un conjunto de k^*puntos $x_1, \ldots, x_{k^*}$ que H no puede separar.

Falso. Al definirse el punto de ruptura como $m_H(k^*) < 2^{k^*}$ y ser $m_H(k^*)$ el máximo número de patrones binarios generados por H, con que exista sólo un conjunto $x_1, \ldots, x_{k^*}$ que sea separable k^* no sería punto de ruptura.

d) Mostrar que H no puede separar ningún conjunto de k^* puntos.

Correcto.

e) Mostrar que $m_H(k^*) = 2^{k^*}$

Falso. Para que sea punto de ruptura, $m_H(k^*) < 2^{k^*}$

9. Para un conjunto H ***con*** $d_{vc} = 10$***, ¿que tamaño muestral se necesita (según la cota de generalización) para tener un 95% de confianza*** (δ) ***de que el error de generalización*** (ϵ) ***sea como mucho 0.05?***

Aplicamos la ecuación implícita $N \geq \frac{8}{\epsilon^2} \log\left(\frac{4((2N)^{d_{vc}} + 1)}{1 - \delta}\right)$ y resolvemos iterativamente:

$$
\begin{aligned}
N = 1000 \quad &\Rightarrow \quad N \geq 257251{,}36 \\
N = 257251{,}36 \quad &\Rightarrow \quad N \geq 434853{,}08 \\
N = 434853{,}08 \quad &\Rightarrow \quad N \geq 451651{,}62 \\
N = 451651{,}62 \quad &\Rightarrow \quad N \geq 452864{,}52 \\
N = 452864{,}52 \quad &\Rightarrow \quad N \geq 452950{,}34 \\
N = 452950{,}34 \quad &\Rightarrow \quad N \geq 452956{,}40 \\
N = 452956{,}40 \quad &\Rightarrow \quad N \geq 452956{,}83
\end{aligned}
$$

$$452956{,}83 - 452956{,}40 = 0{,}43 < 1$$

$$N \geq 452956{,}83 \quad \Rightarrow \quad N = 452957$$

10. Considere que le dan una nuestra de tamaño N de datos etiquetados $\{-1, +1\}$ ***y le piden que encuentre la función que mejor ajuste dichos datos. Dado que desconoce la verdadera función*** f***, discuta los pros y contras de aplicar los principios de inducción ERM y SRM para lograr el objetivo. Valore las consecuencias de aplicar cada uno de ellos.***

ERM:

Pros:

- Si la $d_{vc}(H)$ es finita, H se puede aprender mediante un algoritmo (Buena generalización).
- Con una cantidad suficiente de datos en el aprendizaje y/o un modelo menos complejo, podemos hacer $E_{in} \approx E_{out}$.
- Aumentando la complejidad de H, $E_{in} \approx 0$

Contras:

- Con datos insuficientes, reducir E_{in} a 0 implica un E_{out} mayor.
- La cantidad de datos necesaria para un aprendizaje satisfactorio es muy alta.
- Reducir d_{vc} para conseguir $E_{in} \approx E_{out}$ no garantiza que hayamos aprendido algo.

SRM:

Pros:

- Al trabajar sobre conjuntos más pequeños evita el sobreajuste y es de gran utilidad cuando $N/d_{vc} < 20$, ya que aplicando $O\left(\sqrt{(d_{vc} ln N - ln\delta)/N}\right)$ no se garantiza un $E_{out} \approx 0$

Contras:

- La aplicación es más difícil, ya que la división de H en subclases de menor a mayor complejidad es bastante costosa.