Aprendizaje Automático (2017-2018)
Doble Grado en Ingeniería Informática y Matemáticas
Universidad de Granada

Cuestionario 2

Simón López Vico

29 de mayo de 2018

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.

   Solución:

   El aprendizaje por inducción es una forma de razonamiento que parte de examinar una serie de casos específicos, conduciendo a una conclusión.

   Una condición imprescindible para aplicar inducción es que la muestra de datos presente patrones los cuales sean posibles de diferenciar sobre el resto de datos; por tanto, si estamos trabajando ante una muestra generada de manera aleatoria, no tiene por qué haber una serie de diferencias entre los datos que nos permitan aplicar inducción sobre ellos.

   Otra condición para aplicar inducción de manera correcta es que tengamos suficientes datos con distintas condiciones; por ejemplo, en el caso del *Inductivist Turkey*, el pavo comprueba que le den de comer los días lluviosos, los días soleados, los fines de semana... para así poder inferir inductivamente con mayor precisión.

   Aún así, el método de inducción siempre está sujeto al fallo, por muchos datos que comprobemos que cumplan los patrones que hemos encontrado; no importa cuántos casos enumeremos durante nuestro razonamiento inductivista, nada garantiza que el siguiente caso esté en la inferencia que hemos deducido de nuestras observaciones, ya que los posibles experimentos y observaciones son infinitos.

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.

   Solución:

   El jefe de investigación del que se habla, al analizar los datos de los resultados de los algoritmo usados, está realizando una predicción por inducción. De esta manera, al elegir dicho algoritmo y clase de funciones, está inferenciando a partir de los datos obtenidos mediante la observación de los resultados.

   Por tanto, fijar una clase de funciones y un algoritmo puede beneficiar a la empresa siempre y cuando se haya hecho un profundo estudio sobre los datos obtenidos en anteriores problemas. Aún

así, por el *No-Free-Lunch Theorem*, nuestra decisión siempre estará sujeta a una probabilidad donde la elección que hemos hecho fallará.

Así, dicha decisión será más correcta y más beneficiosa cuanto más a fondo estudiemos los resultados obtenidos en anteriores problemas.

3. Supongamos un conjunto de datos $\mathcal{D}$ de 25 ejemplos extraídos de una función desconocida $f : \mathcal{X} \to \mathcal{Y}$ donde $\mathcal{X} = \mathbb{R}$ e $\mathcal{Y} = \{-1, +1\}$. Para aprender f usamos un conjunto muy simple de hipótesis $\mathcal{H} = \{h_1, h_2\}$, donde h_1 es la función constante igual a $+1$ 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 deliberadamente la otra hipótesis.

   a) ¿Puede S producir una hipótesis que garantice mejor comportamiento que la aleatoria sobre cualquier punto fuera de la muestra? Justificar la respuesta.
   Solución:
   En el caso de que todos los datos pertenecientes a $\mathcal{D}$ estuvieran etiquetados con el mismo valor, sería posible que S garantizase un mejor comportamiento que la hipótesis aleatoria. De esta manera, S escogería la hipótesis h_1 (si todos los datos fueran +1) o h_2 (si todos los datos fueran -1), haciendo que C escogiera la hipótesis errónea para todos los datos.

4. Con el mismo enunciado de la pregunta 3:

   a) Asumir desde ahora que todos los ejemplos en $\mathcal{D}$ tienen $y_n = +1$. ¿Es posible que la hipótesis que produce C sea mejor que la hipótesis que produce S? Justificar la respuesta.
   Solución:
   Sí es posible; que todos los datos extraídos valgan +1 no significa que todos los datos en $\mathcal{D}$ valgan +1, y puede dar la casualidad de que los 25 datos extraídos de la muestra sean los únicos etiquetados con +1 que se encuentran en $\mathcal{D}$. Por tanto, puede que C elija una hipótesis mejor que la elegida por S.

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

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

   a) Dar una expresión explícita para $\delta(\epsilon, N, |\mathcal{H}|)$.
   Solución: La función δ se corresponderá con:

$$\delta(\epsilon, N, |\mathcal{H}|) = 2|\mathcal{H}|e^{-2\epsilon^2 N}$$

   b) Si fijamos $\epsilon = 0{,}05$ y queremos que el valor de δ sea como máximo 0,03, ¿cuál será el valor más pequeño de N que verifique estas condiciones cuando $|\mathcal{H}| = 1$?
   Solución:[1] Simplemente, sustituyamos los valores en la función δ y despejemos el valor de N:

$$\epsilon = 0{,}05, \delta \leq 0{,}03, |\mathcal{H}| = 1 \implies 0{,}03 \geq 2 * 1 * e^{-2*0{,}05^2*N}$$

$$\ln(0{,}03) \geq \ln(2 * 1 * e^{-2*0{,}05^2*N}); \quad \ln(0{,}03) \geq \ln(2) + \ln(e^{-2*0{,}05^2*N});$$

$$\ln(0{,}03) - \ln(2) \geq -2 * 0{,}05^2 * N * \ln(e); \quad N \geq \frac{\ln(0{,}03) - \ln(2)}{-2 * 0{,}05^2} \implies N \geq 839{,}941$$

   Por tanto, para verificar estas condiciones, necesitamos que $N \geq 840$, es decir, que el tamaño de la muestra sea al menos 840.

[1] Los cálculos han sido realizadas mediante la plataforma *Wolfram Alpha*.

c) Repetir para $|\mathcal{H}| = 10$ y $|\mathcal{H}| = 100$.

Solución: Resolviendo de la misma manera que en el apartado anterior, para $|\mathcal{H}| = 10$:

$$N \geq \frac{\ln(0{,}03) - \ln(20)}{-2 * 0{,}05^2} \implies N \geq 1300{,}46$$

y para $|\mathcal{H}| = 100$:

$$N \geq \frac{\ln(0{,}03) - \ln(200)}{-2 * 0{,}05^2} \implies N \geq 1760{,}98$$

¿Qué conclusiones obtiene?

Solución: Podemos concluir que cuanto mayor sea el tamaño de la clase de funciones $|\mathcal{H}|$, más datos necesitaremos para obtener el mismo error fijado; es decir, si fijamos el tamaño N, cuanto mayor sea el cardinal de la clase de funciones, menor será el error cometido.

Aún así, hay que tener en cuenta que cuantas más funciones tengamos, mayor será el tiempo de procesamiento, por lo que tendremos que decidir qué preferencias elegimos para nuestro problema de predicción.

6. Considere la cota para la probabilidad del conjunto de muestras de error $\mathcal{D}$ 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?
   Solución: Cualquier algoritmo que aproxime el valor de E_{in} y E_{out} a 0, y que cuanto mayor sea el tamaño del conjunto de datos, menor sea la probabilidad de que E_{in} y E_{out} difieran.

   b) Si elegimos g de forma aleatoria ¿seguiría verificando la desigualdad?
   Solución: Sí, pues el algoritmo debe de actuar para un conjunto de funciones $\mathcal{H}$ que contenga a dicha función g.

   c) ¿Depende g del algoritmo usado?
   Solución: No, la función g que mejor ajuste los datos será siempre la misma para cualquier algoritmo que usemos, pues en todos los algoritmos debe de minimizar el error de entrada y salida.

   d) ¿Es una cota ajustada o una cota laxa?
   Solución: Una cota ajustada.

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

   Solución:

   Cuando el cardinal de $\mathcal{H}$ es finito y mayor que 1, la función de hipótesis g debe de estar fijada antes de conocer el conjunto de datos, pero el algoritmo de aprendizaje usará los datos de entrenamiento para encontrar dicha g.

   Si tomamos el conjunto $\{\mathcal{S} : |E_{in}(g) - E_{out}(g)| > \epsilon\} = \bigcup_{h_i \in \mathcal{H}} \{\mathcal{S} : |E_{in}(h_i) - E_{out}(h_i)| > \epsilon\}$, por la subaditividad de $\mathbb{P}$, la probabilidad de la unión será menor o igual que la suma de todas las probabilidades hasta $|\mathcal{H}|$, es decir, $|\mathcal{H}|$ veces la parte derecha de la desigualdad de Hoeffding, obteniendo la fórmula:

$$\mathbb{P}(\mathcal{S} : |E_{in}(g) - E_{out}(g)| > \epsilon) < 2|\mathcal{H}|e^{-2\epsilon^2 N}$$

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

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

   b) Mostrar que $\mathcal{H}$ puede separar cualquier conjunto de k^* puntos.

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

   d) Mostrar que $\mathcal{H}$ no puede separar ningún conjunto de k^* puntos .

   e) Mostrar que $m_{\mathcal{H}}(k) = 2^{k^*}$

   Solución:

   Por definición, k^* es un punto de ruptura para una clase de funciones $\mathcal{H}$ si ningún conjunto de datos de tamaño k^* puede ser separado por $\mathcal{H}$. Por tanto, la única afirmación que podríamos usar para mostrar que k^* es un punto de ruptura será la opción *d*; el resto de afirmaciones podrían llegar a ser necesarias pero no suficientes.

9. Para un conjunto $\mathcal{H}$ con $d_{VC} = 10$, ¿qué 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?

   Solución:

   Utilizaremos iterativamente la desigualdad $N \geq \frac{8}{\epsilon^2} \ln(\frac{4((2N)^{d_{VC}}+1)}{\delta})$, la cual convergerá al tamaño muestral necesario, usando los valores aportados por el ejercicio (notar que un 95 % de confianza se corresponde con $\delta = 0{,}05$).

   Por tanto, calculemos[2] el tamaño muestral usando como valor inicial $N = 1000$:

$$N \geq \frac{8}{0{,}05^2} \ln(\frac{4((2N)^{10}+1)}{0{,}05}) \underset{N=1000}{\Longrightarrow} N \geq \frac{8}{0{,}05^2} \ln(\frac{4((2*1000)^{10}+1)}{0{,}05}) = 257251{,}36$$

$$\underset{N=257251{,}36}{\Longrightarrow} N \geq \frac{8}{0{,}05^2} \ln(\frac{4((2*257251{,}36)^{10}+1)}{0{,}05}) = 434853{,}08$$

$$\underset{N=434853{,}08}{\Longrightarrow} N \geq \frac{8}{0{,}05^2} \ln(\frac{4((2*434853{,}08)^{10}+1)}{0{,}05}) = 451651{,}63$$

$$\underset{N=451651{,}63}{\Longrightarrow} N \geq \frac{8}{0{,}05^2} \ln(\frac{4((2*451651{,}63)^{10}+1)}{0{,}05}) = 452864{,}52$$

$$\underset{N=452864{,}52}{\Longrightarrow} N \geq \frac{8}{0{,}05^2} \ln(\frac{4((2*452864{,}52)^{10}+1)}{0{,}05}) = 452950{,}34$$

$$\underset{N=452950{,}34}{\Longrightarrow} N \geq \frac{8}{0{,}05^2} \ln(\frac{4((2*452950{,}34)^{10}+1)}{0{,}05}) = 452956{,}40$$

$$\underset{N=452956{,}40}{\Longrightarrow} N \geq \frac{8}{0{,}05^2} \ln(\frac{4((2*452956{,}40)^{10}+1)}{0{,}05}) = 452956{,}83$$

$$\underset{N=452956{,}83}{\Longrightarrow} N \geq \frac{8}{0{,}05^2} \ln(\frac{4((2*452956{,}83)^{10}+1)}{0{,}05}) = 452956{,}86$$

   Finalmente, la desigualdad convergerá a 452956.86, es decir, el tamaño muestral para obtener un 95 % de confianza de que el error de generalización sea como mucho 0.05 será $N = 452957$.

[2] Los cálculos han sido realizadas mediante la plataforma *Wolfram Alpha*.

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.

    Solución: