Trabajo 2 - Teoría
Aprendizaje Automático

Francisco Solano López Rodríguez

April 28, 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:

   Las dos condiciones imprescindibles son que las muestras sean independientes e identicamente distribuidas. Si se da esto podemos usar la desigualdad de Hoeffding, que viene dada por:

   $$\mathbb{P}(\mathcal{D} : |\mu - v| > \epsilon) \leq 2e^{-2\epsilon^2 N} \quad \forall \epsilon > 0$$

   Podemos ver que la expresión $2e^{-2\epsilon^2 N}$ no depende de μ ni del tamaño del espacio, solo depende de N. Luego cuanto más datos tengamos en nuestra muestra, la probabilidad de equivocarnos estará acotada por un valor cada vez más bajo.

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:

   No, considero que no es correcta y que no va a beneficiar a la empresa. Por un lado por el teorema de no free lunch tenemos que no existe ningún algoritmo que sea el mejor sobre todos los posibles conjuntos de datos, lo cual hace necesario que debamos tener en cuenta muchos algoritmos diferentes para la aproximación y elegir el más adecuado en cada caso, lo cual es una de las tareas más difíciles en problemas de predicción de datos. Un algoritmo dado puede funcionar mejor que los demás para un conjunto particular de datos, pero otros algoritmos podrían funcionar mejor en un conjunto similar pero de diferentes de datos. Esto hace que sea necesario explotar el conocimiento específico. Además será conveniente elegir la clase de funciones más adecuada para cada problema e imponer restricciones particulares de cada problema.

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 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:

Sí, podría resultar que la hipótesis producida coincide con f, es decir todos los datos (tanto dentro como fuera de la muestra) se encuentran etiquetados por el mismo valor (+1 ó -1) en cuyo caso la función constante producida en la hipótesis es la ideal y en dicho caso S siempre garantizaría un mejor comportamiento sobre cualquier punto de la muestra, ya que nunca se equivocaría, al contrario que la hipótesis dada por C que siempre daría una respuesta incorrecta. (Este es el único caso posible que hace verdadero el enunciado, ya que en cualquier otro caso si S se equivoca un punto x, entonces se tendría que C acertaría en dicho punto).

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

   Solución:

   Sí es posible. Solo sabemos que la hipótesis que produce S es buena para los datos de la muestra, fuera de la muestra no sabemos que puede pasar. Podría darse por ejemplo el caso de que $P\left[f(x) = +1\right] = 0.1$, pero ha dado la enorme casualidad de que todos los ejemplos tenían etiqueta +1, era poco probable pero ha ocurrido, en cuyo caso la hipótesis de S sería h_1 (función constante igual a +1), pero el error fuera de la muestra sería muy alto, a diferencia de la hipótesis dada por C.

   El problema que tenemos es que solamente conocemos los datos y no sabemos quien es P, es decir no conocemos cual es la distribución que siguen los datos, luego no podemos decir nada sobre fuera de la muestra.

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,

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

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

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

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

$$2e^{-2\cdot 0.05^2 N} \leq 0.03 \Rightarrow N \geq -\frac{\ln(0.03/2)}{2 \cdot 0.05^2} = 839.94$$

   de donde obtenemos que el valor más pequeño para N es 840.

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

   - $\mathcal{H} = 10$

$$2 \cdot 10 \cdot e^{-2\cdot 0.05^2 N} \leq 0.03 \Rightarrow N \geq -\frac{\ln(0.03/20)}{0.05^2} = 1300.45$$

     de donde obtenemos que el valor más pequeño para N es 1301.

   - $\mathcal{H} = 100$

$$2 \cdot 100 \cdot e^{-2\cdot 0.05^2 N} \leq 0.03 \Rightarrow N \geq -\frac{\ln(0.03/200)}{0.05^2} = 1760.97$$

     de donde obtenemos que el valor más pequeño para N es 1761.

   ¿Que conclusiones obtiene?

   Vemos que conforme nuestra clase de funciones es más amplia, vamos necesitando que nuestra muestra de datos sea mayor para obtener un mismo valor de delta.

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,

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

   (a) ¿Cuál es el algoritmo de aprendizaje que se usa para elegir g?
   Un algoritmo que minimice el valor de E_{in}, ya que conforme se aumente el tamaño de la muestra menor será la probabilidad de que E_{out} difera de E_{in}, luego si el valor de E_{in} es cercano a 0, al aumentar el tamaño de la muestra suficientemente, el valor de E_{out} será también cercano a 0 con alta probabilidad.

   (b) Si elegimos g de forma aleatoria, ¿seguiría verificando la desigualdad?
   Sí, la desigualdad no depende de que g cojamos, lo que dice es que cuanto mayor sea el tamaño de la muestra, menor es la cota superior de la probabilidad de que $E_{in}(g)$ y $E_{out}(g)$ difieran. Es decir que podemos acotar la probabilidad de que E_{in} y E_{out} difieran, dando un tamaño de la muestra suficientemente grande.

   (c) ¿Depende g del algoritmo usado?
   Evidentemente sí, el algoritmo podría ser dar un g aleatorio de la clase de funciones y no tiene porqué coincidir con el g que se obtendría en el descrito en el apartado a). Así que cada algoritmo podrá dar como resultado una g diferente.

   (d) ¿Es una cota ajustada o una cota laxa?
   Es una cota ajustada, ya que es una definición rigurosa que nos proporciona un valor que sabemos que nuestra probabilidad no puede superar y podemos conseguir reducir el valor de δ aumentando el tamaño de la muestra.

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:

   No es aplicable de forma directa porque cuando aplicabamos la desigualdad de Hoeffding la hipótesis g era fijada antes de saber la muestra de datos.

   Una solución para poder aplicar una desigualdad similar, que nos proporcione una cota, es considerar como conjunto todas las hipótesis de $\mathcal{H}$ y utilizar la propiedad de sub-aditividad de la medida: $P(\cup_{i=1}^{|\mathcal{H}|} B_i) \leq \sum_{i=1}^{|\mathcal{H}|} P(B_i)$, de donde se deduce facilmente la expresión:

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

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:

   Veamos primero la definición de punto de ruptura:

   Definición. Si ningún conjunto de datos de tamaño k puede ser separado por $\mathcal{H}$, entonces k se dice punto de ruptura para $\mathcal{H}$.

La única afirmación que nos sirve es la d), que es la que cumple la definición de punto de ruptura.

La a) y la b) evidentemente no nos sirven, ya que no se cumpliría la definición.

La c) no es suficiente, la definición dice que no se pueda separar ningún conjunto de tamaño k por $\mathcal{H}$, luego no nos basta solo con uno.

La e) tampoco ya que si k es un punto de ruptura en dicho caso se tendría $m_{\mathcal{H}}(k) < 2^{k^*}$.

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?

   Para calcular dicho tamaño muestral hacemos uso de la siguiente desigualdad:

$$N \geq \frac{8}{\epsilon^2} \ln\left(\frac{4\big((2N)^{d_{VC}} + 1\big)}{\delta}\right)$$

   Como la N aparece a ambos lados de la desigualdad no podemos hacer el cálculo directamente. Para ello realizamos el cálculo por medio de métodos iterativos. A continuación se muestra el código realizado en python para realizar dichos cálculos.

```
import numpy as np

def calcularN(epsilon, delta, dvc, N):
   return (8/epsilon**2)*np.log(4*((2*N)**dvc+1)/delta)

N = 1

N_old = N
N = calcularN(0.05, 0.05, 10, N)

while np.abs(N-N_old) > 10**-10:
   N_old = N
   N = calcularN(0.05, 0.05, 10, N)

print(N)
```

   Tras la ejecución del programa vemos que el tamaño muestral necesario es 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**
      - Si $\frac{N}{d_{VC}}$ es pequeño (menor de 20), entonces el intervalo de confianza es grande y aún teniendo un error de entrada igual a cero la probabilidad de error fuera puede ser grande. Podría ser útil para muestras suficientemente grandes.
      - Surgen el problema del sobreajuste.
      - Los métodos paramétricos basados ??en el principio inductivo ERM utilizan un conjunto de funciones aproximadas de complejidad fija conocida
    - **SRM**
      - SRM al minimizar el riesgo estructural se favorece estructuras más simples.
      - Sirve como solución al problema de sobre ajuste que había en ERM.

– Proporciona un mecanismo formal para elegir una complejidad de modelo óptima para la muestra finita.

Bonus

1. 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 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. Suponga que hay una distribución de probabilidad sobre $\mathcal{X}$, y sea $P[f(x) = +1] = p$

   (a) Si $p = 0,9$ ¿Cual es la probabilidad de que S produzca una hipótesis mejor que C?

   S producirá una hipótesis mejor que C si el número de datos etiquetados como +1 es mayor que el número de datos etiquetados como -1, como el número de ejemplos extraídos es 25 esto se dará si hay más de 12 datos etiquetados como +1.

   La probabilidad de obtener +1 es de 0.9, y el experimento lo hemos realizado 25 veces, entonces la variable aleatoria que nos dice cual es la probabilidad de que haya x datos etiquetados como +1, sigue una distribución binomial $B(25, 0.9)$, donde la probabilidad de obtener x "aciertos" viene dada por la expresión:

$$f(x) = \binom{25}{x} 0.9^x \ 0.1^{n-x}$$

   Como nosotros queremos que haya más de 12 datos etiquetados como +1, tendremos que calcular esa probabilidad:

$$F(12 < x) = \sum_{i=13}^{25} \binom{25}{i} 0.9^i \ 0.1^{n-i} = 0.999999837916$$

   (b) ¿Existe un valor de p para el cual es más probable que C produzca una hipótesis mejor que S?

   No, si $p > 0.5$ se tiene que la probabilidad de que haya más datos etiquetados como +1 que como -1 en la muestra de 25 datos, es mayor de 0.5, luego en la mayoría de los casos S elegirá la hipótesis h_1 acertadamente. Si $p < 0.5$ ocurrirá lo mismo pero con la hipótesis h_2 es decir en la mayoría de los casos S elegirá la hipótesis h_2 acertadamente. Si hacemos $p = 0.5$ y hacemos los mismos cálculos del ejercicio anterior tenemos:

$$F(12 < x) = \sum_{i=13}^{25} \binom{25}{i} 0.5^i \ 0.5^{n-i} = 0.5$$

   Luego lo máximo a lo que puede aspirar C es que la probabilidad de producir una hipótesis mejor que S sea de 0.5, y esto se da justo cuando hacemos $p = 0.5$.

2. Consideremos el modelo de aprendizaje "M-intervalos"donde la clase de funciones H está formada por $H : \mathbb{R} \to \{-1, +1\}$, con $h(x) = +1$ si el punto está dentro de uno de m intervalos arbitrariamente elegidos y -1 en otro caso. Calcular la dimensión de Vapnik-Chervonenkis para esta clase de funciones.

   Solución:

   Calculemos primero el punto de ruptura. Evidentemente si tenemos $2M$ puntos siempre podremos encontrar M intervalos que separen los puntos. El peor escenario que podríamos encontrarnos con $2M$ puntos es aquel en el que tenemos M puntos con etiqueta +1 y M con -1 y además alternadamente es decir $(-1, 1, -1, 1, \cdots)$, pero se pueden separar facilmente tomando m intervalos de forma que cada 1 caiga dentro de uno de ellos y los -1 fuera.

   Pero, ¿qué pasa si tomamos $2M + 1$ puntos? Supongamos que $M + 1$ puntos etiquetados con +1 y M etiquetados con -1, y además se encuentran repartidos como antes $(1, -1, 1, -1, \cdots, 1, -1, 1)$. Vemos que en este caso resulta imposible separar los puntos con M intervalos, luego tenemos que el punto de ruptura es $2M + 1$.

Como $d_{VC} = k - 1$, donde k es el punto de ruptura, obtenemos que la dimensión de Vapnik-Chervonenkis es igual a $2M$.