Universidad de Granada

Ingeniería Informática

Computación y Sistemas Inteligentes

Cuestionario 2

Autor: José Antonio Ruiz Millán
Asignatura: Aprendizaje Automático
27 de abril 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.**

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.**

   Bueno, como todo, esto tiene sus pros y sus contras, aunque ahora veremos que más contras que pros.

   Si lo hace de esta forma a la hora de una vez progamado el algoritmo, hacerle modificaciones es relativamente fácil y rápido ya que sólo disponemos de 1 algoritmo y esas mofificaciones no pueden ser de muy elevado coste.

   El problema lo tenemos en que lo verdaderamente importante es que este algoritmo dé soluciones con calidad ya que finalmente es lo que a la empresa le interesa. Si únicamente utilizamos un algoritmo para todos los tipos de problemas futuros, estan cometiendo un grave error ya que cada problema es distinto, cada problema puede funcionar muy bien o muy mal con un determinado algoritmo. Debemos estudiar cada tipo de problema y utilizar el algoritmo que mejor se adapte a cada tipo de problema.

   Por último, respecto a una única clase de funciones, tenemos un caso parecido al anterior. Si sólo utilizamos una clase de funciones, verdaderamente no estamos aprendiendo, no estamos comparando entre distintas clases de funciones para escojer la que mejor se ajuste a nuestro problema, con una única clase de funciones sólamente estamos comprobando como de buena o mala es esa clase pero no estamos aprendiendo realmente nada.

   Como conclusión puedo decir que aunque respecto a la limpieza, modificación y creación del codigo la empresa ahorraría tiempo y esfuerzo, no les iva a salir rentable ya que la mayoría de problemas que tuviesen no ivan a dar unas soluciones de calidad y esto haría finalmente que la empresa no fuese bien ya que si no dan buenos resultados a los problemas, ya habría otra empresa que si lo haría y esta desaparecería.

3. **Supongamos un conjunto de datos D de 25 ejemplos extraidos de una funcion 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 +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**

   No, no podemos garantizar el buen comportamiento de nuestra hipótesis final fuera del conjunto de entrenamiento.

Para el problema de aprendizaje, necesitamos plantearlo desde un punto probabilístico ya que nunca podremos asegurar que ocurrirá realmente fuera de la muestra. Genereamos una hipótesis con el objetivo de aproximarnos a una función desconocida que a partir de una muestra intentaremos aproximarnos a ella, pero siempre utilizando la probabilidad. Podremos estar muy cerca de la funcion objetivo basándonos en alguna distribución de probabilidad con un error lo más pequeño posible, pero nunca podremos garantizar con exactitud que nuestra hipótesis tenga un buen comportamiento fuera de la muestra.

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 hipotesis que produce C sea mejor que la hipótesis que produce S ?. Justificar la respuesta**

   Partimos de que $Y_n = +1$ lo que nos dice que S siempre escogerá h_1 ya que con esta hipótesis consigue que $E_{in}(h_1) = \frac{1}{N}\sum_{n=1}^{N}[[h_1(x_n) \neq f(x_n)]] = 0$ lo que significa que clasifica correctamente toda la muestra.

   En el caso de C, es todo lo contrario, tenemos que $E_{in}(h_2) = \frac{1}{N}\sum_{n=1}^{N}[[h_1(x_n) \neq f(x_n)]] = 1$ lo que significa que todos los elementos de la muestra están mal clasificados.

   Ahora, respondiendo a la pregunta, desde el punto de vista del aprendizaje, tanto S como C nos están enseñando lo mismo, ya que C está clasificando todos los puntos de la muestra igual que S sólo que con distinta etiqueta, pero al final para aprender, hace el mismo tipo de clasificación. Resumiendo, tanto la hipótesis de S como la hipótesis de C en este caso son iguales de buenas ya que desde el punto de vista del apredizaje están clasificando de la misma forma.

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** $\delta(\epsilon, N, |H|)$

   La expresion[1] sería $\delta(\epsilon, N, |H|) = 2|H|e^{-2\epsilon^2 N}$, donde $|H|$ es el tamaño del conjunto H. Con esta expresión conseguimos tener el valor dependiendo tanto de ϵ como de N y de $|H|$.

   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?.**

   Tenemos entonces que: $2 \cdot 1e^{-2 \cdot 0,05^2 N} \leq 0{,}03$

   - Pasamos el 2 que está multiplicando en la izquerda, a la parte derecha dividiendo y resolvemos el exponente de e.

$$e^{-0,005N} \leq 0{,}015$$

   - Aplicamos logaritmos para eliminar la e.

$$-0{,}005N \leq \log(0{,}015)$$

- Ahora pasamos el -0.005 a la parte derecha dividiendo. Como el signo es negativo tenemos que cambiar el sentido de la desigualdad.

$$N \geq -\frac{\log(0{,}015)}{0{,}005}$$

Por lo que tenemos que finalmente $N \geq 839{,}94$. Como $N \in \mathbb{N}$ tenemos finalmente que:

$$N \geq 840$$

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

1) **H = 10**

Tenemos entonces que: $20e^{-0{,}005N} \leq 0{,}03$

- Pasamos el 20 que está multiplicando en la izquerda, a la parte derecha dividiendo.

$$e^{-0{,}005N} \leq 0{,}0015$$

- Aplicamos logaritmos para eliminar la e.

$$-0{,}005N \leq \log(0{,}0015)$$

- Ahora pasamos el -0.005 a la parte derecha dividiendo. Como el signo es negativo tenemos que cambiar el sentido de la desigualdad.

$$N \geq -\frac{\log(0{,}0015)}{0{,}005}$$

Por lo que tenemos que finalmente $N \geq 1300{,}46$. Como $N \in \mathbb{N}$ tenemos finalmente que:

$$N \geq 1301$$

2) **H = 100**

Tenemos entonces que: $200e^{-0{,}005N} \leq 0{,}03$

- Pasamos el 200 que está multiplicando en la izquerda, a la parte derecha dividiendo.

$$e^{-0{,}005N} \leq 0{,}00015$$

- Aplicamos logaritmos para eliminar la e.

$$-0{,}005N \leq \log(0{,}00015)$$

- Ahora pasamos el -0.005 a la parte derecha dividiendo. Como el signo es negativo tenemos que cambiar el sentido de la desigualdad.

$$N \geq -\frac{\log(0{,}00015)}{0{,}005}$$

Por lo que tenemos que finalmente $N \geq 1760{,}97$. Como $N \in \mathbb{N}$ tenemos finalmente que:

$$N \geq 1761$$

¿Que conclusiones obtiene?

Como conclusión tanto de las soluciones de este ejercicio como lo visto en teoría y en [1], vemos que la probabilidad depende del tamaño de N que siendo un valor pequeño obtendremos una probabilidad más grande de error mientras que si tenemos un valor muy grande de N obtendremos un valor más pequeño de probabilidad de error, lo que hace que la diferencia entre $E_{in}(h)$ y $E_{out}(h)$ sea cercana a 0. **A mayor tamaño de muestra, menor diferencia entre error en la muestra y error fuera de la muestra.**

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?**

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

   c) **Depende g del algoritmo usado?.**

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

   Justificar las respuestas.

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.**

   Una de las propiedades fundamentales de la desigualdad de Hoeffding es que tenemos que fijar la hipótesis final g antes de conocer la muestra de los datos. Cuando el número de hipótesis $|H|$ es mayor que 1, la función de nuestro algoritmo es elegir, de entre las distintas $h_i \in H$, la que mejor resultado obtenga, es decir, la que minimice más $E_{in}(h_i) = \frac{1}{N}\sum_{n=1}^{N}[[h_i(x_n) \neq f(x_n)]]$. Finalmente obtendremos nuestra funcion g con la mejor $h_i \in H$. Por lo que llegamos a una contradicción, ya que para obtener la función g estamos utilizando los datos de la muestra, cosa que acabamos de decir que el requisito de la desigualdad de Hoeffding era todo lo contrario. Por esa razón no podemos aplicar directamente la desigualdad de Hoeffding cuando $|H| > 1$.

   Para solucionarlo se utilizan otro métodos vistos en el temario[1] y pasamos de $\mathbb{P}[|E_{in}(h) - E_{out}(h)| > \epsilon] \leq 2e^{-2\epsilon^2 N}$ a $\mathbb{P}[|E_{in}(g) - E_{out}(g)| > \epsilon] \leq 2|H|e^{-2\epsilon^2 N}$.

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 $x_1, ..., x_{k^*}$ que H puede separar ("shatter").**

   Si puede separar un conjunto de puntos de tamaño k^*, entonces $m_H(k^*) = 2^{k^*}$. Por lo que estaríamos diciendo por definición que k^* **no** es un punto de ruptura. Por lo que esta afirmación no se cumple.

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

   Si H puede separar cualquier conjunto de k^* puntos, estaríamos demostrando que k^* **no** es un punto de ruptura. Volveríamos al mismo caso del apartado anterior. Por lo

que esta afirmación no nos sirve.

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

No es suficiente para decir que sea punto de ruptura, ya que puede existir un conjunto distinto de k^* que si pueda separar. Por lo tanto esta afirmación tampoco nos sirve.

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

EN este caso, esta afirmación si nos sirve para decir que es un punto de ruptura. Al no poder separar ningún conjunto, sabemos que el número de dicotomías que puede generar H es menor que 2^{k^*}. Por lo tanto, $m_H(k^*) < 2^{k^*}$. Concluimos que en este caso, es un punto de ruptura.

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

Como he indicado justo en el apartado anterior, por definición[1], para que sea un punto de ruptura debe cumplirse que $m_H(k^*) < 2^{k^*}$. En este caso nos dicen que $m_H(k^*) = 2^{k^*}$, lo que hace falso $m_H(k^*) < 2^{k^*}$. Por lo que esta afirmación no nos sirve para decir que es un punto de ruptura.

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?**

En el propio temario o en[1] podemos ver que:

$$\epsilon \geq \sqrt{\frac{8}{N} \log \frac{4((2N)^{dvc}+1)}{\delta}}$$

donde podemos ver que la funcion depende de todos los elementos que nos pide el ejercicio.

Ahora, aplicando el cuadrado para eliminar la raíz y pasando el ϵ^2 a la derecha y la N a la izquerda, conseguimos la siguiente funcíon:

$$N \geq \frac{8}{\epsilon^2} \log \frac{4((2N)^{dvc}+1)}{\delta}$$

Sustituimos ahora los datos que nos da el ejercicio y obtenemos el siguiente resultado:

$$N \geq \frac{8}{0{,}05^2} \log \frac{4((2N)^{10}+1)}{0{,}05}$$

El siguiente paso a realizar sería inicialmente darle un valor a N y evaluar la función para ver el resultado que obtenemos. Iterativamente, sustituir el N que nos devuelve de nuevo en la función hasta que consigamos que la función converja.

- $N = 50000$

$$\frac{8}{0{,}05^2} \log \frac{4((2\cdot 50000)^{10}+1)}{0{,}05} = 382436{,}1$$

- $N = 382436{,}1$

$$\frac{8}{0{,}05^2} \log \frac{4((2\cdot 382436{,}1)^{10}+1)}{0{,}05} = 447541{,}3$$

- $N = 447541{,}3$

$$\frac{8}{0{,}05^2} \log \frac{4((2\cdot 447541{,}3)^{10}+1)}{0{,}05} = 452572$$

- $N = 452572$

$$\frac{8}{0{,}05^2}\log\frac{4((2\cdot 452572)^{10}+1)}{0{,}05} = 452929{,}7$$

- $N = 452929{,}7$

$$\frac{8}{0{,}05^2}\log\frac{4((2\cdot 452929{,}7)^{10}+1)}{0{,}05} = 452954{,}9$$

- $N = 452954{,}9$

$$\frac{8}{0{,}05^2}\log\frac{4((2\cdot 452954{,}9)^{10}+1)}{0{,}05} = 452954{,}9$$

Y encontramos el valor $N = 452954{,}9$ que hace converger esta función. Por lo que necesitamos tener $N \geq 452954{,}9$ para que se cumplan los requisitos que nos piden en el enunciado.

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.**

BONUS

1. **Supongamos que tenemos un conjunto de datos *D* de 25 ejemplos extraidos de una funcion 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 +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 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?**

   Sabemos por el enunciado, que S cojerá la hipótesis que mejor ajusta los datos, por lo que teniendo $P[f(x) = +1] = 0{,}9$, S siempre cojerá h_1 y C cojerá h_2.

   Con esto tenemos que la probabilidad de que S acierte es 0,9 y la probabilidad de que C acierte es 0,1 ya que es lo contrario a S.

   Con estos datos, podemos calcular cuál es la probabilidad de que S produzca una hipótesis mejor que C, ya que este valor se calcula restando las probabilidades de acierto de ambos. Así que tenemos que $p - (1 - p) = 0{,}9 - 0{,}1 = 0{,}8 = 80\,\%$ lo que quiere decir que tenemos un $80\,\%$ de probabilidad de que S produzca una mejor hipótesis que C

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

   No. Basándonos en los cálculos anteriores y sabiendo que S siempre coje la mejor hipótesis, S siempre tendrá un porcentaje de acierto mayor que C. De este modo, lo único que podemos deducir es que sí que podemos conseguir un p donde C pueda producir un hipótesis igual de buena que S, que sería con $p = 0{,}5$, pero nunca mejor que S.

2. **(1 puntos) Consideremos el modelo de aprendizaje "M-intervalos" donde la clase de funciones H está formada por $h : R \to \{-1, +1\}$, 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.**

Referencias

[1] Learning from Data Short Course, Yser S. Abu-Mostafa; Malik Magdon-Ismail; Hsuan-Tien Lin, `https://libsvm.com/data/learn.from.data.pdf`, Accedido el 27 de abril de 2018.