Aprendizaje automático. Cuestionario 2

Antonio Jesús Heredia Castillo
76069518P

---

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.

   Respuesta:
   Una de las dos condiciones es que la muestras de X sean independientes e idénticamente distribuidas. Esto se puede ver usando la desigualdad de Hoeffding's
   La otra es que exista la misma distribución de probabilidad en entrenamiento y test de la muestra respecto a los datos fuera de la muestra. La justificación matemática es la siguiente:

   $$P\left(\mathcal{D} : |\mathrm{E}_{in}(g) - \mathrm{E}_{out}(g)| > \epsilon\right) < 2|\mathcal{H}|e^{-\epsilon^2 N} \text{ for any } \epsilon > 0$$

   Donde se ve que todo depende de N. Por lo tanto cuanto al tener mas datos (siempre que sean i.d.d) pues tendremos que la cota superior es cada vez mas pequeña.

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.

   Respuesta:
   No creo acertada dicha decisión. La propia intuición nos dice que cada problema es diferente y por lo tanto necesitara un modelo diferente. Esto nos lo puede confirmar el teorema de "No free lunch". Ya que para algoritmo existe un P en el que falla, aun cada P puede aprenderse con éxito por otro. Ademas de que en promedio va a ser para todos igual de eficiente, haciendo que para unos vaya a ir muy bien, pero puede ser que para otros muy mal. Por lo tanto si podemos tener diferentes algoritmos que se adapte de forma diferente para cada problema, teniendo así siempre la máxima eficiencia (al menos la que nosotros somos capaces de tener), es un poco absurdo usar una que no siempre nos va a dar los mejores resultados posibles.

   El problema de usar siempre la misma clase de funciones es que nos se va a poder ajustar de forma adecuada para todos los tipos de datos que tenemos. Ya que la distribución de

estos según su clase puede ser diferente. Esto se pudo observar en el ejercicio 1 apartado 3 de la practica 2.

3. ¿Que se entiende por una solución PAC a un problema de aprendizaje? Identificar el porqué de la incertidumbre e imprecisión.

   Respuesta:

   Dada una técnica de aprendizaje que recibe muestras, debe elegir una función de clase (llamado la hipótesis).Y esto tiene por objetivo que con una alta probabilidad.

   $$P\left(\mathcal{D} : |\mathrm{E}_{\mathrm{out}}(h) - \mathrm{E}_{\mathrm{in}}(h)| < \epsilon\right) \geq 1 - \delta$$

   La imprecisión se obtiene de la diferencia entre el error fuera de la muestra y el error dentro de la muestra y por otro lado la incertidumbre que sera 1 menos el epsilón (cuanto menos sea la garantía+, menor sera la precisión).

4. Suponga un conjunto de datos $\mathcal{D}$ de 25 ejemplos extraidos de una función desconocida $f : \mathcal{X} \to \mathcal{Y}$, donde $\mathcal{X} = \mathbb{R}$, 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.

   Respuesta:
   No se puede garantizar, ya que existe la posibilidad que de forma aleatoria se consigan mas aciertos.

   Aunque si es posible que pase (suponiendo que el test esta escogido con la misma proporción al resto), ya que el de forma aleatoria se va a conseguir en media un 50 % de aciertos y con el algoritmo S siempre va a ser un $> 50\,\%$ porque elige la clase que mas veces aparece.

5. Con el mismo enunciado de la pregunta 4:

   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.

   Respuesta:
   Si es posible, todo depende de los datos fuera de la muestra. Si los datos de la muestra tienen mas $y_n = -1$ que $y_n = +1$.

6. Considere la cota para la probabilidad de la hipótesis solución g de un problema de aprendizaje, a partir de la desigualdad generalizada de Hoeffding para una clase finita de hipótesis,

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

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

Respuesta:
Para elegir g elegiremos cualquier algoritmo de la las hipótesis que mas minimice el error de la muestra.

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

Respuesta:

c) ¿Depende g del algoritmo usado?

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

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

   Respuesta:
   Cuando tienes varias hipótesis en $\mathcal{H}$, el algoritmo de aprendizaje elige la hipótesis g que mejor se adapta, a partir del set de datos. Y esto no debe ser asi, ya que una de las propiedades principales de la desigualdad de Hoeffding es que hay que fijar la hipótesis final antes de conocer la muestra de datos.

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

   Respuesta:
   No. Por definición un punto de ruptura se da cuando $m_{\mathcal{H}}(k) < 2^k$. Y según aparece en las diapositivas, cuando $\mathcal{H}$ puede separar al conjunto de puntos $x_1, \cdots, x_{k^*}$, tenemos que $m_{\mathcal{H}}(N) = 2^N$

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

   Respuesta:
   No.Nos ya que este caso engloba al anterior.

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

   Respuesta:
   No. Por que aunque no pueda separar el conjunto de puntos $x_1, \cdots, x_{k^*}$, no dice que no pueda separar cualquier otro conjunto.

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

   Respuesta:
   Si. Como no separa ningún conjunto podemos afirmar que $m_H(k^*) < 2^{k^*}$.

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

Respuesta:
No. Ya que como hemos dicho anteriormente $m_{\mathcal{H}}(k) < 2^{k^*}$ de forma estricta, no nos valdría que fuera igual.

9. Para un conjunto $\mathcal{H}$ 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?

   Respuesta:
   Según los propios apuntes tenemos que:

$$N \geq \frac{8}{\varepsilon^2} \ln\left(\frac{4\left((2N)^d \mathrm{Vd} + 1\right)}{\delta}\right) \rightarrow N \geq \frac{8}{0{,}05^2} \ln\left(\frac{4\left((2N)^{10} + 1\right)}{0{,}05}\right)$$

   Ahora tenemos que evaluar la inecuación de forma iterativa, hasta que consigamos que converja. Empezare por ejemplo con N=30000

$$N \geq \frac{8}{0{,}05^2} \ln\left(\frac{4(2 \cdot 30000)^{10} + 4}{0{,}05}\right) = 366089{,}68$$

   N= 366089.68

$$N \geq \frac{8}{0{,}05^2} \ln\left(\frac{4(2 \cdot 366089{,}68)^{10} + 4}{0{,}05}\right) = 446143{,}47$$

   N= 446143.47

$$N \geq \frac{8}{0{,}05^2} \ln\left(\frac{4(2 \cdot 446143{,}47)^{10} + 4}{0{,}05}\right) = 452471{,}86$$

   N=452471.86

$$N \geq \frac{8}{0{,}05^2} \ln\left(\frac{4(2 \cdot 452471{,}859999999986)^{10} + 4}{0{,}05}\right) = 452922{,}58$$

   En este punto podemos decir que que con un $N \geq 452922{,}58$ necesitamos un tamaño muestral para satisfacer las condiciones expuestas 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.

Referencias