Trabajo 2: Cuestiones de Teoría

Ignacio Vellido Expósito

May 1, 2019

Todas las preguntas tienen el mismo valor

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.

   Para que se permita el aprendizaje partimos de dos condiciones:

   - La muestra debe ser independiente e idénticamente distribuida. Esto indica que los datos no tienen dependencias entre sí y que la forma de construir la muestra es aparentemente aleatoria, por tanto se permite la inducción de conocimiento.

   - La muestra y la población deben seguir la misma distribución. Si no lo fueran implicaría que la muestra no representa al conjunto de la población y por tanto no es posible el aprendizaje.

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.

   Partir de un único algoritmo puede no ser una mala idea si se tuviese la seguridad de que funciona de manera correcta independientemente del problema, pero como hemos visto por el teorema NFL esto no es cierto.

   Además, el hecho de trabajar con una única clase de funciones puede limitar la capacidad de aprendizaje si la función objetivo no se parece a las que tenemos dentro de la clase.

   Sabemos que debemos seleccionar una clase de funciones sin haber observado los datos, pero eso no quita que no podamos repetir el aprendizaje con una clase nueva si vemos que la otra no nos da buenos resultados.

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

   Una solución a un problema de aprendizaje se dice que es un resultado PAC si sigue la ecuación:

   $$P(D :\mid E_{out}(h) - E_{in}(h) \mid > \epsilon) \leq 2e^{-2\epsilon^2 N} \qquad \text{Para cada } \epsilon > 0 \qquad (1)$$

   Esta inecuación se mide con una incertidumbre δ y una precisión ϵ, que representan la complejidad de la clase de funciones. Esta complejidad es la que nos dice si el aprendizaje es posible siguiendo las reglas ERM, y sigue la inecuación:

   $$m_H(\epsilon, \delta) \leq \lceil \frac{2}{\epsilon} log \frac{\mid H \mid}{\delta} \rceil \qquad (2)$$

4. Suponga un conjunto de datos $\mathcal{D}$ de 25 ejemplos extraidos de una funcion 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

   La hipótesis que usa S es bastante simple, pero al ser f totalmente desconocida la hipótesis seleccionada puede clasificar de forma increíblemente buena o increíblemente mala. Por tanto existe la posibilidad de que tenga un mejor resultado que aquella que clasifica de forma aleatoria.

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

   De cara al conjunto de entrenamiento no, pues C siempre clasifica de forma contraria a S, y por lo dicho en el enunciado S elige entre las dos hipótesis aquella que clasifica mejor los datos. Pero si hablamos de la población en general, al ser f desconocida, se puede dar que C clasifique mejor que S.

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,

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

   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

   a) No es un algoritmo concreto, se busca uno que devuelva una hipótesis lo más cercana posible a la mejor dentro de la clase (con una cierta probabilidad).

   b) Si, pues la inecuación es válida para cualquier elemento de la clase, por eso decimos que es una cota laxa.

   c) Sí, diferentes algoritmos podrán elegir hipótesis diferentes, pero el valor de la inecuación no varía.

   d) Esta cota es bastante laxa puesto que estamos asegurándonos de que se cumple la desigualdad para cualquier función dentro de la clase de funciones.

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.

   Porque la clase de funciones se selecciona antes de aplicar la desigualdad. Si es de tamaño uno, equivaldría a elegir una hipótesis directamente, pero nosotros queremos elegir una clase de funciones con un número suficientemente grande de ellas para poder aplicar un algoritmo de aprendizaje.

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 $x_1, \ldots, x_{k^*}$ 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 $x_1, \ldots, x_{k^*}$ que $\mathcal{H}$ no puede separar
   d) Mostrar que $\mathcal{H}$ no puede separar ningún conjunto de k^* puntos
   e) Mostrar que $m_{\mathcal{H}}(k) = 2^{k^*}$

   Un punto de ruptura es un valor k a partir del cuál la clase de funciones no puede separar un conjunto de ese tamaño. Sabiendo esto:

   a) y b) no nos valen, pues solo indican que el punto de ruptura es mayor que ese k^*.

   c) y d) tampoco ya que dicen que que el punto de ruptura es menor que ese k^*.

   e) si nos sirve ya que nos indica que $k^* + 1$ es un punto de ruptura, pues:

$$m_h(k^* + 1) < 2^{k^*+1} \tag{3}$$

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?

   Se tiene que cumplir la ecuación:

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

   Sustituimos los valores de d_{VC}, ϵ y δ y probamos iterativamente con diferentes N hasta que obtenemos:

$$456.118 \geq \frac{8}{0,05^2} ln(\frac{4((2 \times 500.000)^{10} + 1)}{0,05}) \tag{5}$$

   Y vemos que necesitamos un tamaño de muestra $N \approx 500.000$

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.

    Sabemos que ERM no garantiza el aprendizaje si $\frac{N}{d_{VC}} < 20$, por tanto corremos el riesgo de hacer un mal clasificador. Si utilizamos SRM no corremos este riesgo, pero tendremos un mayor gasto de recursos y tiempo.

    Por tanto, si tuviésemos claro que se cumplen las condiciones de aprendizaje con ERM, lo podemos utilizar. En otro caso, SRM es nuestra mejor opción.

BONUS

Los BONUS solo serán tenidos en cuenta si en el cuestionario obligatorio se ha consiguido al menos un 75 % de los puntos totales. Justificar correctamente todas las contestaciones

1. (1 punto)Considere que le dan un conjunto de datos yque tras echarles un vistazo observa que son separables linealmente. Por tanto ajusta un modelo perceptron y obtiene un error zero sobre los datos de aprendizaje. Entonces desea obtener una cota de generalización para lo cual mira la dimension de VC del modelo ajustado y ve que es d+1. Por tanto usa esa cota para obtener una cota del error del modelo.
   a) Hay algún problema con la cota elegida - es correcta?
   b) Conocemos la cota de VC para el modelo que hemos usado realmente.
   c) Si la cota no fuera correcta, ¿cúal deberíamos haber usado?

   Pregunta no contestada

2. (1 punto) Suponga un conjunto de datos y extrae de él 100 muestras que no serán usados en entrenamiento sino que serán usados para seleccionar una de las tres hipótesis g_1, g_2, g_3 producidas por tres algoritmos diferentes que serán entrenados con el resto de los datos. Cada algoritmo trabaja con una clase diferente $\mathcal{H}$ de 500 funciones. Queremos caracterizar la precisión de la estimación de $E_{\text{out}}(g)$ sobre la hipótesis final seleccionada a partir de las 100 muestras.
   a) Que valor de M debería de usarse en la expresión, $2Me^{-2N\epsilon^2}$, de la desigualdad de Hoeffding generalizada, ?
   b) ¿Compare el nivel de contaminación de estas 100 muestras con el caso donde estas muestras hubieran participado en entrenamiento en lugar del proceso de selección?

   Pregunta no contestada

3. (1 punto) Considere las siguientes situaciones:
   a) Suponga que $\mathcal{H}$ esta fijada y aumentamos la complejidad de f. ¿En general subirá o bajará el ruido determinístico?. ¿Habrá mayor o menor tendencia a sobreajustar?
   b) Suponga f fija y decrementamos la complejidad de $\mathcal{H}$. ¿En general subirá o bajará el ruido determinístico?. ¿Habrá mayor o menor tendencia a sobreajustar?

   Pregunta no contestada