Cuestionario 2

Aprendizaje Automático

José Manuel Pérez Lendínez

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

Las dos condiciones son las siguientes:

1. Los ejemplos de la muestra de entrenamiento sea i.i.d (independientes e idénticamente distribuidas) y obtenidos por la misma distribución de probabilidad.

2. Se tenga la misma distribución de probabilidad para los datos de entrenamiento y los de test, puesto que si fueran dados por la misma distribución de probabilidad no se podría inferir nada.

Esto es necesario ya que nos asegura que el E_{out} sera próximo al valor de E_{in} al usar la misma distribución de probabilidad y para usar la desigualdad de Hoeffding

$$P(\mathcal{D} : |E_{in} - E_{out}| > \epsilon) \leq 2e^{-2\epsilon^2 N}$$

necisitaremos que:

1. E_{in} sea próximo a E_{out}.

2. E_{in} sea próximo a 0

La primara ya explicamos que se cumple anteriormente y como vemos en la formula $2e^{-2\epsilon^2 N}$ depende de N que es el tamaño de la muestra. Si la muestra es grande el E_{in} sera próximo a 0 y se cumplen las dos.

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 va a seleccionar un único algoritmo y una única clase de funciones con la que aproximar todas las soluciones a sus problemas de presentes y futuros. ¿Considera que dicha decisión es correcta y beneficiará a la empresa? Argumentar la respuesta usando los resultados teóricos estudiados.

No es una buena decisión para la empresa. Según el teorema de No Free Lunch nos dice que no existe un algoritmo que pueda hacer frente a todo tipos de problemas y conseguir buenos resultados en estos. Por tanto no seria buena decisión elegir un único algoritmo para solucionar todos los problemas de la empresa. Quedarnos con una única clase de función tampoco seria bueno puesto que cada conjunto de datos puede ser ajustado mejor por una clase u otra siendo un problema limitar nuestra elección de la mejor clase de función a una única. Esto haría que solo tengamos buenos resultados en los datos que se ajusten a la clase elegida y malos para el resto.

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

Se reciben muestras y debe seleccionar una hipotesis (función de generalización) de una clase segura de funciones posibles. Se busca que la con una alta probabilidad, la función seleccionada tendrá un error de generalización bajo. La hipótesis tiene que cumplir que $E_{in} - E_{out}$ sea menor que ϵ (imprecisión) y con una probabilidad de 1 - δ (incertidumbre).

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

La imprecisión nos marca la diferencia máxima que podremos tener entre el E_{in} y E_{out}. La incertidumbre nos dará la precisión que aceptaremos.

4 Suponga un conjunto de datos D de 25 ejemplos extraídos de una función desconocida $f : \mathcal{X} \to \mathcal{Y}$ donde $\mathcal{X} = \mathcal{R}$ e $y = \{-1, +1\}$. Para aprender f usamos un conjunto simple de hipótesis $\mathcal{H} = \{h_1, h_2\}$ donde h_1 es la funcion constate igual a +1 y h_2 la funcion constante igual a -1. Consideremos dos algoritmos de aprendizaje, S(smar) y C(crazy). S elige la hispótesis que mejor ajusta a los datos y c elige deliberadamente la otra hipótesis.

1. **¿Puede S producir una hipótesis que garantice mejor comportamiento que la aleatoria sobre cualquier punto fuera de la muestra? Justificar la respuesta.**

No puedes garantizar que fuera de los datos de entrenamiento tu hipótesis sea mejor que un aleatorio en todos los casos. Si la muestra no es i.i.d de la población el aleatorio podría mejorar los resultados de S al tener un 50% de posibilidades de acertar.

5 Con el mismo enunciado de la pregunta 4: Asumir desde ahora que todos los ejemplos D tienen $y_n = +1$. ¿Es posible que la hipótesis que produce C sea mejor que la hipótesis que produce S? Justificar la respuesta.

Si es posible, puesto que en las etiquetas fuera de la muestra se podrían tener mas -1 que +1 y en ese caso la hipótesis que produce C clasificaría mejor.

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,

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

1. **¿Cuál es el algoritmo de aprendizaje que se usa para elegir g?**

   Se elige el algoritmo que minimize el error en E_{in}, puesto que si tenemos un gran tamaño en la muestra D, si minimizamos el error E_{in}, el error E_{out} se reducirá también.

2. **Si elegimos g de forma aleatoria ¿seguiría verificando la desigualdad?**

3. **¿Depende g del algoritmo usado?**

4. **Es una cota ajustada o una cota laxa?**

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

No es aplicable porque una de las propiedades de la desigualdad de Hoeffding es tener que fijar la hipótesis final para la función g sin utilizar la muestra de datos. Cuando tenemos un numero de hipótesis mayor a 1, tenemos que elegir la que mejor se adapte a la muestra. Con esto estamos contradiciendo la propiedad explicada anteriormente y no se podría utilizar directamente.

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:

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

   No. Puesto que que si es existe un k^* puntos $x_1, \ldots, x_{k^*}$ estamos diciendo que $m_{\mathcal{H}}(k) = 2^{k^*}$ y para ser punto de ruptura $m_{\mathcal{H}}(k) < 2^{k^*}$

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

   No. Si puede separar cualquier conjunto de k^* puntos se tendra el mismo caso que el ejemplo anterior.

3. **Mostrar un conjunto de k^* puntos $x_1, \ldots, x_{k^*}$ que H no puede separar. No.** Podría darse el caso de que otro conjunto de k^* puntos $x_1, \ldots, x_{k^*}$ si lo pudiera separar.

4. **Mostrar que H no puede separar ningún conjunto de k^* puntos. Si.** Puesto que si H no separa ningún punto k^* tenemos que $m_{\mathcal{H}}(k) < 2^{k^*}$ y en ese caso es un punto de ruptura

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

   No. Por la definición de break point $m_{\mathcal{H}}(k) < 2^{k^*}$

9 Para un conjunto H con $d_{vc} = 10$, ¿que tamaño muestra 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 eso sustituimos los valores dados en el enunciado en la formula

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

y obtenemos la siguiente:

$$N \geq \frac{8}{0.05^2} \ln\left(\frac{4\left((2N)^{10} + 1\right)}{0.05}\right)$$

Ahora tendríamos que ir comprobando con N de forma iterativa hasta que la función converja.

$$\frac{8}{0.05^2} \ln\left(\frac{4\left((2 * 2000)^{10} + 1\right)}{0.05}\right) = 279432,07$$

$$\frac{8}{0.05^2} \ln\left(\frac{4\left((2 * 279432,07)^{10} + 1\right)}{0.05}\right) = 437499,65$$

$$\frac{8}{0.05^2} \ln\left(\frac{4\left((2 * 437499,65)^{10} + 1\right)}{0.05}\right) = 451845,7929$$

$$\frac{8}{0.05^2}\ln\left(\frac{4\left((2*451845,7929)^{10}+1\right)}{0.05}\right)=452878,2745$$

$$\frac{8}{0.05^2}\ln\left(\frac{4\left((2*452878,2745)^{10}+1\right)}{0.05}\right)=452951,3121$$

$$\frac{8}{0.05^2}\ln\left(\frac{4\left((2*452951,3121)^{10}+1\right)}{0.05}\right)=452956,4724$$

$$\frac{8}{0.05^2}\ln\left(\frac{4\left((2*452956,4724)^{10}+1\right)}{0.05}\right)=452956,837$$

$$\frac{8}{0.05^2}\ln\left(\frac{4\left((2*452956,837)^{10}+1\right)}{0.05}\right)=452956,8628$$

$$\frac{8}{0.05^2}\ln\left(\frac{4\left((2*452956,8628)^{10}+1\right)}{0.05}\right)=452956,8646$$

Por tanto N tiene que ser mayor o igual a 452957.

10 Considere que le dan una muestra 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

1. ERM: