Universidad De Granada

Grado ingeniería informática (2017 – 2018)

Aprendizaje Automático

Trabajo 2: Cuestiones de Teoría

Trabajo realizado por Antonio Miguel Morillo Chica.

---

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.**

Dos de las condiciones que se exigen para que un problema pueda ser aproximado por predicción sería:

- Los datos tienen que ser muestras independientes e idénticamente distribuidas de una probabilidad P. Si siempre cogemos las mismas muestras y no varían, no podemos aprender nada. Podemos sacar valores repetidos, pero deben variar de un conjunto a otro.
- Necesitamos un tamaño de muestras, N, grande, pero no demasiado (mirar formula ejercicio 5) cuanto más grande sea M mayor será N.

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 asus problemas 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 decisión correcta, por el teorema de Non-Free-Lunch, que dice que para cada (A,H), existe una probabilidad P que falla, aunque esa P pueda ser aprendida con éxito por otro modelo. Por otra parte, todos los modelos son equivalentes en termino medio sobre todas las posibles funciones objetivo f.

Debemos, pues, aplicar cada modelo en su distribución de clases de probabilidad P en las que pueda aprender, solo así tendremos algo de éxito. Si la empresa llegara un problema diferente al que han trabajado hasta ahora y el

algoritmo no está preparado para manejar los datos del nuevo problema, entonces el algoritmo no les serviría de nada.

3. **Supongamos un conjunto de datos D de 25 ejemplos extraidos de una funcion desconocida f : X → Y, donde X = R e Y = {−1, +1}. Para aprender f usamos un conjunto simple de hipótesis H = {h1, h2} donde h1 es la función constante igual a +1 y h2 la función constante igual a -1. Consideramos dos algoritmos de aprendizaje, S(smart) y C(crazy). Se 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 podemos asegurar nada puesto que la hipótesis que hace S será el valor de la etiqueta que represente un mayor número de individuos de la muestra. Sin embargo si tomamos las muestras de una parte del espacio donde predominan puntos de una clase, y sin embargo la distribución feneral podría asemejarse a una aleatoria.

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.**

Claro, ya que como en el apartado anterior podríamos haber tomado sólo los 25 puntos con una sola de la etiquetas.

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,**

$$\mathbf{P[|E_{out}(g) - E_{in}(g)| > \varepsilon) < \delta(\varepsilon, N, |H|)}$$

a) **Dar una expresión explícita para δ(e, N, |H|)**

Según las transparencias de teoría vistas en teoría:

$$P[|E_{out}(g) - E_{in}(g)| > e) <= 2He^{-2\varepsilon^2 N} \text{ para cualquier } \varepsilon > 0.$$

b) **Si fijamos ε = 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?.**

Lo que hay que hacer es simplemente despejar por lo que llamaremos alpha a la cota de error.

$$2He^{-2\varepsilon^2 N} <= \alpha ;$$

$$-2\varepsilon^2 N <= \log(\alpha/2H)$$

$$N >= -\log(\alpha/2H)/2\varepsilon^2$$

$$N >= 839{,}941$$

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

Como sería unicamente sustituir los valores de H quedaría como:

- H = 10 : N >= 1300,46
- H = 100 : N >= 1760,98

¿Que conclusiones obtiene?

La conclusión es que mientras la cota de error sea más pequeña necesitaremos más datos para minimizar los errores.

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,**

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

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

Aquel algoritmo que de las clases de hipotesis, H, minimize al máximo posible el error.

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

Sí, pero no nos aseguramos que sea el mejor de todos ya que g, como se ve en el siguiente ejercicio ha de ser el mejor la menor de todas.

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

En el libro "Learning from Data" encontramos la siguiente apartado:

The hypothesis g is not fixed ahead of time before generating the data, beca use which hypothesis is selected to be g depends on the data. [...] The way to get around this is to try to bound $P[|E_{in}(g) - E_{out}(g)| > \varepsilon]$ in a way that does not depend on which g the learning algorithm picks. There is a simple but crude way of doing that [...] y continua explicando el procedimiento que se muestra en el sigiente ejercicio.

Lo que viene a decir que no depende de cual algoritmo usemos ya que como se ve cita en teoría y el propio libro g depende de D, no del algoritmo que se use.

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

Es una cota laxa cuando usamos un conjunto de hipotesis. La desventaja de la tesis de estimación uniforme es que al introducir el factor M este es más flojo que el límite de una única hipótesis, y solo será significativo si M es finito.

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.**

Como se dice en el libro "Learning From Data": If you are allowed to change h after you generate the data set, the assumptions that are needed to prove the Hoeffding Inequality no longer hold.

Con multiples hipótesis el algoritmo selecciona la hipotesis final basada en D, es decir "P[$|E_{in}(h_m)$ - $E_{out}(h_m)| > \varepsilon$] es pequeño". Si el conjunto H es mayor que uno la desigualdad se aplica para todas es decir se aplica la desigualdad para M terminos como se puede ver:

$$\begin{aligned}\mathbb{P}[\ |E_{\text{in}}(g) - E_{\text{out}}(g)| > \epsilon\ ] \quad \leq \quad & \mathbb{P}[\quad |E_{\text{in}}(h_1) - E_{\text{out}}(h_1)| > \epsilon \\ & \quad \textbf{or}\ |E_{\text{in}}(h_2) - E_{\text{out}}(h_2)| > \epsilon \\ & \quad \cdots \\ & \quad \textbf{or}\ |E_{\text{in}}(h_M) - E_{\text{out}}(h_M)| > \epsilon\ ] \\ \leq \quad & \sum_{m=1}^{M} \mathbb{P}\left[|E_{\text{in}}(h_m) - E_{\text{out}}(h_m)| > \epsilon\right].\end{aligned}$$

Con lo que la expresión final sería:

$$\mathbb{P}[|E_{\text{in}}(g) - E_{\text{out}}(g)| > \epsilon] \leq 2Me^{-2\epsilon^2 N}$$

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 se puede separar el conjunto de puntos entonces $m_H(k^*) = 2^{k^*}$ por lo que no se cumple la definición de punto de ruptura.

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

Si divide a cualquier conjunto entonces dividirá a uno, podemos reducir este apartado al anterior, luego, no es un punto de ruptura.

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, puede existir un conjunto distinto de k_* que si pueda separar. Por tanto no nos sirve para saber si es o no punto de ruptura.

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

Si es suficiente para decir que es 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 tanto, $m_H(k^*) < 2^{k^*}$, por ser $m_H(k^*)$ el máximo número de dicotomías posible en una muestra, cumpliendo la definición de punto de ruptura.

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

Si $m_H(k) = 2^{k^*}$, entonces es falso $m_H(k_*) < 2^{k^*}$, y por tanto no cumple la definición. Luego no 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 (e) sea como mucho 0.05?**

Para realizar el calculo del tamaño de la media muestral, usamos el ejemplo que tenemos en la transparencia 16 de la sesión 4, tal y como se hace a continuación:

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

como:

- $\delta = 1 - 95\% = 0{,}05$,
- $\varepsilon = 0{,}05$

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

Tenemos que calcularlo de forma iterativa (probabando) y obetenemos N>=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.**

Si utilizamos el principio de inducción ERM, al minimizar el error obtenemos un conjunto de pesos en función de las clasificaciones que hagamos en cada dato a partir de las etiquetas. Vamos comprobando si la etiqueta está mal colocada o no partiendo de una recta que estará por encima si la etiqueta es -1 y por debajo si esta a 1, con cada clasificación alteramos los pesos del vector, que nos dejarán formar más tarde el hiperplano, (algoritmo Perceptron). Una vez tenemos estos pesos, podemos sacar el hiperplano con un par de valores y ya tendríamos el ajuste prácticamente hecho.

Podemos casi afirmar que el principio ERM va mucho mejor con datos que con funciones.

Si tenemos un conjunto de entrenamiento, ERM va bien para esos datos y si son muchos datos. Ademas puede ser útil cuando tenemos varias dimensiones. No obstante, cuando tenemos un número de muestras pequeño, respecto al número de parámetros efectivo, ERM no garantiza que se pueda aprender. Es aquí donde aparece SRM.

Con SRM aproximamos el error e un conjunto de H´s y minimizamos la dimensión de VC, en otras palabras, minimizan un límite superior al riesgo esperado. Así si tenemos un conjunto de datos pequeño, sería adecuado usar SRM, pues ERM necesita un gran conjunto de entrenamiento para aprender.

Bonus

1. **Supongamos que tenemos un conjunto de datos D de 25 ejemplos extraidos de una funcion desconocida f : X → Y, donde X = R e Y = {−1, +1}. Para aprender f usamos un conjunto simple de hipótesis H = {h1, h2} donde h1 es la función constante igual a +1 y h2 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?**

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

2. **Consideremos el modelo de aprendizaje "M-intervalos"donde la clase de funciones H está formada por h : R → {−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.**

El más pequeño punto de ruptura es 2m + 1, puesto que hasta 2m podemos separar cualquier partición de los datos sin más que situar un intervalo en cada conjunto de puntos contiguos de la clase +1. El problema llega cuando hay m+1 conjuntos de puntos contiguos de la clase +1, y lógicamente el menor número de puntos para que esta circunstancia pueda darse es cuando estos conjuntos son de únicamente un punto, y los conjuntos de puntos que los separan son también de un punto.