Segundo cuestionario de teoría de Aprendizaje Automático

UNIVERSIDAD DE GRANADA

Antonio Manuel Fresneda Rodríguez
antoniomfr@correo.ugr.es
77447672-W

Índice

Ejercicio 1.....3
Ejercicio 2.....3
Ejercicio 3.....4
Ejercicio 4.....4
Ejercicio 5.....5
Ejercicio 6.....6
Ejercicio 7.....6
Ejercicio 8.....7
Ejercicio 9.....8
Ejercicio 10.....8

Ejercicio 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.

Que la muestra sea:

1. Independiente
2. Idénticamente distribuida

Si estas dos condiciones se cumplen, podemos asegurar que la media de la muestra es cercana a la de la población con una probabilidad (teorema del limite central).[1]

Ejercicio 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.

La solución por la que ha optado el jefe de investigación de la empresa es una mala solución.
El teorema NFL dice que para todos los algoritmos existe un conjunto de datos en el que el algoritmo vaya a fallar, incluso si ha sido aprendido correctamente. También dice que todos los algoritmos son equivalentes en media para todas las posibles funciones objetivo.
Por tanto, la decisión del jefe de investigación no es buena, ya que puede darse el caso en el que tenga unos datos y el algoritmo que seleccionó no funcione bien.

Ejercicio 3.

Supongamos un conjunto de datos D de 25 ejemplos extraídos de una función 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.

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

No se puede saber lo que ocurre fuera de la muestra, por lo que no podemos asegurar el comportamiento de C o de S.
Puede ocurrir que escojamos una muestra no representativa de la población. Además, si introducimos conocimiento acerca de la distribución de la muestra a la hora de escoger la función con la que vamos a intentar aproximar f, cometeremos uno de los mayores errores, el cual es que basaremos nuestro aprendizaje en la distribución en una muestra de la que no tenemos garantías de que sea representativa de la población y vamos a tener un error bajo dentro de la muestra, pero no garantizaríamos el buen funcionamiento fuera de la muestra.

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

Vamos a trasladar este ejemplo al del recipiente con bolas verdes y rojas.
Imaginemos que en nuestro bol hay pocas bolas verdes y la muestra que hemos cogido solo tiene esas bolas verdes (la muestra no es representativa de la población), entonces el algoritmo seguramente funcione mal fuera de la muestra, ya que no tenemos una buena muestra para aprender. Pero si nos ponemos en el caso contrario, en el que solo haya bolas verdes en el bol, la muestra pasa a ser representativa de la población y el algoritmo funcionará bien, pero nunca sabemos la verdadera distribución de las bolas que están dentro del bol.
Esto lo que nos da a entender es que si solo miras la muestra, no puedes asegurar nada de lo que ocurre fuera de ella, ya que existe la posibilidad (suele ser baja) de que obtengas una muestra que no es representativa de la población.

Ejercicio 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,

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

a) Dar una expresión explícita para $\delta(\varepsilon,N,|H|)$

La expresión es: $\delta=2|H|e^{-2\varepsilon^2 N}$

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?.

Usando la expresión del apartado anterior:

$$\delta=2|H|e^{-2\varepsilon^2 N}=\delta \rightarrow \frac{\delta}{2|H|}=e^{-2\varepsilon^2 N} \rightarrow \ln\frac{\delta}{2|H|}=\ln e^{-2\varepsilon^2 N} \rightarrow \frac{\ln\frac{\delta}{2|H|}}{-2\varepsilon^2}=N$$

$$N=\frac{\ln\frac{\delta}{2|H|}}{-2\varepsilon^2} \rightarrow \frac{\ln\frac{0,03}{2}}{-2\,0,05^2}=839,94$$

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

Usando la expresión anterior:

1. H=10 → N= 1.300,45
2. H=100 → N= 1.760,98

¿Que conclusiones obtiene?

Que a mayor complejidad de la clase de funciones, tenemos que aumentar el tamaño de la muestra.

Ejercicio 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,

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

a) ¿Cuál es el algoritmo de aprendizaje que se usa para elegir g?
Utilizaremos el algoritmo ERM.
Si escribimos ε en función de N, δ y |H| tenemos que:
Con probabilidad 1-δ:

$$E_{out}(g)\leq E_{in}(g)+\sqrt{\frac{1}{2N}\ln\frac{2|H|}{\delta}}$$

Para clases finitas, la desigualdad de Hoeffding se verifica, ya que |H|es finito, y entonces podemos asegurar que podemos aprender usando ERM.
b) Si elegimos g de forma aleatoria ¿seguiría verificando la desigualdad?
Si, solo si conocemos |H|, siendo H la clase de funciones a la que pertenece g. En el caso de que la clase H sea finita podríamos usarla, pero en el caso de que sea infinita no, ya que el termino de la derecha se iría a infinito
c) ¿Depende g del algoritmo usado?.
Si. Si escogemos ERM g va a pertenecer a una clase finita, mientras que si cogemos SRM puede que la clase de funciones no sea finita.
d) Es una cota ajustada o una cota laxa?
Es una cota laxa.

Ejercicio 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.

Hay que remarcar que la hipótesis más importante que asumimos cuando usamos la desigualdad de Hoeffding es que h debemos de escogerla **antes** de generar el dataset.
Esta hipótesis es critica para que se verifique la desigualdad.
Si cambiamos h después de haber generado el dataset, la hipótesis mencionada anteriormente no se validaría.
Cuando tenemos múltiples hipótesis en H el algoritmo selecciona la hipótesis final **después** de haber generado el dataset.

Ejercicio 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 x1, . . . , xk^* que *H* puede separar ("shatter").
b) Mostrar que *H* puede separar cualquier conjunto de k^* puntos.
c) Mostrar un conjunto de k^* puntos x_1 , . . . , x_{k*} que *H* no puede separar
d) Mostrar que *H* no puede separar ningún conjunto de k^* puntos.
e) Mostrar que mH (k) = 2^{k*}

Nos servirían las afirmaciones:

1. c): Según lo visto en teoría, si para un valor k, $m_H(K)<2^k$, entonces k es un punto de ruptura para H, lo que significaría que H no puede separar una muestra de tamaño k.

Ejercicio 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?

Voy a usar la expresión:

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

Esta ecuación hay que resolverla de forma iterativa. Para ello, he implementado un pequeño código en Python para hacer la resolución. El código es el siguiente:

```
import math

epsilon=0.05
delta=0.05
dvc=10

def calcularN(n):
    r= 8 / (epsilon**2)
    r*= math.log(4*((2*n)**dvc + (dvc+1))/delta)
    return r

N=1000
N2=calcularN(N)
while abs(N2-N) > 1e-5:
    N=N2
    N2=calcularN(N)

print (N)
```

El resultado es: 452956.864, → $N \geq 452957$

Ejercicio 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.

La principal ventaja de escoger el SRM frente al ERM es que podemos usar un mayor numero de clases de funciones (en el ERM estamos sesgados a usar clases de funciones que tienen una d_{vc} finita mientras que con el SRM si podemos usarlas).
Si tenemos un gran número de muestras y usamos una clase de funciones con una d_{vc} finita que ajuste bien los datos (cosa que en la realidad es difícil) podríamos decir que es preferible el ERM frente al SRM.
En casos reales, SRM tiene más interés, aunque el precio a pagar es que es mas complejo, ya que consideramos la clase de funciones infinita como la unión de infinitas clases cada una de ellas con una d_{vc} finita. Encontrar la clase que mejor se ajuste a nuestro objetivo puede llegar a ser difícil ya que tenemos que buscar un buen error en la muestra pero sin tener una clase muy compleja

Bibliografía.

Me he apoyado en las transparencias de teoría, en los libros que hay en la plataforma, en el libro Learning from Data y dudas con algunos compañeros de clase.

[1]https://support.minitab.com/es-mx/minitab/18/help-and-how-to/statistics/basic-statistics/supporting-topics/data-concepts/about-the-central-limit-theorem/