CUESTIONARIO 2.

JULIO A. FRESNEDA – 49215154F

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.

Para formar un modelo a partir de un conjunto finito de datos, es decir, para aproximar un problema de predicción por inducción desde una muestra de datos, necesitamos, entre otras cosas:

Un conocimiento a priori, usado para imponer restricciones sobre el potencial de una función de la clase de funciones elegida para ser solución. Habitualmente, dicho conocimiento a priori proporciona, explícita o implícitamente, el ordenamiento de las funciones de acuerdo con alguna medida de flexibilidad para ajustarse a los datos.

Un principio inductivo, o método de inferencia, es decir, una prescripción general para combinar el conocimiento a priori con los datos de entrenamiento disponibles para producir una estimación de la verdadera dependencia (desconocida). Un principio inductivo especifica qué se necesita hacer, pero no cómo hacerlo. Por ejemplo, un principio inductivo sería el principio inductivo de penalización, Early Stopping Rules o Structural Risk Minimization.

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.

Sí, la decisión es correcta. Lo que el jefe de investigación ha hecho es, a partir de un training data (los problemas sobre los que la empresa ha trabajado, los algoritmos de aprendizaje usados y sus resultados), estimar una única clase de funciones y desarrollar un único algoritmo que pueda resolver los problemas futuros.

El jefe de investigación ha observado que las clases de funciones y algoritmos usados, es decir, el training data, son ejemplos de una distribución de probabilidad P, por lo que se puede estimar un modelo genérico para futuros problemas a partir del training data mediante inductive learning.

Como ejemplo tenemos el modelo BIN (Bolas rojas y verdes), el jefe de investigación sabe que no se puede garantizar nada de cómo van a ser los algoritmos para problemas futuros (al igual que en el modelo BIN no podemos garantizar que, si la muestra de bolas es verde, el total de bolas del cubo no sea rojo en su mayoría) pero sí que podemos confiar en lo que el jefe obtenga como modelo estimado, pues el peor caso, que el training data no represente a todos los datos, es posible, pero no es probable.

Este nuevo algoritmo muy probablemente podrá adaptarse a futuros problemas con un margen de error muy pequeño, pues la inecuación de Hoeffding nos dice que muy probablemente si el error de Ein (error respecto al training data de la función estimada gracias al training data) es cercano a 0, el Eout (error de la función estimada para cualquier problema) será también cercano a 0.

Esto quiere decir que el algoritmo único que el jefe de investigación ha desarrollado para los problemas futuros muy probablemente tendrá un error cercano a 0 (asumimos que funciona con error cercano a 0 para los problemas sobre los que la empresa ya ha trabajado).

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

Suponiendo que la hipótesis aleatoria es la que elige C (crazy):
No, no puede.

En todo caso puede producir una hipótesis que muy posiblemente funcione mejor que la aleatoria, pero no puede garantizarlo pues no sabemos la función real f.

No podemos garantizar nada sobre la función real fuera de la muestra que tenemos, pero sí que podemos confiar (sin total seguridad) en que S produzca una hipótesis que tenga mejor comportamiento que la aleatoria, pues el caso en que no produzca mejor hipótesis es posible, pero es poco probable.

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

Sí, es posible.

Es muy poco probable, pero podría darse el caso. Para justificarlo, voy a mostrar un ejemplo:

Si queremos estimar una función g que se asemeje a la siguiente función f (la cual no sabemos):

$$f(x) = \begin{cases} -1 \; si \; x \; es \; primo \\ +1 \; si \; x \; no \; es \; primo \end{cases}$$

Y con muy mala suerte, para el training data, hemos cogido 25 ejemplos cuyos valores de x son casualmente todos primos, S estimará de forma perfecta la función para estos 25 ejemplos, pero para la mayoría de puntos fuera de la muestra funcionará muy mal.
S creerá que la hipótesis h2 se asemeja muchísimo a f, lo cual es incorrecto. El Ein será de 0, pero el Eout será cercano a 1.

En cambio, C, al usar la hipótesis opuesta, tendrá un error para la muestra de 1, mientras que el Eout será cercano a 0.

Por lo tanto, es posible.

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,

$$\mathbb{P}[|Eout(g) - \; Ein(g)| \; > \; \epsilon) \; < \; \delta(\epsilon, N, |\mathcal{H}|)$$

a) Dar una expresión explícita para $\delta(\epsilon, N, |\mathcal{H}|)$
b) Si fijamos ϵ = 0.05 y queremos que el valor de δ sea como máximo 0.03
¿Cuál será el valor más pequeño de N que verifique estas condiciones cuando H = 1?
c) Repetir para H = 10 y para H = 100 ¿Qué conclusiones obtiene?

a) Una expresión explícita puede ser:

$$\delta(\epsilon, N, |\mathcal{H}|) = 2|\mathcal{H}|e^{-2\epsilon^2 N}$$

b) Vamos a resolver la ecuación.

$$\delta(\epsilon, N, |\mathcal{H}|) = 0.03$$
$$2 \cdot 1 \cdot e^{-2\cdot(0.05)^2 \cdot N} = 0.03$$
$$2e^{-0.005N} = 0.03$$
$$\ln(2e^{-0.005N}) = \ln(0.03)$$
$$\ln(2) + \ln(e^{-0.005N}) = \ln(0.03)$$
$$-0.005\mathrm{N}\ln(e) \approx -4.19970$$
$$N \approx -4.19970/-0.005$$
$$N \approx 840$$

El valor más pequeño de N que verifique estas condiciones es de 840 aproximadamente.

c) Resolvamos para H=10:

$$\delta(\epsilon, N, |\mathcal{H}|) = 0.03$$
$$2 \cdot 10 \cdot e^{-2\cdot(0.05)^2 \cdot N} = 0.03$$
$$20e^{-0.005N} = 0.03$$
$$\ln(20e^{-0.005N}) = \ln(0.03)$$
$$\ln(20) + \ln(e^{-0.005N}) = \ln(0.03)$$
$$-0.005\mathrm{N}\ln(e) \approx -6.50229$$
$$N \approx 1301$$

Para H=100:

$$\delta(\epsilon, N, |\mathcal{H}|) = 0.03$$
$$2 \cdot 100 \cdot e^{-2\cdot(0.05)^2 \cdot N} = 0.03$$
$$200e^{-0.005N} = 0.03$$
$$\ln(200e^{-0.005N}) = \ln(0.03)$$
$$\ln(200) + \ln(e^{-0.005N}) = \ln(0.03)$$
$$-0.005\mathrm{N}\ln(e) \approx -8.8048$$
$$N \approx 1761$$

La conclusión que obtenemos es que cuanto más compleja sea nuestra clase de funciones H, más ejemplos de muestra vamos a necesitar para obtener buenos resultados.

La justificación matemática es la siguiente.

Sabemos que la desigualdad es:

$$\mathbb{P}[|Eout(g) - Ein(g)| > \epsilon) < 2|\mathcal{H}|e^{-2\epsilon^2 N}$$

Vemos que a mayor $|\mathcal{H}|$, más posibilidades hay de que la diferencia de error Eout respecto Ein sea mayor (es decir, de que sobreajuste el training data).

Tenemos que $2|\mathcal{H}|e^{-2\epsilon^2 N}$ es un número constante, el cual debemos mantener pequeño. Por tanto, al aumentar $|\mathcal{H}|$, para compensar y que el número no crezca demasiado, debemos aumentar el número N, para que disminuya $e^{-2\epsilon^2 N}$ y $2|\mathcal{H}|e^{-2\epsilon^2 N}$ no crezca demasiado.

La demostración teórica es la siguiente.

Cuanto más grande sea $|\mathcal{H}|$, más grande debe ser el tamaño de la muestra porque cuantas más funciones tenga H, más probabilidades hay de que alguna función se ajuste bien a los datos de entrenamiento. Es decir, un $|\mathcal{H}|$ grande ayuda a que $Ein \approx 0$ más fácil que un $|\mathcal{H}|$ pequeño. La consecuencia de esto es que corremos el riesgo de sobreajuste, es decir, que ajuste demasiado bien el ruido de la muestra. Es por ello que necesitamos una muestra más grande que si $|\mathcal{H}|$ fuese pequeño, para disminuir este sobreajuste.

En resumen, si queremos que $Ein \approx 0$ y tenemos un $|\mathcal{H}|$ pequeño, podemos usar una muestra pequeña. Pero si tenemos un $|\mathcal{H}|$ grande, nuestro tamaño N de muestra deberá ser mayor.

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,

$$\mathbb{P}[|Eout(g) - Ein(g)| > \epsilon) < \delta$$

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?

a) Cualquier algoritmo vale, la desigualdad de Hoeffding no depende de cómo se ha obtenido g.
b) Sí, se seguiría verificando pues la desigualdad de Hoeffding no depende de cómo se ha obtenido g.
c) Sí, sí que depende. Según qué algoritmo usemos, obtendremos una g u otra.
d) La cota es laxa, pues depende de muchas variables flexibles.

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.

Porque la desigualdad de Hoeffding como tal usa una función específica h, asumiendo que H=1, y no usa el conjunto H, ni tiene en cuenta el número de hipótesis de H.
La inecuación es la siguiente.

$$\mathbb{P}[\mathcal{S}: |Eout(h) - Ein(h)| > \epsilon) \leq 2e^{-2\epsilon^2 N}$$

Donde Ein(h) representa el error que hemos obtenido en la muestra de entrenamiento con la función h, y Eout(h) es el error que probablemente obtengamos fuera de la muestra con h.

Como vemos, el número de hipótesis no se considera en la inecuación, si no que se asume que solo hay una hipótesis h y a partir de ahí desarrollamos la inecuación.

Para aplicar la desigualdad de Hoeffding cuando tengamos varias posibles hipótesis h (|H| > 1), debemos modificar un poco la desigualdad:

En vez de la hipótesis h, usaremos una hipótesis genérica g.
$\{\mathcal{S}: |Eout(g) - Ein(g)| > \epsilon\}$ es la unión de usando cada h perteneciente a H. Es decir:

$$\{\mathcal{S}: |Eout(g) - Ein(g)| > \epsilon\} = \bigcup_{hi \, \epsilon \, H} (\mathcal{S}: |Eout(hi) - Ein(hi)| > \epsilon)$$

Por lo tanto, la desigualdad, ajustada para H mayor que 1, sería la siguiente:

$$\mathbb{P}[\mathcal{S}: |Eout(g) - Ein(g)| > \epsilon) \leq 2|\mathcal{H}|e^{-2\epsilon^2 N}$$

Esta desigualdad sí que se puede aplicar a casos donde el número de hipótesis de H sea mayor que 1, pero podemos aplicar la desigualdad inicial de forma directa, hay que ajustarla al hecho de que tengamos varias hipótesis en H.

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 x1, …, xk* que H no puede separar
d) Mostrar que H no puede separar ningún conjunto de k* puntos
e) Mostrar que $m_H(k^*) = 2^{k^*}$

El punto de ruptura indica que cualquier conjunto a partir de k* puntos, H no puede separar de ninguna forma.
A) No. Si H puede separar algún conjunto de k* puntos, el punto de ruptura debe ser mayor que k*.
B) No. Si H puede separar cualquier conjunto de k* puntos, el punto de ruptura debe ser mayor que k*.
C) No. No basta, para que k* sea punto de ruptura, ningún conjunto (y no sólo alguno) de k* puntos se puede separar por H.

D) Sí. Sí que nos serviría, suponiendo que H sí que pueda separar k*-1 puntos. Si no, tampoco nos serviría: Por ejemplo, si nuestro punto de ruptura es k=5, y tenemos un conjunto de 700 puntos, H no puede separar esos 700 puntos (puede separar 4 como máximo), pero 700 no es un break point.
E) Si $m_H(k^*) = 2^{k^*}$ significa que el break point (si hay) es mayor que k*, al igual que en los casos A) y B).

9. Para un conjunto H con dVC = 10, ¿qué tamaño muestral se necesita (según la cota de generalización) para tener un 95% de confianza de que el error de generalización sea como mucho 0.05?

Sabemos que $\delta = 0.05$, $\varepsilon = 0.05$ y $d_{VC} = 10$. También sabemos que $\sqrt{\frac{8}{N} ln \frac{4m_H(2N)}{\delta}} \leq \varepsilon$, es decir:

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

Sustituyendo:

$$N \geq \frac{8}{0.05^2} \ln\left(\frac{4((2N)^{10} + 1)}{0.05}\right) \overset{N=1000}{\Longrightarrow} N \geq 257257.36$$

Redondeando, el tamaño muestral de N debería ser de N = 300000 aproximadamente.

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.

Usando ERM obtenemos que con una probabilidad de al menos $1 - \delta$, $Eout(g) \leq Ein(g) + \sqrt{\frac{1}{2N} \ln \frac{2|H|}{\delta}}$

Por lo tanto, la ventaja de usar ERM cuando el tamaño N es grande, es que se puede conseguir que $Eout \approx 0$ con alta probabilidad (siempre que Ein se acerque a 0).

Otra ventaja de usar ERM es que la desigualdad anterior no depende de $\mathcal{X}$, P(x), f o de cómo se ha estimado g.

Sin embargo, una desventaja de ERM es que no funciona tan bien con muestras pequeñas. Para muestras pequeñas, es mejor usar el Structural Risk Minimization.

El SRM evita que usemos clases de funciones demasiado complejas, clases que seguramente sobreajusten los datos de entrenamiento, ya que tenemos muy pocos.

El funcionamiento de SRM consiste en mantener el error empírico constante y pequeño, e ir reduciendo la dimensión VC.

Las ventajas de usar SRM son que disminuimos el riesgo de sobreajuste, y podemos estimar una buena función con Eout pequeño a partir de pocos datos.

La desventaja de SRM es que el Eout que obtendremos difícilmente será cercano a 0, pues la función que estimamos no suele ser suficientemente compleja para ello.