UNIVERSIDAD
DE GRANADA

Aprendizaje Automático
grado en ingeniería informática

Trabajo 2

Cuestiones de Teoría

Autor
Vladislav Nikolov Vasilev

Rama
Computación y Sistemas Inteligentes

Escuela Técnica Superior de Ingenierías Informática y de
Telecomunicación

Curso 2018-2019

Índice

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.

Solución

Para que un problema de predicción pueda ser aproximado por inducción desde una muestra de datos, necesitamos que se den las siguientes condiciones:

- Que la muestra de datos sea i.i.d. (independiente e identicamente distribuida). Esto significa que los elementos de la muestra no se influyen entre sí (independiente) y que cada elemento de la muestra es escogido de la misma distribución de probabilidad (idénticamente distribuido).
- Que la distribución de probabilidad de los datos de entrenamiento sea la misma que de los de test.

Si no se dan estas condiciones, no se puede asegurar una correcta aproximación por inducción.

En el caso de la primera condición, por ejemplo, si escogemos los datos de forma arbitraria (no i.i.d.) no podríamos decir nada sobre la población, ya que el análisis probabilístico realizado con la desigualdad de Hoeffding nos dice que, para una muestra escogida de forma aleatoria, se tiene que:

$$\mathbb{P}(\mathcal{D} : \mid \mathrm{E}_{in}(h) - \mathrm{E}_{out}(h) \mid \; > \varepsilon) \leq 2e^{-2\varepsilon^2 N}$$

es decir, que escogiendo un tamaño de muestra N lo suficientemente grande y un ε error razonable, podemos decir que muy probablemente $\mathrm{E}_{in}(h)$ y $\mathrm{E}_{out}(h)$ disten como mucho entre sí un valor ϵ, y que por tanto $\mathrm{E}_{in}(h) \approx \mathrm{E}_{out}(h)$. Así que, escogiendo datos de forma arbitraria sería como trabajar a ciegas, sin ningún tipo de información.

En el caso de la segunda condición, si escogemos datos de entrenamiento de una distribución de probabilidad P_1 y luego escogemos datos de test de otra distribución de probabilidad P_2, por mucho que que con nuestro algoritmo de aprendizaje hayamos conseguido hacer que $\mathrm{E}_{in} \approx 0$, no podríamos afirmar que $\mathrm{E}_{in}(h) \approx \mathrm{E}_{out}(h)$, ya que en este caso los datos de entrenamiento y de test provienen de distribuiciones de probabilidad diferentes, con lo cuál podrían no ser nada parecidos.

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.

Solución

Se puede considerar que la decisión tomada no es la correcta. Al haber escogido una única clase de funciones y un único algoritmo se está restringiendo mucho la cantidad de problemas que se pueden resolver. Puede suceder incluso que no resuelva bien los problemas futuros, ya que la naturaleza de estos no es conocida a priori.

Para intentar justificar por qué no es buena idea restringirse a un único algoritmo, podemos hacer referencia al teorema de **No-Free-Lunch**, que dice que para cada algoritmo $\mathcal{A}$ existe una distribución de probabilidad P en la que dicho algoritmo falla, pero que puede ser aprendida por otro. Por tanto, puede ser que llegue un nuevo problema cuya distribución de probabilidad sea una en la que el algoritmo que se haya escogido en la empresa falle, y por tanto, la no obtendrá unos resultados que satisfagan a los clientes, lo cuál se podría traducir en una mala situación para la empresa.

Por otro lado, si se limita la clase de funciones a una que por ejemplo sea muy pequeña, si llega un nuevo problema puede suceder que la clase de funciones se quede muy corta, y los valores de los errores obtenidos tanto en la muestra de entrenamiento proporcionada como en la muestra de test sean muy malos, debido a que la función no tenga la capacidad de explicar correctamente los datos o de generalizar bien. Esto también sería un problema para la empresa, ya que nadie quiere tener un resultado pésimo que no pueda utilizar luego.

En conclusión, por mucho que en el pasado se hayan usado una serie de algoritmos y clases de funciones, no existe nada que nos indique que éstos funcionen correctamente para nuevos problemas. Es muy importante explotar el conocimiento específico del problema para obtener los mejores resultados, y al imponer límites de lo que se va a utilizar en el problema de aprendizaje se limita la capacidad de decidir qué técnicas utilizar para resolverlo. No existe ninguna clase de funciones ni algoritmo que resuelvan todos los problemas, y por tanto, para cada problema, hay que realizar un buen análisis para determinar cuáles serían los más adecuados.

Ejercicio 3

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

Solución

En el ámbito del aprendizaje, una solución PAC significa que es *Probably Approximately Correct*, lo cuál traducido al español vendría a ser algo así como "correcta probablemente aproximada". Veamos qué significa todo esto sobre la desigualdad de Hoeffding aplicada al problema de aprendizaje:

$$\mathbb{P}(\mathcal{D} : \mid \mathrm{E}_{in}(h) - \mathrm{E}_{out}(h) \mid \; > \varepsilon) \leq 2e^{-2\varepsilon^2 N} \tag{1}$$

- La parte de "probablemente" hace referencia a una probabilidad. Esto se puede ver en la expresión dada por (1) como la probabilidad de que algo malo suceda. Este evento malo es que la diferencia entre los valores de $\mathrm{E}_{in}(h)$ y $\mathrm{E}_{out}(h)$ sea mayor que un ε dado, o lo que es lo mismo, que los errores disten mucho entre sí. Como en la expresión de la parte derecha nos encontramos con un exponencial negativo, con los valores adecuados de ε y N podemos hacer que la probabilidad de que ese mal evento suceda sea pequeña, y por tanto, que la probabilidad del caso contrario (que la diferencia entre los valores de $\mathrm{E}_{in}(h)$ y $\mathrm{E}_{out}(h)$) sea muy probable (tenga una probabilidad más alta).
- La parte de "aproximada" indica que $\mathrm{E}_{in}(h)$ no es exactamente igual que $\mathrm{E}_{out}(h)$, pero que ambos valores están muy próximos. Esta aproximación viene dada por el valor de ε.

La **incertidumbre** viene dada por la probabilidad. Nunca se puede tener la certeza de que el resultado sea 100 % correcto, pero se puede afirmar con una alta probabilidad de que así sea (por eso es PAC). La **imprecisión**, por otro lado, viene dada por el valor de ε. Es decir, los valores de $\mathrm{E}_{in}(h)$ y $\mathrm{E}_{out}(h)$, al estar aprendiendo de una muestra la cuál puede tener un tamaño no lo suficientemente grande o no ser muy representativa de la población, van a ser diferentes. Si pudiésemos aprender de toda la población directamente, en ese caso ε sería 0, ya que los dos errores serían iguales, pero habría que pagar muchos costes de tiempo, potencia de cómputo y almacenamiento. Por tanto, al estar siempre aprendiendo de una muestra y no de la población entera nos vamos a encontrar con estos dos problemas.

Ejercicio 4

Suponga un conjunto de datos $\mathcal{D}$ de 25 ejemplos extraídos de una función 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.

Solución

No se puede garantizar. Se puede intentar afirmar que, con alta probabilidad, S sea capaz de generalizar mejor que C debido a que escoge aquella hipótesis que mejor ajusta los datos. Ahora bien, garantizar que pueda tener un mejor comportamiento sobre cualquier punto fuera de la muestra es algo muy difícil, por no decir casi imposible. Partiendo de que la muestra puede no representar lo suficientemente bien la población, S puede equivocarse entonces y escoger una hipótesis con la que obtenga muy buenos resultados en la muestra, pero que al generalizar luego, obtenga un error fuera de la muestra muy elevado, haciendo por tanto que no sea mejor sobre cualquier punto fuera de la muestra. Además, en el aprendizaje siempre nos encontramos con incertidumbre, ya que nunca se puede afirmar con toda seguridad que los resultados obtenidos son perfectos y posteriormente no se cometerán errores al predecir nuevos datos. Y además, no existe ningún algoritmo que nos garantice un mejor aprendizaje que otro en todos los casos, ya que según el teorema de **No-Free-Lunch**, existirá una distribución de probabilidad P en la que el algoritmo de aprendizaje $\mathcal{A}$ falle, pero en la que otro obtendrá unos buenos resultados.

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

Solución

Puede suceder que la hipótesis producida por C sea mejor que la escogida por S. Para verlo más claro, siguiendo lo que se nos ha dicho en el enunciado sobre los y_n en $\mathcal{D}$, podemos suponer un caso extremo en el que las etiquetas fuera de la muestra sean todas $y_n = -1$. En este caso, S escogería una hipótesis que funciona muy bien en la muestra, ya que clasifica todos los datos bien, mientras que C sería pésimo y no clasificaría nada bien. Sin embargo, debido a que la muestra no es lo suficientemente representativa de la población (de hecho, es una muy mala representación), al intentar generalizar nos encontraríamos que S tiene un rendimiento pésimo ya que no clasificaría nada bien, mientras que C, al haber escogido la hipótesis en las que todas las etiquetas son $y_n = -1$ estaría acertando todos los casos, siendo por tanto la hipótesis que tiene un rendimiento mejor fuera de la muestra de entrenamiento.

Ejercicio 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:

$$\mathbb{P}(\mid \mathrm{E}_{in}(g) - \mathrm{E}_{out}(g) \mid \ > \varepsilon) < \delta$$

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

Solución

El algoritmo de aprendizaje que se utiliza para elegir la función g es indiferente, ya que este va a ir recorriendo el conjunto de hipótesis $\mathcal{H}$ (las cuáles están prefijadas de antes) y escogerá una hipótesis final g, la cual será la mejor de entre todas las hipótesis.

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

Solución

Aún habiendo escogido una función aleatoria como g, la desigualdad se seguiría cumpliendo, ya que la cota es aplicable sobre cualquier hipótesis de la clase de funciones $\mathcal{H}$. Esta cota, δ, viene dada por la siguiente expresión:

$$\delta = 2 \left|\mathcal{H}\right| e^{-2\varepsilon^2 N}$$

Como se puede ver, la cota no depende de la hipótesis final (g no aparece en ningún lado de la expresión), si no que depende de todas las hipótesis de $\mathcal{H}$, entre las cuales, eso sí, se encuentra g. Como g se elige del conjunto de hipótesis, si esta condición se cumple de forma genérica para cualquiera de las hipótesis, entonces también se cumplirá para g, sea cuál sea la forma en la que se escoge.

c) ¿Depende g del algoritmo usado?

Solución

La hipótesis final g depende del algoritmo utlizado, ya que cada uno va a recorrer el conjunto de hipótesis de una forma diferente. Por ejemplo, habrá algoritmos como el Gradiente Descendente que irán recorriendo iterativamente las hipótesis hasta dar con una que sea buena. Otros, como las Ecuaciones Normales darán directamente con la solución al resolver un sistema de ecuaciones.

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

Solución

La cota es laxa (pesimista). Esta cota se obtiene suponiendo que si se da que la diferencia de errores con la hipótesis g es mayor que un ε dado es porque alguna de las hipótesis del conjunto tiene una diferencia de errores superior a ese valor. La cota se obtiene mediante la desigualdad de Boole, ya que se supone que la probabilidad de la unión de los eventos (que para alguna hipótesis la diferencia supere el valor de ε) es menor o igual que la sumatoria de las probabilidades individuales, lo cuál resulta en la expresión $2 \left|\mathcal{H}\right| e^{-2\varepsilon^2 N}$. Este valor, sin embargo, tiene algunos problemas. Al acotar la unión de esta forma se supone que todas las hipótesis son disjuntas; es decir, que no tienen una interesección, cuando en la realidad puede ser que las hipótesis se solapen, y por tanto, el valor real de la unión de las probabilidades de los eventos sea mucho menor que el de la sumatoria.

Ejercicio 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.

Solución

Como tal, la desigualdad de Hoeffding sigue siendo aplicable a cada función de la clase de forma individual, pero para aplicarla sobre el conjunto de funciones necesitamos algo más.

Cada hipótesis $h_i \in \mathcal{H}$ se fija **antes** de generar el conjunto de datos, y de entre todas las funciones de la clase, el algoritmo de aprendizaje escoge aquella función g (hipótesis final) que sea la mejor **una vez generados los datos**, no antes, haciendo imposible además modificar h_i, ya que si no, no se podría probar la desigualdad de Hoeffding. Por tanto, como esa función g es una de las $h_1, h_2, \ldots, h_M$ funciones de la clase, queremos que la probabilidad dada por:

$$\mathbb{P}(\mathcal{D} : \mid \mathrm{E}_{in}(h) - \mathrm{E}_{out}(h) \mid \; > \varepsilon)$$

esté acotada por una expresión que tenga en cuenta todos los elementos de la clase $\mathcal{H}$ y no solo uno, que es lo que pasaba ahora. Con esto en mente, es necesario modificar la desigualdad de Hoeffding para considerar la función g escogida y que funcione para los casos en los que $|\mathcal{H}|$ sea mayor que 1 pero finito.

Para empezar, la expresión de la parte izquierda, una vez escogida la hipótesis final g, quedaría de la siguiente forma:

$$\mathbb{P}(\mathcal{D} : \mid \mathrm{E}_{in}(g) - \mathrm{E}_{out}(g) \mid \; > \varepsilon) \tag{2}$$

Ahora lo que queremos hacer es encontrar una cota para la probabilidad de (2) que tenga en cuenta todas las posibles funciones de la clase, y que además no dependa de la g escogida (puede ser cualquiera dentro de la clase).

Sabemos mediante la teoría de probabilidad que si ha sucedido que $\mid \mathrm{E}_{in}(g) - \mathrm{E}_{out}(g) \mid \; > \varepsilon$ (evento A) es porque se ha dado que para alguna otra función $h_i \in \mathcal{H}$, ha sucedido que $\mid \mathrm{E}_{in}(h_i) - \mathrm{E}_{out}(h_i) \mid \; > \varepsilon$ (evento B). Esto en probabilidad y en lógica se conoce como implicación, lo cuál se puede expresar como $A \Rightarrow B$. Se sabe que si $A \Rightarrow B$ es que $\mathbb{P}(A) \leq \mathbb{P}(B)$.

Como queremos acotar la probabilidad de nuestro evento A, podemos utilizar la desigualdad de Boole, la cuál nos dice que para un conjunto finito de eventos, la probabilidad de que al menos uno suceda (en este caso, que algún $\mathbb{P}(\mathcal{D} : \mid \mathrm{E}_{in}(h_i) - \mathrm{E}_{out}(h_i) \mid \; > \varepsilon)$) es menor o igual que la suma de las probabilidades de los eventos

(debido a que algunos eventos pueden no ser disjuntos). Dicho de otra forma:

$$\mathbb{P}\left(\bigcup_{h_i \in \mathcal{H}} \mathbb{P}(\mathcal{D} : \mid \mathrm{E}_{in}(h_i) - \mathrm{E}_{out}(h_i) \mid \; > \varepsilon)\right) \leq \sum_{i=1}^{|\mathcal{H}|} \mathbb{P}(\mathcal{D} : \mid \mathrm{E}_{in}(h_i) - \mathrm{E}_{out}(h_i) \mid \; > \varepsilon) \leq 2 \, |\mathcal{H}| \, e^{-2\varepsilon^2 N} \tag{3}$$

Así que, combinando las expresiones (2) y (3), obtenemos que:

$$\mathbb{P}(\mathcal{D} : \mid \mathrm{E}_{in}(g) - \mathrm{E}_{out}(g) \mid \; > \varepsilon) \leq 2 \, |\mathcal{H}| \, e^{-2\varepsilon^2 N} \tag{4}$$

Esta expresión ya sí que puede ser aplicada para clases con una o más hipótesis, siempre y cuando el número de éstas sea finito, ya que se tiene en cuenta la cardinalidad de la clase de funciones $\mathcal{H}$.

Ejercicio 8

Si queremos mostrar que k^* es un punto de ruptura para una clase de funciones $\mathcal{H}$ cuáles 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").

Solución

No nos sirve, ya que el punto de ruptura, por definición, es justamente un conjunto de puntos de tamaño k^* que no puede separar, y aquí se está planteando la situación contraria.

b) Mostrar que $\mathcal{H}$ puede separar cualquier conjunto de k^* puntos.

Solución

De nuevo, esto tampoco nos sirve, ya que el punto de ruptura es para demostrar que hay un conjunto de puntos de tamaño k^* que $\mathcal{H}$ no puede separar. Si se intenta demostrar que se puede separar cualquier conjunto de k^* puntos, entonces se está haciendo justo lo contrario.

c) Mostrar un conjunto de k^* puntos $x_1, \ldots, x_{k^*}$ que $\mathcal{H}$ no puede separar.

Solución

De nuevo, tampoco nos serviría, ya que a lo mejor un conjunto de k^* puntos no puede ser separado, pero existe otra disposición de k^* puntos (configurados de otra forma en el espacio) en el que se puedan conseguir todas las posibles dicotomías, y por tanto, ese conjunto sería separable.

d) Mostrar que $\mathcal{H}$ no puede separar ningún conjunto de k^* puntos.

Solución

Esto sí que nos serviría, ya que si se demuestra que ningún conjunto de k^* puntos se puede separar ("shatter") es que en ninguna disposición de puntos se pueden separar las 2^{k^*} posibles dicotomías, con lo cuál se obtiene que para cualquier conjunto de k^* puntos, $m_{\mathcal{H}}(k^*) < 2^{k^*}$.

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

Solución

De nuevo, no nos serviría. Que k^* sea un punto de ruptura no significa que, a partir de entonces, $\forall k > k^*$ el número máximo de dicotomías que se puedan separar satisfactoriamente sea 2^{k^*}, si no que el número máximo de dicotomías separables viene dado por $m_{\mathcal{H}}(k) < 2^k$. Puede haber más de 2^{k^*} dicotomías separables, pero nunca 2^k.

Ejercicio 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?

Solución

Para obtener el tamaño de la muestra mínimo con la que se obtenga una buena generalización podemos calcular la complejidad de la muestra, la cuál depende de

ε y δ. La expresión es la siguiente:

$$N \geq \frac{8}{\varepsilon^2} \ln \left(\frac{4\big((2N)^{d_{\text{VC}}} + 1\big)}{\delta} \right) \tag{5}$$

Ahora hay que fijar los valors de los parámetros, sabiendo que $\varepsilon = 0.05$, $\delta = 0.05$ y $d_{VC} = 10$. Como la resolución se tiene que ir haciendo de forma iterativa, se va a fijar que, inicialmente, $N = 1000$.

Antes de empezar con la resolución iterativa, vamos a tomar en cuenta algunas consideraciones. Es importante destacar que la expresión de la parte izquierda resultante se va a redondear para evitar tener decimales. La resolución se realizará escogiendo un N e intentando ver si con ese valor se cumple la desigualdad. En caso de que se cumpla, se parará. En caso contrario se escogerá como nuevo N el valor obtenido en la parte derecha y se repetirá el cálculo. Una vez dicho esto, veamos el proceso:

$$1000 \geq \frac{8}{0.05^2} \ln \left(\frac{4\big((2 \cdot 1000)^{10} + 1\big)}{0.05} \right) \Rightarrow 1000 \ngeq 257251$$

$$257251 \geq \frac{8}{0.05^2} \ln \left(\frac{4\big((2 \cdot 257251)^{10} + 1\big)}{0.05} \right) \Rightarrow 257251 \ngeq 434853$$

$$434853 \geq \frac{8}{0.05^2} \ln \left(\frac{4\big((2 \cdot 434853)^{10} + 1\big)}{0.05} \right) \Rightarrow 434853 \ngeq 451651$$

$$451651 \geq \frac{8}{0.05^2} \ln \left(\frac{4\big((2 \cdot 451651)^{10} + 1\big)}{0.05} \right) \Rightarrow 451651 \ngeq 452864$$

$$452864 \geq \frac{8}{0.05^2} \ln \left(\frac{4\big((2 \cdot 452864)^{10} + 1\big)}{0.05} \right) \Rightarrow 452864 \ngeq 452950$$

$$452950 \geq \frac{8}{0.05^2} \ln \left(\frac{4\big((2 \cdot 452950)^{10} + 1\big)}{0.05} \right) \Rightarrow 452950 \ngeq 452956$$

$$452956 \geq \frac{8}{0.05^2} \ln \left(\frac{4\big((2 \cdot 452956)^{10} + 1\big)}{0.05} \right) \Rightarrow 452956 \ngeq 452957$$

$$452957 \geq \frac{8}{0.05^2} \ln\left(\frac{4\big((2 \cdot 452957)^{10} + 1\big)}{0.05}\right) \Rightarrow 452957 \geq 452957$$

Con el resultado obtenido, podemos afirmar con un 95 % de confianza que con una muestra de tamaño $N = 452957$ tendremos un error de generalización ε que como mucho será 0.05.

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.

Solución

Al aplicar ERM

Referencias

[1] Texto referencia
`https://url.referencia.com`