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 alta probabilidad. En la expresión mostrada en (1), se puede ver una probabilidad de que algo malo suceda, es decir, 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 esa probabilidad de que algo malo pase sea pequeña. Por tanto, la probabilidad de que la diferencia sea menor que ε vendría dada por:

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

  la cuál sí que tendría un valor muy alto, siendo por tanto más "probable" que esa diferencia sea menor que ε.

- 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 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). Para obtener una cota 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 $\mathcal{H}$.

Referencias

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