Clasificadores bayesianos. El algoritmo Naïve Bayes

Constantino Malagón Luque

14 de mayo de 2003

Resumen

Este trabajo es una introducción al aprendizaje automático basado en el algoritmo Naïve Bayes.

1 Introducción

Las redes bayesianas, junto con los árboles de decisión y las redes neuronales artificiales, han sido los tres métodos más usados en aprendizaje automático durante estos últimos años en tareas como la clasificación de documentos o filtros de mensajes de correo electrónico. Es un método importante no sólo porque ofrece un análisis cualitativo de las atributos y valores que pueden intervenir en el problema, sino porque da cuenta también de la importancia cuantitativa de esos atributos. En el aspecto cualitativo podemos representar cómo se relacionan esos atributos ya sea en una forma causal, o señalando simplemente de la correlación que existe entre esas variables (o atributos). Cuantitativamente (y ésta es la gran aportación de los métodos bayesianos), da una medida probabilística de la importancia de esas variables en el problema (y por lo tanto una probabilidad explícita de las hipótesis que se formulan). Esta es quizá una de las diferencias fundamentales que ofecen las redes bayesianas con respecto a otros métodos -como puedan ser los árboles de decisión y las redes neuronales-, que no dan una medida cuantitativa de esa clasificación. Además de estas consideraciones, el aprendizaje basado en redes bayesianas es especialmente adecuado en ciertas tareas como puede ser la clasificación de textos, siendo incluso más eficiente que los otros métodos ya reseñados, y ofrece una medida para el estudio y comprensión de éstos otros métodos [1].
Entre las características que poseen los métodos bayesianos en tareas de aprendizaje se pueden resaltar las siguientes:

- Cada ejemplo observado va a modificar la probabilidad de que la hipótesis formulada sea correcta (aumentándola o disminuyéndola). Es decir, una hipótesis que no concuerda con un conjunto de ejemplos más o menos grande no es desechada por completo sino que lo que harán será disminuir esa probabilidad estimada para la hipótesis.

- Estos métodos son robustos al posible ruido presentes en los ejemplos de entrenamiento y a la posibilidad de tener entre esos ejemplos de entrenamiento datos incompletos o posiblemente erróneos.
- Los métodos bayesianos permiten tener en cuenta en la predicción de la hipótesis el conocimiento a prior o conocimiento del dominio en forma de probabilidades. El problema puede surgir al tener que estimar ese conocimiento estadístico sin disponer de datos suficientes. Esta dificultad ha sido estudiada por Kanehmann y Tversky [3], que analizaron los sesgos que se producen en los sujetos en la estimación subjetiva de las probabilidades de un suceso.

2 Clasificación de patrones

Cualquier sistema de clasificación de patrones se basa en lo siguiente: dado un conjunto de datos (que dividiremos en dos conjuntos de entrenamiento y de test) representados por pares <atributo, valor>, el problema consiste en encontrar una función f(x) (llamada hipótesis) que clasifique dichos ejemplos.
La idea de usar el teorema de Bayes en cualquier problema de aprendizaje automático (en especial los de clasificación) es que podemos estimar las probabilidades a posteriori de cualquier hipótesis consistente con el conjuntro de datos de entrenamiento para así escoger la hipótesis más probable. Para estimar estas probabilidades se han propuesto numerosos algoritmos, entre los que cabe detacar el algoritmo Naïve Bayes.

2.1 Clasificador basado en el algoritmo Naïve Bayes

Dado un ejemplo x representado por k valores el clasificador naïve Bayes se basa en encontrar la hipótesis más probable que describa a ese ejemplo. Si la descripción de ese ejemplo viene dada por los valores $< a_1, a_2, .., a_n >$, la hipótesis más probable será aquella que cumpla:

$$v_{MAP} = argmax_{v_j \in V} P(v_j | a_1, ...a_n)$$

es decir, la probabilidad de que conocidos los valores que describen a ese ejemplo, éste pertenezcan a la clase v_j (donde v_j es el valor de la función de clasificación $f(x)$ en el conjunto finito V). Por el teorema de Bayes:

$$v_{MAP} = argmax_{v_j \in V} \frac{P(a_1, ...a_n | v_j) p(v_j)}{P(a_1, ...a_n)} = argmax_{v_j \in V} P(a_1, ...a_n | v_j) p(v_j)$$

Podemos estimar $P(v_j)$ contando las veces que aparece el ejemplo v_j en el conjunto de entrenamiento y dividéndolo por el número total de ejemplos que forman este conjunto. Para estimar el término $P(a_1, ...a_n | v_j)$, es decir, las veces en que para cada categoría aparecen los valores del ejemplo x, debo recorrer todo el conjunto de entrenamiento. Este cálculo resulta impracticable para un número suficientemente grande de ejemplos por lo que se hace necesario siplificar la expresión. Para ello se recurre a la hipótesis de independencia condicional con el objeto de poder factorizar la probabilidad. Esta hipóteis dice lo siguiente:

Los valores a_j que describen un atributo de un ejemplo cualquiera x son independientes entre sí conocido el valor de la categoría a la que pertenecen
Así la probabilidad de observar la conjunción de atributos a_j dada una categoría a la que pertenecen es justamente el producto de las probabilidades de cada valor por separado: $P(a_1, ...a_n|v_j) = \prod_i P(a_i|v_j)$

Referencias

[1] Mitchell, Tom, "Machine Learning", Ed. McGraw-Hill (1997).