UNIVERSIDAD ICESI
Analítica de Datos

Taller de Naïve Bayes

Objetivo

El propósito de este taller es entender las bases teóricas del modelo de clasificación Naïve Bayes.

Indicaciones:

Vamos a tratar el problema de clasificación de los clientes potenciales de un negocio dado, a partir del desarrollo de un modelo de Naïve Bayes. Pero para poder lograrlo, vamos a empezar con el entendimiento de las bases teóricas de la probabilidad (aunque de manera simplificada).

En Excel, vamos a analizar las relaciones que existen entre la variable objetivo (con dos valores posibles "cliente potencial" y "no interesado") y dos variables predictivas, el género (hombre|mujer) y estudiante (no|si).

Parte 1. Relación entre género y categoría de cliente

1.1 Descargue el archivo de Excel "**01 - Taller NaiveBayes.xlsx**" y determine los valores de las probabilidades establecidas describiendo la relación entre género y categoría de clasificación.
1.2 ¿Si llega un cliente hombre, cuál es la probabilidad de que se convierta en un cliente? ¿Y si se trata de una mujer?

Parte 2. Relación entre estudiante y categoría de cliente

2.1 Olvídese que dispone de la información acerca de la relación entre la variable género y considere en cambio la información que describe la relación entre estudiante y categoría de clasificación. Determine los valores de probabilidades pedidos.
2.2 ¿Si llega un cliente estudiante, cuál es la probabilidad de que se convierta en un cliente? ¿Y si se trate de un no estudiante?

Parte 3. Relación entre estudiante y categoría de cliente

Ahora considere que tiene la información que relaciona género y clase de cliente, por un lado, y estudiante y clase de cliente, por otro lado.

La regla de Bayes establece lo siguiente:

Probabilidad posterior

$$P(Clase = c|Genero = x, Estudiante = y) = \frac{P(Clase = c, Genero = x, Estudiante = y)}{P(Genero = x, Estudiante = y)}$$

Para poder calcular la probabilidad posterior, necesitamos conocer la distribución de probabilidad conjunta de las variables predictivas con la variable objetivo. A partir de la regla de la cadena podemos establecer lo siguiente:

$$P(Clase = c, Genero = x, Estudiante = y) = P(Clase = c) * P(Genero = x|Clase = c) * P(Estudiante = y|Genero = x, Clase = c)$$

Si se conocen las probabilidades relacionando las variables predictivas y la variable categórica, pero no se conoce la relación entre las variables predictivas género y estudiante, no se puede despejar el último término.

Naïve Bayes propone una suposición muy fuerte, pero que permite eliminar esta restricción: se supone ciegamente que las variables predictivas son independientes con respecto a la categoría de clasificación. Esto implica que:

$$P(Clase = c, Genero = x, Estudiante = y) = P(Clase = c) * P(Genero = x|Clase = c) * P(Estudiante = y|Clase = c)$$

De esta manera llegamos a la regla de cálculo de probabilidades de Naïve Bayes:

Probabilidad posterior — Verosimilitudes — A priori

$$P(Clase = c|Genero = x, Estudiante = y) = \frac{P(Genero = x|Clase = c) * P(Genero = x|Clase = c) * P(Clase = c)}{P(Genero = x, Estudiante = y)}$$

Cabe anotar que el denominador se calcula a partir de los numeradores para todas las categorías de la clase:

$$P(Genero = x, Estudiante = y) = \sum_{c_i \in Clase} P(Genero = x|Clase = c_i) * P(Estudiante = y|Clase = c_i) * P(Clase = c_i)$$

En la fórmula de Naïve Bayes, los datos que tenemos nos permiten calcular numéricamente todos los términos, para poder establecer para un caso dado (un género dado y un valor de estudiante) las probabilidades de todos los valores de la clase.

Pero, si sólo nos interesa la clasificación, cuya regla establecería que la categoría de cliente sería la que tenga un valor máximo de probabilidad, entonces el denominador no es necesario.

3.1 Utilizando Naïve Bayes establezca las probabilidades de los dos valores de clase para una persona mujer, no estudiante.
3.2 Utilizando Naïve Bayes establezca las probabilidades de los dos valores de clase para una persona hombre, estudiante.

Parte 4. Relación entre edad y categoría de cliente

Se tiene ahora una nueva información acerca de la relación entre la edad y la categoría de cliente. Se sabe que los clientes potenciales tienen una media de 40 años, con una desviación estándar de 6 años, mientras que los no interesados tienen una media de 20 años, con una desviación estándar de 3 años. Utilizar la función “DISTR.NORM.N” de Excel para obtener las densidades de probabilidades condicionales entre la edad y la categoría de cliente para la edad dada.

4.1 Utilizando Naïve Bayes establezca las probabilidades de los dos valores de clase para una persona mujer, no estudiante de 32 años.
4.2 Utilizando Naïve Bayes establezca las probabilidades de los dos valores de clase para una persona hombre, estudiante de 32 años
4.3 Utilizando Naïve Bayes establezca las probabilidades de los dos valores de clase para una persona hombre, estudiante de 35 años