

Almacenes y Minería de Datos

Proyecto Final

Entregable 1

Adolfo Marín Arriaga
Juan Carlos López López
Luis Rodrigo Rojo Morales

15 de octubre de 2016

Algoritmo 1: Naive Bayes

Objetivo

El objetivo de Naive Bayes o Clasificador bayesiano ingenuo es la construcción de clasificadores en base a datos que ya se conocen por ejemplo clasificar si una persona es hombre o mujer basándose en los datos que se tienen de su altura, peso y tamaño del pie. Se usa basándose en el teorema de Bayes y asume que las variables que se usan para predecir son independientes

Descripción

Lo que tenemos es un objeto que es lo que se quiere clasificar, las variables que se van a usar para predecir que son los datos que ya conocemos y distintas clases a donde podría ir cada objeto después de la clasificación. Se usa la regla de Bayes que dice que si se tienen 2 eventos A y B se saca la probabilidad de A dado B como:

$$P(A | B) = \frac{P(B | A) P(A)}{P(B)}$$

- $P(A | B)$ Es la probabilidad de que se cumpla A dado B, es la probabilidad a posteriori.
- $P(B | A)$ Es la probabilidad de B dado que se cumple A.
- $P(A)$ y $P(B)$ Son las probabilidades que se cumplan A y B respectivamente, estas son la probabilidad a priori

Para el clasificador bayesiano ingenuo hay distintas clases en las que puede caer un objeto después de ser clasificado y distintas variables predictivas, entonces si tenemos un conjunto de variables predictivas $X = \{x_1, \dots, x_n\}$ y m clases para la clasificación C_1, \dots, C_m , el clasificador dice que la probabilidad de que un objeto pertenezca a la clase C_k se calcula como:

$$p(C_k | X) = \frac{1}{Z} p(C_k) \prod_{i=1}^n p(x_i | C_k)$$

- X es el conjunto de variables predictivas $X = \{x_1, \dots, x_n\}$.
- C_k es una clase de C_1, \dots, C_m
- Z es igual a un factor que depende solo de x_1, \dots, x_n , es una constante si sus valores son conocidos

Ejemplo:

Un ejemplo es si queremos clasificar frutas, las clases son platano, naranja y otras, se hizo un conjunto de entrenamiento con las variables de longitud, sabor y color y nos dio que:

Fruta	Alargada	No Alargada	Dulce	No Dulce	Amarilla	No Amarilla	Total
Platano	400	100	350	150	450	50	500
Naranja	0	300	150	150	300	0	300
Otras	100	100	150	50	50	150	200
Total	500	500	650	350	800	200	1000

Ahora queremos predecir en que clase cae una nueva fruta que llega y es amarilla, alargada y dulce. Calculamos las probabilidades:

- $P(\text{platano}) = 500/1000 = 0.5$
- $P(\text{platano} | \text{alargada}) = \frac{P(\text{alargada}|\text{platano}) P(\text{platano})}{P(\text{alargada})} = \frac{400/500 \cdot 500/1000}{500/1000} = 0.8$
- $P(\text{platano} | \text{dulce}) = \frac{P(\text{dulce}|\text{platano}) P(\text{platano})}{P(\text{dulce})} = \frac{350/500 \cdot 500/1000}{650/1000} = 0.5384$
- $P(\text{platano} | \text{amarilla}) = \frac{P(\text{amarilla}|\text{platano}) P(\text{platano})}{P(\text{amarilla})} = \frac{450/500 \cdot 500/1000}{800/1000} = 0.5625$
- $P(\text{naranja}) = 300/1000 = 0.3$
- $P(\text{naranja} | \text{alargada}) = \frac{P(\text{alargada}|\text{naranja}) P(\text{naranja})}{P(\text{alargada})} = \frac{0/300 \cdot 300/1000}{500/1000} = 0$
- $P(\text{naranja} | \text{dulce}) = \frac{P(\text{dulce}|\text{naranja}) P(\text{naranja})}{P(\text{dulce})} = \frac{150/300 \cdot 300/1000}{650/1000} = 0.2307$
- $P(\text{naranja} | \text{amarilla}) = \frac{P(\text{amarilla}|\text{naranja}) P(\text{naranja})}{P(\text{amarilla})} = \frac{300/300 \cdot 300/1000}{800/1000} = 0.375$
- $P(\text{otras}) = 200/1000 = 0.2$
- $P(\text{otras} | \text{alargada}) = \frac{P(\text{alargada}|\text{otras}) P(\text{otras})}{P(\text{alargada})} = \frac{100/200 \cdot 200/1000}{500/1000} = 0.2$
- $P(\text{otras} | \text{dulce}) = \frac{P(\text{dulce}|\text{otras}) P(\text{otras})}{P(\text{dulce})} = \frac{150/200 \cdot 200/1000}{650/1000} = 0.2307$
- $P(\text{otras} | \text{amarilla}) = \frac{P(\text{amarilla}|\text{otras}) P(\text{otras})}{P(\text{amarilla})} = \frac{50/200 \cdot 200/1000}{800/1000} = 0.0625$

Ahora para el clasificador con su fórmula:

- $p(\text{platano}|\text{alargada}, \text{dulce}, \text{amarilla}) = \frac{1}{Z} p(\text{platano}) p(\text{platano}|\text{alargada}) p(\text{platano}|\text{dulce}) p(\text{platano}|\text{amarilla})$
 $= \frac{1}{Z} 0.5 \cdot 0.8 \cdot 0.5384 \cdot 0.5625 = \frac{1}{Z} 0.121147$
- $p(\text{naranja}|\text{alargada}, \text{dulce}, \text{amarilla}) = \frac{1}{Z} p(\text{naranja}) p(\text{naranja}|\text{alargada}) p(\text{naranja}|\text{dulce}) p(\text{naranja}|\text{amarilla})$
 $= \frac{1}{Z} 0.5 \cdot 0 \cdot 0.2307 \cdot 0.375 = \frac{1}{Z} 0$
- $p(\text{otras}|\text{alargada}, \text{dulce}, \text{amarilla}) = \frac{1}{Z} p(\text{otras}) p(\text{otras}|\text{alargada}) p(\text{otras}|\text{dulce}) p(\text{otras}|\text{amarilla})$
 $= \frac{1}{Z} 0.2 \cdot 0.2 \cdot 0.2307 \cdot 0.0625 = \frac{1}{Z} 0.00057675$

Conocemos la probabilidad de que lo que recibamos sea dulce, alargado y amarillo entonces tomamos Z como:

$$\blacksquare p(\text{alargada}) p(\text{dulce}) p(\text{amarilla}) = \frac{500}{1000} \cdot \frac{650}{1000} \cdot \frac{800}{1000} = 0.26$$

Entonces reemplazando Z

- $p(\text{platano}|\text{alargada}, \text{dulce}, \text{amarilla}) = \frac{0.121147}{0.26} = 0.46595$
- $p(\text{naranja}|\text{alargada}, \text{dulce}, \text{amarilla}) = \frac{0}{0.26} = 0$
- $p(\text{otras}|\text{alargada}, \text{dulce}, \text{amarilla}) = \frac{0.00057675}{0.26} = 0.002218$

La probabilidad mas grande es la de que sea platano, entonces si llega una fruta que sea alargada, dulce y amarilla la clasificamos como que va a ser un platano.

Bibliografía

https://en.wikipedia.org/wiki/Naive_Bayes_classifier

https://en.wikipedia.org/wiki/Bayes%27_theorem