



Asesoría 1

Martinez Santiago Victor



Pregunta/Comentario 1

Por ejemplo en el 1 dice "defina", de qué manera se puede entregar ¿puede ser con diagramas o con pseudocódigo?

1. Sea T un documento de texto. Se desea encontrar la palabra más frecuente y la menos frecuente. Defina:
 - (a) La función *Map*
 - (b) La función *Reduce*



Pregunta/Comentario 2

- Algunos ejemplos de Hadoop mapreduce, el día que lo vimos hubo varios problemas y hay cositas que simplemente me gustaría ver como ejemplo.
- Al final lo termine haciendo en consola y no fue lo mas practico
jaja



Contexto

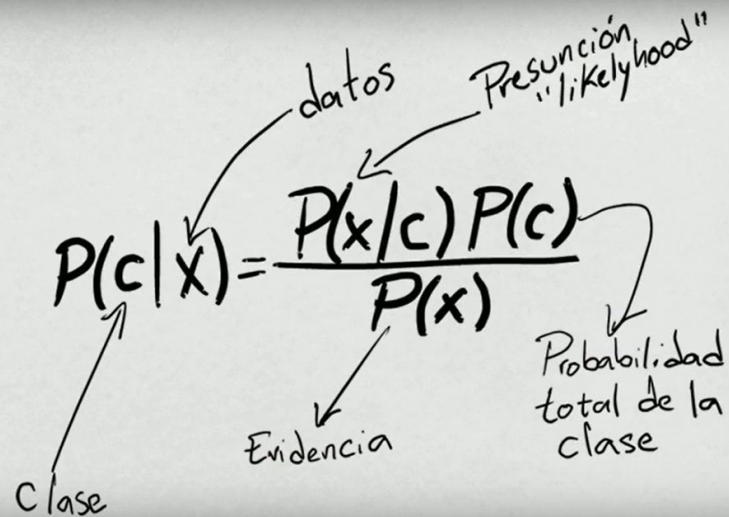
¿Buenas tardes, a mi me causa duda lo referente a Naive Bayes?

3. Sea X una matrix de $m \times n$, donde m es el número de ejemplos en el conjunto de datos, cada uno descrito por n características, y y un vector binario con las etiquetas de clase de cada ejemplo. Se desea construir el clasificador de Naïve Bayes bajo el paradigma MapReduce. Realice los siguiente:
 - (a) Defina la función *Map*
 - (b) Defina la función *Reduce*
 - (c) Usando la librería MRJob, implemente la versión distribuida del Naïve Bayes y compárela usando tres conjuntos de datos con la versión de scikit-learn.
 - (d) Implemente el Naïve Bayes en PySpark y compare usando tres conjuntos de datos con la versión de scikit-learn y la implementada en MapReduce.



TEOREMA DE BAYES

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



A handwritten diagram of Bayes' theorem on a piece of paper. The formula is $P(c|x) = \frac{P(x|c)P(c)}{P(x)}$. Annotations include: an arrow from 'Clase' to $P(c|x)$; an arrow from 'datos' to $P(x|c)$; an arrow from 'Presunción "likelyhood"' to $P(c)$; an arrow from 'Evidencia' to $P(x)$; and an arrow from 'Probabilidad total de la clase' to the denominator $P(x)$.

$$P(c|x) = \frac{P(x|c)P(c)}{P(x)}$$

Clase

datos

Presunción "likelyhood"

Evidencia

Probabilidad total de la clase

Problema: Dada la información climática, decidir si es un buen día para jugar golf si Cielo = Lluvioso, Temperatura = Templado, Humedad = Normal, Viento = Sí

Cielo	Temperatura	Humedad	Viento	Se jugó
Lluvia	Calor	Alta	No	No
Lluvia	Calor	Alta	Sí	No
Nublado	Calor	Alta	No	Sí
Soleado	Templado	Alta	No	Sí
Soleado	Frío	Normal	No	Sí
Soleado	Frío	Normal	Sí	No
Nublado	Frío	Normal	Sí	Sí
Lluvia	Templado	Alta	No	No
Lluvia	Frío	Normal	No	Sí
Soleado	Templado	Normal	No	Sí
Lluvia	Templado	Normal	Sí	Sí
Nublado	Templado	Alta	Sí	Sí
Nublado	Calor	Normal	No	Sí
Soleado	Templado	Alta	Sí	No

$$P(Sí) = \frac{9}{14}$$

$$P(No) = \frac{5}{14}$$



Tablas de frecuencia

		Jugar golf	
		Sí	No
Cielo	Soleado	3 (3/9)	2 (2/5)
	Nublado	4 (4/9)	0 (0/5)
	Lluvioso	2 (2/9)	3 (3/5)

		Jugar golf	
		Sí	No
Temperatura	Calor	2 (2/9)	2 (2/5)
	Templado	4 (4/9)	2 (2/5)
	Frío	3 (3/9)	1 (1/5)

		Jugar golf	
		Sí	No
Humedad	Soleado	3 (3/9)	4 (4/5)
	Lluvioso	6 (6/9)	1 (1/5)

		Jugar golf	
		Sí	No
Viento	Calor	6 (6/9)	2 (2/5)
	Frío	3 (3/9)	3 (3/5)

Tablas de frecuencia

		Jugar golf	
		Sí	No
Cielo	Soleado	3 (3/9)	2 (2/5)
	Nublado	4 (4/9)	0 (0/5)
	Lluvioso	2 (2/9)	3 (3/5)

		Jugar golf	
		Sí	No
Temperatura	Calor	2 (2/9)	2 (2/5)
	Templado	4 (4/9)	2 (2/5)
	Frío	3 (3/9)	1 (1/5)

		Jugar golf	
		Sí	No
Humedad	Soleado	3 (3/9)	4 (4/5)
	Lluvioso	6 (6/9)	1 (1/5)

		Jugar golf	
		Sí	No
Viento	Calor	6 (6/9)	2 (2/5)
	Frío	3 (3/9)	3 (3/5)

$$C = S_1$$

$$P(x|S_1) = P(\text{Cielo} = \text{Lluvia} | S_1) \cdot P(\text{Temp} = \text{Templado} | S_1)$$

$$\cdot P(\text{Humedad} | S_1) \cdot P(\text{Viento} = S_1 | S_1)$$

$$= \frac{2}{9} \cdot \frac{4}{9} \cdot \frac{6}{9} \cdot \frac{3}{9}$$

$$P(x|S_1) P(S_1) = \frac{2}{9} \cdot \frac{4}{9} \cdot \frac{6}{9} \cdot \frac{3}{9} \cdot \frac{9}{14} \approx 0.0141$$

Tablas de frecuencia

		Jugar golf	
		Sí	No
Cielo	Soleado	3 (3/9)	2 (2/5)
	Nublado	4 (4/9)	0 (0/5)
	Lluvioso	2 (2/9)	3 (3/5)

		Jugar golf	
		Sí	No
Temperatura	Calor	2 (2/9)	2 (2/5)
	Templado	4 (4/9)	2 (2/5)
	Frío	3 (3/9)	1 (1/5)

		Jugar golf	
		Sí	No
Humedad	Soleado	3 (3/9)	4 (4/5)
	Lluvioso	6 (6/9)	1 (1/5)

		Jugar golf	
		Sí	No
Viento	Calor	6 (6/9)	2 (2/5)
	Frío	3 (3/9)	3 (3/5)

$C=Sí$

$$P(x|Sí) = P(\text{Cielo}=Luvia|Sí) \cdot P(\text{Temp}=Templado|Sí) \cdot P(\text{Humedad}|Sí) \cdot P(\text{Viento}=Sí|Sí)$$

$$= \frac{2}{9} \cdot \frac{4}{9} \cdot \frac{6}{9} \cdot \frac{3}{9}$$


$$P(x|Sí)P(Sí) = \frac{2}{9} \cdot \frac{4}{9} \cdot \frac{6}{9} \cdot \frac{3}{9} \cdot \frac{9}{14} \approx 0.0141$$

$C=No$

$$P(x|No) = P(\text{Cielo}=Luvia|No) \cdot P(\text{Temp}=Templado|No) \cdot P(\text{Humedad}=Normal|No) \cdot P(\text{Viento}=Sí|No)$$

$$= \frac{3}{5} \cdot \frac{2}{5} \cdot \frac{1}{5} \cdot \frac{3}{5}$$

$$P(x|No)P(No) = \frac{3}{5} \cdot \frac{2}{5} \cdot \frac{1}{5} \cdot \frac{3}{5} \cdot \frac{5}{14} \approx 0.0102$$


$$P(x) = \sum_{i=0}^n P(C_i | x) P(C_i) = P(S_i' | x) P(S_i') + P(N_o | x) P(N_o)$$

$$= 0.0141 + 0.0102 = 0.0243$$

$$\therefore P(S_i' | x) = \frac{0.0141}{0.0243} \approx 0.5783$$

$$P(N_o | x) = \frac{0.0102}{0.0243} \approx 0.4216$$



OTROS COMENTARIOS



HINT:

La covarianza se puede escribir como

$$r = r_{xy} = \frac{n \sum x_i y_i - \sum x_i \sum y_i}{\sqrt{n \sum x_i^2 - (\sum x_i)^2} \sqrt{n \sum y_i^2 - (\sum y_i)^2}}.$$

La varianza se puede escribir como

$$\frac{1}{n-1} \left(\sum_{i=1}^n x_i y_i - n \bar{x} \bar{y} \right)$$