Naïve Bayes

DRA. CONSUELO VARINIA GARCÍA MENDOZA

Clasificador Bayesiano

- Clasificador probabilístico
- Dada la probabilidad de que una instancia pertenezca a una clase es posible clasificar nuevas instancias

$$p(y|X)$$

Instance	Probability of class 1	Probability of class 2
1	0.8	0.6
2	0.4	0.7
3	0.6	0.6

Teorema de Bayes

- Describe la probabilidad de un evento, basado en conocimiento a priori de condiciones que podrían estar relacionadas con el evento

$$p(y|X) = p(y)\frac{p(X|y)}{p(X)}$$

donde:

- $p(y|X)$ es la probabilidad condicional. Es la probabilidad de que una instancia pertenezca a la clase y dadas las características X
- $p(y)$ es la probabilidad a priori. La probabilidad de la clase y antes de considerar las características X
- $p(X|y)$ es la verosimilitud. Probabilidad de las características, cuando la instancia pertenece a la clase y
- $p(X)$ es la probabilidad marginal. La probabilidad de las características X

Ejemplo

Clasificación de manzanas de acuerdo a su tamaño

Instance	Features (X)	Class (y)
	Size	
1	big	pos
2	small	pos
3	small	pos
4	small	pos
5	small	neg
6	big	neg
7	big	neg
8	big	neg
9	small	pos
10	big	pos

Instance	Features (X)	Class (y)
1		pos
2		pos
3		pos
4		pos
5		neg
6		neg
7		neg
8		neg
9		pos
10		pos

- Teorema de Bayes

$$p(y|X) = p(y)\frac{p(X|y)}{p(X)}$$

- Probabilidad a priori $p(y)$
  - $p(pos) = \frac{N_{pos}}{N_{all}} = \frac{6}{10} = 0.6$
  - $p(neg) = \frac{N_{neg}}{N_{all}} = \frac{4}{10} = 0.4$

Instance	Features (X)	Class (y)
	Size	
1	big	
2	small	
3	small	
4	small	
5	small	
6	big	
7	big	
8	big	
9	small	
10	big	

- Teorema de Bayes

$$p(y|X) = p(y)\frac{p(X|y)}{p(X)}$$

- Probabilidad marginal $p(X)$
  - $p(big) = \frac{N_{big}}{N_{all}} = \frac{5}{10} = 0.5$
  - $p(small) = \frac{N_{smalll}}{N_{all}} = \frac{5}{10} = 0.5$

Instance	Features (X)	Class (y)
	Size	
1	big	pos
2	small	pos
3	small	pos
4	small	pos
5	small	neg
6	big	neg
7	big	neg
8	big	neg
9	small	pos
10	big	pos

- Teorema de Bayes

$$p(y|X) = p(y)\frac{p(X|y)}{p(X)}$$

- Verosimilitud $p(X|y)$
  - $p(big|pos) = \frac{N_{big \cap pos}}{N_{pos}} = \frac{2}{6} = 0.33$
  - $p(small|pos) = \frac{N_{small \cap pos}}{N_{pos}} = \frac{4}{6} = 0.66$
  - $p(big|neg) = \frac{N_{big \cap neg}}{N_{neg}} = \frac{3}{4} = 0.75$
  - $p(small|neg) = \frac{N_{small \cap neg}}{N_{neg}} = \frac{1}{4} = 0.25$

- $p(pos) = 0.6$
- $p(neg) = 0.4$
- $p(big) = 0.5$
- $p(small) = 0.5$
- $p(big|pos) = 0.33$
- $p(big|neg) = 0.75$
- $p(small|pos) = 0.67$
- $p(small|neg) = 0.25$

- Probabilidad condicional

$$p(y|X) = p(y)\frac{p(X|y)}{p(X)}$$

- $p(pos|big) = p(pos)\frac{p(big|pos)}{p(big)} = 0.6 \cdot \frac{0.33}{0.5} = 0.396$
- $p(neg|big) = p(neg)\frac{p(big|neg)}{p(big)} = 0.4 \cdot \frac{0.75}{0.5} = 0.6$
- $p(pos|small) = p(pos)\frac{p(small|pos)}{p(small)} = 0.6 \cdot \frac{0.67}{0.5} = 0.804$
- $p(neg|small) = p(neg)\frac{p(small|neg)}{p(small)} = 0.4 \cdot \frac{0.25}{0.5} = 0.2$

Teorema de Bayes

$$p(y|X) = p(y)\frac{p(X|y)}{p(X)}$$

- $p(pos|big) = p(pos)\frac{p(big|pos)}{p(big)} = 0.4$
- $p(neg|big) = p(neg)\frac{p(big|neg)}{p(big)} = 0.6$
- $p(pos|small) = p(pos)\frac{p(small|pos)}{p(small)} = 0.8$
- $p(neg|small) = p(neg)\frac{p(small|neg)}{p(small)} = 0.2$

$$p(y|X) = p(y)p(X|y)$$

- $p(pos|big) = 0.2$
- $p(neg|big) = 0.3$
- $p(pos|small) = 0.4$
- $p(neg|small) = 0.1$

Teorema de Bayes

$$p(y|X) = p(y)p(X|y)$$

Cuando las instancias tienen más de una característica, cada característica contribuye a la clasificación

Para considerar la contribución de cada característica aplicamos la probabilidad conjunta

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

El teorema de Bayes para n características a través de la probabilidad conjunta se expresa como

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

Para aplicar esta regla, debemos suponer que cada atributo es mutuamente independiente

Esta suposición no suele estar justificada, por lo que se denomina a este clasificador como **bayesiano ingenuo**

Ejemplo 2

	Instance	Features (X)		Class (y)
		Size	Color	
Train	1	big	green	pos
	2	small	red	pos
	3	small	red	pos
	4	small	red	pos
	5	small	red	neg
	6	big	red	neg
	7	big	green	neg
	8	big	green	neg
	9	small	green	pos
	10	big	red	pos
Test	11	big	yellow	?

- $p(pos) = 0.6$
- $p(neg) = 0.4$
- $p(big|pos) = 0.33$
- $p(big|neg) = 0.75$
- $p(small|pos) = 0.67$
- $p(small|neg) = 0.25$

- $p(red|pos) = \frac{N_{red \cap pos}}{N_{pos}} = 0.67$
- $p(green|pos) = \frac{N_{green \cap pos}}{N_{pos}} = 0.33$
- $p(red|neg) = \frac{N_{red \cap neg}}{N_{neg}} = 0.5$
- $p(green|neg) = \frac{N_{green \cap neg}}{N_{neg}} = 0.5$

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

- $p(pos|big, yellow) = p(pos) \cdot p(big|pos) \cdot p(yellow|pos)$
  $= 0.6 \cdot 0.33 \cdot 1 = 0.2$
- $p(neg|big, yellow) = p(neg) \cdot p(big|neg) \cdot p(yellow|neg)$
  $= 0.4 \cdot 0.75 \cdot 1 = 0.3$

$p(pos|big, yellow) < p(neg|big, yellow) \therefore$ la instancia 11 pertenece a la clase neg

Distribuciones de probabilidad

Se pueden hacer distintas implementaciones del clasificador Naïve considerando diversas distribuciones de probabilidad

- Bernoulli
- Multinomial
- Gaussiana

Distribución de Bernoulli

Una variable aleatoria Bernoulli es una variable aleatoria que sólo puede tomar dos valores posibles, normalmente 0 y 1

Esta variable aleatoria modela experimentos aleatorios que tienen dos posibles resultados, a veces denominados "éxito" y "fracaso“

La variante Bernoulli para el clasificador Naïve Bayes se utiliza cuando las características toman valores binarios o booleanos

- Género (Hombre = 1 o Mujer = 0)
- Tarjeta de crédito (Sí = 1 o No = 0)
- Vivienda propia (Sí = 1 o No = 0)

Distribución multinomial

La distribución multinomial se utiliza para encontrar probabilidades en experimentos en los que hay más de dos resultados

La variante multinomial para el clasificador Naïve Bayes se utiliza cuando las características describen conteos de frecuencia discretos

- Calificaciones de películas (Número de estrellas 1-5)
- Conteo de palabras (número de veces que aparece la palabra "bueno" en la reseña de un producto)

Distribución Gaussiana

La distribución normal o gaussiana es un tipo de distribución de probabilidad continua para una variable aleatoria de valor real

La variante gaussiana para el clasificador Naive Bayes se utiliza cuando las características son continuas

- Peso
- Altura
- Presión arterial

Clasificación de textos

Dado un grupo de documentos, un modelo debe predecir la categoría a la que pertenece cada documento

Dataset	Document	Words	Class
Training	1	Chinese Beijing Chinese	c
	2	Chinese Chinese Shangai	c
	3	Chinese Macao	c
	4	Tokyo Japan Chinese	j
Test	5	Chinese Chinese Chinese Tokyo Japan	?
	6	Tokyo Macao Shangai Beijin Okinawa	?

El teorema de Bayes para n características a través de la probabilidad conjunta se expresa como

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

Distribución multinomial

$$p(w_i|y_j) = \frac{\boldsymbol{count(w_i, y_j) + \alpha}}{\left(\sum_{w \in V} \boldsymbol{count(w, y_j)}\right) + |\boldsymbol{V}|}$$

Bayes y la distribución multinomial

$$p(y_j|X) = p(y_j)\prod_{i=1}^{n} p(w_i|y_j)$$

Distribución Multinomial

$$p(w_i|y_j) = \frac{\boldsymbol{count(w_i, y_j) + \alpha}}{\boldsymbol{(\sum_{w \in V} count(w, y_j)) + |V|}}$$

donde

$count(w_i, y_i)$: número de veces que aparece la palabra w_i en las instancias de entrenamiento que pertenecen a la clase y_j

α: suavizado de Laplace (Lapace smoothing)

$\sum_{w \in V} count(w, y_j)$: conteo de todas las palabras en las instancias de entrenamiento que pertenecen a la clase y_j no importa si se repiten

$|V|$: número total de palabras distintas que parecen en todas las instancias de entrenamiento

Dataset	Document	Words	Class
Training	1	Chinese Beijing Chinese	c
	2	Chinese Chinese Shangai	c
	3	Chinese Macao	c
	4	Tokyo Japan Chinese	j
Test	5	Chinese Chinese Chinese Tokyo Japan	?
	6	Tokyo Macao Shangai Beijin Okinawa	?

$\boldsymbol{p(y_j)}$

- $p(c) = \frac{N_c}{N_{all}} = \frac{3}{4}$
- $p(j) = \frac{N_j}{N_{all}} = \frac{1}{4}$

Dataset	Docu-ment	Words	Class
Training	1	Chinese Beijing Chinese	c
	2	Chinese Chinese Shangai	c
	3	Chinese Macao	c
	4	Tokyo Japan Chinese	j
Test	5	Chinese Chinese Chinese Tokyo Japan	?
	6	Tokyo Macao Shangai Beijin Okinawa	?

$$\boldsymbol{p(w_i|y_j)} = \frac{\boldsymbol{count(w_i,y_j)+\alpha}}{(\sum_{w\in V}\boldsymbol{count(w,y_j)})+|V|}$$

$$p(Chinese|c) = \frac{count(Chinese,c)+\alpha}{(\sum_{w\in V}\boldsymbol{count(w,c)})+6}$$

$$= \frac{5+1}{8+6} = \frac{6}{14} = \frac{3}{7}$$

$$p(Beijing|c) = \frac{count(Beijing,c)+\alpha}{(\sum_{w\in V}\boldsymbol{count(w,c)})+6} =$$

$$p(Shangai|c) = \frac{count(Shangai,c)+\alpha}{(\sum_{w\in V}\boldsymbol{count(w,c)})+6} =$$

$$p(Macao|\ ) = \frac{count(Macao,c)+\alpha}{(\sum_{w\in V}\boldsymbol{count(w,c)})+6} =$$

$$p(Tokyo|\ ) = \frac{count(Tokio,c)+\alpha}{(\sum_{w\in V}\boldsymbol{count(w,c)})+6} =$$

$$p(Japa\ |\ ) = \frac{count(Japan,c)+\alpha}{(\sum_{w\in V}\boldsymbol{count(w,c)})+6} =$$

$$p(w_i|y_j) = \frac{count(w_i,y_j)+\alpha}{(\sum_{w\in V} count(w,y_j))+|V|}$$

Dataset	Document	Words	Class
Training	1	Chinese Beijing Chinese	c
	2	Chinese Chinese Shangai	c
	3	Chinese Macao	c
	4	Tokyo Japan Chinese	j
Test	5	Chinese Chinese Chinese Tokyo Japan	?
	6	Tokyo Macao Shangai Beijin Okinawa	?

$$p(Chinese|j) = \frac{count(Chinese,j)+\alpha}{(\sum_{w\in V} count(w,j))+6} = \frac{1+1}{3+6} = \frac{2}{9}$$

$$p(Beijing|j) = \frac{count(Beijing,j)+\alpha}{(\sum_{w\in V} count(w,j))+6} = \frac{0+1}{3+6} = \frac{1}{9}$$

$$p(Shangai|j) = \frac{count(Shangai,j)+\alpha}{(\sum_{w\in V} count(w,j))+6} =$$

$$p(Macao|j) = \frac{count(Macao,j)+\alpha}{(\sum_{w\in V} count(w,j))+6} =$$

$$p(Tokyo|j) = \frac{count(Tokio,j)+\alpha}{(\sum_{w\in V} count(w,j))+6} =$$

$$p(Japan|j) = \frac{count(Japan,j)+\alpha}{(\sum_{w\in V} count(w,j))+6} =$$

Bayes y la distribución multinomial

$$p(y_j|d_k) = p(y_j)\prod_{i=1}^{n} p(w_i|y_j)$$

Dataset	Document	Words	Class
Training	1	Chinese Beijing Chinese	c
	2	Chinese Chinese Shangai	c
	3	Chinese Macao	c
	4	Tokyo Japan Chinese	j
Test	5	Chinese Chinese Chinese Tokyo Japan	?
	6	Tokyo Macao Shangai Beijin Okinawa	?

$p(c) = \frac{3}{4}$

$p(Chine \quad | \ ) = \frac{3}{7}$

$p(Beijing|c) = \frac{1}{7}$

$p(Shangai|c) = \frac{1}{7}$

$p(Macao|c) = \frac{1}{7}$

$p(Tokyo|c) = \frac{1}{14}$

$p(Japan|c) = \frac{1}{14}$

$p(j) = \frac{1}{4}$

$p(Chinese|j) = \frac{2}{9}$

$p(Beijing|j) = \frac{1}{9}$

$p(Shang \quad | \ ) = \frac{1}{9}$

$p(Mac \quad | \ ) = \frac{1}{9}$

$p(Toky \ | \ ) = \frac{2}{9}$

$p(Japa \ | \ ) = \frac{2}{9}$

$$p(c|d_5) = p(c)\cdot p(Chinese|c)\cdot p(Chinese|c)\cdot p(Chinese|c)\cdot p(Tokyo|c)\cdot p(Japan|c) = \frac{3}{4}\cdot\frac{3}{7}\cdot\frac{3}{7}\cdot\frac{3}{7}\cdot\frac{1}{14}\cdot\frac{1}{14} = 0.0003 = 3\times 10^{-4}$$

$$p(j|d_5) = p(j)\cdot p(Chinese|j)\cdot p(Chinese|j)\cdot p(Chinese|j)\cdot p(Tokyo|j)\cdot p(Japan| \ ) = \frac{1}{4}\cdot\frac{2}{9}\cdot\frac{2}{9}\cdot\frac{2}{9}\cdot\frac{2}{9}\cdot\frac{2}{9} = 0.0001 = 1\times 10^{-4}$$

$$p(c|d_5) > p(j|d_5) \therefore d_5 \in c$$

$p(c|d_5)$ =3×10^{-4} $p(j|d_5)$ =1×10^{-4}

Logaritmo es una función monótona, lo que significa que si $a > b$ entonces $\log a > \log b$

$$log(p(y_j|d_k)) = log\left(p(y_j)\prod_{i=1}^{n} p(x_i|y_j)\right)$$
$$log(p(y_j|d_k)) = log\, p(y_j) + \sum_{i=1}^{n} log\, p(x_i|y_j)$$

$log(p(c|d_5))$= -8.1

$log(p(j|d_5))$=-9.21

Bayes y la distribución multinomial

$$p(y_j|d_k) = p(y_j)\prod_{i=1}^{n} p(w_i|y_j)$$

$p(c) = \frac{3}{4}$

$p(Chinese|c) = \frac{3}{7}$

$p(Beijing|c) = \frac{1}{7}$

$p(Shangai|c) = \frac{1}{7}$

$p(Macao|c) = \frac{1}{7}$

$p(Tokyo|\ ) = \frac{1}{14}$

$p(Japan|\ ) = \frac{1}{14}$

$p(j) = \frac{1}{4}$

$p(Chinese|\ ) = \frac{2}{9}$

$p(Beijing|\ ) = \frac{1}{9}$

$p(Shangai|j) = \frac{1}{9}$

$p(Ma\quad|\ ) = \frac{1}{9}$

$p(Tokyo|\ ) = \frac{2}{9}$

$p(Japa\ |\ ) = \frac{2}{9}$

Dataset	Document	Words	Class
Training	1	Chinese Beijing Chinese	c
	2	Chinese Chinese Shangai	c
	3	Chinese Macao	c
	4	Tokyo Japan Chinese	j
Test	5	Chinese Chinese Chinese Tokyo Japan	c
	6	Tokyo Macao Shangai Beijin Okinawa	c

$$p(c|d_6) = p(c)\cdot p(Tokyo|c)\cdot p(Macao|c)\cdot p(Shangai|c)\cdot p(Beijing|c) = \frac{3}{4}\cdot\frac{1}{14}\cdot\frac{1}{7}\cdot\frac{1}{7}\cdot\frac{1}{7} = \frac{3}{19208} = 0.00016 = 1.6\times 10^{-4}$$

$$p(j|d_6) = p(j)\cdot p(Tokyo|\ )\quad p(Macao|j)\cdot p(Shangai|j)p(Beijing|j) = \frac{1}{4}\cdot\frac{2}{9}\cdot\frac{1}{9}\cdot\frac{1}{9}\cdot\frac{1}{9} = \frac{1}{13122} = 0.000076 = 7.6\times 10^{-5}$$

$$p(c|d_6) > p(j|d_6) \therefore d_6 \in c$$

Dataset	Document	Words	Class
Training	1	Chinese Beijing Chinese	c
	2	Chinese Chinese Shangai	c
	3	Chinese Macao	c
	4	Tokyo Japan Chinese	j
Test	5	Chinese Chinese Chinese Tokyo Japan	c
	6	Tokyo Macao Shangai Beijin Okinawa	c

$p(c|d_5) = p(c) \cdot p(Chinese|c) \cdot p(Chinese|c) \cdot p(Chinese|c) \cdot p(Tokyo|c) \cdot p(Japan|c) = 3 \times 10^{-4}$ $log(p(c|d_5)) = -8.1$

$p(j|d_5) = p(j) \cdot p(Chinese|j) \cdot p(Chinese|j) \cdot p(Chinese|j) \cdot p(Tokyo|j) \cdot p(Japan| \ ) = 1 \times 10^{-4}$ $log(p(j|d_5)) =$-8.9

$p(c|d_5) > p(j|d_5)$ y $log(p(c|d_5)) > log(p(j|d_5))$

$p(c|d_6) = p(c) \cdot p(Tokyo| \ ) \ \ p(Macao|c) \cdot p(Shangai|c) \cdot p(Beijing|c) = 1.6 \times 10^{-4}$ $log(p(c|d_6)) = -8.76$

$p(j|d_6) = p(j) \cdot p(Tokyo|j) \cdot p(Macao|j) \cdot p(Shangai|j) \cdot p(Beijing|j) = 7.6 \times 10^{-5}$ $log(p(j|d_6)) = -9.48$

$p(c|d_6) > p(j|d_6)$ y $log(p(c|d_6)) > log(p(j|d_6))$