

Índice

1. Objetivos y Descripción de los contenidos	1
2. Materiales disponibles	1
3. Tarea	2
4. Anexo	2

1. Objetivos y Descripción de los contenidos

Estos ejercicios, pretenden ser un complemento a la clase teórica, y tienen como objetivo afianzar los conceptos adquiridos en clase sobre el algoritmo de Naïve Bayes para llevar a cabo una tarea de clasificación.

Las **competencias** que el alumno deberá haber adquirido tras realizar este ejercicio son:

1. Capacidad para calcular probabilidades a priori, condicionadas y a posteriori con atributos discretos.
2. Capacidad para aplicar el suavizado (smoothing) de Laplace, cuando nos enfrentamos a probabilidades con valor 0.
3. Capacidad para hacer los mismos cálculos con atributos continuos (aplicar Gauss).

2. Materiales disponibles

El siguiente material será empleado para el desarrollo de la tarea que se propone y que tiene como objetivo alcanzar las competencias enumeradas anteriormente.

1. Son 3 los ejercicios, el ejercicio de los compradores de ordenadores, el clásico ejercicio del Tenis, y el ejercicio de las frutas:
 - a) Conjunto de datos de entrenamiento (train): Una tabla con 14 instancias que caracterizan posibles compradores de ordenadores.
 - b) Conjunto de datos de entrenamiento (train): Una tabla con 14 instancias que caracterizan posibles días para jugar al tenis.

Los datos están en eGela en formato .csv para poder visualizarlos en Dataiku y en el Anexo, al final del documento.

- c) Conjunto de datos formado por 1000 instancias de frutas. En este caso no se proporcionar un .csv sino una tabla resumen con las frecuencias ya calculadas.

3. Tarea

1. Clasificar instancias empleando Naïve Bayes como clasificador.

Compra de ordenadores

- a) youth, low, no, fair, ?
- b) senior,low, no, fair,?

Tenis

- a) Overcast, 64, Normal, False, ?
- b) Rainy, 65, High, True,?

Frutas

- a) Larga, Brillante, Amarilla, ?
- b) Consultar en el Anexo el enunciado y los datos.

Es importante recordar que cuando nos encontramos frente a atributos numéricos continuos o bien discretizamos o aplicamos Gauss.

- Discretizarlos
- Emplear la fórmula de la distribución normal: $P(x|Clase) = \frac{1}{\sigma\sqrt{2\pi}}e^{-(x-\mu)^2/2\sigma^2}$;

4. Anexo

id	age	income	student	credit_rating	Buy_Computer
1	youth	high	no	fair	no
2	youth	high	no	excellent	no
3	middle.age	high	no	fair	yes
4	senior	medium	no	fair	yes
5	senior	low	yes	fair	yes
6	senior	low	yes	excellent	no
7	middle.age	low	yes	excellent	yes
8	youth	medium	no	fair	no
9	youth	low	yes	fair	yes
10	senior	medium	yes	fair	yes
11	youth	medium	yes	excellent	yes
12	middle.age	medium	no	excellent	yes
13	middle.age	high	yes	fair	yes
14	senior	medium	no	excellent	no

 Cuadro 1: **Datos de Compras de Ordenadores**

Outlook	Temperature	Humidity	Windy	PlayTennis
Sunny	85	High	False	No
Sunny	80	High	True	No
Overcast	83	High	False	Yes
Rainy	70	High	False	Yes
Rainy	68	Normal	False	Yes
Rainy	65	Normal	True	No
Overcast	64	Normal	False	Yes
Sunny	72	High	False	No
Sunny	69	Normal	False	Yes
Rainy	75	Normal	False	Yes
Sunny	75	Normal	True	Yes
Overcast	72	High	False	Yes
Overcast	81	Normal	False	Yes
Rainy	71	High	True	No

 Cuadro 2: **Datos del Tenis**

Digamos que tenemos 1000 frutas que podrían ser 'plátano', 'pomelo' u 'otras'. Estas son las 3 clases posibles de la variable Y. Supongamos que disponemos de una cinta por la que van pasando las frutas y a través de sensores, por cada fruta podemos saber su longitud, su brillo y su color. Supongamos que cada instancia viene representada por esos tres atributos cuyos valores son binarios (1 o 0).

Las primeras instancias del conjunto de entrenamiento son las siguientes:
En aras de calcular las probabilidades, agregaremos los datos de entrenamiento en forma de frecuencias.

Fruta	Largo(x1)	Brillo (x2)	Amarillo (x3)
Pomelo	0	1	0
Banana	1	0	1
Banana	1	0	0
Otras	1	1	0
...

Fruta	Largo(x1)	NoLargo(x1)	Brillo (x2)	NoBrillo (x2)	Amarillo (x3)	NoAmarillo (x3)
Banana	400	100	350	150	450	50
Pomelo	0	300	150	150	300	0
Otras	100	100	150	50	50	150

Cuadro 3: **Datos resumidos de las frutas**

Entonces, el objetivo del clasificador es predecir si una fruta dada es un 'Plátano' o 'Pomelo' u 'Otro' dados esos 3 atributos (largo, brillo y amarillo). Digamos que te dan una fruta que es: Larga, Brillante y Amarilla, ¿puedes predecir qué fruta es suponiendo que hayas entrenado un Naïve Bayes con los datos de las frecuencias que aparecen en la tabla?

Referencias

- [1] Datos del Compras de ordenadores: Kaggle Computer Buying Dataset
- [2] Datos del Tenis: UCI Machine Learning Repository. Tennis Dataset

1

id	age	income	student	credit_rating	Buy_Computer
1	youth	high	no	fair	no
2	youth	high	no	excellent	no
3	middle_age	high	no	fair	yes
4	senior	medium	no	fair	yes
5	senior	low	yes	fair	yes
6	senior	low	yes	excellent	no
7	middle_age	low	yes	excellent	yes
8	youth	medium	no	fair	no
9	youth	low	yes	fair	yes
10	senior	medium	yes	fair	yes
11	youth	medium	yes	excellent	yes
12	middle_age	medium	no	excellent	yes
13	middle_age	high	yes	fair	yes
14	senior	medium	no	excellent	no

Compra de ordenadores

a) youth, low, no, fair, ?

b) senior, low, no, fair, ?

No = 5

Yes = 9

	youth	m-age	senior	high	medium	low	no	yes	fair	excellent	
Buy	9	2	4	3	2	4	3	3	6	6	3
!Buy	5	3	0	2	2	2	1	4	1	2	3
	14	5	4	5	4	5	4	7	7	5	6

a)

$$P(\text{Buy} \mid \text{youth, low, no, fair})$$

$$P(\text{youth} \mid \text{Buy}) \cdot P(\text{low} \mid \text{Buy}) \cdot P(\text{no} \mid \text{Buy}) \cdot P(\text{fair} \mid \text{Buy})$$

$$\frac{2}{9} \cdot \frac{3}{9} \cdot \frac{3}{9} \cdot \frac{6}{9} = \underline{0,0165}$$

$$P(\text{!Buy} \mid \text{youth, low, no, fair})$$

$$P(\text{youth} \mid \text{!Buy}) \cdot P(\text{low} \mid \text{!Buy}) \cdot P(\text{no} \mid \text{!Buy}) \cdot P(\text{fair} \mid \text{!Buy})$$

$$\frac{3}{5} \cdot \frac{1}{5} \cdot \frac{4}{5} \cdot \frac{2}{5} = \underline{0,0384}$$

b)

$$P(\text{Buy} \mid \text{senior, low, no, fair})$$

$$P(\text{senior} \mid \text{Buy}) \cdot P(\text{low} \mid \text{Buy}) \cdot P(\text{no} \mid \text{Buy}) \cdot P(\text{fair} \mid \text{Buy})$$

$$\frac{3}{9} \cdot \frac{3}{9} \cdot \frac{3}{9} \cdot \frac{6}{9} = \underline{0,0247}$$

$$P(\text{youth} \mid \text{!Buy}) \cdot P(\text{low} \mid \text{!Buy}) \cdot P(\text{no} \mid \text{!Buy}) \cdot P(\text{fair} \mid \text{!Buy})$$

$$\frac{2}{5} \cdot \frac{4}{5} \cdot \frac{4}{5} \cdot \frac{2}{5} = \underline{0,0256}$$

2

Outlook	Temperature	Humidity	Windy	PlayTennis
Sunny	85	High	False	No
Sunny	80	High	True	No
Overcast	83	High	False	Yes
Rainy	70	High	False	Yes
Rainy	68	Normal	False	Yes
Rainy	65	Normal	True	No
Overcast	64	Normal	False	Yes
Sunny	72	High	False	No
Sunny	69	Normal	False	Yes
Rainy	75	Normal	False	Yes
Sunny	75	Normal	True	Yes
Overcast	72	High	False	Yes
Overcast	81	Normal	False	Yes
Rainy	71	High	True	No

Tenis

- a) Overcast, 64, Normal, False, ?
 b) Rainy, 65, High, True, ?

Play Tennis	!Play Tennis
83	85
70	80
68	65
64	72
69	71
75	
75	
72	
81	

$$\mu = 73$$

$$\sigma = 6,16$$

$$\mu = 74,6$$

$$\sigma = 7,89$$

	Sunny	Overcast	Rainy	High	Normal	True	False	
Play	9	2	4	3	3	6	1	8
! Play	5	3	0	2	4	1	3	2
	14	14	14	14	14	14	14	14

a)

$$P(\text{PT} | \text{Overcast}, 64, \text{Normal}, \text{False}) = \frac{1}{6,16\sqrt{2\pi}} e^{-\frac{(64-73)^2}{2 \cdot 6,16^2}} = 0,038$$

$$P(\text{Overcast} | \text{PT}) \cdot P(64 | \text{PT}) \cdot P(\text{Normal} | \text{PT}) \cdot P(\text{False} | \text{PT}) \cdot P(\text{PT}) \\ 4/9 \cdot ? \cdot 6/9 \cdot 8/9 \cdot 9/14 = 0,00643$$

$$P(\text{!PT} | \text{Overcast}, 64, \text{Normal}, \text{False}) = \frac{1}{7,89\sqrt{2\pi}} e^{-\frac{(64-74,6)^2}{2 \cdot 7,89^2}} = 0,02$$

$$P(\text{Overcast} | \text{!PT}) \cdot P(64 | \text{!PT}) \cdot P(\text{Normal} | \text{!PT}) \cdot P(\text{False} | \text{!PT}) \\ 0/5 \cdot 1 \cdot 1/5 \cdot 2/5 \cdot 5/14 =$$

b)

$$P(\text{PT} | \text{Rainy}, 65, \text{High}, \text{True})$$

3

Digamos que tenemos 1000 frutas que podrían ser 'plátano', 'pomelo' u 'otras'. Estas son las 3 clases posibles de la variable Y. Supongamos que disponemos de una cinta por la que van pasando las frutas y a través de sensores, por cada fruta podemos saber su longitud, su brillo y su color. Supongamos que cada instancia viene representada por esos tres atributos cuyos valores son binarios (1 o 0).

Las primeras instancias del conjunto de entrenamiento son las siguientes:
En aras de calcular las probabilidades, agregaremos los datos de entrenamiento en forma de frecuencias.

Fruta	Largo(x1)	Brillo (x2)	Amarillo (x3)
Pomelo	0	1	0
Banana	1	0	1
Banana	1	0	0
Otras	1	1	0
...

Fruta	Largo(x1)	NoLargo(x1)	Brillo (x2)	NoBrillo (x2)	Amarillo (x3)	NoAmarillo (x3)
Banana	400	100	350	150	450	50
Pomelo	0	300	150	150	300	0
Otras	100	100	150	50	50	150

Entonces, el objetivo del clasificador es predecir si una fruta dada es un 'Plátano' o 'Pomelo' u 'Otro' dados esos 3 atributos (largo, brillo y amarillo). Digamos que te dan una fruta que es: Larga, Brillante y Amarilla, ¿puedes predecir qué fruta es suponiendo que hayas entrenado un Naïve Bayes con los datos de las frecuencias que aparecen en la tabla?

- $P(\text{Pomelo} | \text{Larga, Brillante, Amarilla})$

$$\frac{P(\text{Larga, Brillante, Amarilla} | \text{Pomelo}) \cdot P(\text{Pomelo})}{P(\text{Larga, Brillante, Amarilla})} \quad (\text{Bayes})$$

$$P(\text{Larga, Brillante, Amarilla})$$



$$\frac{P(\text{Larga} | \text{Pomelo}) \cdot P(\text{Brillante} | \text{Pomelo}) \cdot P(\text{Amarillo} | \text{Pomelo}) \cdot P(\text{Pomelo})}{P(\text{Larga, Brillante, Amarillo})} \quad (\text{Naïve})$$



$$(0 \cdot 150 \cdot 300 \cdot 300) / 1000$$



(Laplace)

$$(1 \cdot 151 \cdot 301 \cdot 302) / 1006$$

• $P(\text{Banana} \mid \text{Largo, Brillante, Amarilla})$

$$\underline{P(\text{Largo, Brillante, Amarilla} \mid \text{Banana}) \cdot P(\text{Banana})}$$

$$P(\text{Largo, Brillante, Amarilla})$$



$$\underline{P(\text{Largo} \mid \text{Banana}) \cdot P(\text{Brillante} \mid \text{Banana}) \cdot P(\text{Amarillo} \mid \text{Banana}) \cdot P(\text{Banana})} \quad (\text{Número})$$

$$P(\text{Largo, Brillante, Amarillo})$$



$$(400 \cdot 350 \cdot 450 \cdot 500) / 1000$$

↓ Laplace

$$(401 \cdot 351 \cdot 451 \cdot 502) / 1006$$

• $P(\text{Oz} \mid \text{Largo, Brillante, Amarilla})$

$$\underline{P(\text{Largo, Brillante, Amarilla} \mid \text{Oz}) \cdot P(\text{Oz})}$$

$$P(\text{Largo, Brillante, Amarilla})$$



$$\underline{P(\text{Largo} \mid \text{Oz}) \cdot P(\text{Brillante} \mid \text{Oz}) \cdot P(\text{Amarillo} \mid \text{Oz}) \cdot P(\text{Oz})} \quad (\text{Número})$$

$$P(\text{Largo, Brillante, Amarillo})$$



$$(100 \cdot 150 \cdot 50 \cdot 200) / 1000$$

↓ Laplace

$$(101 \cdot 151 \cdot 51 \cdot 202) / 1006$$