Pecotche Andres
Lucas Carballo

Minería de Datos usando Sistemas Inteligentes

PRACTICA 3 – NAIVE BAYES

4) Aplicación de un modelo Naive Bayes

Dado el siguiente modelo NB para clasificar las frutas en Pera o Manzana en base a los atributos Color y Esfericidad:

Atributo / Valor	Clase **Manzana**	Clase **Pera**
Color = Amarillo	0.8	0.1
Color = Mezcla	0.0	0.6
Color = Rojo	0.2	0.3
Esfericidad (μ)	0.5	0.8
Esfericidad (σ)	0.3	0.2

a) Si ambas clases son equiprobables (las probabilidades de clase a priori son P(Pera)=0.5 y P(Manzana)=0.5), y no se utiliza corrección de Laplace, indicar cómo clasificaría los siguientes 2 ejemplos, incluyendo los cálculos realizados, en la siguiente tabla:

Color	***Esfericidad***	***P(x|pera)***	***P(x|Manzana)***	***P(x|pera) *P(pera)***	***P(x|Manzana) *P(Manzana)***	***Predicción***
Amarillo	***0.6***	*0.120985*	*1.0063552*	*0.06045*	***0.50315***	***Manzana***
Mezcla	***0.8***	*1.19682*	*0.0*	***0.598413***	*0.0*	***Pera***

P(pera)=0.5 ; P(manzana)=0.5 (Datos del problema 4a).

Caso 1

Procedemos a realizar la predicción del primer dato(Amarrillo).

Para realizar las siguientes probabilidades utilizamos Free Statistics Calculator utilizando los datos de esfericidad dados en el cuadro anterior para los dos casos, la media de la manzana es 0.5,y su desviación es 0.3; la media de la pera es 0.8, y su desviación es 0.2.

P(0.6 | pera)= 1.209853 → función densidad de probabilidad, con media = 0.8 y desviación = 0.2

$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}}\, e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

P(0.6 | manzana)= 1.257944

Las siguientes probabilidades son datos.

P(Amarrillo|pera)=0.1

P(Amarillo|manzana)=0.8

Luego procedemos a calcular P(x|pera) y P(x|manzana)

*P(x|pera) = P(0.6|pera) * P (amarillo|pera) = 1.209853 * 0.1 = 0.120985*

*P(x|manzana) = P(0.6|manzana) * P (amarillo|manzana) = 1.257944 * 0.8 = 1.0063552*

*P(x|pera) * P(pera)= 0.120985 * 0.5= 0.06045*

*P(x|manzana) * P(manzana)= 1.0063552 * 0.5= **0.50315***

Caso 1 → MANZANA

Caso 2
P(0.8|Pera)=1.9947114
P(0.8|Manzana)=0.8065690
P(mezcla|pera)= 0.6
P(mezcla|manzana)= 0.0
Acá ya nos damos cuenta que el segundo caso es si o si pera, ya que no hay ninguna posibilidad de que sea manzana. Seguimos avanzando con las cuentas así queda demostrado.
*P(x|pera)=1.9947114 * 0.6= 1.19682*
*P(x|manzana)=0.8065690 * 0.0= 0*
*P(x|pera) * P(pera)= 1.19682 * 0.5 =* ***0.598413***
*P(x|manzana) * P(manzana)= 0.0 * 0.5= 0.0*
Caso 2 → PERA

B) Si las probabilidades de clase a priori son P(Manzana)=0.01 y P(Pera)=0.99, y no se utiliza corrección de Laplace, indicar cómo clasificaría los siguientes 2 ejemplos, incluyendo los cálculos realizados en la siguiente tabla:

Color	***Esfericidad***	***P(x|pera)***	***P(x|Manzana)***	***P(x|pera) *P(pera)***	***P(x|Manzana) *P(Manzana)***	***Predicción***
Amarillo	***0.6***	*0.1209*	*1.0063*	***0.119691***	*0.0100*	***PERA***
Mezcla	***0.8***	*1.19682*	*0.0*	***1.1848***	*0.0*	***PERA***

Utilizo algunos resultados encontrados anteriormente en el inciso a.

Caso 1
*P(x|pera) * P(pera)= 0.1209 * 0.99 =* ***0.119691***
*P(x|manzana) * P(manzana)= 1.0063 * 0.01= 0.0100*
En el caso 1, se cambia la seleccion con respecto a el inciso a, ahora es pera.
CASO 1 → PERA

Caso 2:
*P(x|pera) * P(pera)= 1.19682 * 0.99 =* ***1.1848***
*P(x|manzana) * P(manzana)= 0.0 * 0.01= 0.0*
CASO 2 → PERA

5) Generación de un modelo NB
a) En base a los datos del archivo estrellas.xslx, generar un modelo de NB para clasificar su tipo espectral (F o K), sin utilizar corrección de Laplace. No usar Rapidminer o herramienta similar. Incluir sus cálculos.

Atributo / Valor	Clase F	Clase K
Temperatura μ	6600	3292.85
Temperatura σ	1104.53	21003.18
Habitable = Si	1/4	2/7
Habitable = No	3/4	5/7

Luminosidad μ	11.75	2.6285
Luminosidad σ	7.62807	2.19006

P (clase F)	4/11
P (clase K)	7/11

P (Clase F)= 4/11.
P (Clase K)= 7/11.
Datos del archivo estrellas.xslx

A	B	C	D
Temperatura (°K)	Planeta Habitable Cercano	Luminosidad (Relativa al Sol)	Clase Espectral
6200	No	22	F
7500	No	6	F
3000	Si	0,9	K
6600	Si	3	F
1900	No	0,2	K
2300	No	0,3	K
1400	No	6	K
6100	No	16	F
2500	No	2	K
3450	Si	5	K
8500	No	4	K

Los ordenamos por clase:

	Temperatura (°K)	Planeta Habitable Cercano	Luminosidad (Relativa al Sol)	Clase Espectral
2	6200	No	22	F
3	7500	No	6	F
4	6100	No	16	F
5	6600	Si	3	F
6	1900	No	0,2	K
7	2300	No	0,3	K
8	1400	No	6	K
9	3000	Si	0,9	K
10	2500	No	2	K
11	3450	Si	5	K
12	8500	No	4	K

Media de temperatura:

Clase F:

6200+7500+6100+6600 / 4 = ***6600***

Clase K:

1900+2300+1400+3000+2500+3450+8500 / 7 = ***3292.85***

Desviación de temperatura:

Clase F:

6200-6600= -400 = (-400)² = 160000

7500-6600= 900 = (900)² = 810000

6100-6600= -500 = (-500)² = 250000

6600-6600= 0 = 0² = 0

Dato	*Dato-u*	*Cuadrado*	*Sumatoria*	*Division por n*	*Desviación*
6200	-400	160000	1220000	406666,6667	637,7042157
7500	900	810000			
6100	-500	250000			
6600	0	0			

Clase K:

1900-3292.85 = -1392.85 =(-1392.85)² = 1940031.1

2300-3292.85 = -992.85 =(-992.85)² = 985751.1

1400-3292.85 = -1892.85 =(-1892.85)² = 3582881.1

3000-3292.85 = -292.85 =(-292.85)² = 85761.1

2500-3292.85 = -792.85 =(-792.85)² = 628611.12

3450-3292.85 = 157.15 =(157.15)² = 24696.1225

8500-3292.85 = 5207.15 =(5207.15)² = 27121701.62

Dato	*Dato-u*	*Cuadrado*	*Sumatoria*	*Division por n*	*Desviación*
CLASE K					
1900	-1392,85	1940031,123	34362142,86	5727023,81	2393,120099
2300	-992,85	985751,1225			
1400	-1892,85	3582881,123			
3000	-292,85	85761,1225			
2500	-792,85	628611,1225			
3450	157,15	24696,1225			
8500	5207,15	27114411,12			

Media Luminosidad

Clase F

22+6+16+3 / 4 = ***11.75***

Clase K

0.2+0.3+6+0.9+2+5+4 / 7 = ***2.6285***

Desviación de Luminosidad

Clase F

22-11.75= 10.25

6-11.75= -5.75

16-11.75= 4.25

3-11.75= -8.75

Dato	Dato-u	Cuadrado	Sumatoria	Division por n	Desviación
22	10,25	105,0625	232,75	58,1875	7,628073151
6	-5,75	33,0625			
16	4,25	18,0625			
3	-8,75	76,5625			

Clase K

El procedimiento a utilizar es el mismo descrito anteriormente para la clase F pero con los 7 datos de la clase K. (n=7).

0.2-2.62 →el resultado de esta resta al cuadrado da → 5.8564

0.3-2.62→el resultado de esta resta al cuadrado da →5.3824

6-2.62→el resultado de esta resta al cuadrado da →11.4244

0.9-2.62→el resultado de esta resta al cuadrado da →2.9584

2-2.62→el resultado de esta resta al cuadrado da →0.3844

5-2.62→el resultado de esta resta al cuadrado da →5.6644

4-2.62→el resultado de esta resta al cuadrado da →1.9044

Dato	Dato-u	Cuadrado	Sumatoria	Division por n	Desviación
CLASE K					
0,2	-2,42	5,8564	33,5748	4,7964	2,190068492
0,3	-2,32	5,3824			
6	3,38	11,4244			
0,9	-1,72	2,9584			
2	-0,62	0,3844			
5	2,38	5,6644			
4	1,38	1,9044			

B) Utilizando el modelo anterior, clasifique los 3 primeros ejemplos del conjunto de datos estrellas.xslx. Utilice una tabla como la siguiente para realizar los cálculos. Utilizar notación científica con 2 decimales para escribir las probabilidades para números menores a 0.01. Por ejemplo, 0.0000436 se escribe como 4.36e-5. y 7.3214123e-12 se redondea a 7.32e-12.

temperatura	*habitable*	*luminosidad*	*P(x|F)*	*P(x|K)*	*P(x|F)*P(F)*	*P(x|K)*P(K)*	*Predicción*
6200	*no*	*22*	*8.17E-06*	*0.00E+00*	***2.97E-06***	*0.0*	***F***

7500	*no*	*6*	*6.83E-06*	*1.40E-06*	***2.48E-06***	*8.89E-07*	***F***
3000	*si*	*0.9*	*0.00E+00*	*6.20E-06*	*0.00E+00*	***3.95E-06***	***K***

Cálculos realizados:

Temperatura	*Desviacion K*	*2393,12*	*Desviacion F*	*637,70*	**Luminosidad**	Desviacion K	2,19	Desviacion F	7,62
	Media K	*3292,85*	*Media F*	*6600,00*		Media K	2,62	Media F	11,75

Para calcular las siguientes probabilidades utilizamos la página web → Free Statiscics Calculator con los datos de desviaciones y medias halladas en el inciso anterior, este programa nos facilitará el cálculo de la función de densidad de probabilidad.

DATO 1		**DATO 2**		**DATO 3**	
P(6200,K)	0.00007971	P(7500,K)	0.00003555	P(3000,K)	**0.00016546**
P(6200,F)	0.00051388	P(7500,F)	0.00023109	P(3000,F)	**0**
P(No,K)	0.71	P(No,K)	0.71	P(SI,K)	0.28
P(No,F)	0.75	P(No,F)	0.75	P(SI,F)	0.25
P(22,K)	0.00E+00	P(6,K)	5.54E-02	P(0.9,K)	1.34E-01
P(22,F)	2.12E-02	P(6,F)	3.94E-02	P(0.9,F)	1.90E-02
P(X|K)	0.00E+00	P(X|K)	1.40E-06	P(X|K)	6.20E-06
P(X|F)	8.17E-06	P(X|F)	6.83E-06	P(X|F)	0.00E+00
P(K|6200,NO,22)	0.00E+00	P(K|7500,NO,6)	8.89E-07	P(K|3000,SI,0.9)	**3.95E-06**
P(F|6200,NO,22)	**2.97E-06**	P(F|7500,NO,6)	**2.48E-06**	P(F|3000,SI,0.9)	0.00E+00
Prediccion	**F**	Prediccion	**F**	Prediccion	**K**

Algunas de las cuentas:

P(6200,K) = $$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$ siendo x = 6200 = 0.00007971. La media y la desviacion son las de la tabla de arriba (3292,85 y 2393.12 respectivamente)

P(No,K) = 5/7 ya que en la tabla 5 de las 7 muestras eran “No”

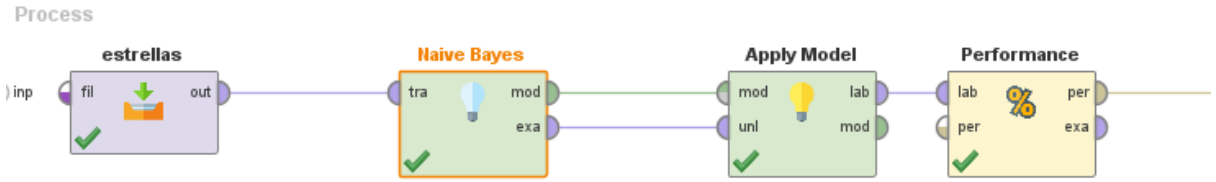
P(22,K) = $$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$ siendo x=22 = 0. La media es 2.62 y la desviación 2.19. (esto explica porque la función densidad de probabilidad evaluada en X = 22 es casi nula)

P(X|K) = P(6200,K) * P(No,K) * P(22,K) = 0

P(K|6200,”No”,22) = P(X|K) * P(K), donde P(K) es la probabilidad de la clase, que es igual a la cantidad de muestras con clase K sobre el total de muestras.

C) Las predicciones del modelo ¿coinciden con las etiquetas del dataset? Calcule las predicciones para el resto de los ejemplos (sin llenar la tabla, utilizando RapidMiner). Luego, calcule el accuracy (porcentaje de ejemplos clasificados correctamente) del modelo para ese conjunto de datos. Por ejemplo, dados 10 ejemplos, si el modelo acierta la clase de 7 de ellos, entonces el accuracy es 7/10=0.7 o 70%.

Las predicciones del modelo coinciden con los 3 primeros valores de la tabla.



accuracy: 90.91%

	true F	true K	class precision
pred. F	3	0	100.00%
pred. K	1	7	87.50%
class recall	75.00%	100.00%	

Como se puede observar la predicción de la clase F fue 3 de 4, siendo que existen 4 valores con clase F, y el modelo predijo 3 de forma correcta. De forma que los 3 que prefijo como F, efectivamente eran clase F, por lo que la precisión de esta clase es el 100%.
La predicción de la clase K por otro lado, no fue perfecta, ya que predijo un valor como K, siendo F. Por lo tanto como acertó 7 de 8 que predijo, la precisión de predicción es de 87.5%

d) Se agrega un ejemplo al conjunto de datos:

Temperatura	Habitable	Luminosidad	Clase
20000	No	35	K

Vuelva a generar el modelo, ahora incluyendo este ejemplo (no es necesario incluir los cálculos). ¿Cómo afecta al modelo de cada atributo o clase?

Atributo / Valor	Clase F	Clase K
Temperatura μ	6600	5381.25
Temperatura σ	637.7	6308.7
Habitable = Si	1/4	2/8
Habitable = No	3/4	6/8

Luminosidad μ	11.75	6.675

Luminosidad σ	8.80	11.65

P (clase F)	4/12
P (clase K)	8/12

El modelo se ve afectado solo en los atributos de la clase K, ya que se agregó un dato de dicha clase, y por tanto todas las medias y desviaciones de cada atributo ahora deberán ser recalculadas considerando el nuevo dato.

e) Vuelva a calcular el accuracy para los datos y el modelo del punto d) ¿Cambió? Si es así, ¿Qué ejemplo cambió su predicción? ¿Por qué?

Ahora el accuracy resulta del 100% para ambas clases:

Row No.	Clase Espec...	prediction(C...	confidence(F)	confidence(K)	Temperatur...	Planeta Habi...	Luminosida...
1	F	F	0.867	0.133	6200	No	22
2	F	F	0.674	0.326	7500	No	6
3	K	K	0.000	1.000	3000	Si	0.900
4	F	F	0.811	0.189	6600	Si	3
5	K	K	0.000	1.000	1900	No	0.200
6	K	K	0.000	1.000	2300	No	0.300
7	K	K	0.000	1.000	1400	No	6
8	F	F	0.856	0.144	6100	No	16
9	K	K	0.000	1.000	2500	No	2
10	K	K	0.000	1.000	3450	Si	5
11	K	K	0.057	0.943	8500	No	4
12	K	K	0	1	20000	No	35

Con los datos originales resultaba:

Row No.	Clase Espec...	prediction(C...	confidence(F)	confidence(K)	Temperatur...	Planeta Habi...	Luminosida...
1	F	F	1.000	0.000	6200	No	22
2	F	F	0.700	0.300	7500	No	6
3	K	K	0.000	1.000	3000	Si	0.900
4	F	K	0.447	0.553	6600	Si	3
5	K	K	0.000	1.000	1900	No	0.200
6	K	K	0.000	1.000	2300	No	0.300
7	K	K	0.000	1.000	1400	No	6
8	F	F	1.000	0.000	6100	No	16
9	K	K	0.000	1.000	2500	No	2
10	K	K	0.000	1.000	3450	Si	5
11	K	K	0.058	0.942	8500	No	4

El dato de la estrella de la columna número 4, antes fue predicho como K, siendo F, ya que la confianza para este segundo valor era ligeramente superior. Con el nuevo dato y el nuevo modelo generado, ahora la confianza para elegir el valor F subió de 0.447 a 0.811, y por tanto, ahora la predicción es acertada.