Data Mining (CSE542)

Homework 09

ID: _____ Name: __조원석___ Date: _2023.05.29_

Task-1 Consider the dataset in Table 18.3. Classify the new point: (Age=23, Car=truck) via the full and naive Bayes approach. You may assume that the domain of Car is given as {sports, vintage, SUV, truck}.

Table 18.3.

	X_1: Age	X_2: Car	Y: Class
$\mathbf{x}_1^T$	25	sports	L
$\mathbf{x}_2^T$	20	vintage	H
$\mathbf{x}_3^T$	25	sports	L
$\mathbf{x}_4^T$	45	suv	H
$\mathbf{x}_5^T$	20	sports	H
$\mathbf{x}_6^T$	25	suv	H

Let. (Age = 23, Car = truck)'s permutation is P((23, truck)),

And it is same P((23, truck) | ALL) == P((23, truck))

Then,P((23, truck) | H) = P(23 | H) × P(truck | H)

P((23, truck) | L) = P(23 | L) × P(truck | L)

For age => $\boldsymbol{D_L = \{x_1, x_3\}, D_H = \{x_2, x_4, x_5, x_6\}}$

P(23 | H) = N(23 | $\boldsymbol{\mu_H = 27.5, \sigma_H = \sqrt{\frac{425}{4}}}$ **=10.31), P(23 | L) = N(23 |** $\boldsymbol{\mu_L = \frac{25+25}{2} = 25, \sigma_L = \sqrt{\frac{0}{2}}}$ **)**

P(sports | H) = $\boldsymbol{\frac{1+1}{4+4} = \frac{2}{8}}$ **, P(sports | L) = =** $\boldsymbol{\frac{2+1}{2+4} = \frac{3}{6}}$

P(vintage | H) = $\boldsymbol{\frac{1+1}{4+4} = \frac{2}{8}}$ **, P(vintage| L) = =** $\boldsymbol{\frac{0+1}{2+4} = \frac{1}{6}}$

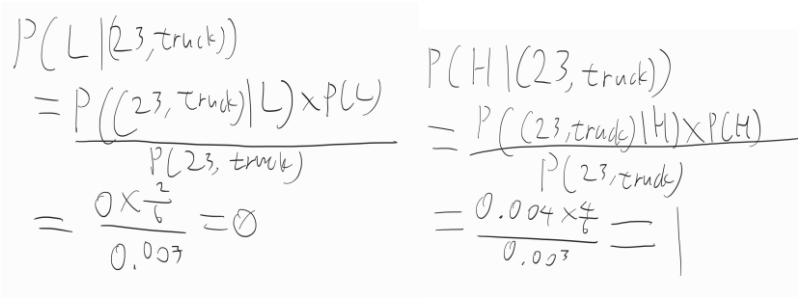
P(suv | H) = $\boldsymbol{\frac{2+1}{4+4} = \frac{3}{8}}$ **, P(suv | L) = =** $\boldsymbol{\frac{0+1}{2+4} = \frac{1}{6}}$

P(truck | H) = $\boldsymbol{\frac{0+1}{4+4} = \frac{1}{8}}$ **, P(truck| L) = =** $\boldsymbol{\frac{0+1}{2+4} = \frac{1}{6}}$

P((23, truck) | H) = P(23 | H) × P(truck | H) = 0.035 ×1/8 = 0.0044

P((23, truck) | L) = P(23 | L) × P(truck | L) = 0 ×1/6 = 0

P (23, truck) = P(23, truck | ALL) = P((23, truck) | H) × P(H) + P((23, truck)|L)×P(L) = 0.0044 ×4/6 + 0 ×2/6 = 0.003



⇨ **Classify (23, truck) as high risk 'H' using Naive Bayes.**

class	Car	Age < 20	Age >= 20	Age >= 25	Age <= 40	Age > 40
H	Vintage	1+1/5	2+2/5	2+3/5	2+4/5	2+1
	Sports	1+1/5	2+2/5	2+3/5	2+4/5	2+1
	Suv	1+1/5	1+1/5	2+3/5	3+4/5	3+1
	Truck	1/5	2/5	3/5	4/5	1
L	Vintage	1+1/5	1+2/5	1+3/5	1+4/5	1+1
	Sports	1+1/5	1+2/5	3+3/5	3+4/5	3+1
	Suv	1/5	2/5	3/5	4/5	5/5
	Truck	1/5	2/5	3/5	4/5	5/5

⇨ **So, P(H | (23, truck)) ∝ P((23, truck) | H) × P(H) / P(23, truck) = 1/5*4/6 = 4/30 = 0.133**

⇨ **So, P(L | (23, truck)) ∝ P((23, truck) | L) × P(L) / P(23, truck) = 1/5 * 2/6 = 2/30 = 0.067**

⇨ **(23, truck) high risk (H)**

⇨ **So, Full Bayes**

Task-2: Given the dataset in Table 18.4, use the naive Bayes classifier to classify the new point(T, F, 1.0).

Table 18.4.

	a_1	a_2	a_3	Class
$\mathbf{x}_1^T$	T	T	5.0	Y
$\mathbf{x}_2^T$	T	T	7.0	Y
$\mathbf{x}_3^T$	T	F	8.0	N
$\mathbf{x}_4^T$	F	F	3.0	Y
$\mathbf{x}_5^T$	F	T	7.0	N
$\mathbf{x}_6^T$	F	T	4.0	N
$\mathbf{x}_7^T$	F	F	5.0	N
$\mathbf{x}_8^T$	T	F	6.0	Y
$\mathbf{x}_9^T$	F	T	1.0	N

$\mathbf{P}(\boldsymbol{a_1 = T \mid Y}) = \frac{3}{4},\ \boldsymbol{P(a_1 = T \mid N)} = \frac{1}{5}, \mathbf{P}(\boldsymbol{a_2 = F \mid N}) = \frac{2}{4}, \boldsymbol{P(a_2 = F \mid Y)} = \frac{2}{5}$

$\boldsymbol{a_3 \Rightarrow Mean}: \ \mu_Y = \mathbf{5.25}, \mathbf{Variance}: \ \boldsymbol{\sigma_Y} = \mathbf{1.71}$

$\boldsymbol{a_3 \Rightarrow Mean}: \ \mu_N = \mathbf{5}, \mathbf{Variance}: \ \boldsymbol{\sigma_N} = \mathbf{2.74}$

Density func. => P(1.0 | Y) = 0.0106 , P(1.0 | N) = 0.0502

P(T, F, 1.0 | Y) = 0.75 * 0.5 * 0.0106 = 0.003975, P(T, F, 1.0 | N) = 0.2* 0.4 * 0.0502 = 0.004016

P(Y | T, F, 1.0) ∝0.003975 * 4/9 = 0.00177

P(N | T, F, 1.0) ∝ 0.004016 * 5/9 = 0.00223