DA Assignment 2019110001

12/11/22

Q.1] In this database, there are four attributes
A = [Day, Season, Fog, Rain) with 20 tuples

The categories of classes are
C = [On Time, Late, Very late, Cancelled]

So we can apply Naive Bayes classification technique to map input tuples into accurate classes.
Therefore we need to find all the prior and posterior probability with the help of dataset.

For prior probabilities for the categories of classes.

$$P[\text{On Time}] = \frac{14}{20}$$

$$P[\text{Late}] = \frac{2}{20}$$

$$P[\text{Very Late}] = \frac{3}{20}$$

$$P[\text{Cancelled}] = \frac{1}{20}$$

From the dataset
The posterior probabilities for the attribute 'Day':

$$P[\text{Weekday} / \text{On Time}] = \frac{9}{14}$$

$$P[\text{Weekday} / \text{Very late}] = \frac{3}{3}$$

$$P[\text{Weekday} / \text{Late}] = \frac{1}{2}$$

$$P[\text{Weekday} / \text{Cancelled}] = \frac{0}{1}$$

Similarly after calculating and tabulating the posterior probabilities for all attributes.

For the attribute 'Day' from the Dataset.

Day	Class: On Time	Late	Very Late	Cancelled
Weekday	9/14	1/2	3/3	0/1
Saturday	2/14	0/2	0/3	1/1
Sunday	1/14	0/2	0/3	0/1
Holiday	2/14	1/2	0/3	0/1

For the attribute 'Season' from the Dataset

Season	Class: On Time	Late	Very late	Cancelled
Spring	4/14	0/2	0/3	1/1
Summer	6/14	0/2	0/3	0/1
Autumn	2/14	0/2	1/3	0/1
Winter	2/14	2/2	2/3	0/1

For attribute 'Fog' -

Fog	On Time	Late	Very Late	Cancelled
	Class			
None	5/14	0/2	0/3	0/1
High	4/14	1/2	1/3	1/1
Normal	5/14	1/2	2/3	0/1

For the attribute 'Rain' from the dataset -

Rain	On Time	Late	Very Late	Cancelled
	Class			
None	6/14	1/2	1/3	0/1
Slight	6/14	1/2	0/3	0/1
Heavy	2/14	0/2	2/3	1/1

For the instance in the question.

<Weekday, winter, High, None>

$P_{NB}(\text{onTime}) = P(\text{OnTime}) \times P[\text{Weekday}/\text{onTime}] \times P[\text{Winter}/\text{onTime}] \times P[\text{High}/\text{onTime}] \times P[\text{None}/\text{onTime}]$

$= \frac{14}{20} \times \frac{9}{14} \times \frac{2}{14} \times \frac{4}{14} \times \frac{6}{14} = 0.0079$

Similarly

$P_{NB}(\text{late}) = \frac{2}{20} \times \frac{1}{2} \times \frac{2}{2} \times \frac{1}{2} \times \frac{1}{2} = 0.0125$

$P_{NB}[\text{Very late}] = \frac{3}{20} \times \frac{3}{3} \times \frac{2}{3} \times \frac{1}{3} \times \frac{1}{3} = 0.0111$

$P_{NB}[\text{Cancelled}] = \frac{1}{20} \times \frac{0}{1} \times \frac{0}{1} \times \frac{44}{1} \times \frac{0}{1} = 0$

P_{NB} [late] is highest, hence correct classification is 'late'

Similarly, we can classify any input tuple into accurate class.

Q2] In this problem we have to test the hypothesis that gender and preferred reading have no correlation between them, and they are independent.

By using x^2 (chi-square) calculation for a contingency table of 2x2 and with degrees of freedom of 1 it will be (2-1) x (2-1)

By using the formula

$$x^2 = \sum_{i=1}^{m} \sum_{j=1}^{n} \frac{[a_{ij} - e_{ij}]^2}{e_{ij}}$$

[m and n denotes the size of the table]

(where o_{ij} = observed frequency

a_{ij} = expected frequency)

$$\therefore x^2 = \frac{[250-90]^2}{90} + \frac{[50-210]^2}{210} + \frac{[200-360]^2}{360} + \frac{[1000-840]^2}{840}$$

$$= 507.93.$$

We can see that the chi-square value 507.93 is very large as compared as to 2.706 (For 1 degrees of freedom, the x^2 valued needed to reject the hypothesis is 2.706 at significance level of 0.1)

Hence, the gender and preffered reading are strongly correlated.

Date:

So we can reject the null hypothesis of independence at a confidence level of 0.1.