

12/11/22

Q.1] In this database, there are four attributes
 $A = [\text{Ray}, \text{Season}, \text{Fog}, \text{Rain}]$ with 20 tuples

The categories of classes are

$C = [\text{OnTime}, \text{Late}, \text{Very late}, \text{Cancelled}]$

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.

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

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

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

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

From the dataset

The posterior probabilities for the attribute 'Ray':

$$P[\text{Weekday} / \text{OnTime}] = \frac{9}{14} \quad P[\text{Weekday} / \text{Cancelled}]$$

$$P[\text{Weekday} / \text{Very late}] = \frac{3}{3} = \frac{0}{1}$$

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

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

For the attribute 'Day' from the Dataset.

Day	On Time	Late	Very Late	Canceled
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	On Time	Late	Very Late	Canceled
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' -

	class			
Fog	OnTime	late	Very late	Cancelled
None	5/14	0/2	0/3	0/1
High				
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 -

	class			
Rain	OnTime	late	Very late	Cancelled
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

$\langle \text{weekday}, \text{winter}, \text{High}, \text{None} \rangle$

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

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

Similarly

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

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

$$P_{NB}(\text{cancelled}) = \frac{1}{20} \times \frac{0}{1} \times \frac{0}{1} \times \frac{4}{1} \times \frac{0}{1} = 0$$

$P_{NB}(\text{late})$ is highest, hence correct classification is 'late'.

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

Q.2] 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 χ^2 (chi-square) calculation for a contingency table of $a \times b$ and with degrees of freedom of 1 it will be $(a-1) \times (b-1)$

By using the formula

$$\chi^2 = \sum_{i=1}^m \sum_{j=1}^n \frac{(a_{ij} - e_{ij})^2}{e_{ij}} \quad [m \text{ and } n \text{ denote the size of the table}]$$

[where a_{ij} = observed frequency
 e_{ij} = expected frequency]

$$\begin{aligned} \therefore \chi^2 &= \frac{(250-90)^2}{90} + \frac{(50-210)^2}{210} + \frac{(200-360)^2}{360} \\ &\quad + \frac{(1000-840)^2}{840} \\ &= 507.93 \end{aligned}$$

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

Hence, the gender and preferred reading are strongly correlated.

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

Conclusion: There is a significant difference between the two groups at the 0.1 level of significance.

Interpretation: The results suggest that the two groups differ significantly in their responses to the treatment.

Assumptions: The assumptions for the chi-square test are met, including independence of observations and expected cell counts.

Limitations: The study has several limitations, including a small sample size and the use of a self-reported measure.

Future Research: Future research should investigate the underlying mechanisms of the observed effects and replicate the study with a larger sample.

References: The following references were consulted during the preparation of this report.

Statistical significance is defined as the probability of observing a result as extreme as the one observed, assuming the null hypothesis is true.

The p-value is the probability of observing a result as extreme as the one observed, assuming the null hypothesis is true.

The significance level is the threshold for rejecting the null hypothesis, typically set at 0.05.

The results of the chi-square test indicate a significant difference between the two groups.

The findings of this study have important implications for the field of research.