ST.FRANCIS INSTITUTE OF TECHNOLOGY

Mount Poinsur, S.V.P. Road, Borivli (West), Mumbai - 400103

Computer Engineering Department
Experiment No. 03

Aim:

To implement classification using Naïve Bayes Algorithm using scikit learn

Theory:

The dataset is divided into two parts, namely, **feature matrix** and the **response vector**.

- Feature matrix contains all the vectors(rows) of dataset in which each vector consists of the value of **dependent features**. In above dataset, features are 'Outlook', 'Temperature', 'Humidity' and 'Windy'.
- Response vector contains the value of **class variable**(prediction or output) for each row of feature matrix. In above dataset, the class variable name is 'Play golf'.

Assumption:

The fundamental Naive Bayes assumption is that each feature makes an:

- Independent
- Equal(None of the attributes is irrelevant and assumed to be contributing equally to the outcome.)

contribution to the outcome.

Note: The assumptions made by Naive Bayes are not generally correct in real-world situations. In-fact, the independence assumption is never correct but often works well in practice.

Bayes' Theorem

Bayes' Theorem finds the probability of an event occurring given the probability of another event that has already occurred. Bayes' theorem is stated mathematically as the following equation:

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}$$

Now, with regards to our dataset, we can apply Bayes' theorem in following way:

$$P(y|X) = \frac{P(X|y)P(y)}{P(X)}$$

where, y is class variable and X is a dependent feature vector (of size n) where:

$$X = (x_1, x_2, x_3, ..., x_n)$$

Naive Assumption

Now, if any two events A and B are independent, then,

P(A,B) = P(A) x P(B)

Applying:

$$P(y|x_1, ..., x_n) = \frac{P(x_1|y)P(x_2|y)...P(x_n|y)P(y)}{P(x_1)P(x_2)...P(x_n)}$$

which can be expressed as:

$$P(y|x_1, ..., x_n) = \frac{P(y)\prod_{i=1}^{n} P(x_i|y)}{P(x_1)P(x_2)...P(x_n)}$$

Now, as the denominator remains constant for a given input, we can remove that term:

$$P(y|x_1, ..., x_n) \propto P(y)\prod_{i=1}^{n} P(x_i|y)$$

ST.FRANCIS INSTITUTE OF TECHNOLOGY

Mount Poinsur, S.V.P. Road, Borivli (West), Mumbai - 400103

Computer Engineering Department
Experiment No. 03

Now, we need to create a classifier model. For this, we find the probability of given set of inputs for all possible values of the class variable y and pick up the output with maximum probability. This can be expressed mathematically as:

$$y = argmax_y P(y) \prod_{i=1}^{n} P(x_i|y)$$

Experiment:

Naïve Bayes Problem Definition:

Classification Problem : Classify the CHAS value into two types as 0.0 and 1.0 given CRIM, ZN, INDUS, CHAS, NOX, RM, AGE, DIS, TAX, PTRATIO, B, LSTAT.

Data Analysis Report :

Confusion Matrix:

138	0
0	14

Accuracy = 100.00 %

Conclusion: The data analysis was performed on IRIS dataset and the following results were identified:

- Using all the features, Naïve Bayes Algorithm gave 100.00 % Accuracy
- No. of mislabeled classification is 0.