Assignment - DA-2

Title: Naive Bayes Algorithm

Problem Statement

Download Pima Indians Diabetes dataset. Use Naive Bayes Algorithm for classification

1) Load the data from csv file and split it into training and test dataset
2) Summarize the properties in the training dataset so that we can calculate probabilities and make predictions
3) Classify samples from a test dataset and a summarised dataset

Objective

1) To learn classification algorithms like Naive Bayes algorithm
2) To implement such algorithms to predict data

Outcome

To be able to learn classification algorithms like Naive Bayes and make prediction using training dataset

Software and hardware requirements

OS: Windows 10 / Ubuntu (64-bit)
Python Scipy libraries / R Studio with R libraries, Gedit editor,
500 GB HDD

Theory related concepts

A] Bayes Theorem

It is a way of finding a probability when we know certain other probabilities

$$P(A/B) = \frac{P(A).\,P(B)}{P(B)}$$

where,

$P(A/B)$ = how often A happens given that B happens

$P(B/A)$ = how often B happens given that A happens

$P(A)$ = how likely A is on its own

$P(B)$ = how likely B is on its own

Example

If dangerous fires rare 1% and smoke is fairly common 10% due to barbeques and 90% of dangerous fires make smoke then,

$$P(fire/smoking) = \frac{P(fire).\,P(smoke/fire)}{P(smoke)}$$

$$= \frac{0.01 \times 0.9}{0.1} = 9\%.$$

∴ Probability of dangerous fire when smoke is 9%.

B] Naive Bayes classification

- It is a simple yet effective and commonly used machine learning classifier
- It is a probabalistic classification that makes classifications using the maximum decision rule in a Bayesian setting. It can be represented using a very simple Bayesian network.
- It is eventually popular for text classification and is a traditional

solution for problems such as spam detection.

c] Applications

1. Real-time predictions
Naive Bayes is an eager learning classifier and it is a very fast. Thus, it could be used to make predictions in real time.

2. Multi-class predictions
This algorithm is well known for multi-class prediction feature. Here, we can predict the probability of multiple classes of target variable.

3. Text classification
It is used to have higher success rate as compared to other algorithms. As a result, it is widely used in spam filtering and sentiment analysis

Test Cases

I/P : Diabetes dataset
O/P : Confusion matrix

$$\begin{array}{ccc} & 0 & 1 \\ 0 & 125 & 67 \\ 1 & 25 & 43 \end{array}$$

Accuracy = 0.7304
Test was 30% of dataset and 73% of predicted values were obtained correctly.

Conclusion
Thus, we successfully learnt and implemented Naive Bayes classification algorithm.