Camlin Page

Assignment

- Title: Naive Bayes Algorithm

- Problem Statement:
  Doonload Prima Indians Oeiabetes dataset. Use Naive-Bayes algo for classification
  1.) Load the data into csv file & split it into training & test datasets
  2.) Summarize properties in the training dataset so that we can calculate probability & make predictions
  3.) Classify samples from the test dataset and a summarized training dataset.

- Objective:
  - To learn classification algorithms like Naive-Bayes
  - To implement such algorithms to predict data

- Outcomes:
  We will be able to -
  - learn classification algorithms
  - make predictions using training datasets

- Requirements:
  - OS: Windows 10 / Ubuntu (64bit)
  - Python Scipy libraries / RStudio with R lib.
  - Gedit
  - RAM: 4GB - HDD: 500GB

- Theory:

A.] Bayes' Theorem:-

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

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

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 is A on it's own

$P(B)$ = how likely is B on it's own

Example:

If dangerous fires are rare (1%) but smoke is fairly common (10%) due to barbeque and 90% of dangerous fires make smoke them

$$P(\text{fire}/\text{smoke}) = \frac{P(\text{fire}) \cdot P(\text{smoke}/\text{fire})}{P(\text{smoke})}$$

$$= \frac{0.01 \times 0.9}{0.1}$$

$$= 9\%$$

∴ Probability of dagerous fires when there is smoke = 9%

B.] Naive-Bayes classification:

- It is a simple, yet effective & commonly used m/c learning classifier
- It is a probabilistic classifier that makes classifications using maximum Aposteriori decision rule in a Bayesian setting. It can be represented using a very simple Bayesian network
- It is especially popular for text classification & it's a traditional soln for problems such as spam detection.

C.] Applications:-

1.) Real Time Predictions:-

Naive-Bayes is an eager learning classifier & it is very fast. Thus, it could be used to make predictions in real life

2.) Multiclass predictions:-

This algorithm is also well known for multiclass prediction feature. Here we can predict the probability of multiple classes of target variable

3.) Text classification:-

It is used to have higher success rates than other algo. As a result, widely used in spam filtering & sentiment analysis.

- Conclusion:-

  Thus we successfully learnt & implemented Naive Bayes classification algorithm.

- Test cases:

  Input → Diabetes dataset I

  Output →

  Confusion Matrix

	0	1
0	125	37
1	25	43

Accuracy – 0.7304

Test set was 30% of dataset & 73% of predicted values are obtained correctly.