Page :
Date : / /

Assignment : C2

- TITLE : Download Pima Indians Diabetes dataset. Use Naive Bayes Algorithm for classification.
  i] Load the data from CSV file and Print
  Split it into training and test datasets.
  ii] Summarizes the property in the training dataset so that, we can calculate probabilities and make predictions.
  iii] Classify sample from test dataset and a summarized training dataset.

- Objective : i] To understand Naive Bayes Algorithm.
  ii] To understand classified samples from a test dataset and summarize training dataset.

- Outcomes : We will be able to :
  i] learn classification algorithm.
  ii] Calculating Probabilities and prediction in the dataset.

- Software and Hardware Requirements : OS : windows / ubuntu distribution.
  Python libraries, Python frameworks.
  4 GB RAM, 500 GB HDD, i3 above CPU
  Pima Indians Diabetes dataset CSV file.
  Jupyter notebook.

- Theory : Naive Bayes classifier ..

In Statistics Naive Bayes classifier are a family of simple "Probablistic classifiers" based on applying Bayes theorem with strong independence assumptions between the features. They are among the simplest Bayesian network models. But they could be coupled with kernel density estimation and achive higher accuracy level. Naive Bayes classifiers are highly scalable. requiring a number of parameters linear in the number of variables in a learning problem. Maximum likelihood training can be done by evaluating a closed form expression, which takes linear time rather than by expensive iterative approximation as used for many other types of classifiers. In the statistics and computer science literature naive bayes models are known under a variety of names, including simple bayes and independence bayes.

- Probablity Model :

Abastracty naive bayes is a conditional probablity model: given a problem instance to be classified represented by a vector :

$X = (x_1, x_2, \ldots, x_n)$ representing some n features (independent variable), it assigns to this instance probabilities.

$$P(C_k \mid x_1, x_2, \ldots, x_n)$$

for each of K possible outcomes of classes C_k

The problem with the above formulation is that if the number of features n is large or if a feature can take on a large number of values, then basing such a model on probability table is infeasible. We therefore reformulate the model to make it more tractable. using Bayes theorem, the conditional probability can be decompose as:

$$P(C_k|x) = \frac{P(C_k)\, P(x|C_k)}{P(x)}$$

In plain english using Bayesian theorem probability terminology the above equation can be written as:

$$\text{Posterior} = \frac{\text{prior} \times \text{likelihood}}{\text{evidence}}$$

The conditional distribution over the class variable C is:

$$P(C_k|x_1, \ldots, x_n) = \frac{1}{Z} P(C_k) \prod_{i=1}^{n} P(x_i|C_k)$$

where the evidence:

$Z = P(x) = \sum_k P(C_k)\, P(x|C_k)$ is a scaling factor dependent only on $x_1, \ldots, x_n$. that is a constant if the value of the features variables are known.

- Pima Indians diabetes dataset:

This dataset is originally from the National Institute of Diabetes and Digestive and Kidney Diseases. The dataset consist of several medical predictor (independent) variables and one target (dependent) variable, outcome. Independent

Variables, outcome. Independent variables include the number of pregnancies the patient has had their BMI, insulin level, age and so on.

- Test cases:

	Discription	Output	Result
i]	accuracy score	0.729166	Pass
ii]	Confusion Matrix	array([[103, 22], [30, 37]])	Pass
iii]	Split train/test dataset	train dataset: 576 rows x 8 columns test dataset: 192 rows x 8 columns	Pass

- Conclusion: Thus we successfully implement the naive bayes classifier and summarized the Pima Indians diabetes dataset.