Roll no. 41310

Page No.:
Date:

* Assignment DA2 *

* Title:- Naive Byes algorithm for classification on pima indians dataset.

* Problem statement:- Download pima Indians diabetes dataset. Use naive byes algorithm for classification
  * Load the data from CSV and split it into training and test datasets.
  * summarise the properties in the training dataset so that we can calculate and make predictions.
  * Classify the samples from test dataset and a summarized training dataset.

* Objective:-
  * Learning Naive Bayes algorithm.
  * Learn to use naive Bayes algorithm classification on given dataset.

* Slw packages and Hlw apparatus used:-
  1. OS:- 64 bit open source linux.
  2. Programming language python 3.

* Outcomes:-
  students will be able to summarise the properties of the dataset split the dataset into training & test data and apply naive bayes algorithm for classification of application.

Related mathematics:-

Mathematical Model

Let s be system set

$S = \{ S; e; x; y; Fme; DD, NDD, FC, SC \}$

where dataset is loaded into the dataframe

S = start state

e = end state i.e. classification of sample from the test dataset.

x = set of inputs.

$x = \{ x_1 \}$

where

x_1 = Pima indians diabetes dataset where,

y = set of outputs.

① splitting of dataset into training & test datasets

② Naive bayes classifier.

Fme is the set of main functions

$Fme = \{ f1, f2, f3 \}$

where

f1 = function to load dataset into dataframe

f2:- function to split dataset into train & test data.

f3:- function to invoke naive bayes classifier

DD:- Deterministic dataset.

PIMA Indians diabetes datasets.

NDD:- Non-deterministic data.

NULL values in dataset.

FC:- Failure Case.

Failed to classify the record into correct class.

Page No.:
Date:

Theory:-

Naive Baye Classifiers are a collection of classification algorithms based on baye theorem. It is not a single algorithm but a family of algorithms where all of them share a common principle. i.e. every pair of features being classified is independent of each other.

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

Feature matrix contains all the vectors (rows) of dataset in which each vector consists of value of dependent features. In datasets features like outlook, temperature, humidity and windy are dependent features.

Response vector matrix contains the value of class variable (prediction or output) for each row of feature matrix.

The fundamental naive bayes assumption is that each feature makes an independent & equal contribution to the outcome.

* Bayes theorem:-

Baye theorem finds the probability of an event occuring given the probability of another event occuring given the probablity of another event that has already occupied Baye's theorem is stated mathematic

as following equation.

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

where A & B are events and P(B) Basically, we are trying to find probability of event A, given the event B is true. Event B is true and also termed as evidence P(A) is the priori of A (the prior probability i.e probability of event before evidence is seen. The evidence is an attribute value of an unknown instance (here it is event B). P(A/B) is a posteriori probability of B, i.e. probability of event after evidence is seen.

Now, with regards to our dataset, we can apply Baye's 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:

$$x = (x_1, x_2, x_3, \ldots, x_n).$$

*Naive Assumption:-

Now, we can put naive assumption to baye's theorem which is independant among the features. So now, we split evidence into the independent parts.

Page No.:
Date:

Now, if any two events A & B are independent.

then

$$P(A,B) = P(A) \cdot P(B)$$

* Test Cases:-

For given dataset

Confusion matrix is $\begin{bmatrix} 96 & 29 \\ 26 & 41 \end{bmatrix}$

and accuracy score is 0.7135416666

* Conclusion:- In this way, naive bayes classifier is used for pima indians dataset analysis.