Assignment : DA

Title :-

Naive Bayer algorithm for classification on piama Indians dataset.

Problem Statement :-

- Download piima indians diabeter dataset use Naive Bayes 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 traing dataset.

Objective;

- learning Naive Bayes Algorithm.
- learn to use 'Naive Bayes Algorithm for classification on given dataset.

Software and hardware apprartus used.

1) os : 64 bit open sours linux
2) programming lang: python / R

Outcomes :-

Students will be able to summarise the properties of the dataset. split the dataset into training and test data and apply

Teacher Signature

Naive Bayes algorithm from classification.
of application.

Related Mathematics:
Mathematical model:
let S be the 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 samples
from the test dataset
x = set of inputs.
$x = \{x_i\}$
where
x_1 = Pima Indians Diabeter dataset
where,
y = set of outputs.
1) Spitting the dataset into training and 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 data set into train & test data
f3 = function to invoke naive bayes classifier
DD = Determistic data
PIMA indians diabeter dataset
NDD = Non-Deterministic data
Fc = failure case
Failed to classify the record into correct classe.

Teacher Signature _____

Theory:-

Naive Bayer classifier are a collection of classification algorithms ~~is~~ based on bayes 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 feature 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 feature. In datasets features like outlook, temprecature, humidity and windy are dependent features.

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

The fundamental Naive Bye Bayer assumption is the each feature makes an independent & equal contribution to outcome

- Bayes Theorem.

Bayer Theorem finds the probablity of an event occuring give probablity of another event occuring given probablity of another event already occured Bayer ~~theorem~~

Teacher Signature

theorem is stated mathematically as. Following equation.

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

Where, A and B are events and $P(B) \neq 0$.
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 posteriori probability of B i.e probability of event after evidence is seen

Now, with regards to our datasets we can apply Bayes algorithm 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:

$x = (x_1, x_2, x_3 \dots x_n)$

- Naive assumption

Now, we put naive assumption to bayes algorithm which is independence among the feature so now we split evidence into independent parts.

Teacher Signature

Now, if any two events A and 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.

Teacher Signature _____