Title :

Naive Bayes algorithm for classification on Pima Indians Diabetes dataset.

Problem Statement :

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

- Load the data from CSV file and split it into training and test datasets.
- Summarize the properties in the training dataset so that we can calculate probabilities and make predictions.
- Classify samples from a test dataset and a summarized training dataset.

Learning Objectives:

- Learn Naive Bayes algorithm
- Learn to use Naive Bayes algorithm for classification on given dataset

Learning Outcome:

Be able to :

Summarize the properties of dataset, split dataset into training and test data, apply Naive Bayes algorithm for classification application.

Software and Hardware Requirements :

- 64 bit CPU
- 4 GB RAM
- Ubuntu 18 OS
- Anaconda Environment
- Jupyter Notebook
- Python Libraries : Pandas, Sklearn, Matplotlib, Numpy

Theory:

The Bayesian Classification represents a supervised learning method as well as a statistical method for classification. Assumes an underlying probabilistic model and it allows us to capture uncertainty about the model in a principled way by determining probabilities of the outcomes. It can solve diagnostic and predictive problems. This Classification is named after Thomas Bayes, who proposed the Bayes Theorem. Bayesian classification provides practical learning algorithms and prior knowledge and observed data can be combined. Bayesian Classification provides a useful perspective for understanding and evaluating many learning algorithms. It calculates explicit probabilities for hypothesis and it is robust to noise in input data.

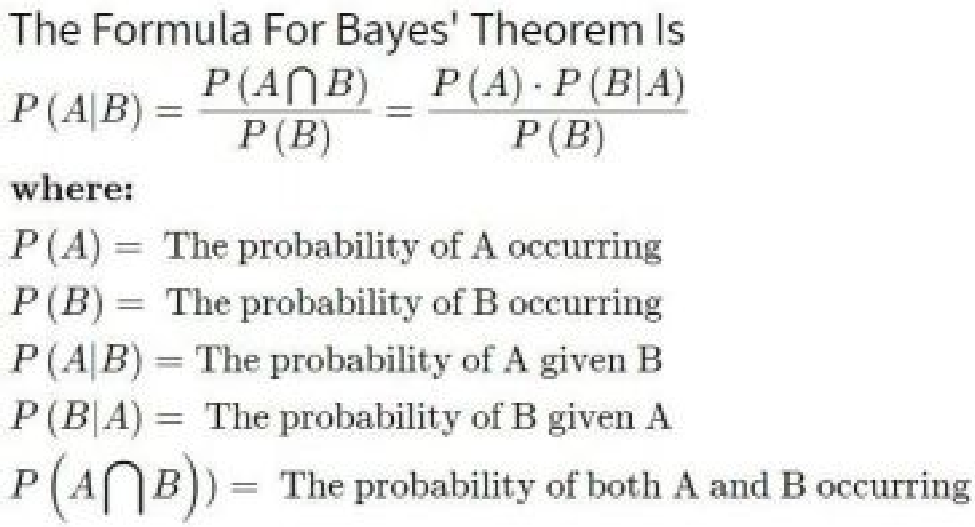
The fundamental Naive Bayes assumption is that each feature makes an:

- independent
- equal

contribution to the outcome.

Bayes’ Theorem

Bayes’ Theorem finds the probability of an event occurring given the probability of another event that has already occurred. Bayes’ theorem is stated mathematically as the following equation:



The Formula For Bayes' Theorem Is

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

where:

$P(A) =$ The probability of A occurring

$P(B) =$ The probability of B occurring

$P(A|B) =$ The probability of A given B

$P(B|A) =$ The probability of B given A

$P\left(A \cap B\right)) =$ The probability of both A and B occurring

Basically, we are trying to find probability of event A, given the event B is true. Event B is 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.

Gaussian Naive Bayes classifier

In Gaussian Naive Bayes, continuous values associated with each feature are assumed to be distributed according to a Gaussian distribution. A Gaussian distribution is also called Normal distribution. When plotted, it gives a bell shaped curve which is symmetric about the mean of the feature values.

Algorithm:

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 = {X1}
where
X1 = Pima Indians Diabetes dataset
where ,
Y=set of outputs
1) Splitting of 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 dataset into training and test datasets
f3 = function to invoke Naive Bayes classifier
DD= Deterministic Data
PIMA Indians diabetes dataset
NDD=Non-deterministic data
null values in the dataset
Fc =failure case:
Failed to classify the record into correct class

Test cases:

Sr No	Input	Actual Output	Expected output	Status
1.	Glucose:199.0 Blood Pressure: 122 Skin Thickness:99 Insulin:846.0 BMI:67.10 DiabetesPedigreeFunction: 2.420 Age:81.0	1	1	Success
2.	Glucose:0.0 Blood Pressure: 0 Skin Thickness:0 Insulin:0 BMI:0 DiabetesPedigreeFunction: 0.078 Age:21	0	0	Success

Conclusion:

Used Naïve Bayes to successfully classify Pima Indians Diabetes dataset and performed summarization on training data.