**Quaid-i-Azam University, Islamabad**



**Data Mining**

**Assignment # 01**

**Implementation of Naïve Bayes Classification algorithm**

**in R Statistical Tool**

**Submitted to:**

**Dr. Khalid Saleem**

**Submitted by:**

Jay Kumar

Kiran Shafique

(M.Phil.)

**Abstract:**

*Classification is an important data mining technique. It has broad applications to classify the various kinds of data used and prediction in nearly every field of our life. Classification is used to classify the item according to the features of the item with respect to the predefined set of classes. This case study report is about diabetes patient dataset classification using Naïve Bayes algorithm in R statistical tool.*

Keywords**:** Classification Algorithms, Data Mining, Diabetes, statistical tool.

**Introduction:**

Classification algorithms are widely used in various applications i-e medical etc. Data classification is a two phase process in which first step is the training phase where the classifier algorithm builds classifier with the training set of tuples and the second phase is classification phase where the model is used for classification. One of the simplest classification algorithms is Naïve Bayes Algorithm.

**Naive Bayes Algorithm:**

Bayesian Classifiers are statistical classifiers based on *Bayes theorem*. Bayesian classification is very simple and it shows high accuracy and speed when applied to large data sets. It works on one assumption that is the effect of an attribute value on a given class is independent of the values of the other attributes. This assumption is called *class conditional independence.*

Bayesian classification can predict class membership probabilities, such as probability that a given tuple belongs to a particular class. The Naïve Bayesian classification predicts that the tuple X belongs to the class Ci. Using the formula

**P (Ci /X) = (P (X/ Ci) P (Ci))/ P(X)**

Where P (Ci /X) is ***maximum posteriori hypothesis*** for the class Ci.

As P(X) is constant for all classes, only P(X/ Ci) P (Ci) needed to be maximized.

If the class prior probabilities are not known, then it is commonly assumed that the classes are equally likely, that is,

**P (C1) = P (C2) =….. = P (Cm).P (Ci /X) = P (Xj/ Ci).**

Otherwise **P (Ci /X) = P(X/Ci) P(Ci)**

The class prior probabilities may be estimated by P (Ci) =|Ci, D|/|D|

Where |Ci, D| is the number of training tuples of class Ci in D.

Given datasets with many attributes, it would be extremely computationally expensive to compute P(X/Ci). In order to reduce computation in evaluating P(X/Ci), the naive assumption of class conditional independence is made. This presumes that the values of the attributes are conditionally independent of one another, given the class label of the tuple i.e., that there are no dependence relationships among the attributes.

Thus,

= P (X1|Ci) × P (X2|Ci) ×… P (Xn|Ci)

Probabilities P(X1/Ci), P(X2/Ci), …. are easily estimated from the training tuples.

Here refers to the value of attribute for tuple X which may be categorical or continuous valued.

**Training Data:** Dataset used for training purpose is “**Pima Indians Diabetes Database**”.

**Sources**: The sample data of data set which we use for classification is Naïve Bayes algorithm is taken from link.[1]

1. **Original owners:** National Institute of Diabetes and Digestive and Kidney Diseases
2. **Donor of database:** [Vincent Sigillito](mailto:vgs@aplcen.apl.jhu.edu) Research Center, RMI Group Leader Applied Physics Laboratory, The Johns Hopkins University Laurel, MD 20707 (301) 953-6231
3. **Date received:** 9 May 1990

**Past Usage:**

Smith,~J.~W., Everhart,~J.~E., Dickson,~W.~C., Knowler,~W.~C., \Johannes,~R.~S. (1988). Using the ADAP learning algorithm to forecast the onset of diabetes mellitus. In {\it Proceedings of the Symposium on Computer Applications and Medical Care} (pp. 261--265). IEEE Computer Society Press.[1]

The diagnostic, binary-valued variable investigated is whether the patient shows signs of diabetes according to World Health Organization criteria (i.e., if the 2 hour post-load plasma glucose was at least

200 mg/dl at any survey examination or if found during routine medical Care). The population lives near Phoenix, Arizona, USA.

**Results:** Their ADAP algorithm makes a real-valued prediction between 0 and 1. This was transformed into a binary decision using a cutoff of 0.448. Using 576 training instances, the sensitivity and specificity of their algorithm was 76% on the remaining 192 instances.

**Relevant Information:**

Several constraints were placed on the selection of these instances from a larger database. In particular, all patients here are females at least 21 years old of Pima Indian heritage.

**Number of Instances:** 768

**Number of Attributes:** 8 plus class

**For Each Attribute:** (all numeric-valued)

1. Number of times pregnant
2. Plasma glucose concentration a 2 hours in an oral glucose tolerance test
3. Diastolic blood pressure (mm Hg)
4. Triceps skin fold thickness (mm)
5. 2-Hour serum insulin (mu U/ml)
6. Body mass index (weight in kg/(height in m)^2)
7. Diabetes pedigree function
8. Age (years)
9. Class variable (0 or 1)

**Missing Attribute Values:** None

**Class Distribution:** (class value ‘D’ is interpreted as "tested positive for diabetes")

**Class Value Number of instances**

N 500

D 268

**Solving in R:**

We have trained the classifier according to the given 8 attributes. For working on Naïve Bayes classification we need to install packages ‘caret’ and ‘klaR’. Load the both library in R.

install.packages ("klaR")

install.packages ("caret")

library(klaR)

library(caret)

For training the model, we have load the training dataset from CSV file. Attributes and classes have been passed for training.

myFile <- read.csv("D://pima.csv")

attr <- myFile[,-9]

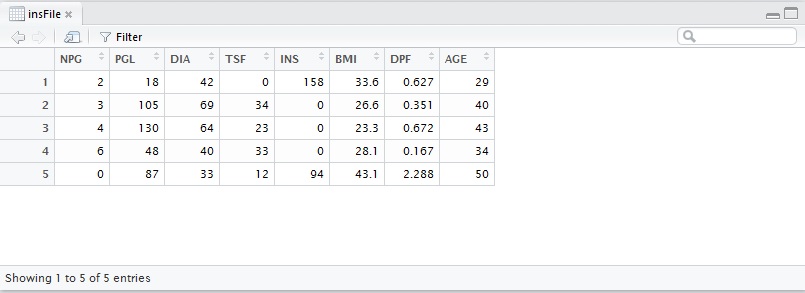
cls <- myFile$Diabet

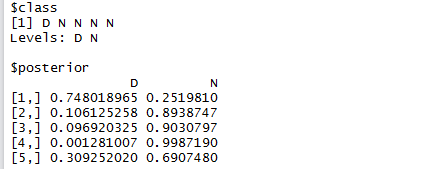
model <- train(attr, cls, ‘nb’)

We have taken 5 instances loaded from file and predicted their classes as shown in figure below.

insFile <- read.csv("D://instances.csv")

view(insFile)





As it the probability of every instance according to both classes is calculated. Only first instance is diabetic and remaining are non-diabetic.

**Conclusion:**

In this case study, popular Naïve Bayes algorithm is used for classification of diabetes patient database using R statistical tool. Classification is done easily and quickly. Error rate in calculation is reduced by using R statistical tool. Therefore, we can say that classification can be done more efficiently with the help of Naïve Bayes algorithm and predicted results are more accurate.

**References:**

**[1]** [**https://github.com/timm/auk/blob/master/data/diabetes**](https://github.com/timm/auk/blob/master/data/diabetes)