**Abstract:**

*Classification is an important data mining technique. Classification is used to classify the item according to the features of the item with respect to the predefined set of classes. This report is about Pima dataset classification using K-Nearest Neighbor algorithm in R statistical tool.*

**Introduction:**

Classification algorithms are widely used in various applications. 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. Here we are discussing the k-nearest neighbor classification algorithm.

**K-Nearest Neighbor:**

In pattern recognition, the *k*-Nearest Neighbors algorithm (*k*-NN) is a non-parametric method used for classification and regression. k-NN is a type of [instance-based learning](https://en.wikipedia.org/wiki/Instance-based_learning), or [lazy learning](https://en.wikipedia.org/wiki/Lazy_learning), where the function is only approximated locally and all computation is deferred until classification. The k-NN algorithm is among the simplest of all [machine learning](https://en.wikipedia.org/wiki/Machine_learning) algorithms.

**Algorithm:**

The training examples are vectors in a multidimensional feature space, each with a class label. The training phase of the algorithm consists only of storing the [feature vectors](https://en.wikipedia.org/wiki/Feature_vector) and class labels of the training samples.

In the classification phase, *k* is a user-defined constant, and an unlabeled vector (a query or test point) is classified by assigning the label which is most frequent among the *k* training samples nearest to that query point. A commonly used distance metric for [continuous variables](https://en.wikipedia.org/wiki/Continuous_variable) is [Euclidean distance](https://en.wikipedia.org/wiki/Euclidean_distance). For discrete variables, such as for text classification, another metric can be used.

**Training Data:** .

The dataset is taken of ‘pima’ which is related to those attributes which are related to diabetes. There are 769 instances for training set. This dataset is for predicting a person either person is diabetic or normal. There are eight attributes in attribute set.

|  |  |
| --- | --- |
| NPG | Normal Pressure Glaucoma |
| PGL | Plasma Glucose Level |
| DIA | Diazoxide |
| TSF | Trans-sphincteric flow |
| INS | Insulin |
| BMI | Bicuculline methiodide |
| DPF | Diffuse Pulmonary Fibrosis |
| AGE | Age of the Person |

**Solving in R:**

We have to train the classifier according to the given 8 attributes. Load the dataset csv file. Store the class attribute in separate variable and remaining in other. The ninth column is of class attribute.

pima <-read.csv("D:\\MPhil\\SLecture\\DM\\assignment\_2016\\KNN\\pima\_n.csv")

pima.trainingClass <- pima[,9] #class column from dataset

pima <- pima[,-9]

For K-NN we require a default package ‘class’. Then we will load the testing dataset from different file. Actually it was the part of the original dataset but we separate it in different file. Here we have take value of k is 9.

pima.testing <- read.csv("D:\\Study and Documents\\MPhil\\Slides-Lecture\\DM\\assignment\_2016\\KNN\\instances\_n.csv")

require(class)

m1 <- knn(train = pima, test = pima.testing, cl = pima.trainingClass, k=9)

