**Practical No. 3**

**Aim**: To implement K nearest neighbor classification algorithm

**Prerequisite:**

1. Working of KNN classification algorithm
2. Understanding of fundamental programming constructs in C/C++/Java
3. Basic features of WEKA tool

**Outcome:** After successful completion of this experiment students will be able to

* Understand the working of KNN algorithm through implementation and the various measures used for checking its performance
* Use Classifier tab in WEKA, apply KNN classification algorithm and analyze the model created.
* Build KNN classifier using Knowledge Flow in WEKA

**Theory:**

Classification is a data mining function that assigns items in a collection to target categories or classes. The goal of classification is to accurately predict the target class for each case in the data. For example, a classification model could be used to identify loan applicants as low, medium, or high credit risks.

A classification task begins with a data set in which the class assignments are known. For example, a classification model that predicts credit risk could be developed based on observed data for many loan applicants over a period of time.

In addition to the historical credit rating, the data might track employment history, home ownership or rental, years of residence, number and type of investments, and so on. Credit rating would be the target, the other attributes would be the predictors, and the data for each customer would constitute a case.

Classifications are discrete and do not imply order. Continuous, floating-point values would indicate a numerical, rather than a categorical, target. A predictive model with a numerical target uses a regression algorithm, not a classification algorithm.

The simplest type of classification problem is binary classification. In binary classification, the target attribute has only two possible values: for example, high credit rating or low credit rating. Multiclass targets have more than two values: for example, low, medium, high, or unknown credit rating.

K nearest neighbors is a simple algorithm that stores all available cases and classifies new cases based on a similarity measure (e.g., distance functions). KNN has been used in statistical estimation and pattern recognition already in the beginning of 1970’s as a non-parametric technique.

**Algorithm**

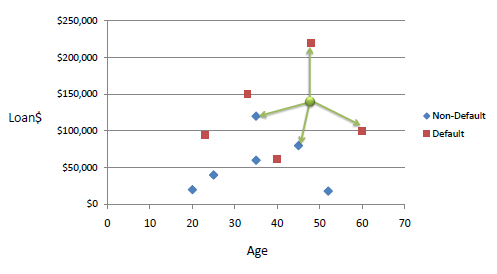
A tuple is classified by a majority vote of its neighbors, with the case being assigned to the class most common amongst its K nearest neighbors measured by a distance function. If K = 1, then the case is simply assigned to the class of its nearest neighbor.



It should also be noted that all three distance measures are only valid for continuous variables. In the instance of categorical variables the Hamming distance must be used. It also brings up the issue of standardization of the numerical variables between 0 and 1 when there is a mixture of numerical and categorical variables in the dataset.

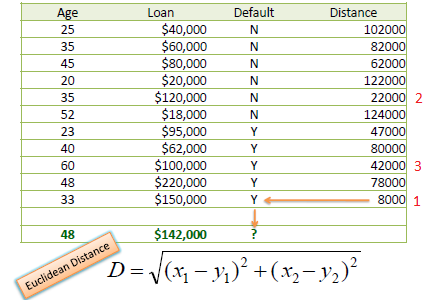
Example:

Consider the following data concerning credit default. Age and Loan are two numerical variables (predictors) and Default is the target.



We can now use the training set to classify an unknown case (Age=48 and Loan=$142,000) using Euclidean distance. If K=1 then the nearest neighbor is the last case in the training set with Default=Y.

D = Sqrt[(48-33)^2 + (142000-150000)^2] = 8000.01 >> Default=Y



With K=3, there are two Default=Y and one Default=N out of three closest neighbors. The prediction for the unknown case is again Default=Y.

\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*

(TO BE COMPLETED BY STUDENTS)

|  |  |
| --- | --- |
| Roll No. | Name: |
| Class: | Batch: |
| Date of Practical: | Date of Submission: |
| Grade: |  |

A production unit of a company conducts survey to find out people’s opinions about the tissue paper that they manufacture. Based on the questionnaire that was taken, the following data set was created. The factory now wants to test the new tissue paper produced by them with the attribute of acid durability = 3 and strength being 7.

|  |  |  |
| --- | --- | --- |
| Acid durability | Strength | Classification |
| 7  7  3  1  2  6  2  5 | 7  4  4  4  3  3  5  4 | Bad  Bad  Good  Good  Good  Bad  Good  Bad |

1. Preprocess and implement a KNN classifier model with value of k being 3 and use Euclidean as the distance measure.
2. Using WEKA tool: For the same data set, build a KNN model and classify the tuple,

With attribute of acid durability = 4 and strength being 6

1. Use Knowledge flow in WEKA and build KNN classifier for the same data set.

**Questions to be answered:**

1. Why is KNN algorithm also called as lazy learning and Nonparametric learning algorithm?

**\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\***