UCAIP210: Machine Learning Algorithms Practicals
Lab Manual
Practical Teacher : Dr Monika Y. Dangore

Experiment No: 3

Aim:

Implement K-Nearest Neighbour Classification using Python

Introduction to K-Nearest Neighbors (KNN) Algorithm

The **K-Nearest Neighbors (KNN) algorithm** is a supervised machine learning method used to solve classification and regression problems. Mostly it is employed for classification problems.

Let X be the training dataset with n data points, where each data point is represented by a d-dimensional feature vector and Y be the corresponding labels or values for each data point in X. Given a new data point x, the algorithm calculates the distance between x and each data point in X using a distance metric, such as Euclidean distance.

The algorithm selects the K data points from X that have the shortest distances to x. For classification tasks, the algorithm assigns the label y that is most frequent among the K nearest neighbors to x. For regression tasks, the algorithm calculates the average or weighted average of the values y of the K nearest neighbors and assigns it as the predicted value for x.

How to choose the value of k for KNN Algorithm?
The value of k is very crucial in the KNN algorithm to define the number of neighbors in the algorithm. The value of k in the k-nearest neighbors (k-NN) algorithm should be chosen based on the input data. If the input data has more outliers or noise, a higher value of k would be better. It is recommended to choose an odd value for k to avoid ties in classification. Cross-validation methods can help in selecting the best k value for the given dataset.

KNN performed best with less number of features (curse of dimension).

Example: Suppose we have a new data point and need to put it in the required category. Consider the below image :

- First, we will choose the number of neighbors as k=5.
- Next, we will calculate the Euclidean distance between the data points.
- By calculating the Euclidean distance, we got the nearest neighbors, as three nearest neighbors in category A and two nearest neighbors in category B. Consider the below image :
- As we can see, the three nearest neighbors are from category A. Hence this new data point must belong to category A.
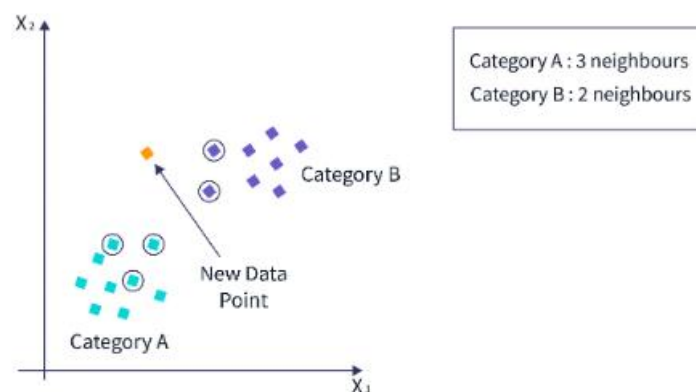
How to Choose the Factor 'K'?

A KNN algorithm is based on feature similarity. Selecting the right K value is a process called parameter tuning, which is important to achieve higher accuracy.

There is not a definitive way to determine the best value of K. It depends on the type of problem you are solving, as well as the business scenario. The most preferred value for K is five. Selecting a K value of one or two can be noisy and may lead to outliers in the model, and thus resulting in overfitting of the model. The algorithm performs well on the training set, compared to its true performance on unseen test data. To choose the value of K, take the square root of n (sqrt(n)), where n is the total number of data points. Usually, an odd value of K is selected to avoid confusion between two classes of data.

Consider the following example below to predict which class the new data point belongs to.

If you take K=3, the new data point is a red square.
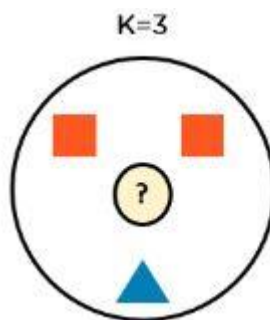
But, if we consider K=7, the new data point is a blue triangle. This is because the amount of red squares outnumbers the blue triangles.
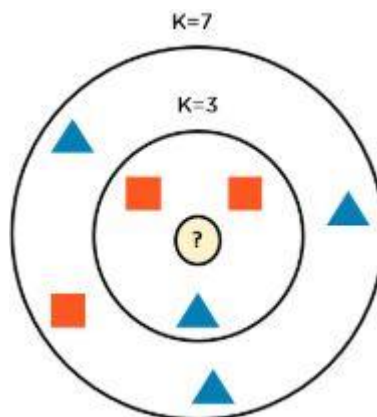
Program:

Attach the printouts of the program.

Result:

The concept of K-Nearest Neighbour Classification is studied and program is executed successfully.