KNN

Epoch IIT Hyderabad

Himani Agrawal
MA22BTECH11008

Contents

Introduction

The k-nearest neighbors algorithm, also known as KNN or k-NN, is a non-parametric, supervised learning classifier, which uses proximity to make classifications or predictions about the grouping of an individual data point. While it can be used for either regression or classification problems, it is typically used as a classification algorithm, working off the assumption that similar points can be found near one another

Distance metrics

1. Euclidean Distance:
   This is nothing but the cartesian distance between the two points which are in the plane/hyperplane. Euclidean distance can also be visualized as the length of the straight line that joins the two points which are into consideration. It is given by:

$$d\left(x,y\right)=\sqrt{\sum_{i=1}^{n}\left(x_i-y_i\right)^2}$$

2. Manhattan Distance:
   This distance metric is generally used when we are interested in the total distance traveled by the object instead of the displacement. This metric is calculated by summing the absolute difference between the coordinates of the points in n-dimensions.it is given by:

$$d\left(x,y\right)=\sum_{i=1}^{n}\left|x_i-y_i\right|$$

3. Minkowski Distance:
   We can say that the Euclidean, as well as the Manhattan distance, are special cases of the Minkowski distance.It is given by:

$$d\left(x,y\right) = \left(\sum_{i=1}^{n}\left(x_i - y_i\right)^p\right)^{\frac{1}{p}}$$

From the formula above we can say that when p = 2 then it is the same as the formula for the Euclidean distance and when p = 1 then we obtain the formula for the Manhattan distance.

Selecting the value of K

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.for this cross validation methods could be used.