KNN

KNN also called K- nearest neighbour is a supervised machine learning algorithm that can be used for classification and regression problems.

K nearest neighbour is also termed as a lazy algorithm .

## Distance Metrics

It is essential to choose the most appropriate distance metrics for a particular dataset. The following are the various distance metrics:

Minkowski Distance-Minkowski distance is calculated where distances are in the form of vectors that have a length and the length cannot be negative.

Manhattan Distance-The distance between two points is the sum of the absolute differences of their Cartesian coordinates.

Euclidean Distance- It is a measure of the true straight line distance between two points in Euclidean space.

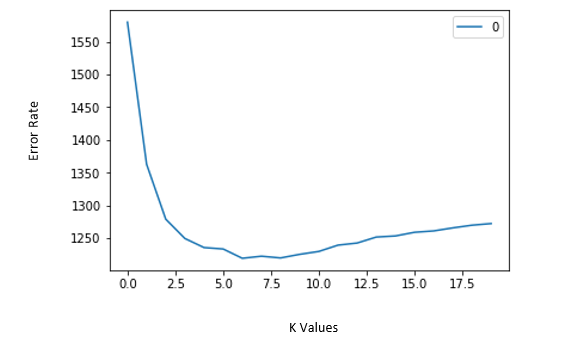
Cosine Distance-It is used to calculate the similarity between two vectors. It measures the direction and uses the cosine function to calculate the angle between two vectors.

Jaccard Distance-It is similar to cosine distance as both the methods compare one type of attribute distributed among all data. The Jaccard approach looks at the two data sets and finds the incident where both values are equal to 1.

**How to choose the K value?**

One of the trickiest questions to be asked is how we should choose the K value.

* One should not use a low value of K= 1 because it may lead to overfitting i,e during the training phase performs good but during the testing phase, the model performs badly. Choosing a high value of K can also lead to underfitting i.e it performs poorly during the training and testing phase.
* We should not use even values of K when classifying binary classification problems. Suppose we choose K=4 and the neighbouring 4 points are evenly distributed among classes i.e 2 data points belong to category 1 and 2 data points belong to category 2. In that case, the data point cannot classify as there is a tie between the classes.
* Choose K value based on domain knowledge.
* Plot the elbow curve between different K values and error. Choose the K value when there is a sudden drop in the error rate.



## Different Algorithms of KNN

**kd\_tree**

**ball\_tree**

**brute**

**auto**

## Advantages & Disadvantages of KNN Algorithm

## Advantages

* It is very easy to understand and implement
* It is an instance-based learning(lazy learning) algorithm.
* KNN does not learn during the training phase hence new data points can be added with affecting the performance of the algorithm.
* It is well suited for small datasets.

## Disadvantages

* It fails when variables have different scales.
* It is difficult to choose K-value.
* It leads to ambiguous interpretations.
* It is sensitive to outliers and missing values.
* Does not work well with large datasets.
* It does not work well with high dimensions.

How does K-NN work?

The K-NN working can be explained on the basis of the below algorithm:

* **Step-1:** Select the number K of the neighbors
* **Step-2:** Calculate the Euclidean distance of **K number of neighbors**
* **Step-3:** Take the K nearest neighbors as per the calculated Euclidean distance.
* **Step-4:** Among these k neighbors, count the number of the data points in each category.
* **Step-5:** Assign the new data points to that category for which the number of the neighbor is maximum.
* **Step-6:** Our model is ready.





Here K is 5 so we take 5 points