The **k-Nearest-Neighbours (kNN)** method is simplest method in machine learning and very simple to understand and implement, this method has seen wide application in many domains, such as in **recommendation systems**, **semantic searching**, and **anomaly detection**

How does it Works

KNN can be used for both classification and regression problems. The algorithm uses '**feature similarity**' to predict values of any new data points. This means that the new point is assigned a value based on how closely it resembles the points in the training set.

Steps to follow

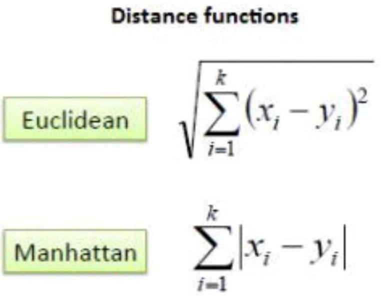
- Compute a distance value between the item to be classified and every item in the training data-set
- Pick the k closest data points (the items with the k lowest distances)
- Conduct a **"majority vote"** among those data points — the dominating classification in that pool is decided as the final classification

Methods of calculating distance between points

The **first step** is to calculate the distance between the new point and each training point. There are various methods for calculating this distance, of which the most commonly known methods are – Euclidian, Manhattan (for continuous) and Hamming distance (for categorical).

1. **Euclidean Distance:** Euclidean distance is calculated as the square root of the sum of the squared differences between a new point (x) and an existing point (y).

2. **Manhattan Distance**: This is the distance between real vectors using the sum of their absolute difference.



Distance functions

Euclidean $\sqrt{\sum_{i=1}^{k}(x_i - y_i)^2}$

Manhattan $\sum_{i=1}^{k}|x_i - y_i|$

3. **Hamming Distance**: It is used for categorical variables. If the value (x) and the value (y) are same, the distance D will be equal to 0 . Otherwise D=1.

$$D_H = \sum_{i=1}^{k}|x_i - y_i|$$

$$x = y \Rightarrow D = 0$$

$$x \neq y \Rightarrow D = 1$$

Once the distance of a new observation from the points in our training set has been measured, the next step is to pick the closest points. The number of points to be considered is defined by the value of k.

Parameter

The **second step** is to select the k value. This determines the number of neighbors we look at when we assign a value to any new observation (K =3/4/5).