Future Connect Media

Machine learning Part B

Part of Future Connect Media’s IT Course

By Abhishek Sharma

Future
Connect
Media

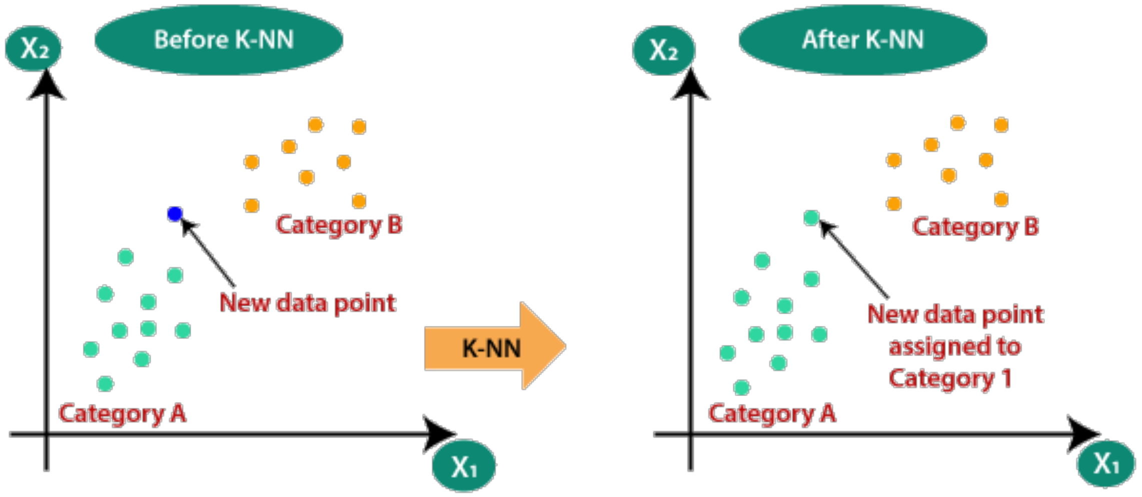
Topics to be covered:

- K-Nearest Neighbor
- Distances in K-NN
- Calculating K

Future Connect Media

K-Nearest Neighbor

- 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.



Distances in K-NN

- Following are the distances used in K-NN:
- Euclidean Distance
- Manhattan Distance
- Minkowski Distance

Euclidean Distance

This is the most used distance measure, and it is limited to real-valued vectors. Using the below formula, it measures a straight line between the query point and the other point being measured.

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

Manhattan Distance

This is also another popular distance metric, which measures the absolute value between two points. It is also referred to as taxicab distance or city block distance as it is commonly visualized with a grid, illustrating how one might navigate from one address to another via city streets.

$$Manhattan\,Distance = \sum_{i=1}^{n} |p_i - q_i|$$

Minkowski distance

This distance measure is the generalized form of Euclidean and Manhattan distance metrics. The parameter, p, in the formula below, allows for the creation of other distance metrics. Euclidean distance is represented by this formula when p is equal to two, and Manhattan distance is denoted with p equal to one.

$$Minkowski\,(A,B) = (\sum_{i=1}^{n} |(fa_i - fb_i)^p|)^{\frac{1}{p}}$$

Calculating K

K can be calculated based on these 2 rules

- Square Root the number of entries in data set
- Making the number Odd by adding or subtracting 1