K-Nearest Neighbour Algorithm

K-NN is a type of supervised learning algorithm that’s is simple and classifies new data based on a similarity measure. It is otherwise known as a ‘lazy learner’ algorithm because there is no model learning phase. All the work happens at the time of prediction as requested. The algorithm assumes that if you are like your neighbour, you are one of them.

The ‘K’ denotes the number of nearest neighbours, which are the voting class of the new data or testing data.

For example, if k = 1, the testing data are given the same label as the closest example in the training data set.

Similarly, if k = 3, the labels of the three closest classes are checked, and the most common label is assigned to the testing data.

KNN is most useful for data classification that requires high accuracy. KNN is also helpful in imputing missing values in a data set.

Application of KNN in the real world.

Recommended system: When you search for a product in Amazon, the system shows you the product you search for and displays a set of items similar or related to the product you are interested in buying.

Concept search: the main problem is to extract a concept from a set of documents as each document contains thousands of information that could be a concept considering the vast amount of data on the web.

Predicting things using KNN:

KNN use the least distance measure to find its nearest neighbours. The common distance measures used include:

1. Euclidean distance = direct distance between two points. Calculates the Sqrt of the sum of the difference between a new point ‘x’ and an existing point ‘y’.
2. Manhattan distance = used to calculate the distance between real vectors using the sum of their absolute difference.

There is no physical or biological way to determine ‘K’’s best values. So, you may have to try out a few values before settling on one. Low values of K (like K = 1 or K = 2) can be noisy and subject to the effects of outliers.