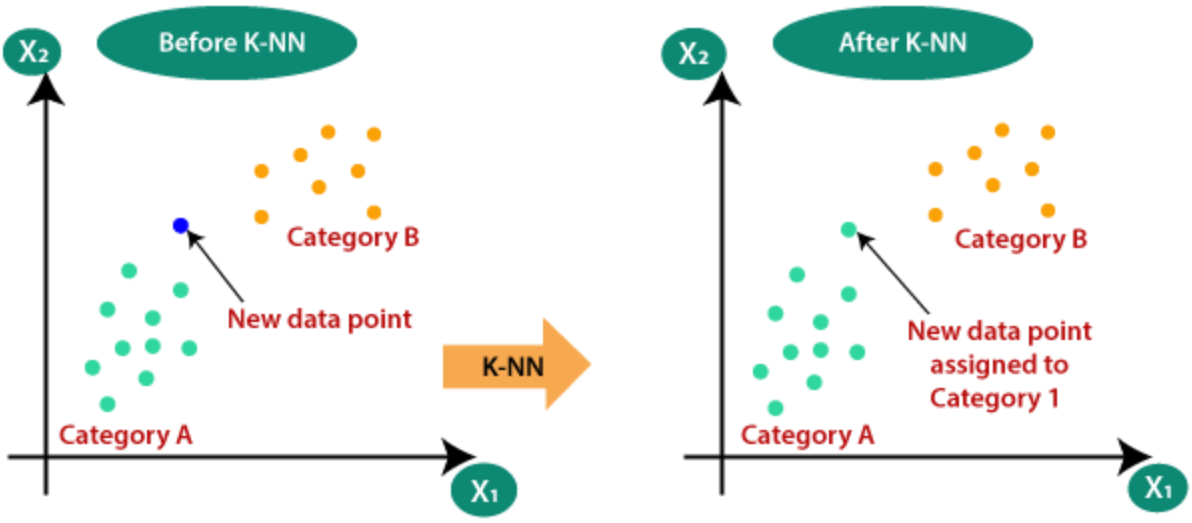
KNN – K – Nearest Neighbors is one of the Simple Supervised Machine Learning Algorithm that used to solve for both classification and regression problem but KNN is mostly used for classification. It classifies a data point on how its neighbors are classified. K in KNN is parameter that refer the number of nearest neighbors. The best K for KNN is called parameter tuning. In addition, K number has effect on algorithm accuracy.

KNN is used when the data is labeled and there is no Nosie on data. Classification problem has a discrete value on its output for example: is the tomorrow going to rain or not. (rain (1), not rain (0)) we have just two class here and there is no middle ground. Another used example is classifying images of Cat and Dog. Suppose we have an image of a creature that looks similar to cat and dog. But we want to know either it is a cat or a dog, so for this classification we use KNN algorithm as it works on similarity measure. The KNN model will find the similar features of the new data set to the cats and dogs images and based on the most similar features it will put it in either cat or dog category.

Why we need a KNN?

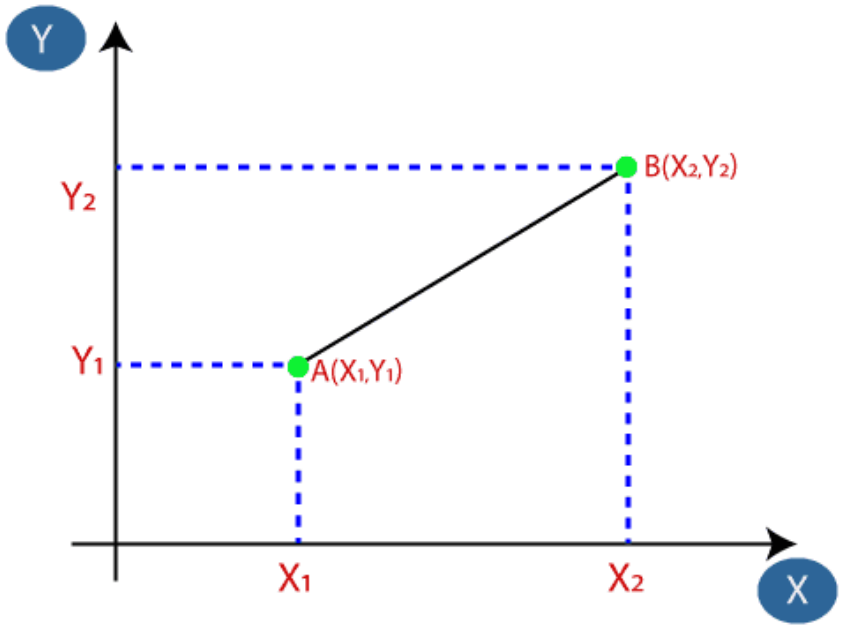
Suppose there are two categories, i.e., Category A and Category B, and we have a new data point x1, so this data point will lie in which of these categories. To solve this type of problems, we need a K-NN algorithm. With the help of K-NN, we can easily identify the category or class of a particular dataset. As you can see the blow diagram.



In the KNN algorithm process:

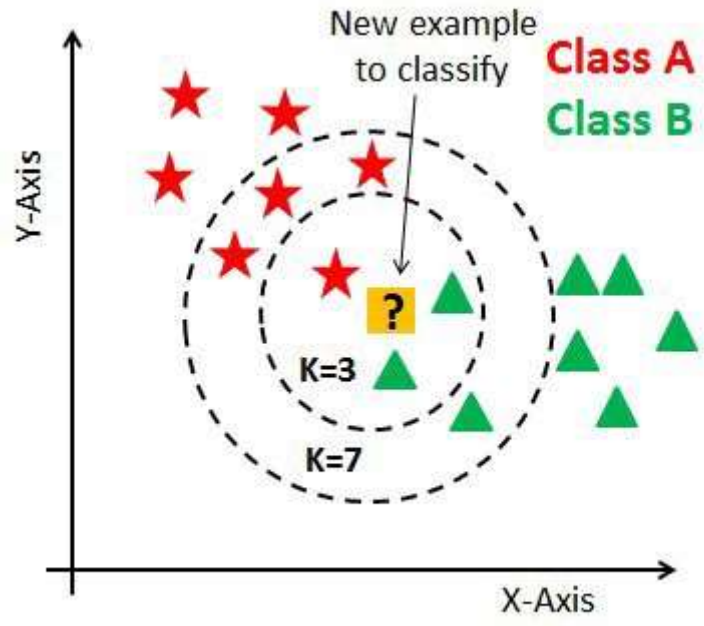
Firstly, we choose the Number of Neighbors for example we choose the K = 5.

Second, we calculate the Euclidean distance between the data points. The Euclidean distance is the distance between two points, which we have already studied in geometry. It can be calculated as:



Euclidean Distance between A_1 and $B_2 = \sqrt{(X_2-X_1)^2+(Y_2-Y_1)^2}$

K value indicates the number of the nearest neighbors. By increasing the K number, it is computationally expensive, and that's why the KNN algorithm are also called lazy learning algorithm. This negative point lead to not used widely in different cases.



For the Selecting the K values there is no pre-defined statistical method to find K value. But it is better to initialize a random K value and start computing.

When use the KNN:

a. We have label data.
b. Data is error free
c. Dataset size is small.

Advantages

1. KNN algorithm is simple and easy to implement.
2. It can be used for classification, regression

Disadvantages

1. The algorithm gets significantly slower as the number of examples and/or predictors/independent variables increase.