K Nearest Neighbour Technique

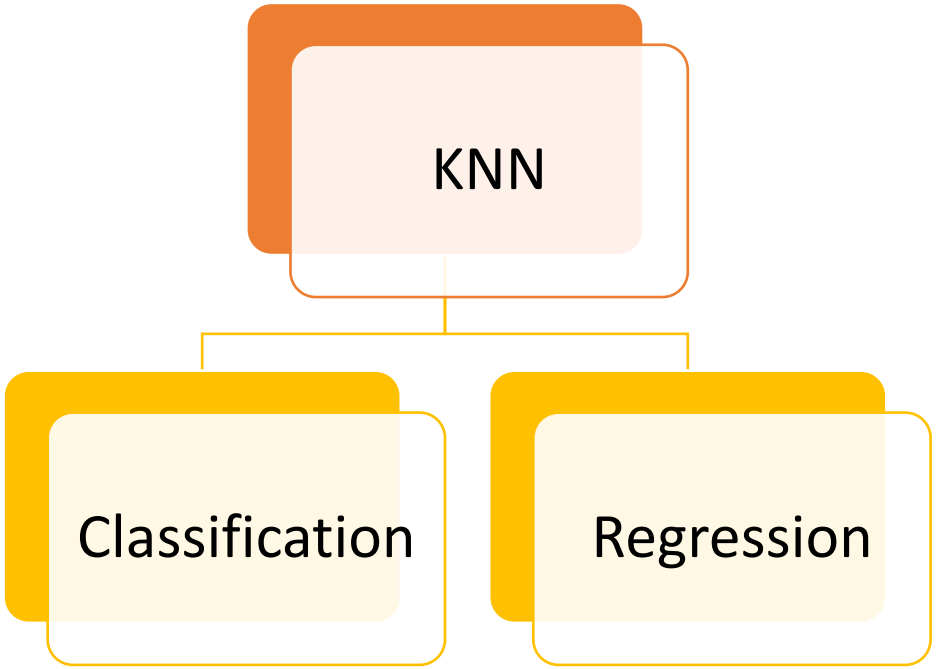
By
Sapna Yadav

Content…..

- **Nearest Neighbor Approach**
- **K Nearest Neighbor algorithm**
- **Distance Metrics used in KNN**
- **Working of KNN with Example**
- **Selection of Nearest Neighbor Parameter K**
- **Why use KNN**
- **Advantages & Disadvantages of KNN**

K Nearest Neighbor algorithm

K Nearest Neighbor algorithm is a Supervised Learning technique, which can be used for classification and regression. Most commonly it is used for the Classification problems.

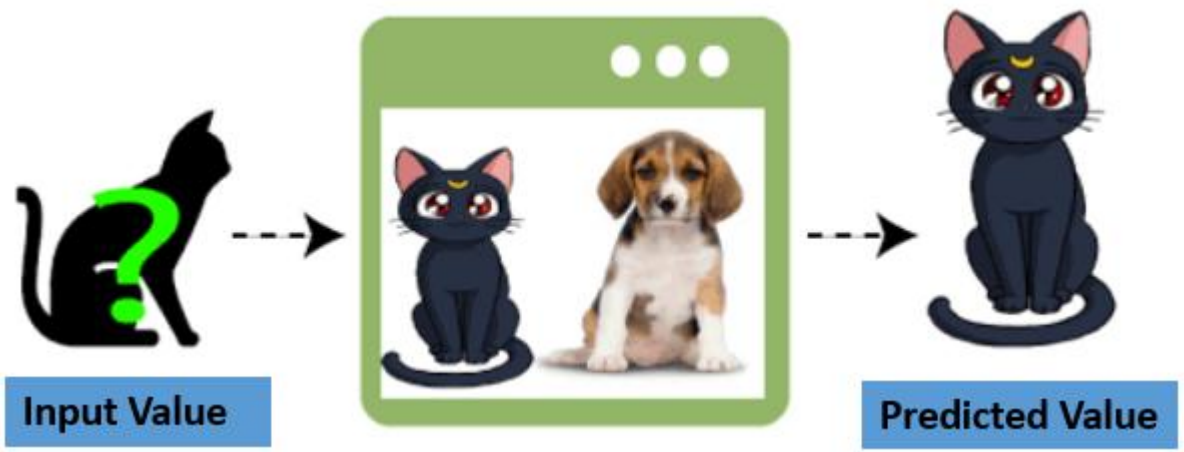


K Nearest Neighbor algorithm

- K Nearest Neighbor examines K Nearest data points to estimate the class or continuous value for a new Data-point, as the name says.

- The K-NN method assumes that the new data and existing dataset are comparable and places the new data in the category that is most similar to the existing categories.

- The K-NN method saves all available data and classifies a new data point based on its similarity to the existing data. This means when new data appears then it can be easily classified into a well suite category by using K- NN algorithm.



KNN's learning

1. **Instance-based learning:** We use whole training instances to predict output for unseen data, rather than learning weights from training data to predict output (as in model-based techniques).

2. **Lazy Learning:** KNN is a lazy learning algorithm because it does not have a specialised training phase and instead uses all of the data for training while classification, i.e. the model is not learned using training data prior to classification and the learning process is postponed until the prediction is demanded on the new instance.

3. **Non –Parametric learning:** There is no preset form of the mapping function in KNN, which means it makes no assumptions about the underlying data.

KNN Algorithm

Step 1: In order to use any machine learning algorithm, a dataset is required. So the first step is to load training and test data.

Step 2: Second step is to choose the integer value K, which is the number of nearest data points needs to consider.

Step 3: For every sample data points in the test data set, perform the following steps:

3.1 – Calculate the distance between the test samples and each data sample of the training set.

3.2 – On the basis of distance value sort them in increasing order.

3.3 – Then select the top K rows from the sorted list.

3.4 – Based on most frequent class of top K rows, the class to the test data will be assigned.

Distance Metrics

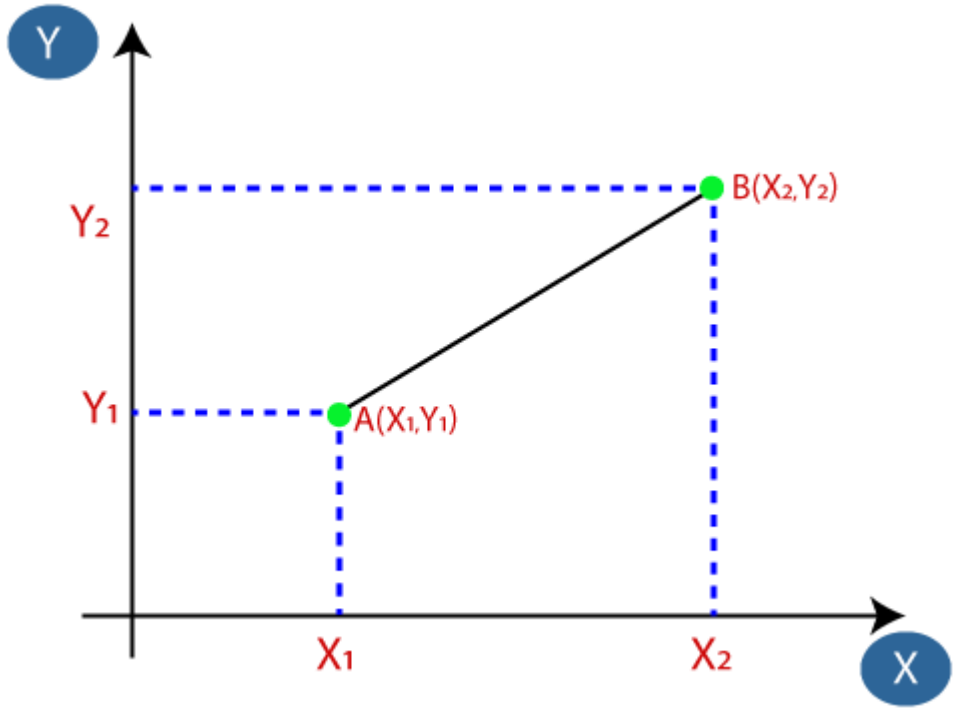
Four distance metrics are used to calculate distance between the test data point and training data sample:

1. **Euclidean distance**,
2. **Manhattan distance**,
3. **Minkowski distance**,
4. **Hamming distance**.

Out of the them, most commonly used distance metric is Euclidean distance.

Euclidian Distance

According to the Euclidean distance formula, the distance between two points in the plane with coordinates (x, y) and (a, b) is given by



$$dist((X1, Y1), (X2, Y2)) = \sqrt{(X1 - X2)^2 + (Y1 - Y2)^2}$$

$$\begin{aligned} dist((2, -1), (-2, 2)) &= \sqrt{(2 - (-2))^2 + ((-1) - 2)^2} \\ &= \sqrt{(2 + 2)^2 + (-1 - 2)^2} \\ &= \sqrt{(4)^2 + (-3)^2} \\ &= \sqrt{16 + 9} \\ &= \sqrt{25} \\ &= 5. \end{aligned}$$

Working with an Example

Given Query → X = (Maths = 5, Science = 7), K = 3

Sr. No.	Maths	Science	Result
1	3	2	Fail
2	5	6	Pass
3	6	7	Pass
4	4	4	Fail
5	7	7	Pass

On the basis of Euclidean distance, we will measure the distance between X and its neighbors, and find the k nearest neighbors.

Working with an Example

For the given data and query:

1. $\sqrt{(5 - 3)^2 + (7 - 2)^2} = \sqrt{29} = 5.38$
2. $\sqrt{(5 - 5)^2 + (7 - 6)^2} = 1$
3. $\sqrt{(5 - 6)^2 + (7 - 7)^2} = 1$
4. $\sqrt{(5 - 4)^2 + (7 - 4)^2} = \sqrt{10} = 3.16$
5. $\sqrt{(5 - 7)^2 + (7 - 7)^2} = 2$

For K=3, we have three nearest neighbors for the test data X, i.e. Second, third, and fifth samples.

For these three samples, we have three Pass and zero fail. So accordingly, the class of unseen test data X is to be predicted as Pass.

Selection of K

- There is no particular method of selecting the value of K. However domain knowledge is also play very important role in selecting value of K.
- The value of K should not be very less (eg. K=1 or K=2). It may add noise and lead to outlier effect on the model.
- And larger K values will produce the better decision boundary. But K should not be too large. Otherwise, sets with less number of data points will always be outvoted by other sets.
- The square root of N is considered as the optimal value of K, Where N is the total number of samples in the dataset.
- The most preferred value of K can be obtained by using error plot or accuracy plot on different values of K.
- For binary classification odd K value should be preferred.

When KNN is Used

We can use KNN for some reasons given below:

Labelled Data

The labelled data is the dataset where the result of the data is already known. And on the basis of which, e try to build the model to predict/ classify the unknown data.

Small Dataset

Dataset should not be much large as KNN is lazy learner and may underperform for large dataset. And also may not learn the discriminative function from the given training dataset.

Noise Free Data

Noise is the data, record, or features which do not contribute in capturing relation in between the features and the target data. Noise can be present in the form of anomalies, irrelevant features or records.

Advantages

- A simple algorithm that is easy to understand.
- The versatile algorithm used for both classification as well as regression.
- Used for nonlinear data, as there is no assumption about the underlying data.
- KNN can work more effectively for the larger training data.
- For the noisy training data, it is considered as robust algorithm.

Disadvantages

- High computation cost as it requires distance calculation of test data with entire training set.
- Slow prediction, when number of training samples are more.
- Required high memory storage.
- Always need to calculate the value of K, which may not be feasible some times.

THANKS