KNN Report

Donal Loitam

July 19, 2022

Contents

1 Introduction

- K-NN algorithm is one of the easiest algorithm which can be used for Regression as well as for Classification but mostly it is used for the Classification problems.
- The KNN algorithm assumes that similar things exist in close proximity. In other words, similar things are near to each other.
- K-NN algorithm stores all the available data and classifies a new data point based on the similarity with the available datas

Example:
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 dog. So for this identification, we can use the KNN algorithm, as it works on a similarity measure. Our 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.

Figure 1: Example

2 Key Points of the Algorithm

Let's take a simple case to understand this algorithm. Suppose there are two categories : Following is a spread of orange circles and green circles . Now, suppose we have a new data point x1, so we have to classify this data point in one of the categories based on it's similarity. We can use K-NN to accomplish this classification task :

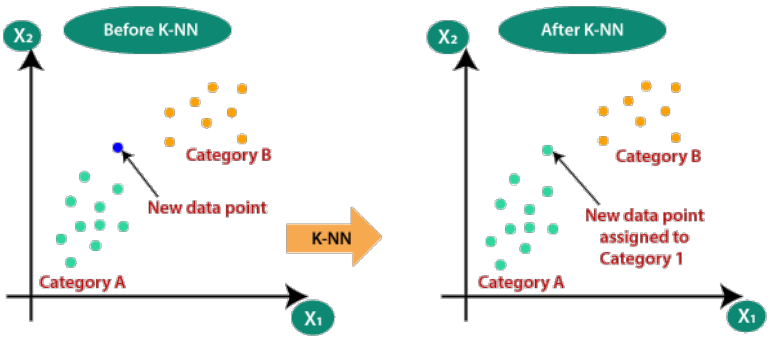


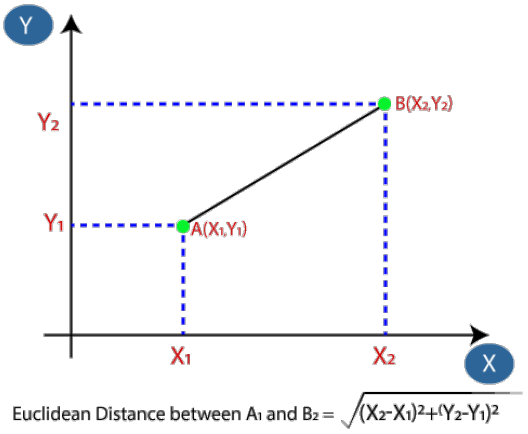
Figure 2: Classification using K-NN

But How does it work ?

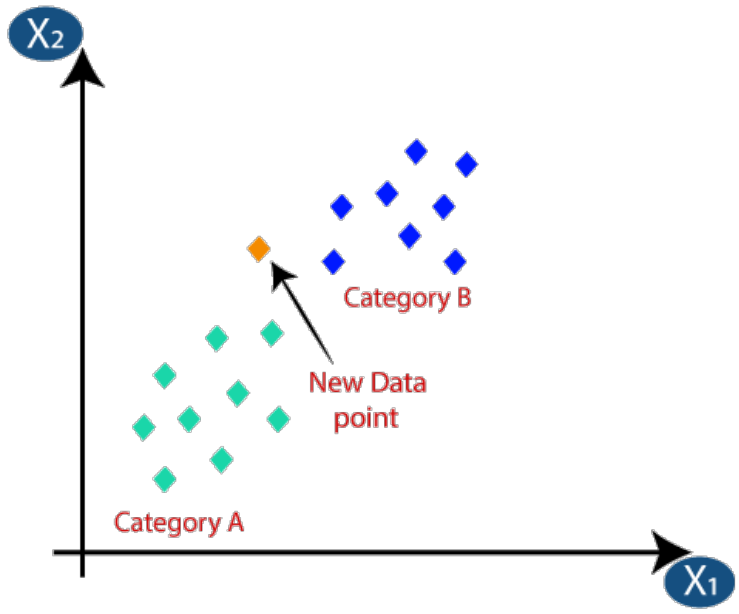
The K-NN working can be explained on the basis of the below algorithm:

- Select the number 'K' of the neighbors.(how to know what to select is discussed later)
- Calculate the Euclidean distance of the new test obseravtion from all the observation of the training dataset
- Take the K nearest neighbors as per the calculated Euclidean distance.
- Among these k neighbors, count the number of the data points in each category.
- Assign the new data points to that category for which the number of the neighbor is maximum.

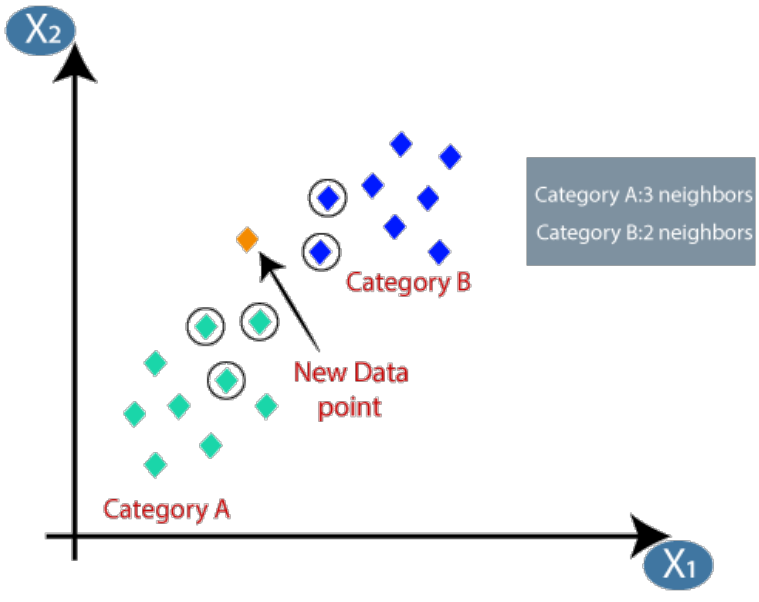
Euclidean distance: The Euclidean distance is the distance between two points calculated as:



Suppose we have a new data point and we need to put it in the required category. Consider the below image:



By calculating the Euclidean distance we got the nearest neighbors, as three nearest neighbors in category A and two nearest neighbors in category B. Consider the below image:



As we can see the 3 nearest neighbors are from category A compared to 2 from category B, hence this new data point must belong to category A.

3 Some Questions

1. How to select the value of K in the K-NN Algorithm?
Ans
2. What are the cons of KNN?
Ans.
3. Why do you need to scale your data for the k-NN algorithm?
Ans
4. Why should we not use the KNN algorithm for large datasets?
Ans.
5. How to handle categorical variables in the KNN Algorithm?
Ans