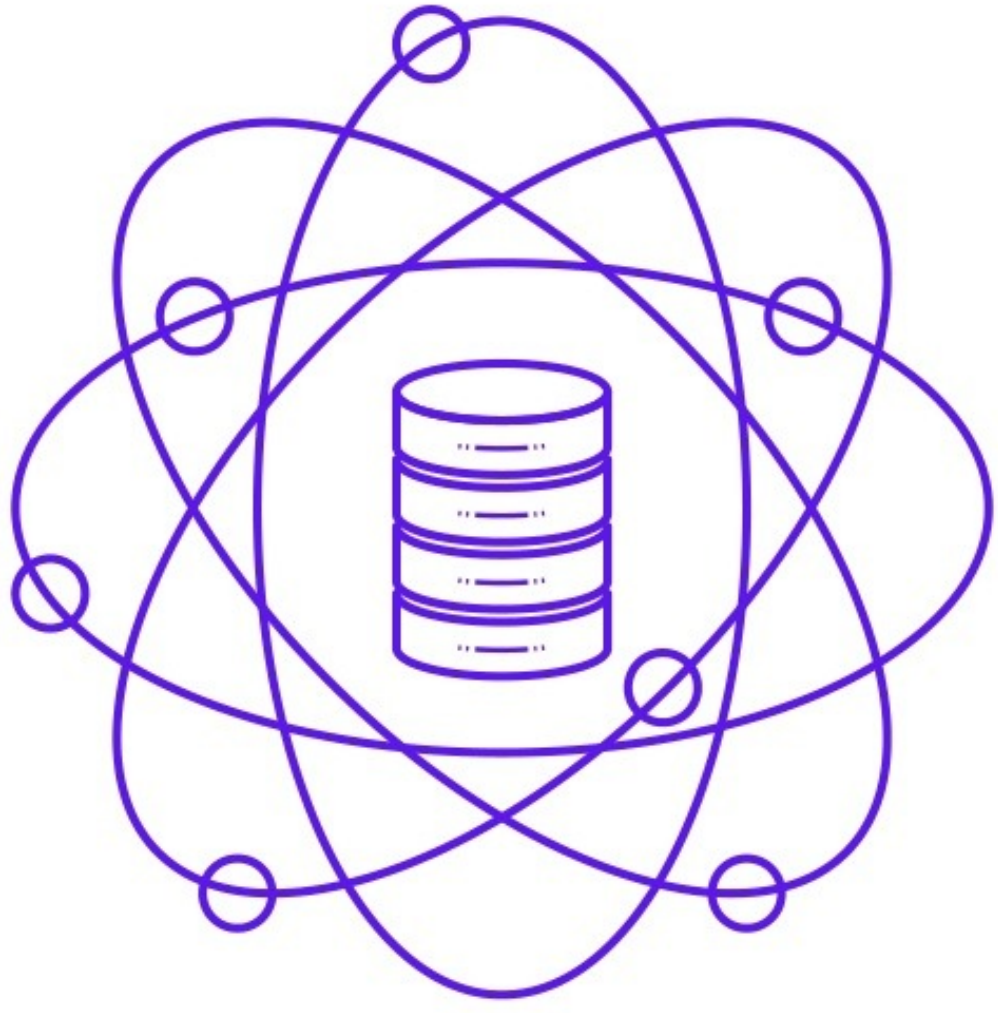
k-nearest neighbor Algorithm

Swipe

k-nearest neighbor

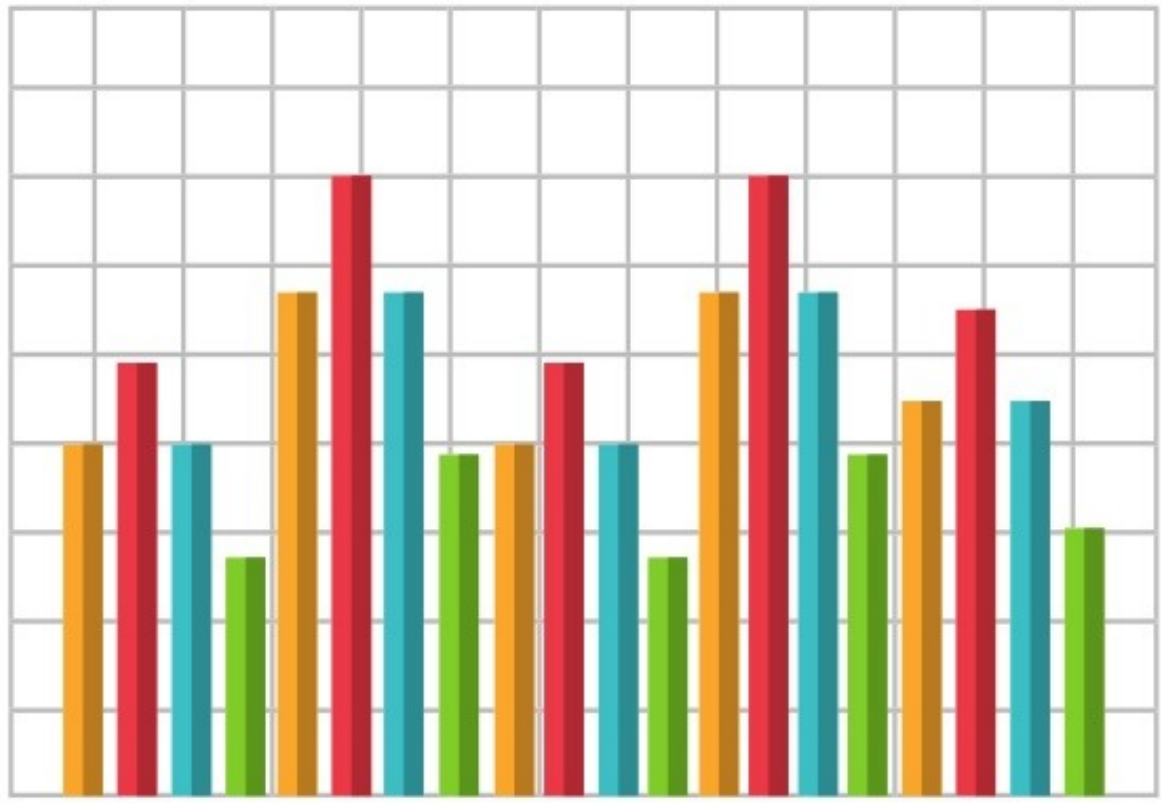
KNN stands for K Nearest Neighbor and is one of the most basic machine learning algorithms. The number K in KNN stands for the number of nearest neighbours we utilised to categorise fresh data points.



k-nearest neighbor Algorithm

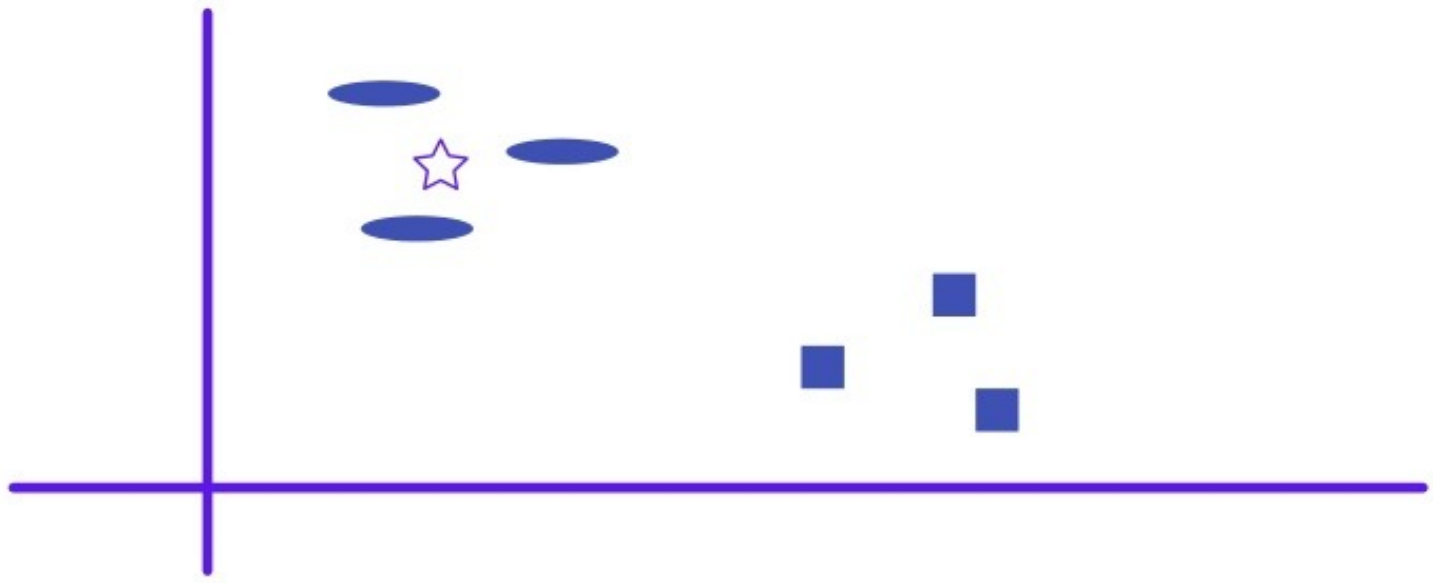
KNN can be used for both classification and regression predictive problems. However, it is more widely used in classification problems in the industry. To evaluate any technique we generally look at 3 important aspects:

- Ease to interpret output.
- Calculation time.
- Predictive Power.



How does the KNN algorithm work?

Let's take a simple case to understand this algorithm. Following is a spread of red circles (RC) and green squares (GS) :



You intend to find out the class of the blue star (BS). BS can either be RC or GS and nothing else. The "K" is KNN algorithm is the nearest neighbor we wish to take the vote from. Let's say K = 3. Hence, we will now make a circle with BS as the center just as big as to enclose only three datapoints on the plane.

How does the KNN algorithm work?

Refer to the following diagram for more details:

The three closest points to BS is all RC. Hence, with a good confidence level, we can say that the BS should belong to the class RC. Here, the choice became very obvious as all three votes from the closest neighbor went to RC. The choice of the parameter K is very crucial in this algorithm. Next, we will understand what are the factors to be considered to conclude the best K.

How do we choose the factor K?

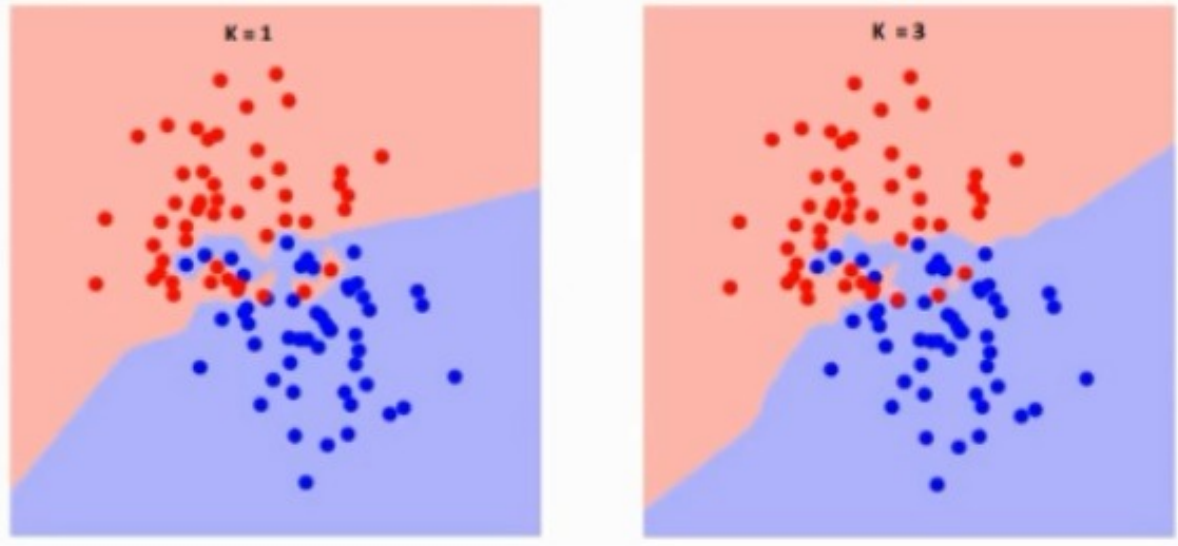
- First let us try to understand what exactly does K influence in the algorithm.

- If we see the last example, given that all the 6 training observation remain constant, with a given K value we can make boundaries of each class.

- These boundaries will segregate RC from GS. In the same way, let's try to see the effect of value "K" on the class boundaries.

- The following are the different boundaries separating the two classes with different values of K.

How do we choose the factor K?



If you watch carefully, you can see that the boundary becomes smoother with increasing value of K. With K increasing to infinity it finally becomes all blue or all red depending on the total majority.

Implementation of kNN in R

- **Step 1: Importing the data**
- **Step 2: Checking the data and calculating the data summary**
- **Step 3: Splitting the Data**
- **Step 4: Calculating the Euclidean Distance**
- **Step 5: Writing the function to predict kNN**
- **Step 6: Calculating the label(Name) for K=1**

Topics for next Post

- Neural Networks
- Similarity learning

Stay Tuned with Learnbay