k-nearest neighbors (KNN)

OVERVIEW & PURPOSE

The k-nearest neighbors (KNN) algorithm is a supervised machine learning algorithm used for classification and regression tasks.

OBJECTIVE

The main objective of the KNN algorithm is to predict the classification of a new sample point based on data points that are separated into several individual classes.

What is KNN (K-Nearest Neighbor) Algorithm?

The K-Nearest Neighbors (KNN) algorithm is a popular machine learning technique used for classification and regression tasks. It relies on the idea that similar data points tend to have similar labels or values.

During the training phase, the KNN algorithm stores the entire training dataset as a reference. When making predictions, it calculates the distance between the input data point and all the training examples, using a chosen distance metric such as Euclidean distance.

Next, the algorithm identifies the K nearest neighbors to the input data point based on their distances. In the case of classification, the algorithm assigns the most common class label among the K neighbors as the predicted label for the input data point. For regression, it calculates the average or weighted average of the target values of the K neighbors to predict the value for the input data point.

The KNN algorithm is straightforward and easy to understand, making it a

popular choice in various domains. However, its performance can be affected by the choice of K and the distance metric, so careful parameter tuning is necessary for optimal results.

Why is it used?

KNN Algorithm 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:

1. Ease of interpreting output

2. Calculation time

3. Predictive Power

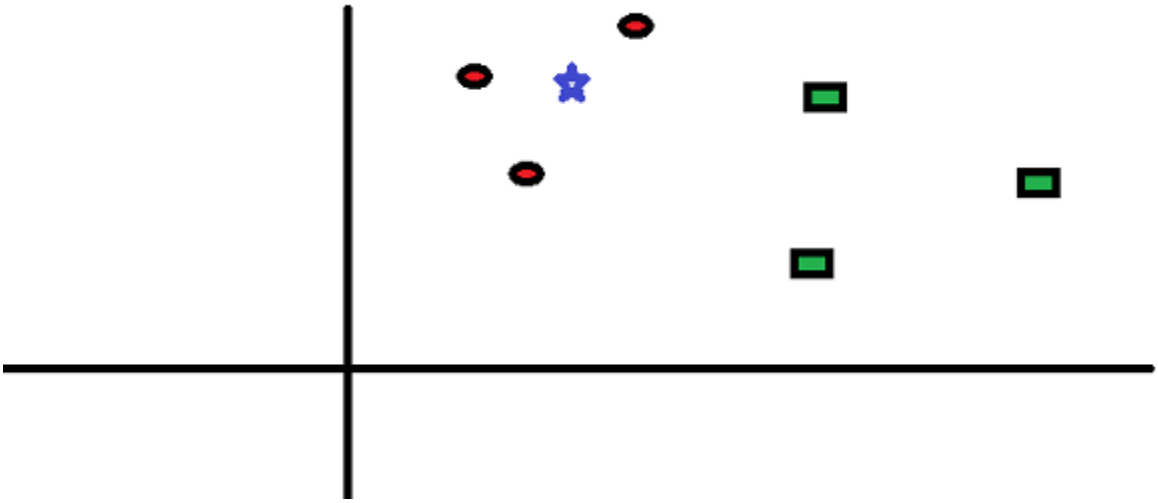
Let us take a few examples to place KNN in the scale :

	Logistic Regression	CART	Random Forest	KNN
1. Ease to interpret output	2	3	1	3
2. Calculation time	3	2	1	3
3. Predictive Power	2	2	3	2

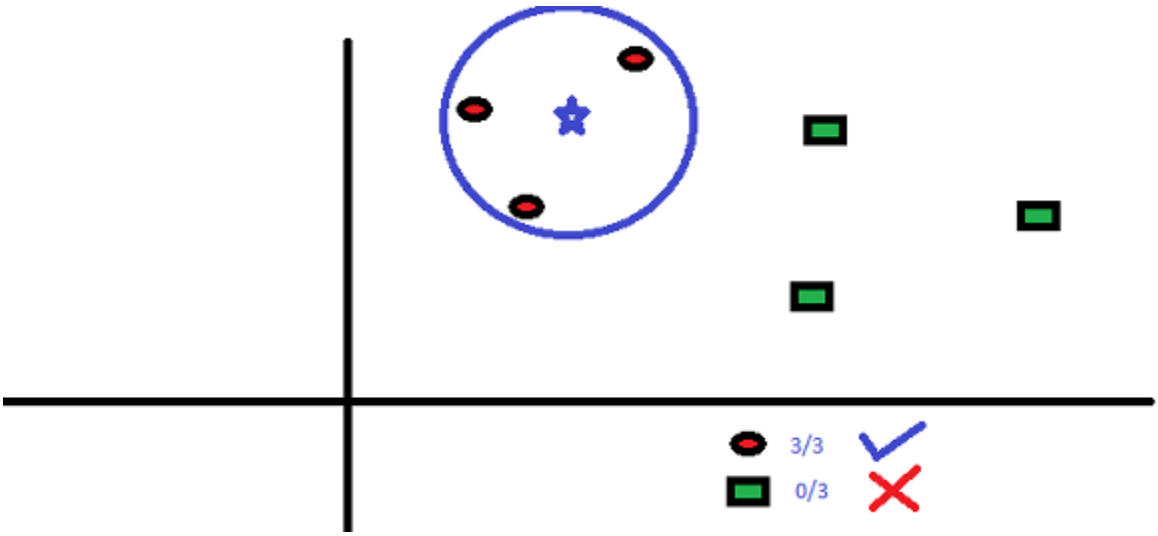
KNN classifier fairs across all parameters of consideration. It is commonly used for its ease of interpretation and low calculation time.

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” in 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 data points on the plane. Refer to the following diagram for more details:



The three closest points to BS are all RC. Hence, with a good confidence level, we can say that the BS should belong to the class RC. Here, the choice became 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 the factors to be considered to conclude the best K.

Reading Material

- https://www.datacamp.com/tutorial/k-nearest-neighbor-classification-scikit-learn
- https://serokell.io/blog/knn-algorithm-in-ml
- https://www.analyticsvidhya.com/blog/2018/03/introduction-k-neighbours-algorithm-clustering/
- https://www.usaii.org/ai-insights/understanding-knn-algorithm-and-its-role-in-machine-learning
- https://towardsdatascience.com/machine-learning-basics-with-the-k-nearest-neighbors-algorithm-6a6e71d01761?gi=d8e3a720665d
- https://www.freecodecamp.org/news/k-nearest-neighbors-algorithm-classifiers-and-model-example/

Resources

- https://www.youtube.com/watch?v=v5CcxPiYSlA
- https://www.youtube.com/watch?v=abnL_GUGub4
- https://www.youtube.com/watch?v=0p0o5cmgLdE