KNN/ K-Nearest Neighbor

Breaking the Supervised Machine Learning Algorithm into it's easiest understanding

A **supervised machine learning** algorithm (as opposed to an unsupervised machine learning algorithm) is one that relies on labeled input data to learn a function that produces an appropriate output when given new unlabeled data.

Imagine a computer is a child, we are its supervisor (e.g. parent, guardian, or teacher), and we want the child (computer) to learn what a pig looks like. We will show the child several different pictures, some of which are pigs and the rest could be pictures of anything (cats, dogs, etc).

When we see a pig, we shout "pig!" When it's not a pig, we shout "no, not pig!" After doing this several times with the child, we show them a picture and ask "pig?" and they will correctly (most of the time) say "pig!" or "no, not pig!" depending on what the picture is. That is supervised machine learning.

Supervised machine learning algorithms are used to solve classification or regression problems.

A **classification problem** has a discrete value as its output. For example, "likes pineapple on pizza" and "does not like pineapple on pizza" are discrete. There is no middle ground. The analogy above of teaching a child to identify a pig is another example of a classification problem.

The KNN algorithm assumes that similar things exist in close proximity. In other words, similar things are near to each other. The KNN algorithm hinges on this assumption being true enough for the algorithm to be useful. KNN captures the idea of similarity (sometimes called distance, proximity, or closeness) with some mathematics.

The distance between two points on the graph is calculated with the help of technique called **Euclidean distance**
Euclidean Distance=sqrt of (X1-Y1)2 + (X2-Y2)2 + (X3-Y3)2 + (XN-YN)2……

Choosing the right value of K
To select the K that's right for your data, we run the KNN algorithm several times with different values of K and choose the K that reduces the number of errors we encounter while maintaining the algorithm's ability to accurately make predictions when it's given data it hasn't seen before.

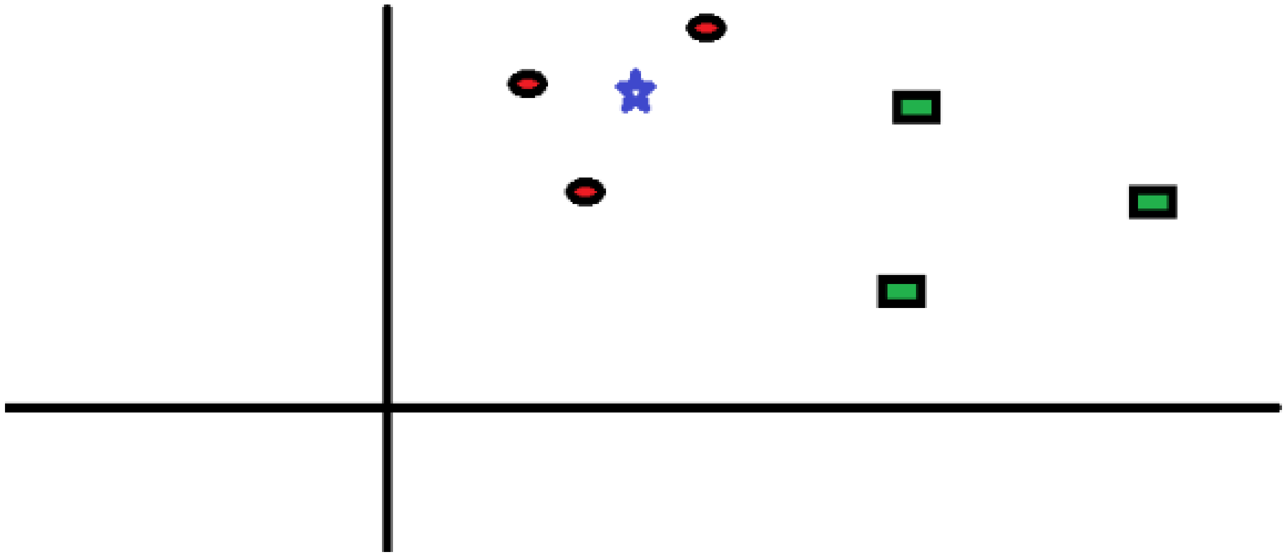
Advantages

1. The algorithm is simple and easy to implement.
2. There's no need to build a model, tune several parameters, or make additional assumptions.
3. The algorithm is versatile. It can be used for classification, regression.

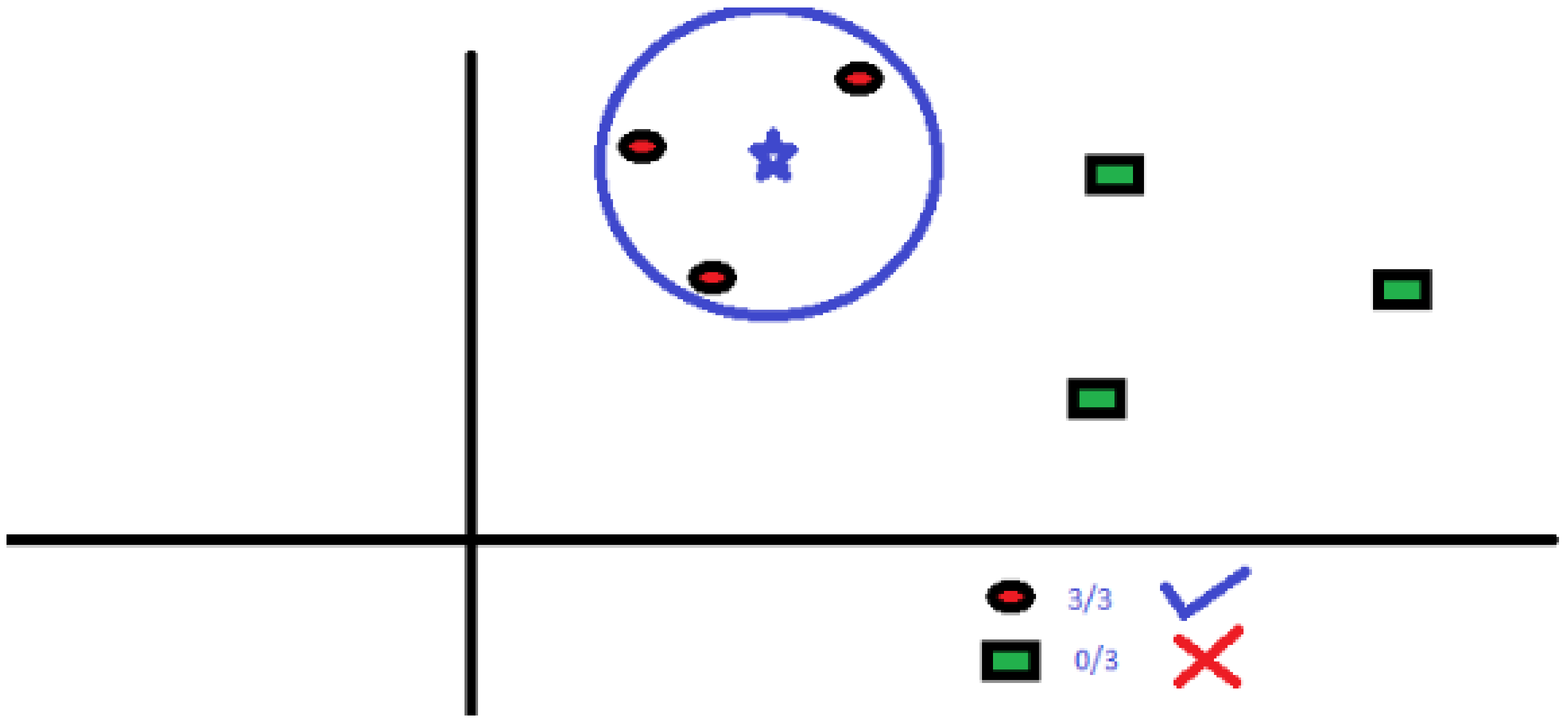
Disadvantages

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

K-Nearest Neighbor Intuition



The “K” is KNN algorithm is the nearest neighbor



In this example we took the K's value as 3 that's why the circle circumferences on 3 datapoints. The three closest points to Star is all Red Circles. Hence, with a good confidence level, we can say that the Star should belong to the class Red Circle. Here, the choice became very obvious as all three votes from the closest neighbor went to Red Circle.