Introduction to KNN (The Game of Thrones way)

Introduction

So there is a famous quote by Atal Bihari Vajpaayee

> "You can change friends but not neighbours"

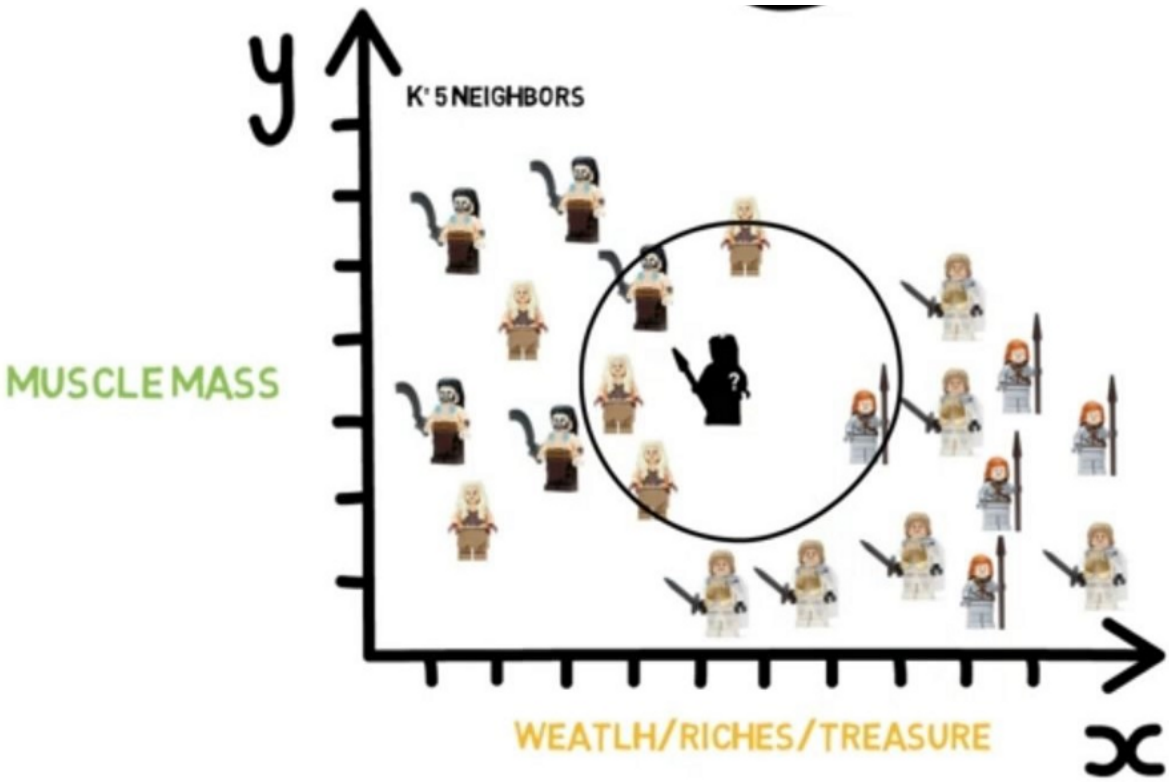
So we are often judged by the the vicinity or the group of people we live with.People belonging to a particular group can be termed similar with the characteristics they possess.

It has been around 5 months and I have been learning few supervised machine learning algorithms like Linear,Logistic,Decision Tree,Random Forest etc.While learning these algorithms I came across a very interesting algorithm named KNN. So what is KNN? **k-nearest neighbour algorithm** (**k-NN**) is a non-parametric method used for classificationand regression.When we say non-parametric,it means the algorithm doesn't make any assumptions based on underlying data distribution.In other words model structure is determined from data.Therefore KNN can be used in scenarios where there is little or no prior knowledge about distribution data.

So how does KNN work?

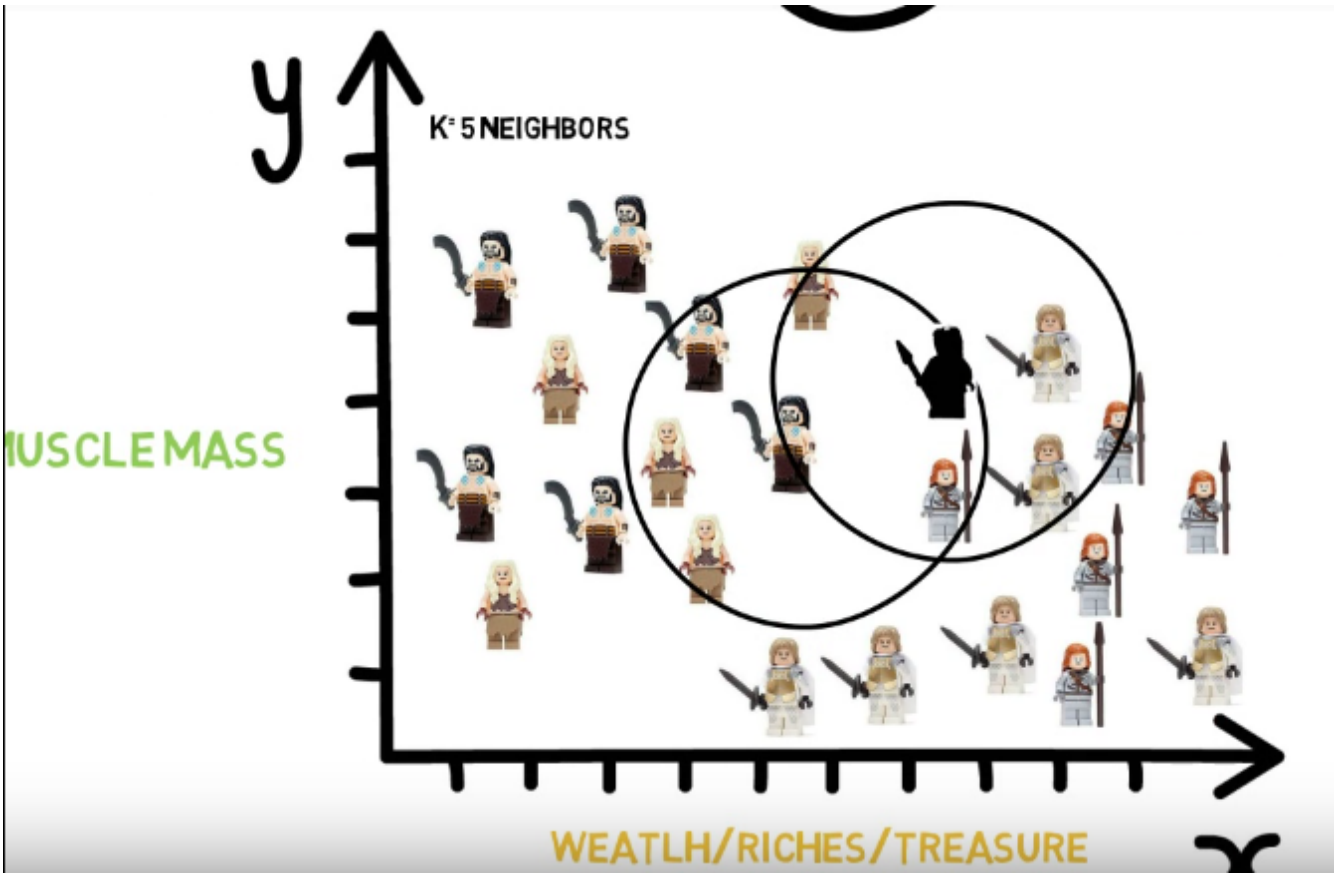
So we read that KNN actually structures the model based on the data.So lets take an example of Game of Thrones to get acquainted with the super classification algorithm.

Suppose we have to design a ultimate classifier for either an unknown person is a Dothraki or a Westerosian.We can use two features to classify or predict which clan the person belongs to.So for example,We can use muscle mass,wealth and riches as our independent variable and features.



So for the bulk of Dothraki lets assume that they have a greater muscle mass and for the westerosians they are high in wealth/riches,but their muscle mass is significantly lower than that of a dothraki.

So for our unknown person we have placed him exactly between the bulk of dothraki and westeronian.And then using k=5 we drew a circle until it has 5 neighbours in the vicinity.Now using nearest proximity we can see that the unknown person has 3 neighbours in the vicinity who belong to Dothraki and 1 neighbour who belong to westerosians.Thus using majority voting we can classify that the person is a Dothraki.



Similarly if we take a second example and draw a circle around the unknown person we can surely make out from the majority that the person is a Westerosian.

There are two important decisions that must be made before making classifications. One is the value of **k** that will be used; this can either be decided arbitrarily, or you can try **cross-validation** to find an optimal value. The next, and the most complex, is the **distance metric** that will be used.

Distance metrics are a method to find the distance between new data point and existing training dataset.

Distance functions

Euclidean $\sqrt{\sum_{i=1}^{k}(x_i - y_i)^2}$

Manhattan $\sum_{i=1}^{k}|x_i - y_i|$

Minkowski $\left(\sum_{i=1}^{k}(|x_i - y_i|)^q\right)^{1/q}$

Euclidean distance is probably the one that you are most familiar with; it is essentially the magnitude of the vector obtained by subtracting the training data point from the point to be classified.Other two popular distance functions are Manhattan and Minkowski.

End Notes

KNN algorithm is one of the simplest classification algorithm. Even with such simplicity, it can give highly competitive results. KNN algorithm can also be used for regression problems. The only difference from the discussed methodology will be using averages of nearest neighbors rather than voting from nearest neighbors.