June 5, 2020

Professional training 2

Salma elsayed 17105510

1. **What is the k value in knn?**

'K' in KNN is a parameter that refers to the number of nearest neighbors to include in the majority of the voting process. Suppose, if we add a new glass of wine in the dataset.

1. **What does the k value stand for in a knn model?**

K actually is the number of neighbors to be considered.

1. **How do you select the value of k in knn?**

* Choice of k is very critical – A small value of k means that noise will have a higher influence on the result. A large value make it computationally expensive and kind of defeats the basic philosophy behind KNN (that points that are near might have similar densities or classes) .A simple approach to select k is set k = n^(1/2).
* The value of **k** is non-parametric and a general rule of thumb in choosing the value of**k**is**k = sqrt(N)/2**, where **N** stands for the **number of samples in your training dataset**. Another tip that I suggest is to try and keep the value of k odd, so that there is no tie between choosing a class but that points to the fact that training data is highly correlated between classes and using a simple classification algorithm such as k-NN would result in poor classification performance.

1. **What is a good k value?**

If training samples of similar classes form clusters, then using k value from 1 to 10 will achieve good accuracy. If data is randomly distributed then one cannot say which k value will give the best results.

1. **How is knn calculated?**
2. Determine parameter k = number of nearest neighbors.
3. Calculate the distance between the query-instance and all the training samples.
4. Sort the distance and determine nearest neighbors based on the k-th minimum distance.
5. Gather the category of the nearest neighbors.
6. Use simple majority of the category of nearest neighbors as the prediction value of the query instance.
7. **k-nearest neighbor algorithm in python:**

**Supervised Learning:**   
It is the learning where the value or result that we want to predict is within the training data (labeled data) and the value which is in data that we want to study is known as Target or Dependent Variable or Response Variable.  
All the other columns in the dataset are known as the Feature or Predictor Variable or Independent Variable.

Supervised Learning is classified into two categories:

1. **Clarification**: Here our target variable consists of the categories.
2. **Regression:** Here our target variable is continuous and we usually try to find out the line of the curve.

This algorithm is used to solve the classification model problems. K-nearest neighbor or K-NN algorithm basically creates an imaginary boundary to classify the data. When new data points come in, the algorithm will try to predict that to the nearest of the boundary line.

Therefore, larger k value means smother curves of separation resulting in less complex models. Whereas, smaller k value tends to over fit the data and resulting in complex models.

1. **How can I improve my knn?**

Therefore rescaling features is one way that can be used to improve the performance of Distance-based algorithms such as KNN.

The steps in rescaling features in KNN are as follows:

* Load the library.
* Load the dataset.
* Sneak Peak Data.
* Standard Scaling.
* Robust Scaling.
* Min-Max Scaling.
* Tuning hyper parameters.

Figure 4

Figure 3

Figure 2

Figure 1







