K Nearest Neighbours on Student Performance Data

Shubham Barudwale
School of Computing Sciences and Engineering, VIT Chennai, Tamilnadu, India 600127
Email: barudwaleshubham.dinesh2015@vit.ac.in

Abstract-

K Nearest Neighbour (KNN) algorithm is the algorithm in which the classification of the test data is done by using its neighbour's properties and calss. In KNN the training part is not consided that is why its only the observe and classify the data.

In KNN algorithm we got the accuracy of 74.3%. We will see the method how KNN works in detail in upcoming section

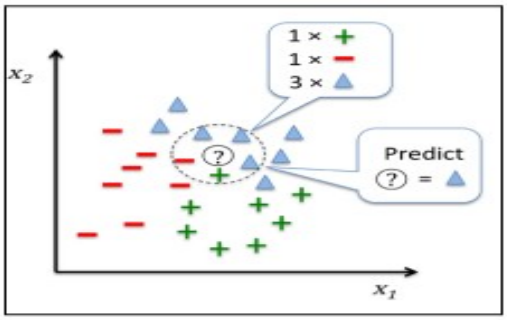
1. Introduction, Algorithm and Methodology

In KNN algorithm following steps takes place:

1. Choose the number of k and a distance metric.
2. Find the k nearest neighbors of the sample that we want to classify.
3. Assign the class label by majority vote.

K in the KNN is always odd. So that the both class numbers are not even.

The point to be classified finds the distance between all the points in the cluster and takes K nearest neighbours and finds the most number of the classes in the k neighbours and point belongs to the class which are more in numbers in k.



Once the data point is classified then that point is also included into the pool of the points to classify the further points. This is called the memory based approch. The main advantage of such a memory-based approach is that the classifier immediately adapts as we collect new training data. However, the downside is that the computational complexity for classifying new samples grows linearly with the number of samples in the training dataset in the worst-case scenario—unless the dataset has very few dimensions (features) and the algorithm has been implemented using efficient data structures such as KD-trees.

The distance is calculated by many distance formulas. But the most used formula is Euclidian distance:

$$d = [\Sigma((x_i - x_j)^2 + (y_i - y_j)^2)\ ]^{1/2}$$

Some of the varients of the KNN uses the weighted distances by giving more weights to some of the attributes. So that the classification takes place correctly.

When we consider the data in the higher dimensions the nearest points might be far away from the point to be classified.

3. Data

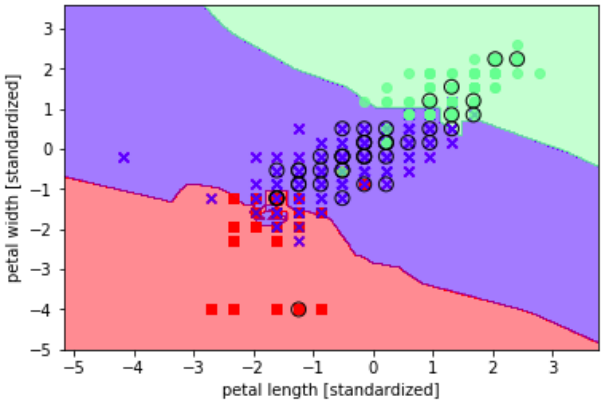
Student performance dataset is the dataset from which we are trying to gain some information from 650 student's academic performance data and various attributes which are affecting the performance. Attributes include sex, age, study time, free time, family information etc.

We will be using the Logistic regerssion method to learn and predict the performance. Out of 650 I have taken 455 data entries for lerning and rest for prediction i.e testing the performance of the algorithm.(70:30). The labels or target functions are mentioned in the 'new' field which is added manually. bad(0-7), medium(8-14), good(15-20) .

5. Experiments

We have used the algorithm for training the students data. And the accuracy on test data

was low as 74.3%. We here have used the Sklearn KNN algorithm. In which we can alter some of the parameters such as K and p to alter the power functions used in calculating the distances. As we increase the distance the classification can go wrong at some places where near the boundary the density of the poits are greater.



6. Conclusion

Algorithm of this algorithm gave fine results. Accuracy was about 74.3% while using KNN Which is the not very good accuracy for given data.

7. Referances

1) Sklearn KnearestNeighbours.

2) Python_Machine_Learning Sebastian Raschka

3) Cbapter 8, Machine Learining Tom Michell