Abdul Rehman Liaqat

Lab Course Machine Learning

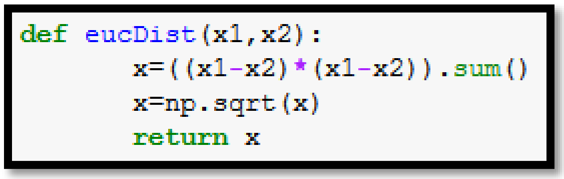
Exercise Sheet 8

Report

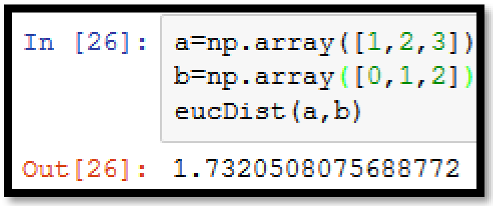
Exercise 1: Implement k-Nearest Neighbor

1- Splitting data into 70% train and 30% test using separate data. Then separated features and labels of the data using "sepxy" function. Features are normalized.

2- Implemented Euclidean Distance using "eucDist" function.

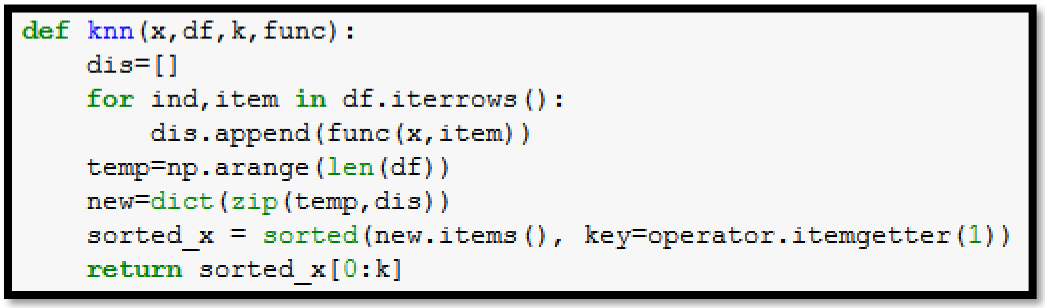


```
def eucDist(x1,x2):
        x=((x1-x2)*(x1-x2)).sum()
        x=np.sqrt(x)
        return x
```



```
In [26]: a=np.array([1,2,3])
         b=np.array([0,1,2])
         eucDist(a,b)

Out[26]: 1.7320508075688772
```

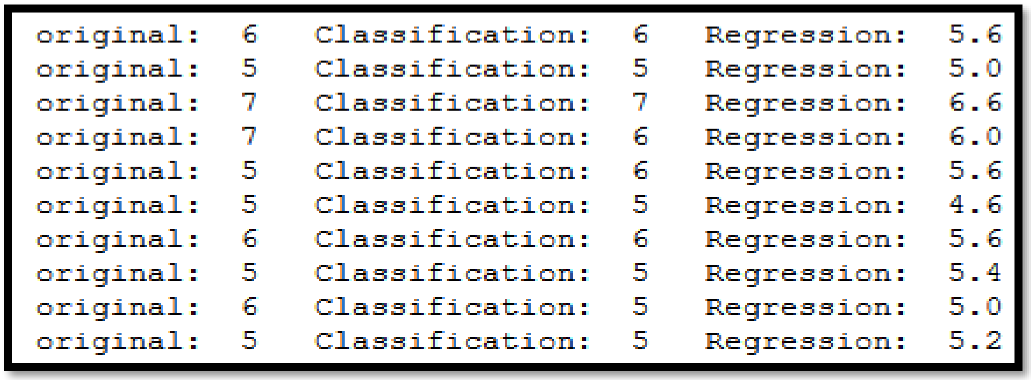
3- Implemented "knn" function to find out K-Nearest Neighbors of any data query and return k nearest neighbor in ascending order.



```
def knn(x,df,k,func):
    dis=[]
    for ind,item in df.iterrows():
        dis.append(func(x,item))
    temp=np.arange(len(df))
    new=dict(zip(temp,dis))
    sorted_x = sorted(new.items(), key=operator.itemgetter(1))
    return sorted_x[0:k]
```

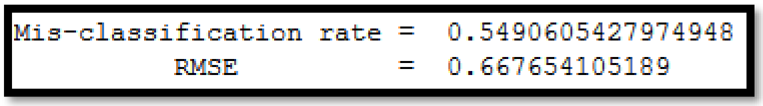
<u>**Abdul Rehman Liaqat**</u>

4- For regression prediction “reg” function and classification prediction “clas” function was implemented. Since Wine quality data was used for this part it was possible to provide both regression and classification prediction of the data. K was 5. Predictions of 10 random data wine qualities are as following:



```
original:  6   Classification:  6   Regression:  5.6
original:  5   Classification:  5   Regression:  5.0
original:  7   Classification:  7   Regression:  6.6
original:  7   Classification:  6   Regression:  6.0
original:  5   Classification:  6   Regression:  5.6
original:  5   Classification:  5   Regression:  4.6
original:  6   Classification:  6   Regression:  5.6
original:  5   Classification:  5   Regression:  5.4
original:  6   Classification:  5   Regression:  5.0
original:  5   Classification:  5   Regression:  5.2
```

5- Quality of classification based prediction is measures with misclassification rate and regression based prediction with RMSE. Both quality measures on 30% of Wine-Quality test data provide us following results:



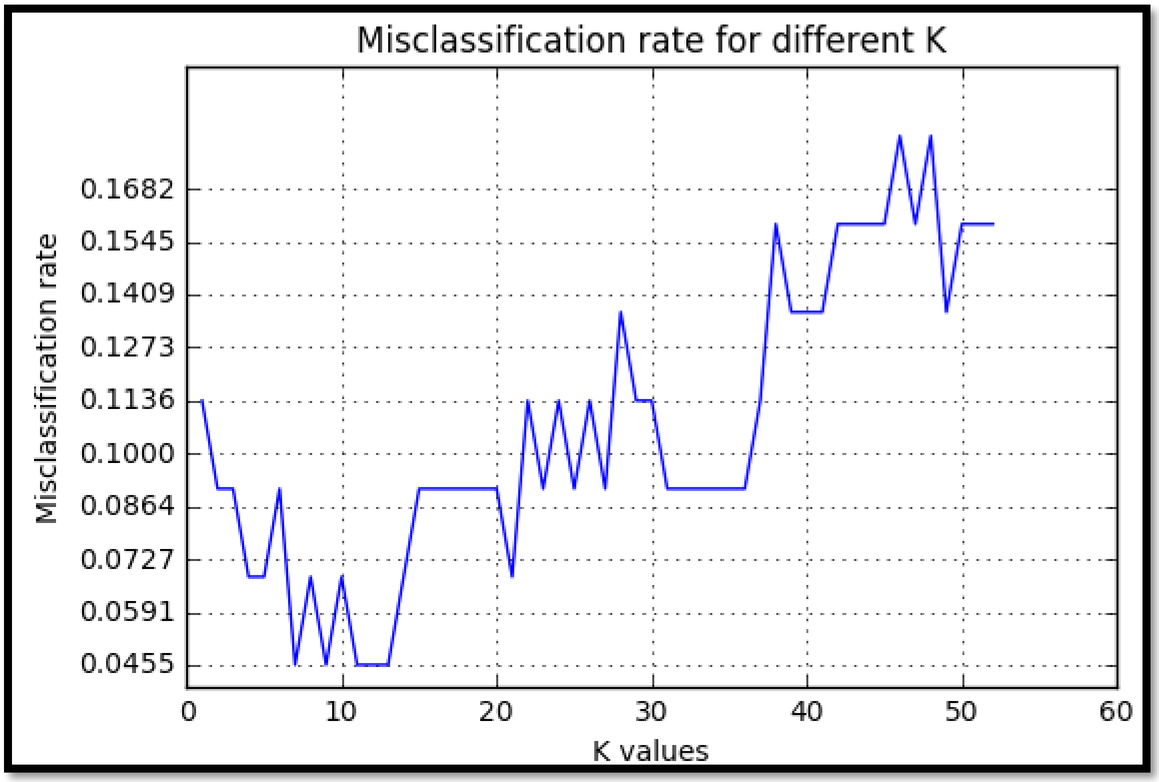
```
Mis-classification rate =  0.5490605427974948
          RMSE          =  0.667654105189
```

<u>**Abdul Rehman Liaqat**</u>

Exercise 2: Optimize and Compare KNN algorithm

<u>Part A: Determine Optimal Value of k in KNN algorithm:</u>

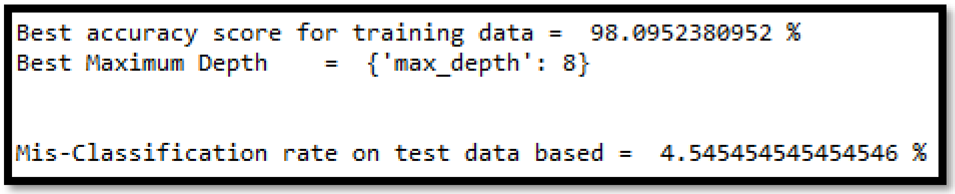
1- Iris data set was used for this part. To choose an optimal value of K a range of K values can be tested and best k will be selected. Quality criterion for the selection of k is misclassification rate hence the k provide least misclassification rate will be selected.
2- A range of K (1 to half of total training data set) was selected. Misclassification rate against each K value was recorded.
3- Since misclassification rate was used as decision criterion it was possible to have more than 1 best k values as shown in the following plot:



Abdul Rehman Liaqat

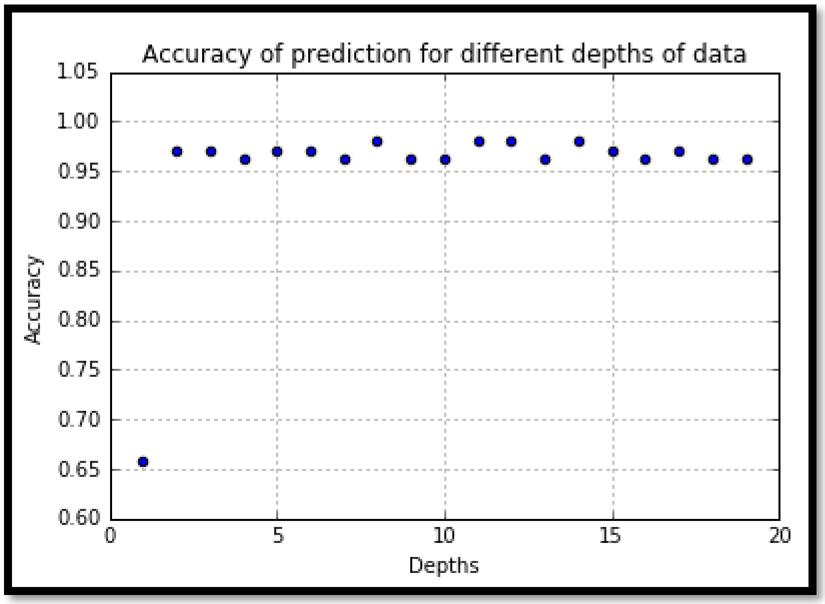
Part B: Compare KNN algorithm with Tree based method:

1- Iris Data set was used in the part as well. First decision tree classifier was implemented using Scikit-learn DecisionTreeClassifier function. Decision criterion was information gain and maximum depth was regularization term and grid search for optimal maximum depth. A range of depth (1 to 20) was selected. After grid search and cross validation following results was obtained:



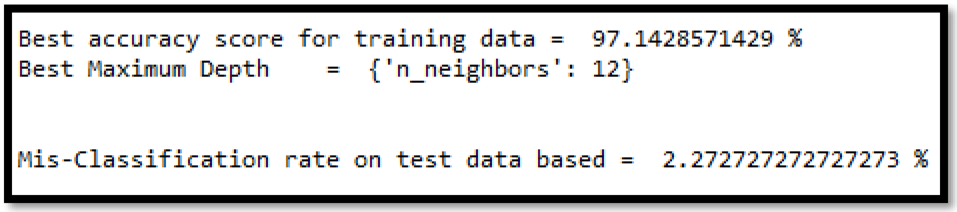
```
Best accuracy score for training data =  98.0952380952 %
Best Maximum Depth     =  {'max_depth': 8}


Mis-Classification rate on test data based =  4.545454545454546 %
```

Accuracy of prediction for all depths provide us following scatter graph:



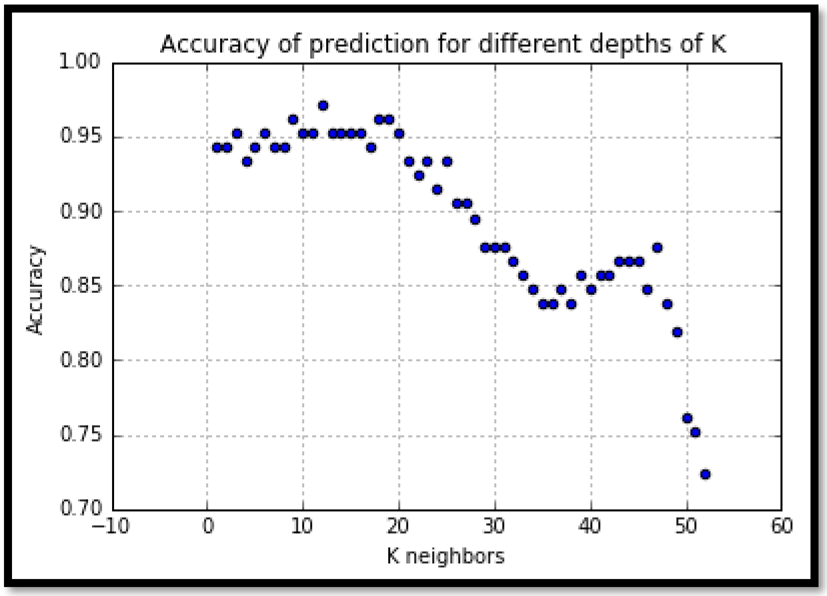
Abdul Rehman Liaqat

Similarly for KNN classifier algorithm was optimized against different K as hyperparameter. The range of K was from 1 to half of total training data points (52). Following results were obtained:



```
Best accuracy score for training data =  97.1428571429 %
Best Maximum Depth     =  {'n_neighbors': 12}


Mis-Classification rate on test data based =  2.272727272727273 %
```

Ultimately KNN classifier was providing better results.



<u>**Abdul Rehman Liaqat**</u>

Bonus: Recommender System using similarity measures