

**Ankur Ugra**

Project report

Data mining projeCT

Task A

On ATNT

We ran 5-fold cross-validation (CV)using all the three classifiers: KNN, Centroid and SVM.  
Each of the K-Fold CV gave one accuracy, below I presented all the 5 accuracies and the average of them.  
Repeated this with 2-Fold CV, 3-Fold CV and 10-Fold CV for all the three classifiers.

1. **KNN**

KFold(n\_splits=2)

KNN Accuracy with 2 Fold: 0.77,0.725

Average accuracy of KNN with all folds: 0.7475

KFold(n\_splits=3)

KNN Accuracy with 3 Fold: 0.850746268657,0.789473684211,0.766917293233

Average accuracy of KNN with all folds: 0.802379082033

KFold(n\_splits=5 )

KNN Accuracy with 5 Fold: 0.925,0.8875,0.8125,0.85,0.9125

Average accuracy of KNN with all folds: 0.8775

KFold(n\_splits=10 )

KNN Accuracy with 10 Fold: 0.875,0.95,0.95,0.85,0.85,0.9,

0.925,0.925,0.925,0.9

Average accuracy of KNN with all folds: 0.905

1. **Centroid**

KFold(n\_splits=2 )

Centroid Accuracy with 2 Fold: 0.86,0.91

Average accuracy of centroid with all folds: 0.885

KFold(n\_splits=3 )

Centroid Accuracy with 3 Fold: 0.910447761194,0.90977443609,0.924812030075

Average accuracy of centroid with all folds: 0.91501140912

KFold(n\_splits=5 )

Centroid Accuracy with 5 Fold: 0.9625,0.95,0.925,0.925,0.9375

Average accuracy of centroid with all folds: 0.94

KFold(n\_splits=10 )

Centroid Accuracy with 10 Fold: 0.95,0.925,0.975,0.925,0.975,

0.85,0.95,0.95,0.9,0.925

Average accuracy of centroid with all folds: 0.9325

1. **SVM**

KFold(n\_splits=2 )

SVM Accuracy with 2 Fold: 0.02,0.02

Average accuracy of SVM with all folds : 0.02

KFold(n\_splits=3 )

SVM Accuracy with 3 Fold: 0.0,0.015037593985,0.0

Average accuracy of SVM with all folds : 0.00501253132832

KFold(n\_splits=5 )

SVM Accuracy with 5 Fold: 0.0,0.0,0.0,0.0,0.0

Average accuracy of SVM with all folds : 0.0

KFold(n\_splits=10 )

SVM Accuracy with 10 Fold: 0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0

Average accuracy of SVM with all folds : 0.0

**Knn Accuracy List :**

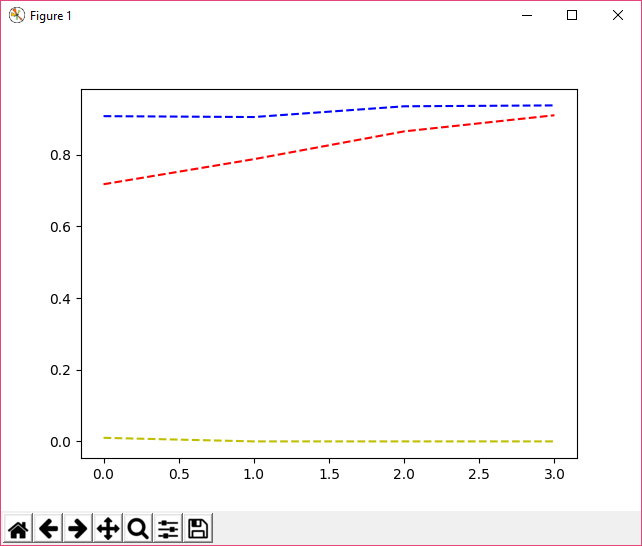
[0.75, 0.80, 0.87, 0.91]

**Centroid Accuracy List :**

[0.89, 0.92, 0.94, 0.94]

**Svm Accuracy List :**

[0.02, 0.0050, 0.0, 0.0]



**Trend :**

a)Blue dash curve shows for the centroid classifier, it is constant  
for all 2,3,5 and 10 value of folds.  
b)Red dash is for KNN classifier which is getting increased with  
increase in the number of folds.  
c)Yellow dash is for SVM classifier and we can see that the data is not essential to train the classifier hence it is not able to predict labels for the test data we are feeding to the classifier.  
d) For ATNT file KNN and Centroid are working fine and able to predict pretty good but for the SVM it is not able to get trained.

Task B

On Hand Written Letter data,,

We ran 5-fold cross-validation (CV)using all the three classifiers: KNN, Centroid and SVM.  
Each of the K-Fold CV gave one accuracy, below I presented all the 5 accuracies and the average of them.  
Repeated this with 2-Fold CV, 3-Fold CV and 10-Fold CV for all the three classifiers.

1. **KNN**

KFold(n\_splits=2)

KNN Accuracy with 2 Fold: 0.684418145957,0.721893491124

Average accuracy of KNN with all folds: **0.70315581854**

KFold(n\_splits=3)

KNN Accuracy with 3 Fold: 0.736686390533,0.769230769231,0.686390532544,

Average accuracy of KNN with all folds: **0.730769230769**

KFold(n\_splits=5)

KNN Accuracy with 5 Fold: 0.733990147783,0.714285714286,0.753694581281,

0.783251231527,0.717821782178

Average accuracy of KNN with all folds: **0.740608691411**

KFold(n\_splits=10)

KNN Accuracy with 10 Fold: 0.754901960784,0.774509803922,0.745098039216,

0.813725490196,0.693069306931,0.752475247525,

0.673267326733,0.792079207921,0.762376237624,

0.693069306931

Average accuracy of KNN with all folds: **0.745457192778**

1. **Centroid**

KFold(n\_splits=2)

Centroid Accuracy with 2 Fold: 0.688362919132,0.706114398422

Average accuracy of centroid with all folds: **0.697238658777**

KFold(n\_splits=3)

Centroid Accuracy with 3 Fold: 0.724852071006,0.683431952663,0.718934911243

Average accuracy of centroid with all folds: **0.709072978304**

KFold(n\_splits=5)

Centroid Accuracy with 5 Fold: 0.71921182266,0.743842364532,0.729064039409

0.724137931034,0.623762376238

Average accuracy of centroid with all folds: **0.708003706775**

KFold(n\_splits=10)

Centroid Accuracy with 10 Fold: 0.696078431373,0.754901960784,0.764705882353,

0.78431372549,0.633663366337,0.722772277228,

0.683168316832,0.762376237624,0.732673267327,

0.722772277228

Average accuracy of centroid with all folds: **0.725742574257**

1. **SVM**

KFold(n\_splits=2)

SVM Accuracy with 2 Fold: 0.696252465483,0.682445759369

Average accuracy of SVM with all folds : **0.689349112426**

KFold(n\_splits=3)

SVM Accuracy with 3 Fold: 0.751479289941,0.733727810651,0.724852071006

Average accuracy of SVM with all folds : **0.736686390533**

KFold(n\_splits=5)

SVM Accuracy with 5 Fold: 0.738916256158,0.753694581281,0.743842364532,

0.79802955665,0.772277227723

Average accuracy of SVM with all folds : **0.761351997269**

KFold(n\_splits=10)

SVM Accuracy with 10 Fold: 0.843137254902,0.774509803922,0.745098039216,

0.78431372549,0.752475247525,0.792079207921,

0.762376237624,0.762376237624,0.752475247525,

0.762376237624

Average accuracy of SVM with all folds : **0.773121723937**

**Knn Average Accuracy List :**

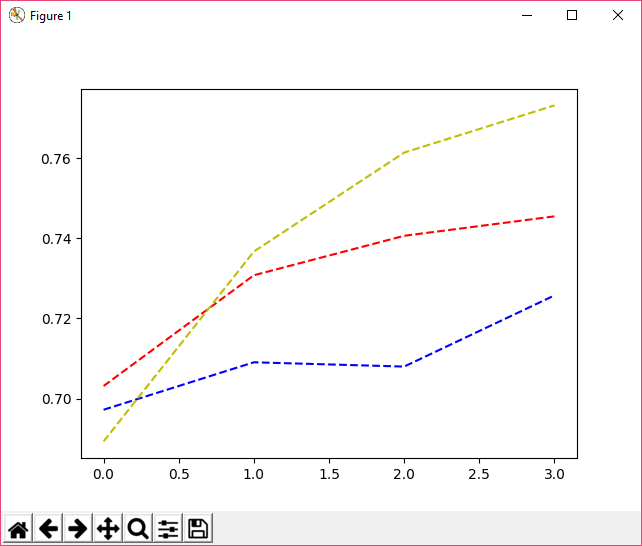
[0.70, 0.73, 0.74, 0.75]

**Centroid Average Accuracy List :**

[0.70, 0.71, 0.71, 0.73]

**Svm Average Accuracy List :**

[0.69, 0.74, 0.76, 0.77]



Trend :

1. SVM Accuracy is getting increase   
   with the increase in number of  
   folds, and very less when 2 or 3.
2. KNN Accuracy is also getting increase  
   with the increase in number of folds, but coming out less than SVM classfier.
3. Centroid is not showing much of variance when fold vale is 2,3 or 5 but it gradually increase when the value of fold is 10.
4. Getting eyes on graph we can say that for all the three classifiers the optimal value on which the data sets should be divided into training and testing sets should be at least 10. Poor accuracy when fold value is less.

Task D

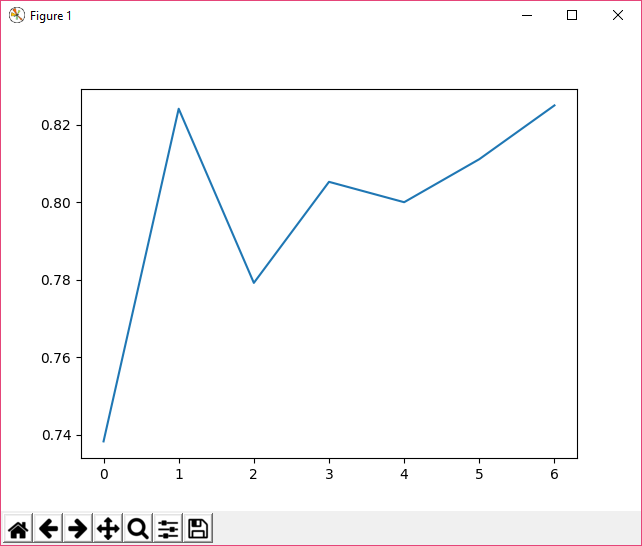
On handwritten letter data, I have used method pickData()routine to generate training and test data files.

This input is given for different splits and for each split we get an accuracy.

I have plotted all these 7 accuracies on a graph which is attached below.

**Fixed classes or input ‘abcdefghij’ which will be converted to [1,2,3,4,5,6,7,8,9,10] from charToInt(input) routine.**

**Input = ‘abcdefghij’**



**1)Split (train=5 test=34)**

Accuracy: 0.738235294118

**2)Split (train=10 test=29)**

Accuracy: 0.824137931034

**3)Split (train=15 test=24)**

Accuracy: 0.779166666667

**4)Split (train=20 test=19)**

Accuracy: 0.805263157895

**5)Split (train=25 test=14)**

Accuracy: 0.8

**6)Split (train=30 test=9)**

Accuracy: 0.811111111111

**7)Split (train=35 test=4)**

Accuracy 0.825

Trend :

1. Looking at the graph we can say that accuracy is getting increased abruptly for the Split(train=5 test=34) and then decreased for the next split.
2. Alternatively increase-decrease graph is getting plotted for increasing size of training datasets.

Task E

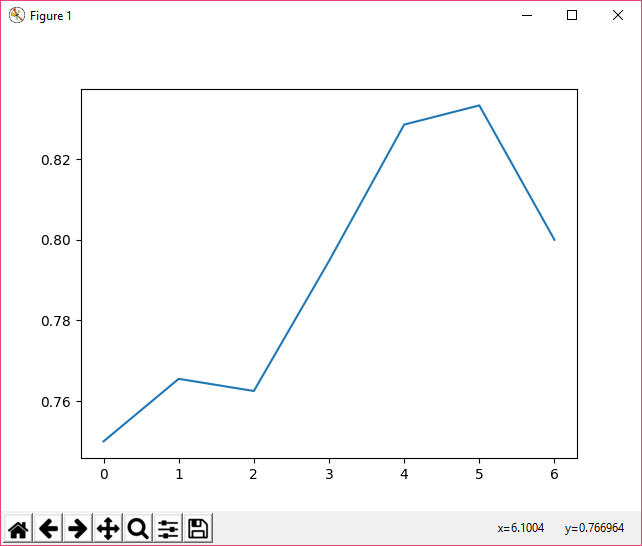
On handwritten letter data, I have used method pickData()routine to generate training and test data files.

This input is given for different splits and for each split we get an accuracy.

I have plotted all these 7 accuracies on a graph which is attached below.

**Fixed classes or input ‘klmnopqrst’ which will be converted to [11,12,13,14,15,16,17,18,19,20] from charToInt(input) routine.**

**Input = ‘klmnopqrst’**



**1)Split (train=5 test=34)**

Accuracy: 0.75

**2)Split (train=10 test=29)**

Accuracy: 0.765517241379

**3)Split (train=15 test=24)**

Accuracy: 0.7625

**4)Split (train=20 test=19)**

Accuracy: 0.794736842105

**5)Split (train=25 test=14)**

Accuracy: 0.828571428571

**6)Split (train=30 test=9)**

Accuracy: 0.833333333333

**7)Split (train=35 test=4)**

Accuracy 0.8

Trend :

1. As looking at the graph we can that when we are increasing the training dataset, the accuracy is also getting increased.
2. But for Split(train=35 test=4), the accuracy dipped for this particular input”klmnopqrst”.
3. We are getting different trend for different set of input. From A to J we are getting increase-decrease accuracy but from K to T we are getting increased accuracy for the increase size of training dataset, the only exception is at the end of the input string.