PROJECT 2

By Swapnil Sagar

Ss3854@g.rit.edu

How to Run the Program:

I have implemented K-Means using Java hence we need JDK and JRE to run the programs. Please extract all the files in the .zip file and do not change the folder structure so that the data can be read easily.

Inorder to run the program, you have to run the Project2.java file. I have placed 2 static path variables which you can use to switch between the Wine and Iris datasets. In line 14 of Project 2, when we are creating the object of kMeans, we have to specify the dType parameter as either “euclidian” or “mahanalobis”, this will specify the type of distance to be used.

Results:

The results can be a little random, in some runs we get great f1 scores while in the others it isn’t as good. This is due to the randomization of the initial centroids and sometimes they can really be in a corner. Due to which they barely get any neighbouring nodes which are close enough to them.

Program might have to be rerun multiple times and average of the results should be taken.

Here are a few snapshots of the results:

```
_________________________
0 50 0
Current State = 1
tp = 19 fp = 51  fn = 31
Results of Cluster No. 1
Accuracy = 38.0%
Precision = 27.142857142857142%
Recall = 38.0%
f1 Score = 31.666666666666664
_________________________
_________________________
0 19 31
Current State = 2
tp = 48 fp = 31  fn = 2
Results of Cluster No. 2
Accuracy = 96.0%
Precision = 60.75949367088607%
Recall = 96.0%
f1 Score = 74.4186046511628
_________________________
_________________________
1 1 48
Current State = 2
tp = 48 fp = 31  fn = 2
Results of Cluster No. 3
Accuracy = 96.0%
Precision = 60.75949367088607%
Recall = 96.0%
f1 Score = 74.4186046511628
_________________________
```

K-Means using Euclidian Distance on Iris Dataset.

```
Clustering
Clustering
Clustering Completed
_________________________

Results of Cluster No. 1
Accuracy = 100.0%
Precision = 100.0%
Recall = 100.0%
f1 Score = 100.0
_________________________
_________________________

Results of Cluster No. 2
Accuracy = 79.36507936507937%
Precision = 79.36507936507937%
Recall = 100.0%
f1 Score = 88.49557522123894
_________________________
_________________________

Results of Cluster No. 3
Accuracy = 100.0%
Precision = 100.0%
Recall = 74.0%
f1 Score = 85.05747126436782
_________________________
PS C:\Programming\RIT\Cyber Analytics\Project2>
```

k-Means using Mahanalobis Distacne on Iris Dataset

```
Clustering
Clustering
Clustering Completed
_________________________

Results of Cluster No. 1
Accuracy = 63.265306122448976%
Precision = 63.265306122448976%
Recall = 52.54237288135593%
f1 Score = 57.407407407407405
_________________________
_________________________

Results of Cluster No. 2
Accuracy = 62.745098039215684%
Precision = 62.745098039215684%
Recall = 90.14084507042253%
f1 Score = 73.98843930635837
_________________________
_________________________

Results of Cluster No. 3
Accuracy = 63.265306122448976%
Precision = 63.265306122448976%
Recall = 52.54237288135593%
f1 Score = 57.407407407407405
_________________________
PS C:\Programming\RIT\Cyber Analytics\
```

K-Means using Euclidian Distance on Wine Dataset

```
Clustering
Clustering
Clustering Completed
_________________________

Results of Cluster No. 1
Accuracy = 97.87234042553192%
Precision = 97.87234042553192%
Recall = 77.96610169491525%
f1 Score = 86.7924528301887
_________________________
_________________________

Results of Cluster No. 2
Accuracy = 72.46376811594203%
Precision = 72.46376811594203%
Recall = 70.4225352112676%
f1 Score = 71.42857142857142
_________________________
_________________________

Results of Cluster No. 3
Accuracy = 46.774193548387096%
Precision = 46.774193548387096%
Recall = 60.416666666666664%
f1 Score = 52.72727272727273
_________________________
PS C:\Programming\RIT\Cyber Analytics\Project2>
```

K-Means using Mahanalobis Distance on Wine Dataset

It was observed that Mahanalobis distance gives better accuracy for a specific class while the accuracy for other classes is low. Euclidian distance however, while specific classes don't get as high of an accuracy, it is observed that on average it is more accurate.

However, this is also due to the fact that we are using a modified version of the Mahanalobis distance which doesn’t use a covariance matrix, instead it uses a positive diagonal matrix which is randomly generated while calculating the distance.