Sai Rahul Dhulipalla

sxd170033@utdallas.edu

Abstract

Implementing K-means, Expectation Maximization, feature selection algorithm , PCA, ICA & RCA on two datasets and experimenting with Various Parameters

Assignment 4

BUAN 6341

**Introduction:**

The goal of this assignment is to implement following clustering algorithms:

1. K-means
2. Expectation Maximization

In addition, implementing the following feature dimensionality reduction algorithms

1. Any one feature selection algorithm (decision tree, forward selection, backward elimination, etc.)

2. PCA

3. ICA

4. Randomized Projections

**Dataset’s:**

The same datasets used in previous assignments were used.

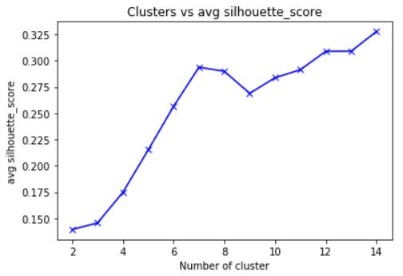
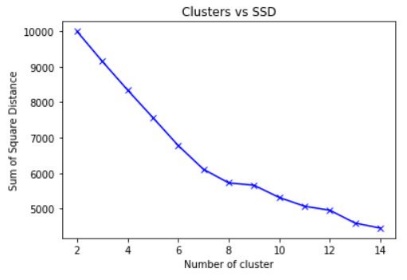
**Task 1: Run the clustering algorithms on both datasets**

**Clustering on Facebook Data:**

**K Means Clustering:**

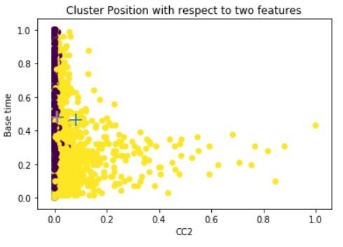
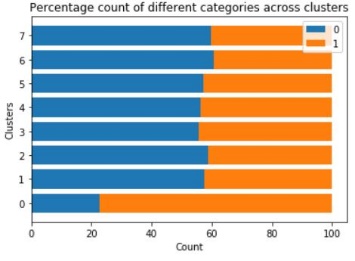
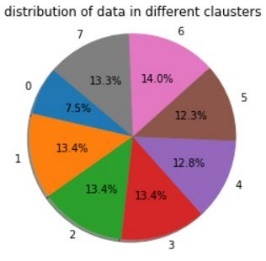
For Clustering Datasets, I choose Sklearn library.

I ran K-mean’s Clustering algorithm on data with different K values and plotted elbow curve (K Vs SSD) and (K Vs avg silhouette) to figure out optimum K values. The curves can be found below.



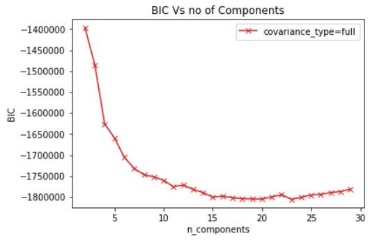
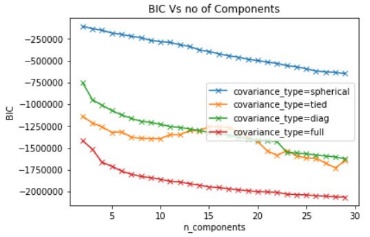
From the curves Elbow can be seen at K = 8. So, I Choose no of Clusters as 8.

* The Pie chart in Left most figure the distribution of data in different clusters. We can see that data is equally distributed in clusters.
* The Horizontal Bar charts in the middle show the distribution of different classes in each cluster. We see that in almost all the clusters approx. 60% is Class-0 and remaining is 40% is Class-1. This show that the clusters did not naturally line up with class labels.
* In the right most figure present at the bottom I plotted the Data on two dimensions with different colors for the two classes. I also plotted the centers of the clusters with different colors and they are marked with “+” sign. From this figure we can see that clusters are very compact packed. From just this visualization we cannot say weather the data is naturally clustered or not. We would need visualization on many other dimensions to comment on this.



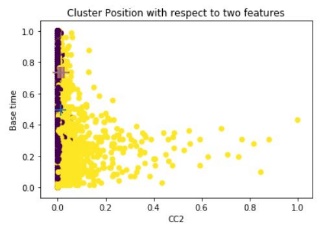
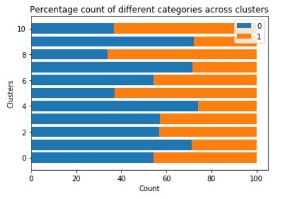
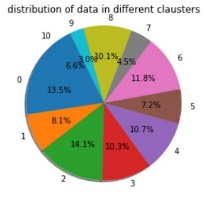
**Expectation Maximization:**

I ran GaussianMixture Clustering algorithm on data with different no of components values and plotted BIC score at each cluster to figure out optimum components. The curves can be found below.



From the curves BIC is low at 11 components and didn’t change much after that. So, I Choose no of Clusters as 11.

* The Pie chart in Left most figure the distribution of data in different clusters. We can see that data is more or less equally distributed in clusters.
* The Horizontal Bar charts in the middle show the distribution of different classes in each cluster. We see that in almost all the clusters approx. ratio of Class-0 to Class-1 is same. This show that the clusters did not naturally line up with class labels.
* In the right most figure present at the bottom I plotted the Data on two dimensions with different colors for the two classes. I also plotted the centers of the clusters with different colors and they are marked with “+” sign. From this figure we can see that clusters are very compact packed. From just this visualization we cannot say weather the data is naturally clustered or not. We would need visualization on many other dimensions to comment on this.

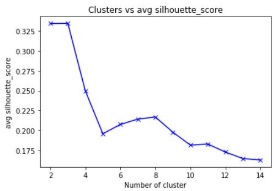
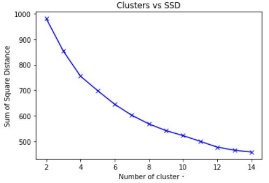


**Clustering on Spotify Data:**

**K Means Clustering:**

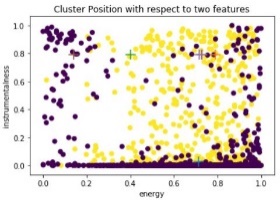
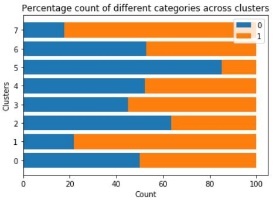
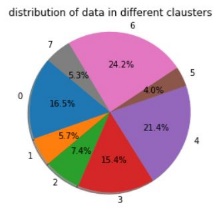
For Clustering Datasets, I choose Sklearn library.

I ran K-mean’s Clustering algorithm on data with different K values and plotted elbow curve (K Vs SSD) and (K Vs avg silhouette) to figure out optimum K values. The curves can be found below.



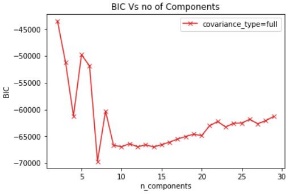
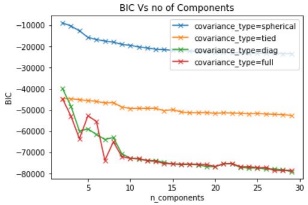
From the curves Elbow can be seen at K = 8. So, I Choose no of Clusters as 8.

* The Pie chart in Left most figure the distribution of data in different clusters. We can see that data is highly distributed in clusters – 6,4,3,0.
* The Horizontal Bar charts in the middle show the distribution of different classes in each cluster. We see that in almost all the clusters either of classes is dominant. This show that the clusters seems to be naturally line up with class labels.
* In the right most figure present at the bottom I plotted the Data on two dimensions with different colors for the two classes. I also plotted the centers of the clusters with different colors and they are marked with “+” sign. From this figure we can see that clusters are not very compact packed. From just this visualization we cannot say weather the data is naturally clustered or not. We would need visualization on many other dimensions to comment on this.



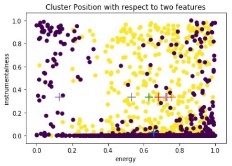
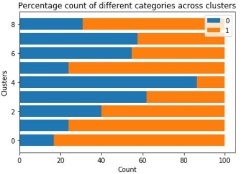
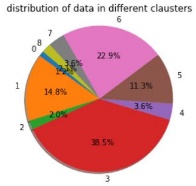
**Expectation Maximization:**

I ran GaussianMixture Clustering algorithm on data with different no of components values and plotted BIC score at each cluster to figure out optimum components. The curves can be found below.



From the curves BIC is low at 9 components and didn’t change much after that. So, I Choose no of Clusters as 9.

* The Pie chart in Left most figure the distribution of data in different clusters. We can see that data is more or less equally distributed in clusters.
* The Horizontal Bar charts in the middle show the distribution of different classes in each cluster. We see that in almost all the clusters either of classes is dominant. This show that the clusters seems to be naturally line up with class labels.
* In the right most figure present at the bottom I plotted the Data on two dimensions with different colors for the two classes. I also plotted the centers of the clusters with different colors and they are marked with “+” sign. From this figure we can see that clusters are not compact packed. From just this visualization we cannot say weather the data is naturally clustered or not. We would need visualization on many other dimensions to comment on this.

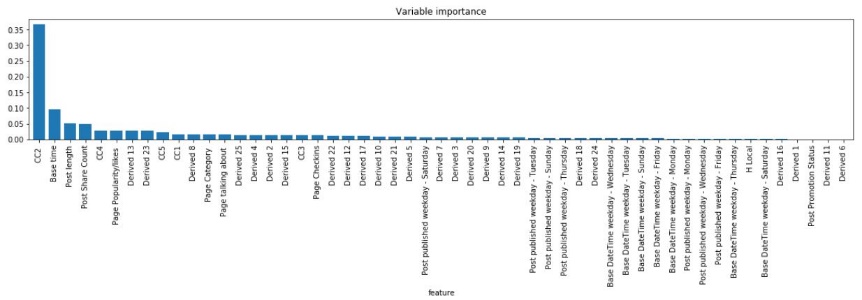


**Task 2: Run the dimensionality reduction algorithms on both datasets**

**Dimension Reduction on Facebook Data:**

**Feature Selection using Decision Trees:**

I used Decision Tree Classifier with Entropy criteria to pick the Features. The plot of Features with their importance can be found below.

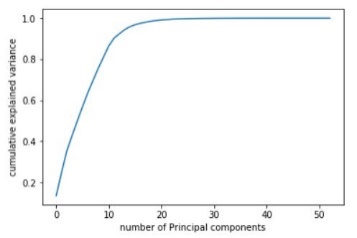
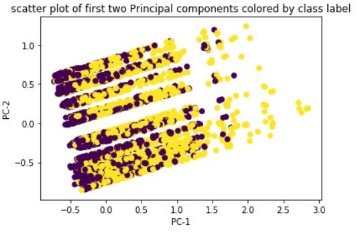


I Picked features that have more than mean importance in splitting the nodes.

So the First 9 Features in above plot were selected and the remaining features were dropped.

**Dimension Reduction using PCA:**

I ran PCA on the Data Set and Plotted the below graph of ‘cumulative explained variance’ Vs ‘number of Principal components’. The plot can be seen below.

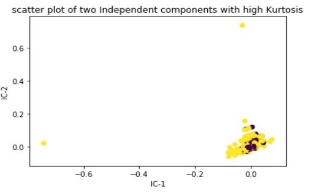
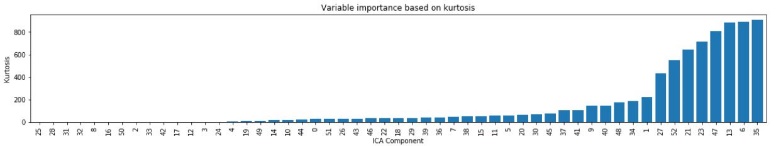
 

* From the Plot above in the right I choose the first 12 PC’s which explain 90% of variation as the reduced features .
* The above Plot in the right is the plot of data points with PC-1 and PC-2 as X and Y variables. The Data in plot Show High Variance in both dimensions.

**Dimension Reduction using ICA:**

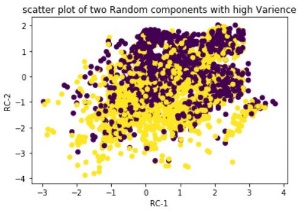
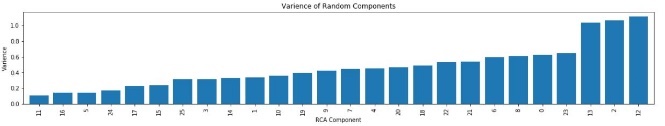
I ran ICA on the Data Set and recovered the Independent Components. To select the important IC’s from all the IC’s generated to reduce the dimensions Plotted the Kurtosis of each IC.

* I choose all the IC’s with Kurtosis > 2 as reduced Dimensions. Now the I got 41 components as reduced components.
* The below right Plot is the plot of data points with IC-1 and IC-2 as X and Y variables here IC1 & IC2 are Ic’s with high Kurtoses values. The Data in plot seemed to be concentrated in one point of the graph.



**Dimension Reduction using RCA:**

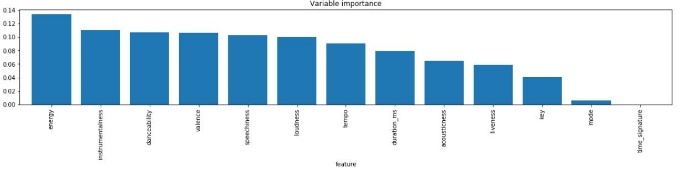
* I ran Randomly projected data into 26 reduced number of dimension. The Bar plots of each dimension variance and plot of data on top two Random components can be found below. The Data in plot Show High Variance in both dimensions.



**Dimension Reduction on Spotify Data:**

**Feature Selection using Decision Trees:**

I used Decision Tree Classifier with Entropy criteria to pick the Features. The plot of Features with their importance can be found below.

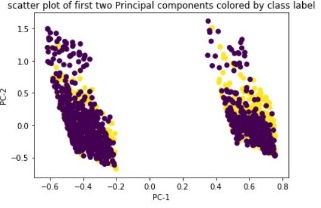
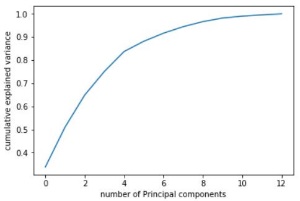


I Picked features that have more than mean importance in splitting the nodes.

So the First 8 Features in above plot were selected and the remaining features were dropped.

**Dimension Reduction using PCA:**

I ran PCA on the Data Set and Plotted the below graph of ‘cumulative explained variance’ Vs ‘number of Principal components’. The plot can be seen below.

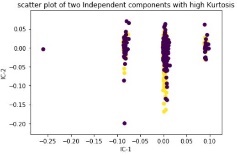
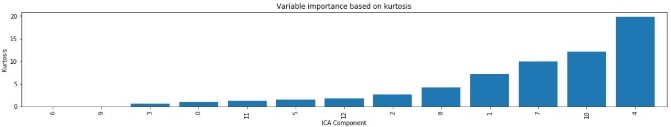


* From the Plot above in the right I choose the first 7 PC’s which explain 90% of variation as the reduced features .
* The above Plot in the right is the plot of data points with PC-1 and PC-2 as X and Y variables. The Data in plot Show High Variance in both dimensions.

**Dimension Reduction using ICA:**

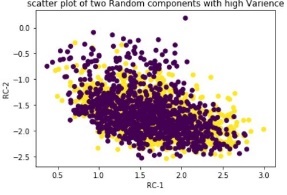
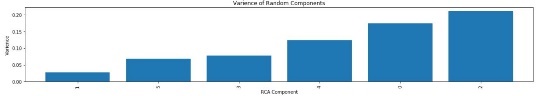
I ran ICA on the Data Set and recovered the Independent Components. To select the important IC’s from all the IC’s generated to reduce the dimensions Plotted the Kurtosis of each IC.

* I choose all the IC’s with Kurtosis > 2 as reduced Dimensions. Now the I got 6 components as reduced components.
* The below right Plot is the plot of data points with IC-1 and IC-2 as X and Y variables here IC1 & IC2 are Ic’s with high Kurtoses values. The Data in plot seemed to be concentrated in 3 point of the graph.



**Dimension Reduction using RCA:**

* I ran Randomly projected data into 6 reduced number of dimension. The Bar plots of each dimension variance and plot of data on top two Random components can be found below. The Data in plot Show High Variance in both dimensions.

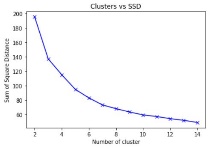
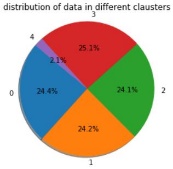
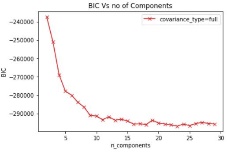
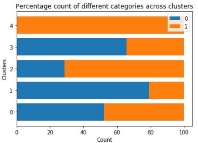
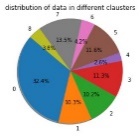
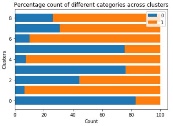


**Task 3: Clustering on Reduced Dimensions**

**Clustering on Reduced Facebook Dimensions:**

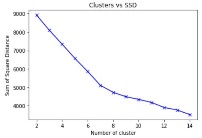
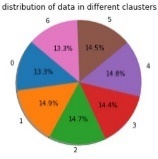
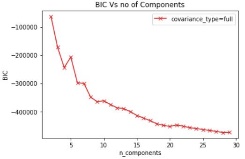
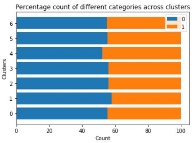
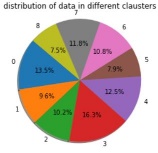
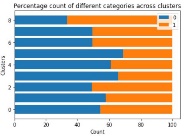
I ran K means and Expectation maximization Clustering on all the reduced data generated from different algorithms. The same plots of clustering produced above can be found below.

**Clustering on Decision Tree Dimensions:**

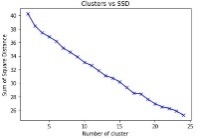
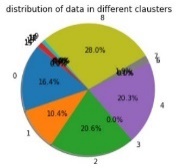
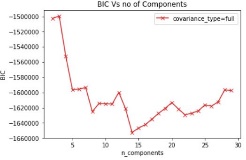
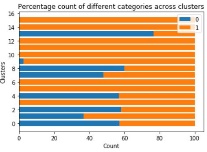
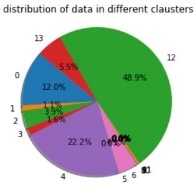
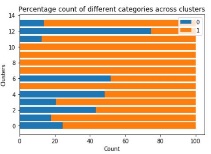
No of K-Mean Clusters: 5, No of EM Clusters: 9, These clusters seem to be different from original data clusters. This might be because the number of dim are reduced.

**Clustering on PCA Dimensions:**

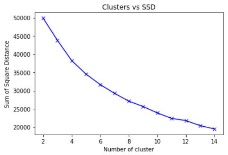
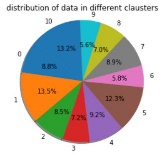
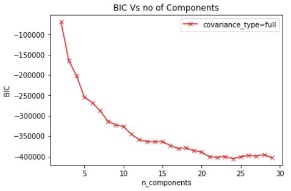
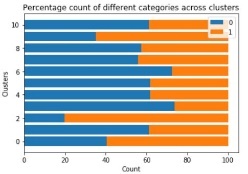
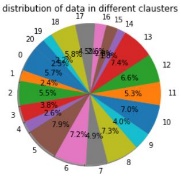
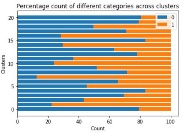
No of K-Mean Clusters: 7, No of EM Clusters: 9, These clusters seem to be different from original data clusters. This might be because the dim are of high variance.

**Clustering on ICA Dimensions:**

No of K-Mean Clusters: 16, No of EM Clusters: 14, These clusters seem to be different from original data clusters. This might be because the dim are of independent.

**Clustering on RCA Dimensions:**

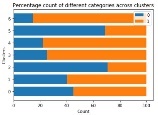
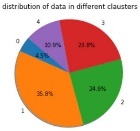
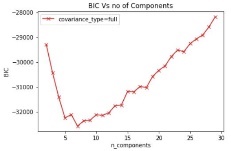
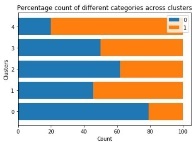
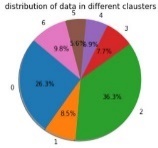
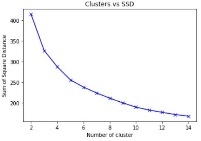
    

No of K-Mean Clusters: 11, No of EM Clusters: 21, These clusters seem to be different from original data clusters

**Clustering on Reduced Spotify Dimensions:**

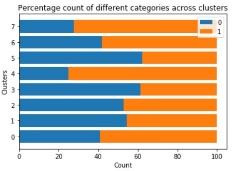
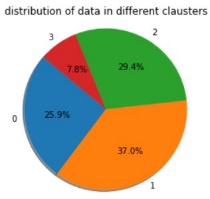
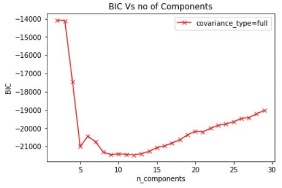
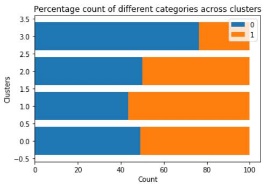
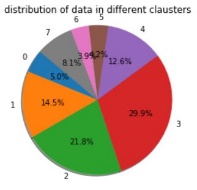
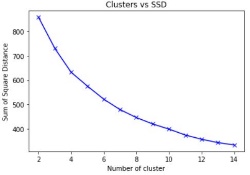
I ran K means and Expectation maximization Clustering on all the reduced data generated from different algorithms. The same plots of clustering produced above can be found below.

**Clustering on Decision Tree Dimensions:**



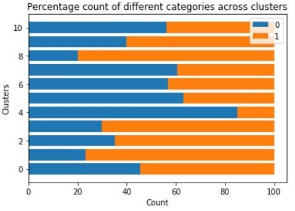
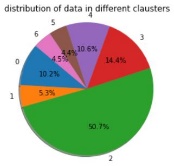
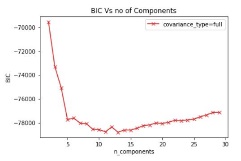
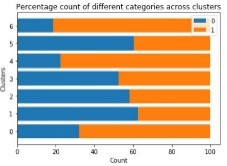
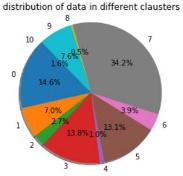
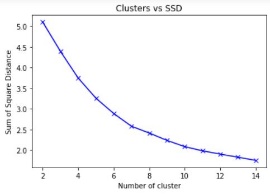
No of K-Mean Clusters: 5, No of EM Clusters: 7, These clusters seem to be different from original data clusters. This might be because the number of dim are reduced.

**Clustering on PCA Dimensions:**



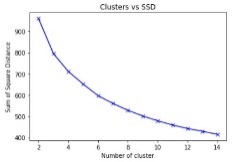
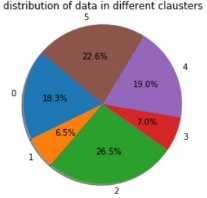
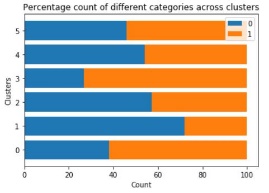
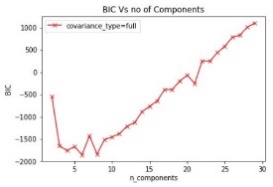
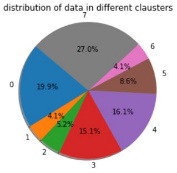
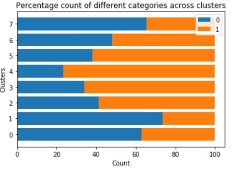
No of K-Mean Clusters: 4, No of EM Clusters: 8, These clusters seem to be different from original data clusters. This might be because the dim are of high variance.

**Clustering on ICA Dimensions:**



No of K-Mean Clusters: 7, No of EM Clusters: 11, These clusters seem to be different from original data clusters. This might be because the dim are of independent.

**Clustering on RCA Dimensions:**

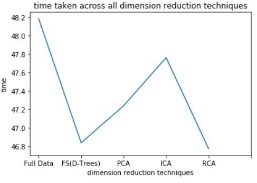
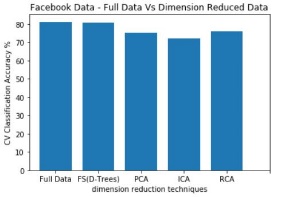


No of K-Mean Clusters: 6, No of EM Clusters: 8, These clusters seem to be different from original data clusters

**Task 4: Neural Nets on Reduced Dimensions**

**Neural Nets on Reduced Facebook Dimensions:**

I ran same Neural Nets as in assignment3 on all the reduced features from all algorithms and their performance on Error and Training Time are plotted below.



* From the Plots above we can rank the different Dimension Reduction algorithms based on Cross validation Accuracy as below

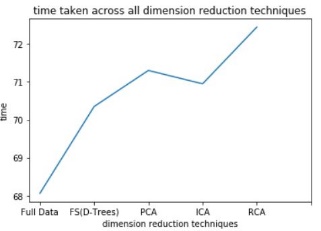
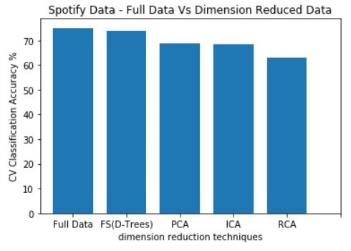
1. Feature Selection using Decision Trees 2) PCA 3) RCA 4) ICA

* From the Plots above we can rank the different Dimension Reduction algorithms based on Cross validation Accuracy as below.

1)RCA 2) Feature Selection using Decision Trees 3) PCA 4) ICA

**Neural Nets on Reduced Spotify Dimensions:**

I ran same Neural Nets as in assignment3 on all the reduced features from all algorithms and their performance on Error and Training Time are plotted below.



* From the Plots above we can rank the different Dimension Reduction algorithms based on Cross validation Accuracy as below

1)Feature Selection using Decision Trees 2) PCA 3) ICA 4) RCA

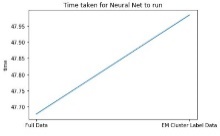
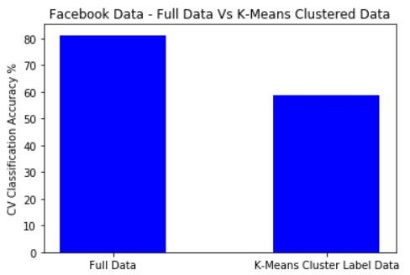
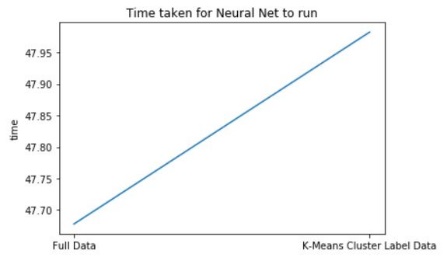
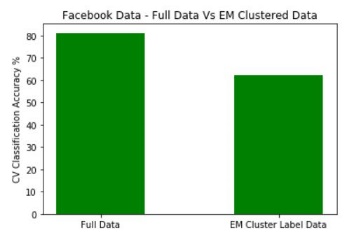
* From the Plots above we can rank the different Dimension Reduction algorithms based on Cross validation Accuracy as below.

1)Feature Selection using Decision Trees 2) ICA 3) PCA 4) RCA

**Task 5: Neural Nets on Clustered Data from task 1**

**Neural Nets on Clustered Facebook Dimensions:**

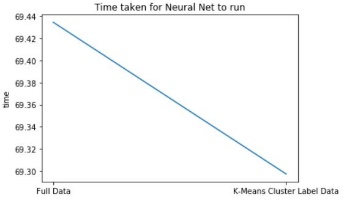
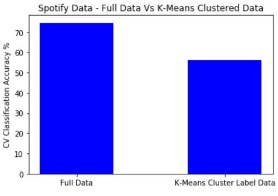
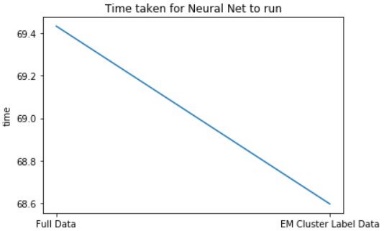
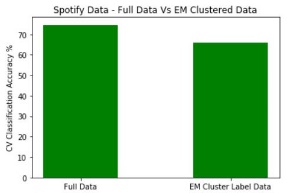
I ran same Neural Nets as in assignment3 on all the Clusters as features from all algorithms and their performance on Error and Training Time are plotted below.

* From the Plots above we can rank the different Clustering algorithms based on Cross validation Accuracy as below

1)Expectation Maximization 2) K-Means

**Neural Nets on Reduced Spotify Dimensions:**

I ran same Neural Nets as in assignment3 on all the Clusters as features from all algorithms and their performance on Error and Training Time are plotted below.

* From the Plots above we can rank the different Clustering algorithms based on Cross validation Accuracy as below

1)Expectation Maximization 2) K-Means