APPLIED MACHINE LEARNING

ASSIGNMENT 4
Srinikethan Pusthay (SXP210162)

Goal: Implement Clustering Algorithms like K-Means and Expectation Maximization on the dataset and observe the clusters behavior. And implement feature dimensionality reduction algorithms like PCA, ICA, Randomized Projections.

Dataset:
The dataset contains count of public bikes rented at each hour in Seoul Bike haring System with the corresponding Weather data and Holidays information.

Tasks:
We have used the data set from previous assignment, and we have modified the features same as previously. To convert the dataset into binary classification problem we have converted the output to class label. We have considered the mean value of bike count 704 as the threshold and transformed the column to 1 if the values are above threshold and 0 if the values are below threshold. The data is then split into train and test by 70% and 30%.

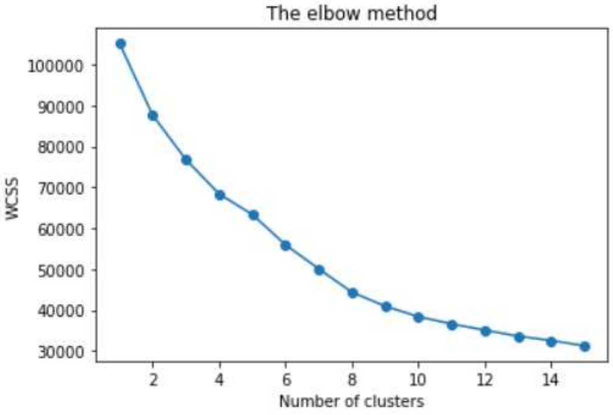
We have used Scikit Learn packages for clustering and dimensionality algorithms.

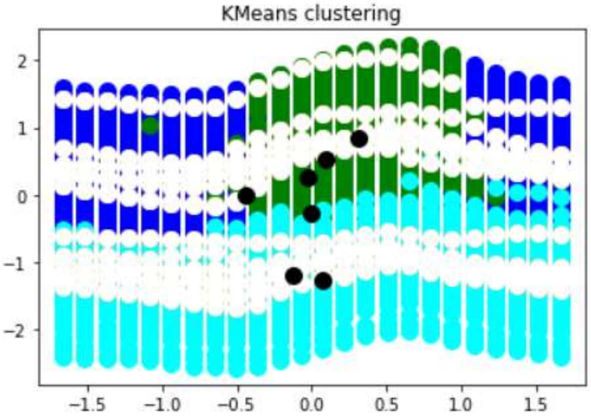
Experiments and Results:

Task1:

We have performed initial clustering using K-Means and Expectation Maximization algorithms using Scikit learn libraries.

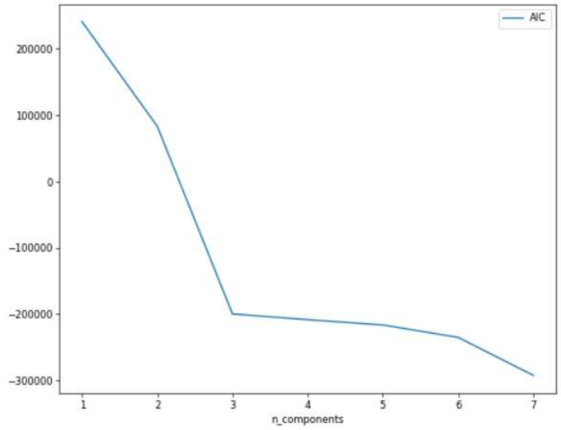
K-Means:

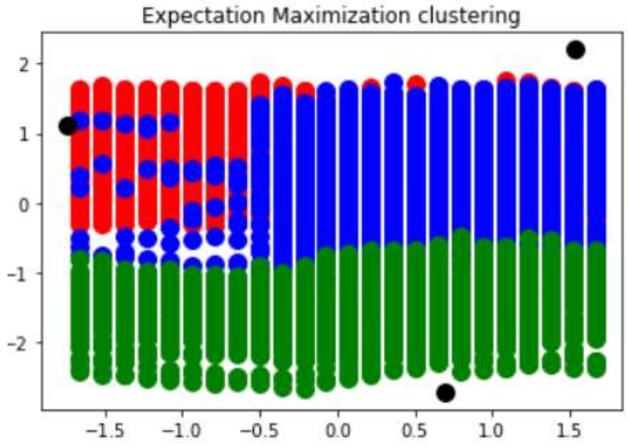




Above are the plots related to K-Means algorithm. To decide on the number of clusters, we have plotted the graph between within cluster sum of squares to number of clusters. By Elbow method, we have considered 7 clusters to be optimal number pf clusters to divide the data. Since we have 12 features, it is not easy to plot the clusters formed. So, we have used two of the most influential features temperature, dew temperature to plot clusters. In the similar way, we can observe the clusters for all the features. As temperature and dew temperature are highly correlated to the output, we are presenting the plot of its clusters. The black dots are the cluster centers of all the clusters. We can observe that clusters are not divided thoroughly and most of the data points are overlapping.

Expectation Maximization:



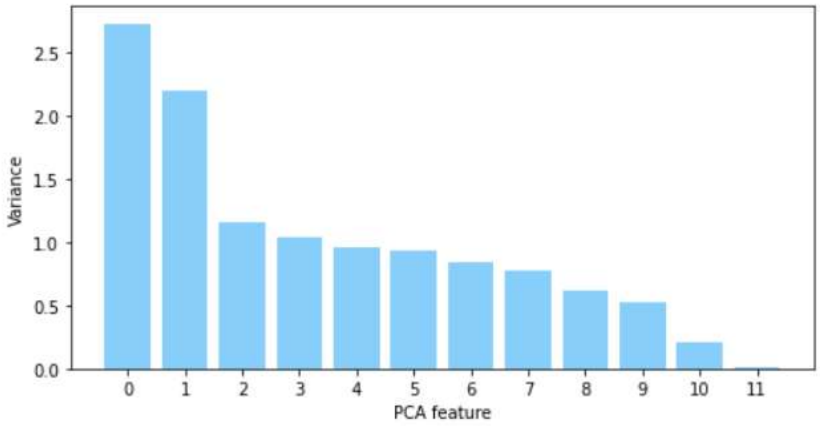


Above are the plots related to Expectation Maximization (EM). We have used Gaussian Mixture object to implement the EM algorithm for fitting mixture of Gaussian models. We have decided on the number of components/clusters for this algorithm by plotting a graph between AIC score and number of components. From the graph, we have a sharp elbow at 3 clusters. So, we have developed EM clustering algorithm using 3 results, and we have plotted the clustering graph using temperature and dew temperature. Here we can clusters are clearly visible in different parts but still there are few overlapping in data points.

Task 2:

For dimensionality reduction techniques, we are Principal Component Analysis (PCA), Independent Component Analysis (ICA), and Randomized Projections. We have used respective functions to implement them from scikit learn libraries.

For selecting the number of components for PCA, we have plotted the graph for the variance of different PCA features. Most of the variance of the data is capture by the first two PCA feature components, so we are moving forward with these two features for clustering.

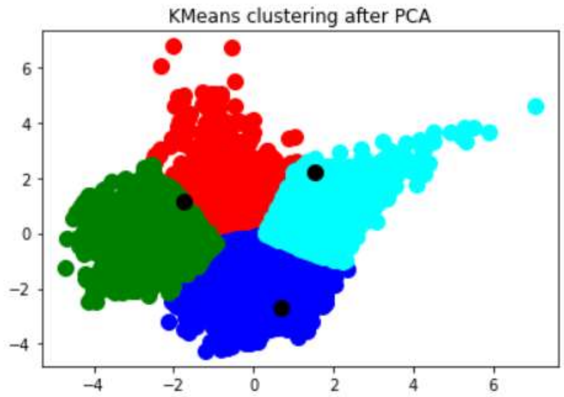


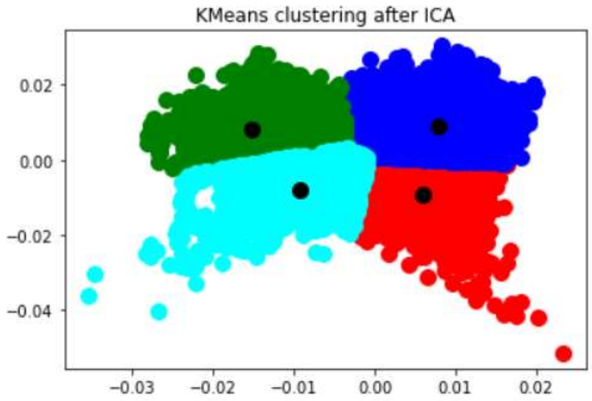
For ICA and RP, we have tried using different number of transformed feature components and we have found two number of features providing better clusters for the data.

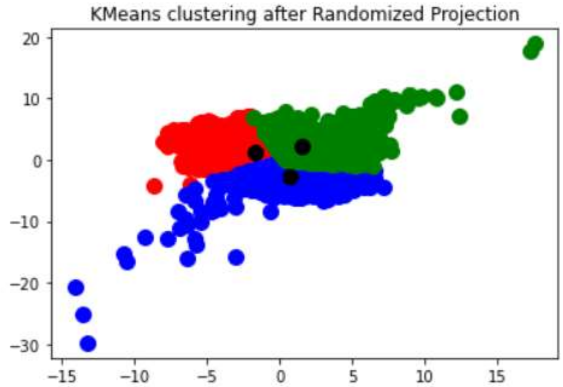
Task 3:

From the above tasks we have got updated features from PCA, ICA, and RP. We have performed clustering analysis again using these features and below are the results.

K-Means after Dimensionality reduction:

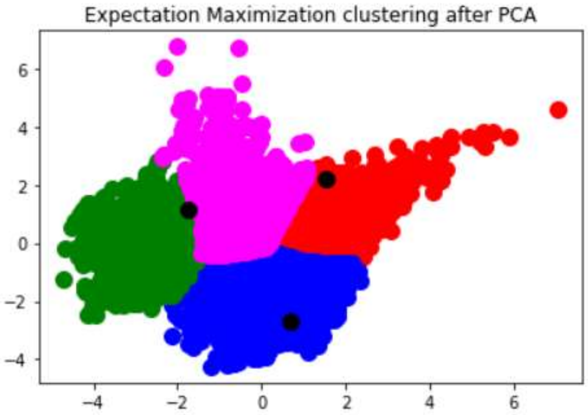


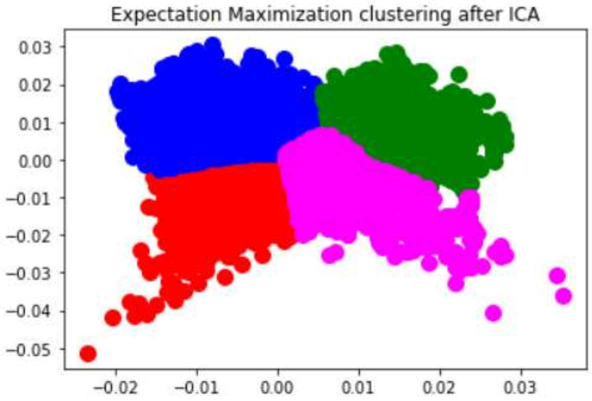


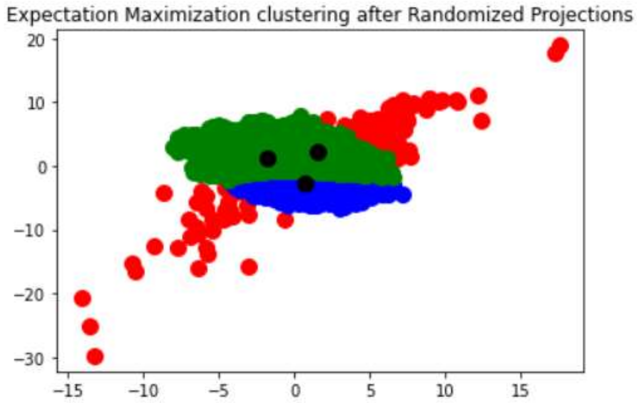


After dimensionality reduction using PCA, ICA and RP, we have performed K-Means algorithm and above are the resulted clusters. For all the three experiments, we have selected the number of clusters using the Elbow Method. Here, we are dealing with two features, and we can clearly observe the clusters divided. There is minimal overlap in the datapoints. Clusters obtained from PCA, and ICA features are similar to each other, but the order is different. Four clusters are formed using PCA and ICA but only three clusters were sufficient to segments RP features. Clusters in RP are more in elliptical shape than the shapes of other clusters. The clusters obtained here are much clearer and more visible than obtained from the original features.

Expectation Maximization after Dimensionality Reduction:


Expectation Maximization clustering after PCA


Expectation Maximization clustering after ICA


Expectation Maximization clustering after Randomized Projections

Similarly, we have performed EM algorithm with updated features from PCA, ICA and RP. Here also we can observe four clusters for PCA and ICA and three clusters for RP. Cluster shapes are like each other for PCA, ICA but they are different in order. For RP, clusters are more in elliptical shape than other clusters.

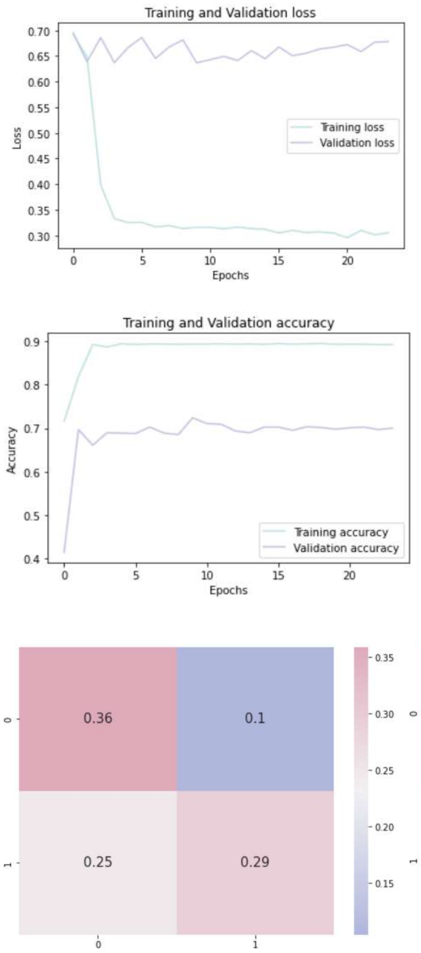
Task 4:

Here, we have performed experiments from Neural Network algorithm developed in previous assignment. But here we are using features obtained from PCA, ICA and RP. The accuracy from the original features from previous assignment is 74%.

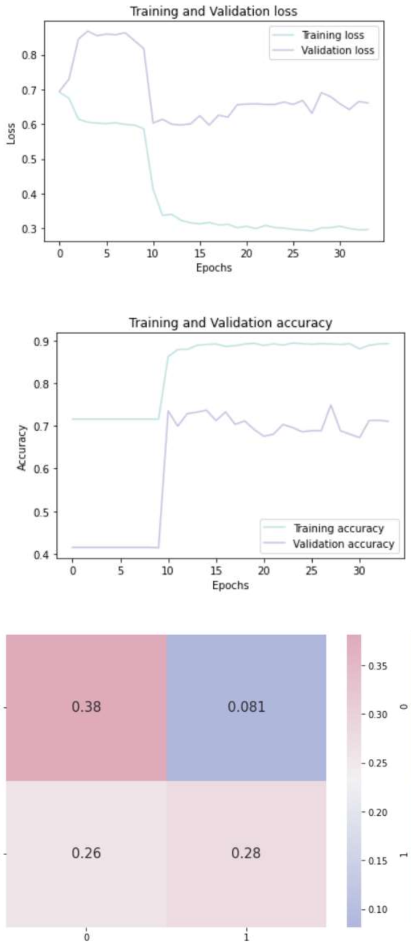
Below we have various plots obtained from the experiments. We have plots of Training and Validation loss, Training and Validation accuracy, and Confusion Matrix respectively for all three experiments. The trend of training loss and validation loss is same for all three. Validation loss is higher than training loss. Training accuracy is higher than validation accuracy. For models of PCA and ICA, accuracies have increased initially and has been stable for increasing number of epochs. But for RP model, validation accuracy has been unstable and decreased with many epochs.

Accuracies for models of PCA, ICA and RP are 65%, 66% and 48% respectively. Dimensionality reduction techniques are helpful in many cases, to decreased test and train time, removal of multi-collinearity etc. Here in our case, it might not be effective but with large data and many features these are very much effective in training and avoiding curse of dimensionality.

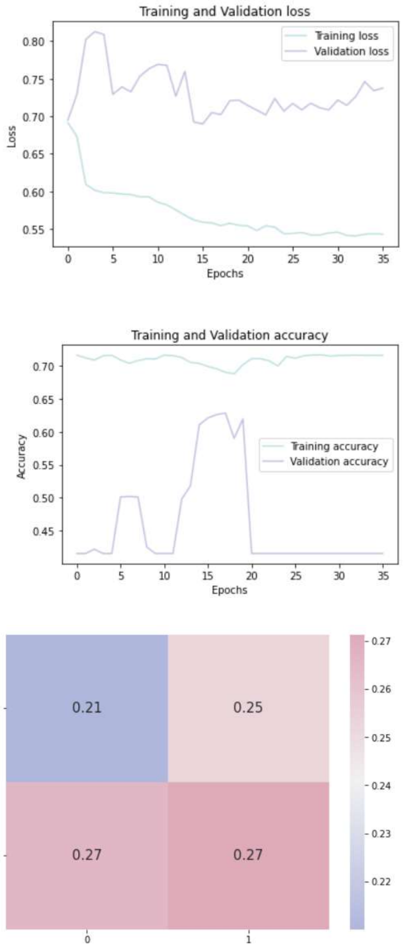
Neural Network after PCA:



After ICA:

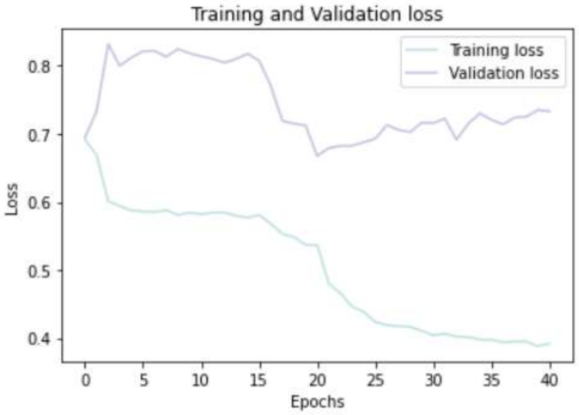


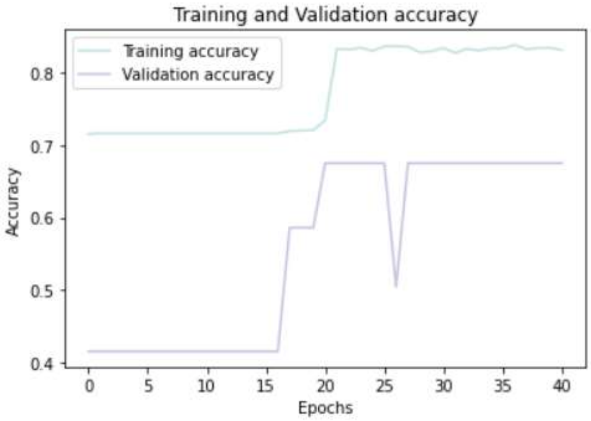
After RP:

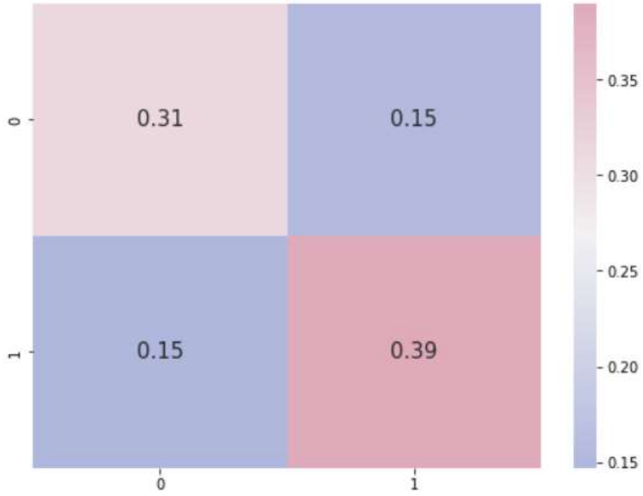


Task 5:

We have used clustering labels from Task 1 as input and class label as output in this task. We have experimented Neural Network with these features and obtained the below results. We have achieved an accuracy of 70% higher than the models from Neural Networks of PCA, ICA and RP. But less than the accuracy of model using original features. There might be many factors for these, neural network design, number of clusters labels, data size etc. The accuracy can always be improved from many further experimentations and optimizing hyperparameters.


Training and Validation loss
Training loss
Validation loss
Loss
0.8
0.7
0.6
0.5
0.4
0
5
10
15
20
25
30
35
40
Epochs


Training and Validation accuracy
Training accuracy
Validation accuracy
Accuracy
0.8
0.7
0.6
0.5
0.4
0
5
10
15
20
25
30
35
40
Epochs


0.31
0.15
0.15
0.39
0
1
0
1
0.35
0.30
0.25
0.20
0.15