CS 7641 – Machine Learning

Spring 2018

Submitted by- Yashovardhan Jallan (GT ID: yjallan3)

1. Introduction

This assignment applies different unsupervised learning algorithms and dimensionality reduction techniques to analyze two distinct datasets: Pima Indians Diabetes Data Set and Wine Quality Data Set, both taken from UCI Machine Learning Repository. In particular, K-Means Clustering and Expectation Maximization are used. For dimensionality reduction, Principal Components Analysis (PCA), Independent Component Analysis (ICA), Randomized Projections and Random Forests are used. These datasets are also used for Neural networks analysis after clustering and dimensionality reduction. The whole analysis is done using R programming language and its various libraries.

2. Datasets Description and why are they interesting?

Wine Quality Data Set

The Red Wine dataset is used from UCI Machine Learning Library. This is a very popular ML dataset. With the growth in social wine drinking, the determination of wine price and quality has become a very important task. I think it is an interesting exercise to apply Unsupervised Learning and Dimensionality Reduction techniques to make these predictions.

On preliminary analysis of the Red wine quality dataset, all the sample points were classified into 6 quality scores (3, 4, 5, 6, 7, 8). Out of 1599 records, there were only a small handful of records having quality score as either 3, 4 or 8. Due to only a small handful of observations with these labels, I dropped them and used a final dataset of 1518 samples.

Pima Indians Diabetes Data Set

Pima Indians are a group of Indigenous Americans living in Arizona that have a high incidence rate of diabetes. The Pima Indian diabetes dataset is a collection of medical diagnostic reports from 768 records of female patients at least 21 years old. There are eight attributes and one class variable with binary values 0 for tested positive or 1 for tested negative.

From machine learning perspective, this dataset is significant. I looked at its existing results and found that for most classifiers, it turns out this dataset is not easily classifiable and performance with most classifiers is generally low. I directly downloaded this dataset which is made available by the R library 'mlbench'.

3. Important Section for my methodology – PLEASE READ THIS FULLY

Throughout this report, I have presented several accuracy plots depicting the training, validation and test accuracies achieved by either the clustering algorithm or neural network. I would like to explain my methodology of obtaining these plots.

3.1 Methodology for predictions using only Clustering (Task 1) and Dimensionality Reduction followed by Clustering (Task 3)

For Task 1, I have used the original dataset itself to apply clustering. For Task 3, I have used different Dimensionality Reduction algorithms to produce a new set of features/components and I have applied clustering on these datasets. To test the accuracy of my clustering algorithms (both Kmeans and EM) in Task 1 and Task 3, I have applied clustering techniques on their respective training set and found 'k' centers or clusters for both 'kmeans' and 'EM'. After having found the centers, I have taken the mode of all the points in a particular center/cluster to assign a label. For example, for wine dataset, if my cluster 'A' has 10 observations out of which 7 have a label of 'quality=5', then I assign a label of 'quality=5' to this cluster. This is the train phase.

Next, I use 'k' centers which I have found in the training phase and then for each observation in the validation and test sets, I map it to one of the 'k' center. This is done by minimizing the Euclidean distance of that particular point (also known as minimizing the L2 Norm) from all of these centers. So, if a certain observation was assigned to cluster 'A' and cluster 'A' was assigned a label of 'quality=5' during the training phase, then my current observation will be labelled as 'quality=5'. This is the predicted value of my clustering technique. This is compared with the true label of the validation and the test set to see the accuracy of the predictions. The validation set is used to fine tune the hyperparameters and finally the best version of the algorithm is applied on the test set.

3.2 Methodology for predictions using Dimensionality Reduction - Neural Networks (Task 4)

To test the accuracy of using Neural Networks after Dimensionality Reduction on the dataset (Task 4), I have independently worked with each dimensionality reduction algorithm and applied neural network on them. For each of the Dimensionality Reduction algorithm, I have used the features/components which they generate as my Xs (features) and the existing labels as my Ys. For ex: on the Wine Quality dataset, I have applied PCA which generates several principal components. Based on literature available on PCA, I used only the required number of components which cumulatively explained at least 85% of the variance in my data. And using these components, I trained my neural network. The training was done on a train set, and I used a validation set for hyperparameter tuning and finally applied the algorithm to test set.

I realize that applying Dimensionality Reduction on a dataset followed by Neural Nets can be done in several ways. Some other ways could be keeping the original features, adding the features given by the Dimensionality Reduction algorithms additionally and then running the NNets. Another variation could be in the approach by which we validate and test the neural net. For testing on a dataset, one way to go about it would be applying Dimensionality Reduction on only test data and then making predictions. Or, we could use both test and train Xs together, apply Dimensionality Reduction on this entire dataset and only then make predictions (I have done this). It requires domain knowledge and other factors to make this determination.

3.3 Methodology for predictions using Dimensionality Reduction – Clustering - Neural Networks (Task 5)

To test the accuracy of using Neural Networks after Dimensionality Reduction and Clustering (Task 5), I have applied the Dimensionality Reduction algorithms to the dataset individually. Using each of these new datasets, I have applied both 'kmeans' and 'EM' algorithms to make predictions as required by Task 3. Now, the key part is this.

Using these predictions, I have created a new dataset which contain my original Xs along with the predictions made by my DR+Clustering technique as new Xs or new features. In essence, this dataset contains the original features, plus information added due to all the Dimensionality Reduction algorithms followed by applying clustering on them. For example, the wine quality dataset originally has 11 Xs and 1 Y. I added a total of 8 new features or Xs. Four of them were predictions made by 'kmeans' on PCA, ICA, RP and Random Forest respectively. The remaining four were predictions made by 'EM' on PCA, ICA, RP and Random Forest respectively

I have made all of this into a single dataset. I could have done it in various other ways to find the best performance. Meaning, I could have taken any combination of these new 8 features to add to my dataset. This in itself could have been a hyperparameter to optimize. Other than this also, we could have followed several other permutations and combinations to capture Dimensionality Reduction and Clustering in Neural Nets. I have experimented with some of these but have not included in the report for the sake of brevity and the 10 page limit.

4. Task One - Clustering Algorithms on both Datasets

The R – ‘kmeans’ library is used for K-Means algorithm and ‘Mclust’ library is used for expectation maximization.

4.1 k-means

k-means algorithm aims to partition the points into ‘k’ groups such that the sum of squares from points to the assigned cluster centers is minimized. I have assigned 10 random restarts to ensure best clustering is found. We are concerned with Within-Cluster-Sum-of-Squares (WCSS). To find the best ‘k’ value, I have used the Elbow-Method using WCSS – his is one of the most common and technically robust methods. This is based on principle that while clustering performance as measured by WCSS increases (i.e. WCSS decreases) with increase in k, rate of increase is usually decreasing. So performance improvement for increasing number of cluster from, say, 3 to 4 is higher than that for increasing from 4 to 5.

This is done with two different types of methods to determine optimal clusters. One is the Forgy method, and one is the Lloyd method. The Lloyd’s method is the simplest method that minimizes within-cluster sum of squares. The Forgy method is similar to Lloyd’s algorithm except that it considers the data distribution continuous instead of discrete.

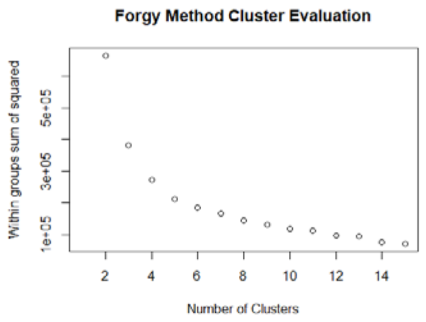

Fig. 1. Wine Data – Kmeans (Forgy)

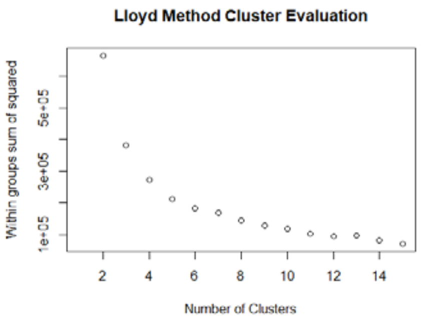

Fig. 2. Wine Data – Kmeans (Llyod)

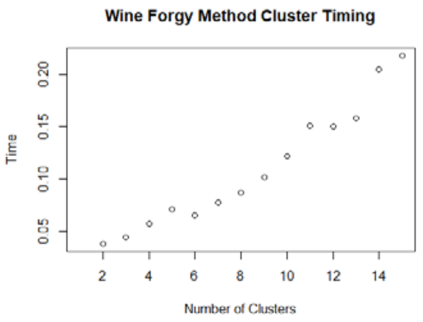

Fig. 3. Wine Data – Kmeans Time Plot

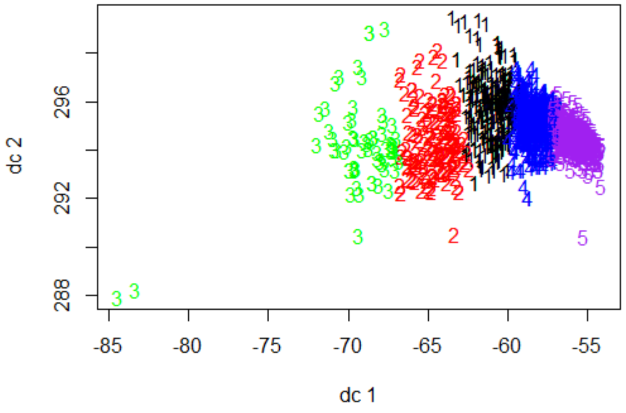

Fig. 4. Wine Data – Cluster Visualization using Kmeans

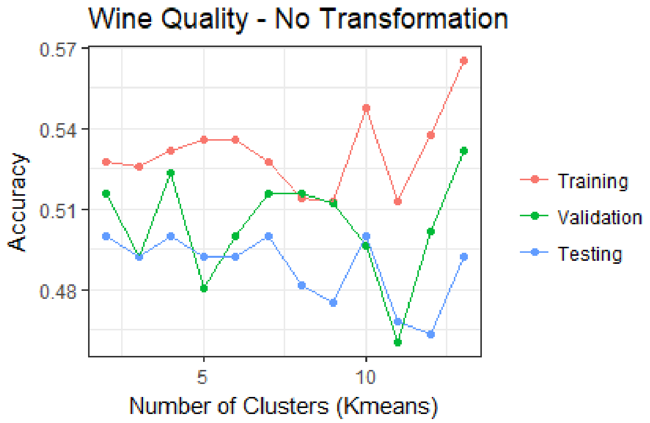

Fig. 5. Wine Data – Kmeans Accuracy Test

Fig. 1-5 show the results obtained from application of k-means algorithm on the Wine dataset. Both ‘forgy’ and ‘llyod’ method are consistent with the elbow method described earlier. From the plots in Fig. 1-2, we can see that k=5 is a good choice. Fig. 4 shows the clustering achieved using ‘kmeans’. Fig. 5 displays the accuracy plot obtained using Kmeans on test set for wine-quality. The best test accuracy of close to 50% was observed.

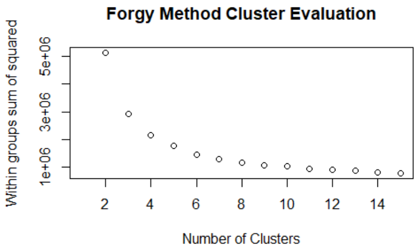

Fig. 6. Pima Data – Kmeans (Forgy)

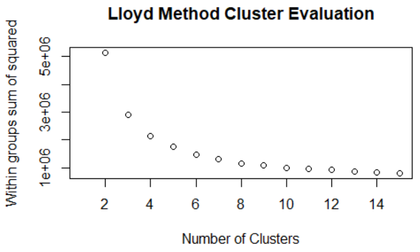

Fig. 7. Pima Data – Kmeans (Llyod)

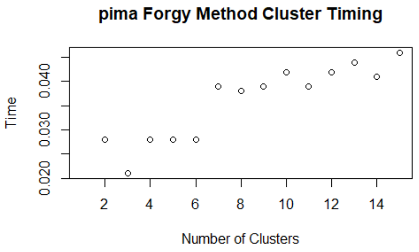

Fig. 8. Pima Data – Kmeans Time plot

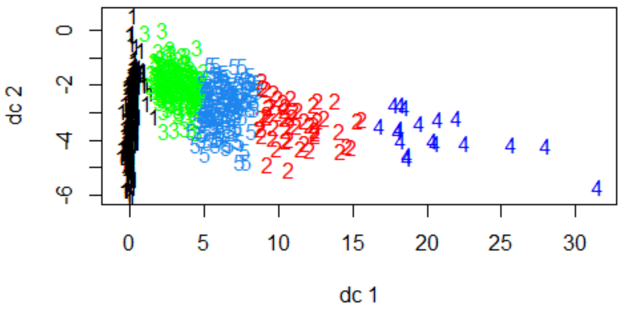

Fig. 9. Pima Data – Cluster Visualization using Kmeans

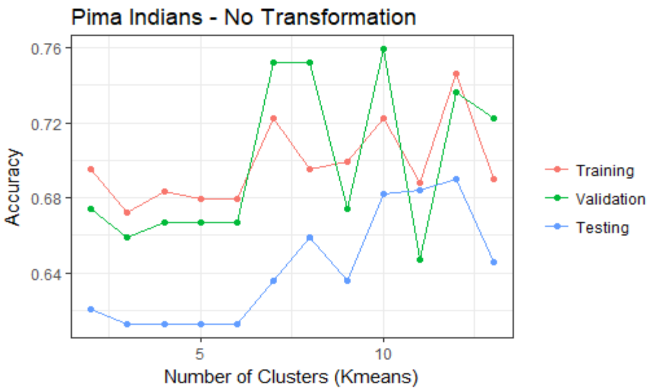

Fig. 10. Pima Data – Kmeans Accuracy Test

Fig. 6-10 show the results obtained from application of k-means algorithm on the Pima dataset. Both 'forgy' and 'llyod' method are consistent with the elbow method described earlier. From the plots in Fig. 6-8, we can see that k=4 is a good choice. Fig. 9 shows the clustering achieved using 'kmeans'. Fig. 10 displays the accuracy plot obtained using Kmeans on test set for Pima Indians dataset. The best test accuracy of close to 69% was observed.

4.2 Expectation Maximization

Expectation Maximization (EM) is a clustering method for estimating maximum likelihood. It is an iterative process that computes log-likelihood for a current posterior, then solves for the maximum likelihood parameters. It is a soft clustering method where instances are given probabilities or scores on whether they are in the cluster or not. The Bayesian information criterion, that approximates Bayes factors that, is an information criteria method that balances model complexity and precision. The key is to maximize this value.

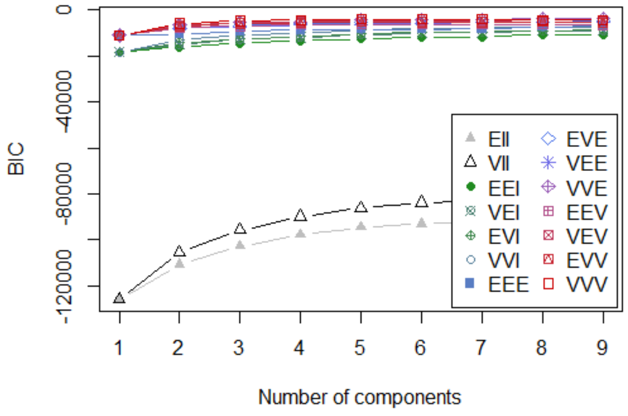

Fig. 11. Wine Data – Various Models of EM

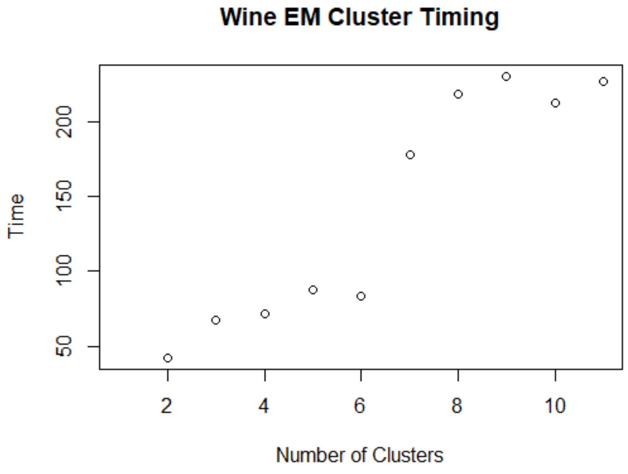

Fig. 12. Wine Data – Runtime of EM vs No. of Clusters

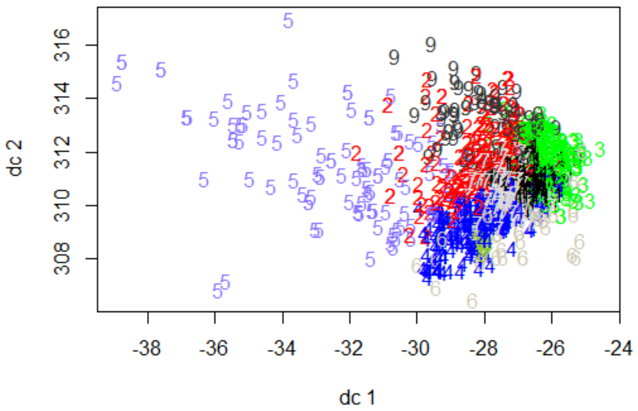

Fig. 13. Wine Data – Cluster Visualization using EM

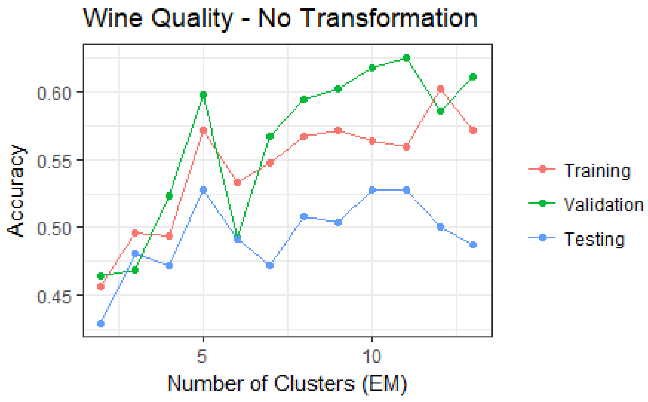

Fig. 14. Wine Data – EM Accuracy Test

Fig 11-14 describe the different analysis plots using Expectation Maximization on Wine Quality Dataset. Fig 11 is a plot showing different models of EM applied to wine dataset. It was found that VVE model with 8 components gave the best score. Fig. 12 shows the computational time increase as the number of clusters increase. Fig. 13 is one representation of clustering the dataset. We can see that EM does not do a great job in this regard. Fig. 14 shows that the best test accuracy was close to 53%.

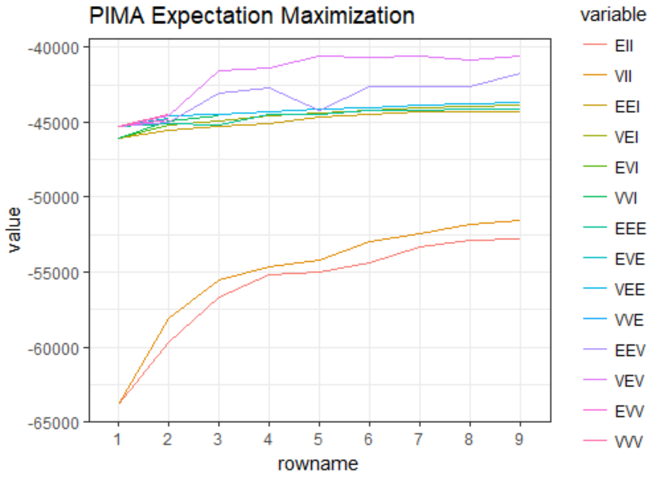

Fig. 15. Pima Data – Various Models of EM

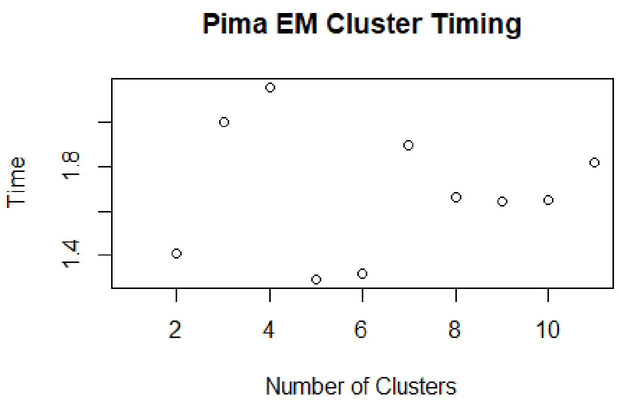

Fig. 16. Pima Data – Runtime of EM vs No. of Clusters

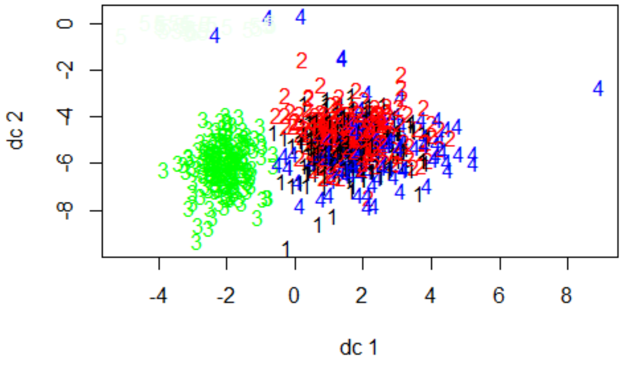

Fig. 17. Pima Data – Cluster Visualization using EM

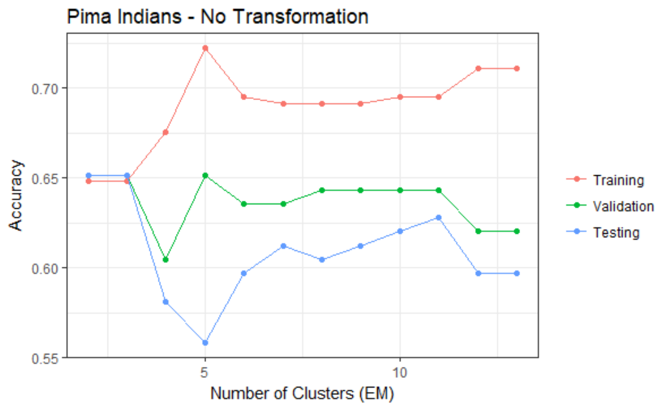

Fig. 18. Pima Data – EM Accuracy Test

Fig 15-18 describe the different analysis plots using Expectation Maximization on Pima Dataset. Fig 15 is a plot showing different models of EM applied to wine dataset. It was found that VEV model with 5 components gave the best score. Fig. 16 shows the computational time increase as the number of clusters increase. Fig. 17 is one representation of clustering the dataset. We can see that EM does not do a great job in this regard. Fig. 18 shows that the best test accuracy was close to 63%.

5. Task Two – Dimensionality Reduction on both Datasets

5.1 Principal Components Analysis

Principal component analysis finds the orthogonal eigenvectors that best explain the maximum amount of variance. It orders components on the value of their eigenvalues and the last few components have relatively small eigenvalues, giving the possibility of removing them to apply classification.

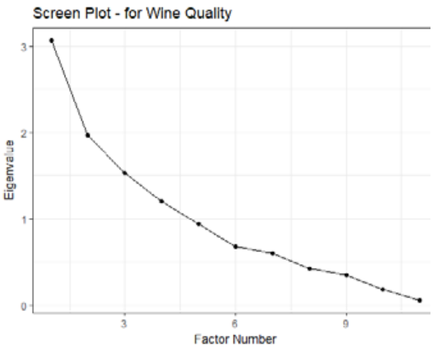

Fig. 19. Wine Data –Component Eigenvalues

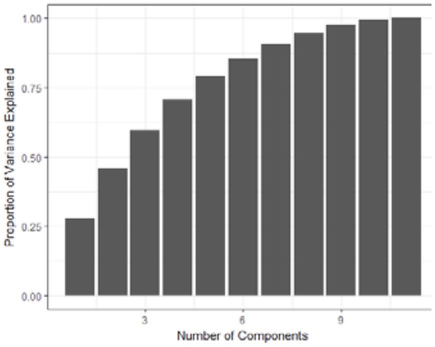

Fig. 20. Wine Data –Variance Explanation

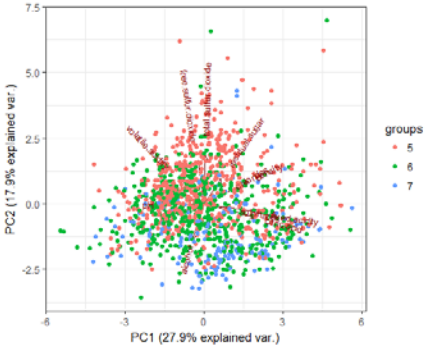

Fig. 21. Wine Data –Visualizing PCA

Fig 19- 21 shows the analysis done using PCA on wine quality dataset. Fig 19. Shows the plot of eigenvalues vs the new components. Fig. 20 plots the proportion of variance explained by the components. Fig. 21 describes the cluster using the first two principal components obtained using PCA. We know that PCA suggest that we can use the minimum number of components which cumulatively describe about 80-85% of the variance in the data. In the case of wine quality, it was found to be roughly 7 components.

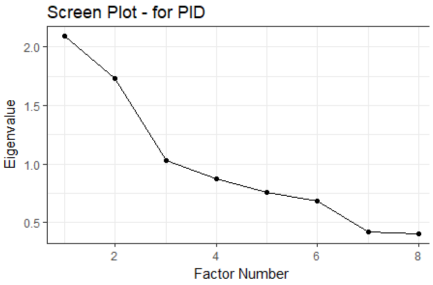

Fig. 22. Pima Data –Component Eigenvalues

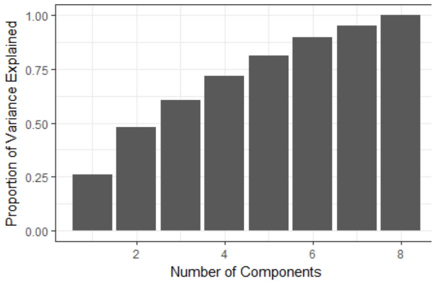

Fig. 23. Pima Data –Variance Explanation

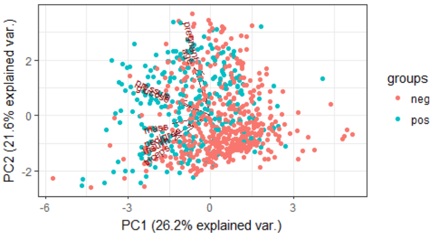

Fig. 24. Pima Data –Visualizing PCA

Fig 22- 24 shows the analysis done using PCA on Pima dataset. Fig 22. Shows the plot of eigenvalues vs the new components. Fig. 23 plots the proportion of variance explained by the components. Fig. 24 describes the cluster using the first two principal components obtained using PCA. In the case of Pima, the required number of components was found to be 6.

5.2 Independent Components Analysis

Independent component analysis tries to reconstruct the data by maximizing the difference between components and find independent components of the original data. I have used fastICA algorithm. The independent components are sorted by kurtosis values from highest to lowest. A Gaussian distribution has a kurtosis value of 3. So the components obtained by ICA which have a kurtosis value of atleast greater than 3 can be considered important.

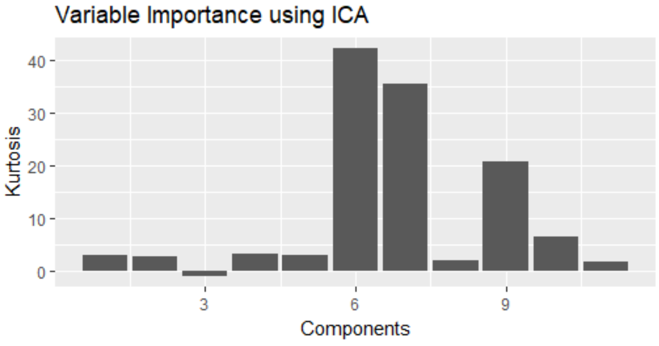

Fig 25. Wine Quality – Kurtosis for the new component

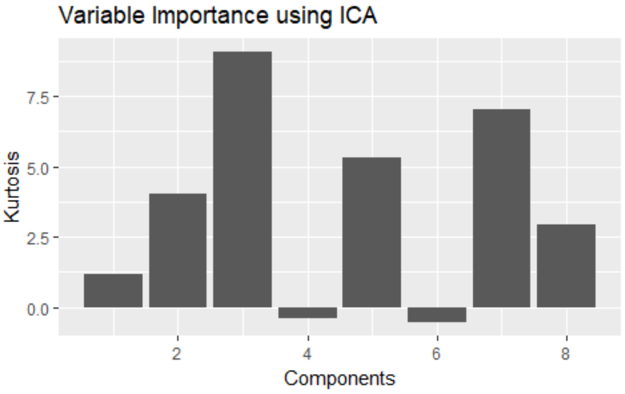

Fig 26. Pima Data – Kurtosis for the new component

Fig. 25 shows the kurtosis plot for wine quality dataset. We see that components number 6, 7 and 9 have a really high kurtosis value and thus ICA algorithm suggests that these are the most important components. Fig. 26 shows the kurtosis plot for Pima dataset. The components number 3, 5 and 7 are seen to be most significant.

5.3 Randomized Projection

Random projection is a dimensionality reduction method that projects the total number attributes to a lower dimensional space. It is a faster algorithm that transforms the data into a random rotation matrix while guaranteeing a maximum distortion. As opposed to PCA, random projection projects the original input space on a randomly generated Gaussian matrix.

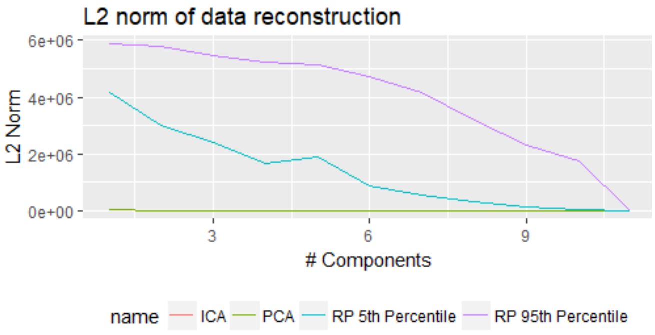

Fig 27. Wine Quality – L2 norm of data reconstruction

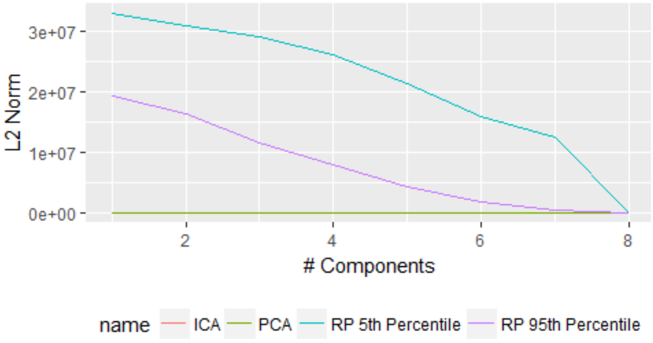

Fig 28. Pima Data– L2 norm of data reconstruction

Fig. 27 and Fig 28 show the comparison of PCA and ICA with 95th percentile Randomized Projection and 5th percentile Randomized Projection for both the datasets. These plots depict the value of the L2 norm vs the number of components.

5.4 Feature Selection using Random Forest

Variable importance might generally be computed based on the corresponding reduction of predictive accuracy when the predictor of interest is removed (with a permutation technique, like in Random Forest). With RF, the Gini importance index is defined as the averaged Gini decrease in node impurities over all trees in the forest (it follows from the fact that the Gini impurity index for a given parent node is larger than the value of that measure for its two daughter nodes.

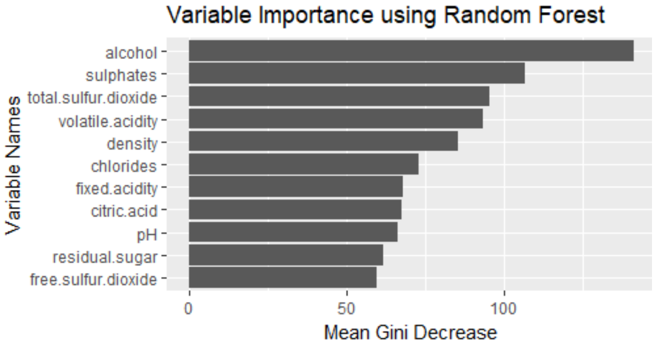

Fig. 29 Wine Quality – Variable Importance

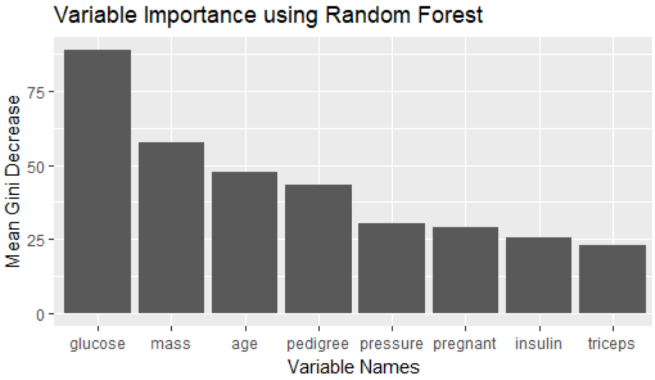

Fig. 30 Pima Data – Variable Importance

Fig 29 and 30 show the variable importance obtained by Mean gini decrease method using Random Forest. For the wine dataset, we see that the feature 'alcohol' is found to be most important and for the Pima dataset, the feature 'glucose' is found to be the most significant. Both of these intuitively make sense.

6. Task Three – Dimensionality Reduction followed by Clustering on both

As described earlier in section 3.1, I have used different Dimensionality Reduction algorithms to produce a new set of features/components and I have applied clustering on these datasets. There are a total of 4 different dimensionality reduction algorithms and two clustering algorithms. So for each dataset, we have 8 different analysis plots.

6.1 Wine Quality

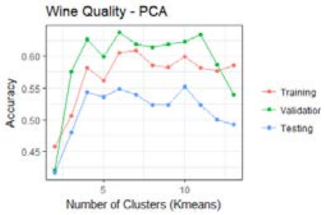

Fig. 31 Wine-PCA-kmeans

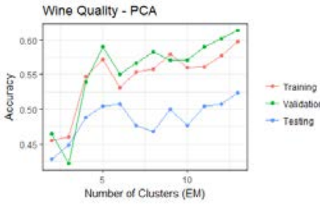

Fig. 32 Wine-PCA-EM

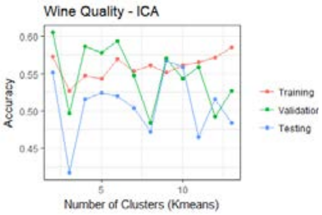

Fig. 33 Wine-ICA-kmeans

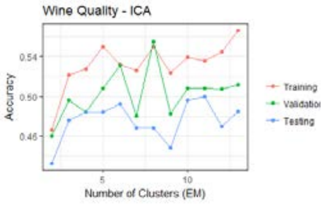

Fig. 34 Wine-ICA-EM

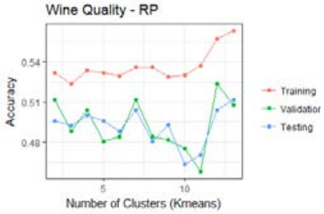

Fig. 35 Wine-RP-kmeans

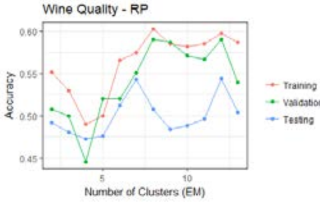

Fig. 36 Wine-RP-EM

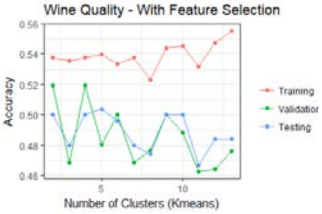

Fig. 37 Wine-RF-kmeans

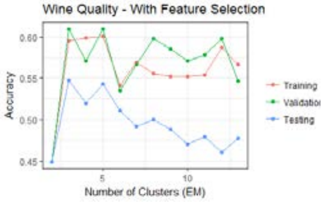

Fig. 38 Wine-RF-EM

For the Wine Quality dataset, the Figs 31-38 show the eight different sets of accuracy plots for various combinations of Dimensionality Reduction methods and clustering algorithms. The best test accuracy was found to be for ICA followed by Kmeans clustering of 57% for k=8.

6.2 Pima Dataset

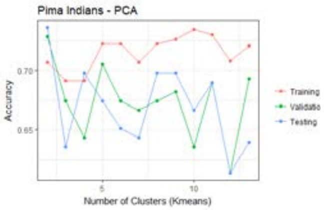

Fig. 39 Pima-PCA-kmeans

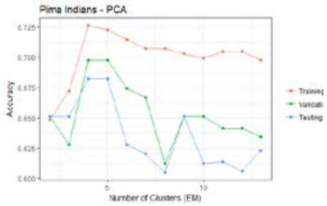

Fig. 40 Pima-PCA-EM

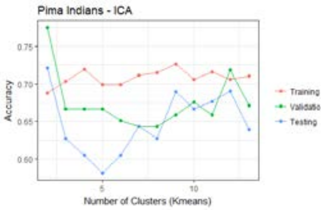

Fig. 41 Pima-ICA-kmeans

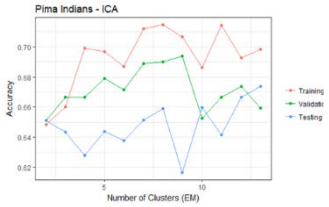

Fig. 42 Pima-ICA-EM

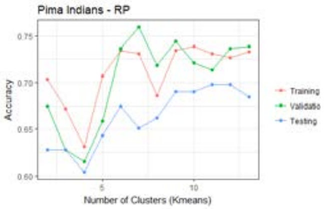

Fig. 43 Pima-RP-kmeans

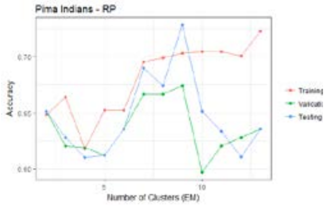

Fig. 44 Pima-RP-EM

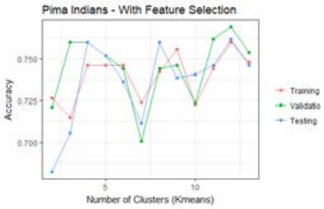

Fig. 45 Pima-RF-kmeans

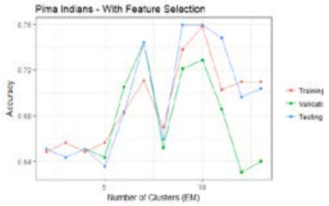

Fig. 46 Pima-RF-EM

For the Pima dataset, the Figs 39-46 show the eight different sets of accuracy plots for various combinations of Dimensionality Reduction methods and clustering algorithms. The best test accuracy was found to be for Random Forest followed by EM clustering of 76%.

7. Repeat from Assignment #1 – Neural Network on Wine Quality

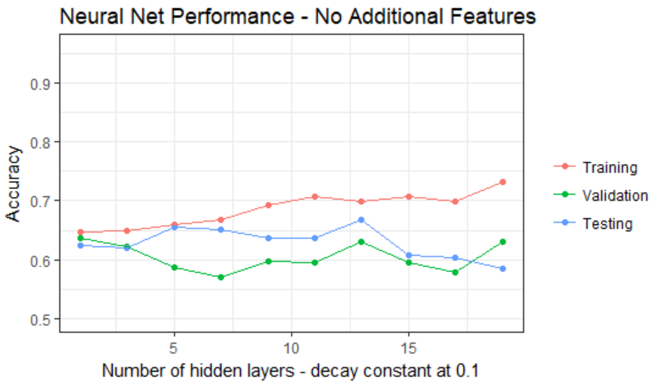

Fig. 47 Wine quality neural Network

Before going ahead with Task 4 and Task 5, we take a look at how the Neural Network algorithm performed on the original dataset without any dimensionality reduction or clustering,

We see that the best test accuracy was observed for # of hidden layers = 13. The test accuracy observed was 66%. In the next two sections we will compare the results obtained using neural networks with this result.

8. Task Four – Dimensionality Reduction followed by Neural Net - Wine Data

As explained earlier in section 3.2, to test the accuracy of using Neural Networks after Dimensionality Reduction on the wine dataset, I have independently worked with each dimensionality reduction algorithm and applied neural network on them. For each of the Dimensionality Reduction algorithm, I have used the features/ components which they generate as my Xs (features) and the existing labels as my Ys.

Fig 48-51 show the accuracy plots for the analysis conducted for Task 4. The best accuracy for the training set was seen to 68% obtained by applying Random Forest feature selection followed by neural networks. This accuracy score is slightly better than the 66% on the vanilla version described in section 7.

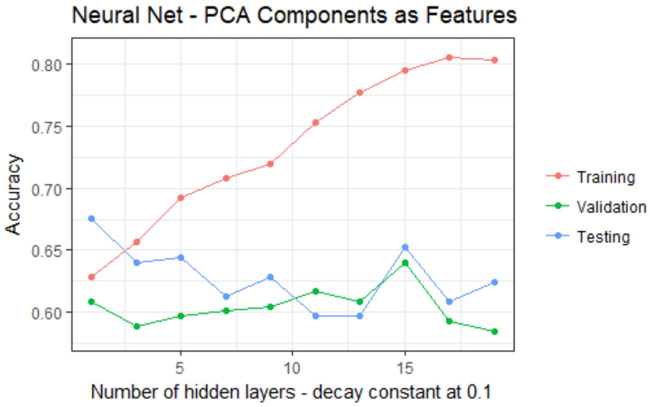

Fig. 48 Wine quality –PCA - Neural Network

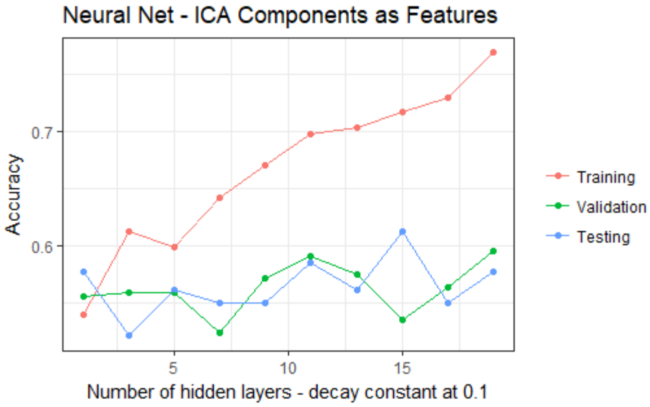

Fig. 49 Wine quality – ICA- Neural Network

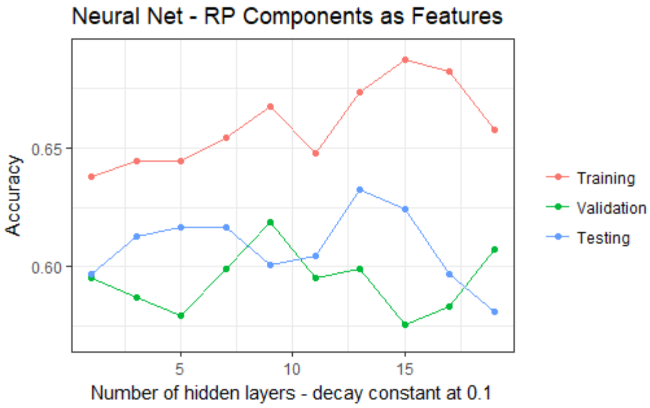

Fig.50 Wine quality –RP - Neural Network

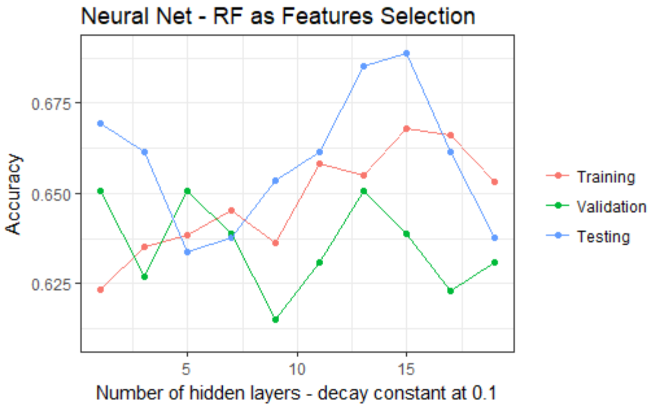

Fig. 51 Wine quality–Random Forest- Neural Network

9. Task Five - Dimensionality Reduction – Clustering - Neural Net - Wine data

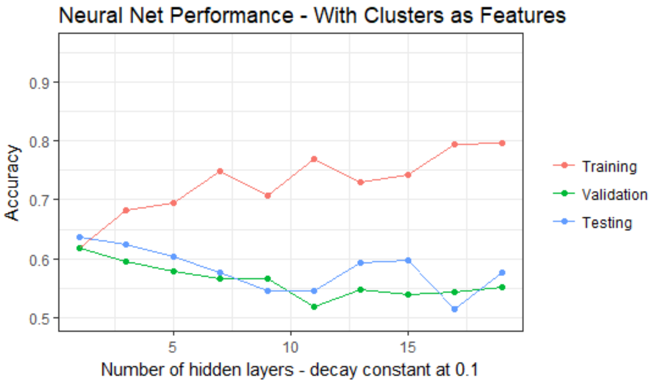

Fig. 52 Wine Quality – DR + Clustering + Neural Nets

As mentioned in section 3.3, for Task 5, I have applied the Dimensionality Reduction algorithms to the dataset individually. Using each of these new datasets, I have applied both ‘kmeans’ and ‘EM’ algorithms to make predictions.

Using these predictions, I have created a new dataset which contain my original Xs along with the predictions made by my DR +Clustering technique as new Xs or new features. From Fig. 52, we can see that the best test accuracy was seen to be 60%.

10. Summary

I have applied 4 different dimensionality reduction techniques and two clustering algorithms on these two datasets to do a wide variety of analysis. Important learning out of this assignment is that we can go about applying these techniques in numerous ways, each giving different results from the other. I am now able to appreciate the large ocean of machine learning and hyperparameter optimization. Domain knowledge is extremely crucial and professor Isbell’s lecture in class stressing its importance now makes a lot of sense. Thanks!