CS7641 Fall 2022 HW3 Unsupervised Learning and Dimensionality Reduction

Fung Yi Yuen
fyuen3@gatech.edu

***Abstract*— Two clustering algorithms (k-means and expectation maximization) and four dimensionality reduction algorithms (PCA, ICA, Randomized Projections and LDA) were implemented on two classification datasets (The Divorce Predictors Dataset and The Contraceptive Method Choice Dataset). The 2 clustering algorithms were then run on the 4 dimensionally reduced datasets and analyzed. Neural networks models were used to run on the 4 dimensionally reduced datasets and 2 clustered datasets for analysis.**

I. DATASETS

1.1. Divorce Predictors Dataset

This is a binary classification dataset with 53 features and 170 entries. The class distribution is balanced (51% married, 49% divorced). All features are interview questions related to couples' daily lives e.g. "I enjoy our holidays with my wife" and "Fights often occur suddenly.". All the interview questions are answered on a scale of 0 to 4 where 0=Never, 1=Seldom, 2=Averagely, 3=Frequently, 4=Always.

1.2. Contraceptive Method Choice Dataset

This is a multi-classes classification dataset with 9 features and 1473 entries. The class (i.e. contraceptive method used) distribution is imbalanced (42.7% no-use, 22.6% long-term, 34.7% short-term). The features are families' information e.g. wife's education, wife's religion and husband's occupation etc. The answers to these features are represented by numerical values e.g. education 1=low, 2,3,4=high, or working now 0=Yes, 1=No.

Comparing two datasets:

Dataset	The Divorce	The Contraceptive
# of features	53	9
# of entries	170	1473
classification type	binary	3 types
balance?	balanced	imbalanced

Choosing these two datasets would be an interesting experiment to see how different clustering algorithms behave on balanced and imbalanced datasets, how will the data points be clustered if the classification problem is binary or multi-classed, what are the difference between dataset which has many features (i.e. The Divorce dataset) vs the one has fewer features (i.e. The Contraceptive dataset).

II. CLUSTERING ALGORITHM

2.1. K-means Clustering

K-means algorithm groups the data points into k clusters. k is the number of centroids defined by elbow method, silhouette analysis or any other means which helps to find optimal k. A centroid is the center of a cluster. By minimizing the sum of squared distance from all the data points to their nearest centroid within a cluster (i.e. repeatedly calculating the sum of squared means within a cluster until no further improvement), the optimal value of k can be found.

2.1.1 Divorce Dataset

The elbow method was used initially. Theoretically, the optimal k should be the point where the within-cluster sum of square error (SSE) ceases to drop even with increasing number of clusters.

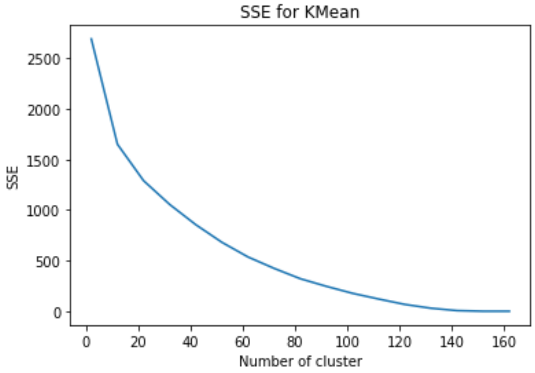


fig.1 SSE vs number of clusters (divorce dataset)

However, fig.1 shows an ambiguous elbow plot where the optimal k cannot be found clearly (i.e. no sharp elbow point). Thus the silhouette method and homogeneity score were then used (fig.2) to further check on the optimal k value.

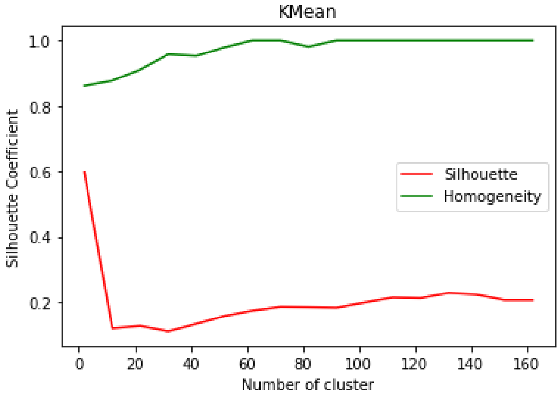
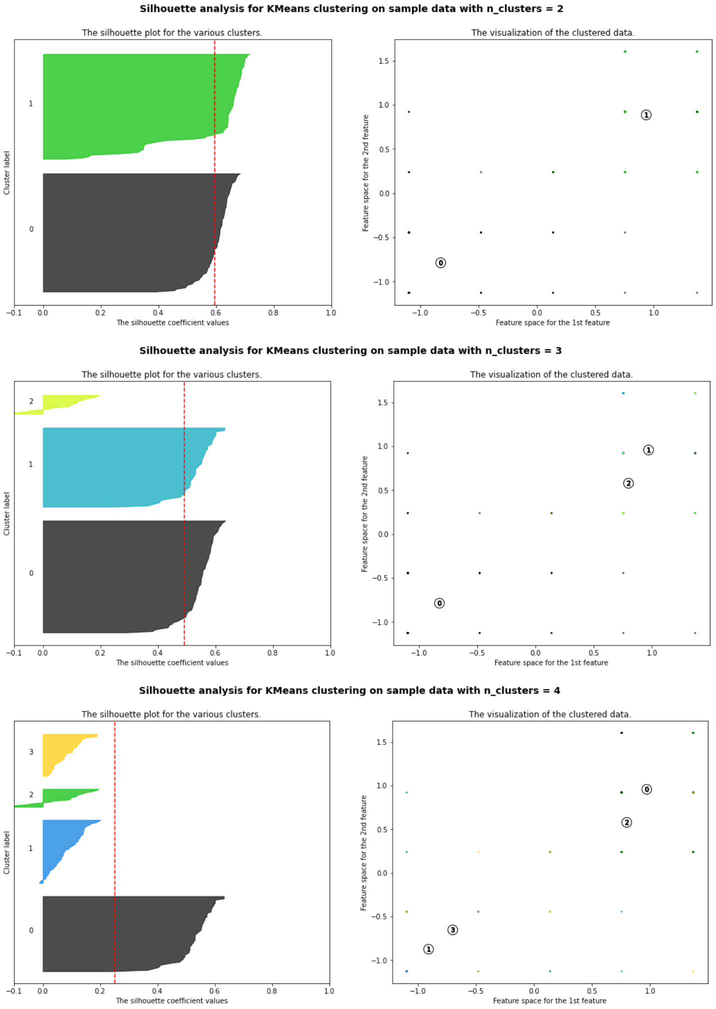


fig.2 Silhouette coeff. and homogeneity score vs #of cluster (divorce dataset)

The silhouette coefficient drops significantly but the homogeneity score only increases slightly with the number of clusters. It means increasing the number of clusters does not help sorting the data points better but making the k-means algorithm overfits the data points (as homogeneity reaches 1).

The optimal k value would be 2 where the silhouette coefficient has not dropped significantly yet while the homogeneity score is high but has not overfitted the data. This conclusion can be verified by looking at silhouette plots (fig.3).

n clusters	average silhouette score
2	0.5969155461245153
3	0.4931020826356653
4	0.2506974555813306


Silhouette analysis for KMeans clustering on sample data with n_clusters = 2

Silhouette analysis for KMeans clustering on sample data with n_clusters = 3

Silhouette analysis for KMeans clustering on sample data with n_clusters = 4

fig.3 Silhouette plots for n_cluster=2, 3, 4 (divorce dataset)

From fig.3, n_cluster=3 and 4 are not good picks since some cluster_lables are below average silhouette score, which means some data points are not similar to their own cluster labels.

n_cluster=2 is a good pick since both cluster_lables meet the average silhouette score and the thickness of both bars are similar, which aligns with the dataset's class distribution (a balanced dataset with 51% married, 49% divorced).

2.1.2. Contraceptive Dataset

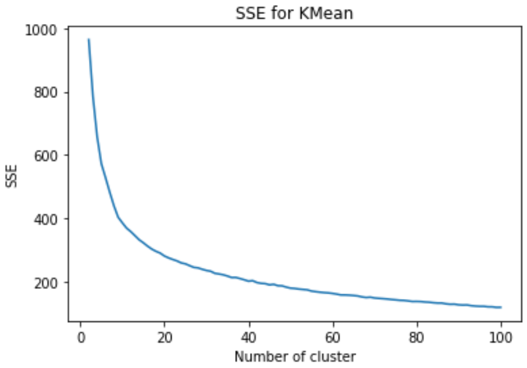


fig.4 SSE vs number of clusters (contraceptive dataset)

Similar to the divorce dataset, fig.4 shows no sharp elbow point. The only noticeable change is the curve starts flattening out at k=20. The optimal k would be somewhere around 20.

Further check on the silhouette coefficient and homogeneity score (fig.5), k=10 seems a better choice since the silhouette coefficient is the highest while the homogeneity is not too low.

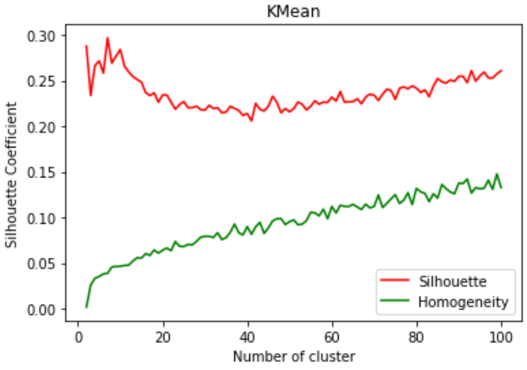
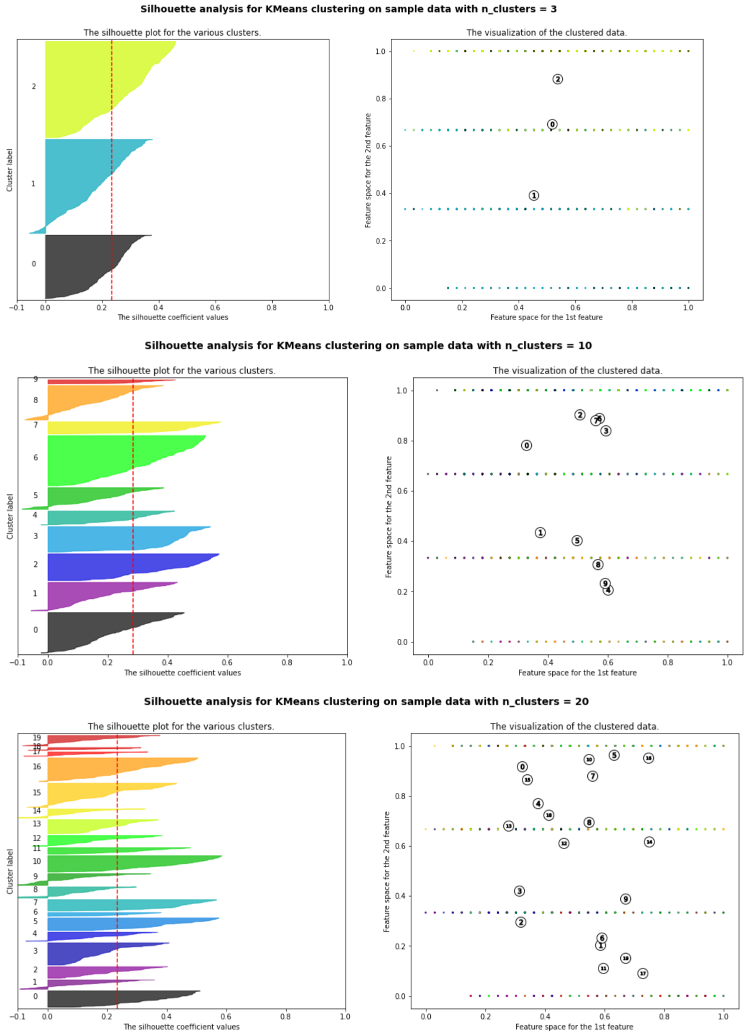


fig.5 Silhouette coeff. and homogeneity score vs #of cluster (contraceptive dataset)

The silhouette plots and their corresponding scores are listed below: (k=10 has the highest silhouette coefficient)

n_clusters	average silhouette_score
3	0.23356985296202146
10	0.28398130966676294
20	0.23422078357763906


Silhouette analysis for KMeans clustering on sample data with n_clusters = 3

Silhouette analysis for KMeans clustering on sample data with n_clusters = 10

Silhouette analysis for KMeans clustering on sample data with n_clusters = 20

fig.6 Silhouette plots for n_cluster=3, 10, 20 (contraceptive dataset)

Fig.6 shows all the cluster labels for n_clusters=3, 10, 20 respectively. They all reach the average silhouette score but contain negative silhouette coefficients for some cluster labels, which means some data points are wrongly clustered. The larger the k, the more the negative cluster labels. k=10 seems a better choice since it has the highest average silhouette score, contains less negative score labels, and less fluctuation in silhouette score among the labels.

The purpose of plotting n_cluster=3 is because this dataset has three class labels (42.7% no-use, 22.6% long-term, 34.7% short-term). The ideal number of clusters should be k=3. But turns out k=10 has a higher silhouette score than k=3, which means the features of this dataset suggest a different grouping from the original dataset's class labels. It may be because some features are not relevant to determine the class label (for example, features like "media exposure" may not be relevant in determining contraceptive methods used.). Thus making the clustering results different from the class label of the dataset.

2.2. Expectation Maximization (EM)

EM is an iterative algorithm which keeps alternating between estimating the log-likelihood of current estimates (E step) and maximizing the expected log-likelihood on that E step (M step) until convergence occurs.

The optimal number of clusters (k) is determined by observing the silhouette coefficient and the homogeneity score. Elbow method is not used here since EM uses probability distribution and covariance metrics to cluster data points and the elbow method does not give distinct results.

2.2.1. Divorce Dataset

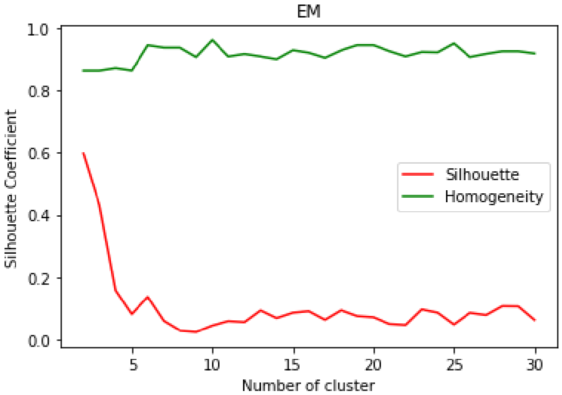


fig.7 Silhouette coeff. and homogeneity score vs #of cluster (divorce dataset)

Fig.7 shows k=2 seems a better choice since the silhouette coefficient is the highest and then drops significantly with increasing number of clusters. The homogeneity score is more or less the same no matter what number of clusters used.

Compared to the original dataset's class labels (51% married, 49% divorced), cluster number k=2 matches with the dataset's class labels (i.e. binary classification). Notice that this Gaussian Mixture Model (GMM) with cluster number k=2 has a f1 score 0.9764412416851441 which is very accurate in predicting labels.

2.2.2 Contraceptive Dataset

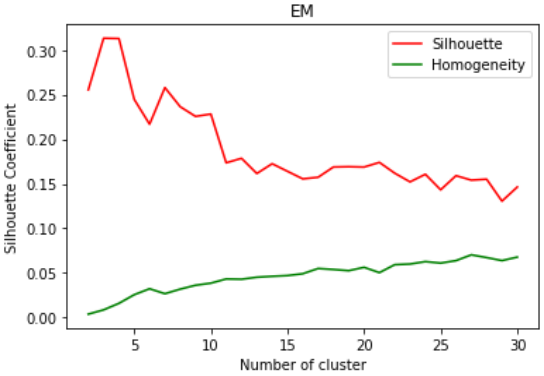


fig.8 Silhouette coeff. and homogeneity score vs #of cluster (contraceptive dataset)

Fig.8 shows the k=3 or 4 has the highest silhouette coefficient while the homogeneity slowly increases with increasing number of clusters. The table below listed the f1 score of cluster k=3, 4 and 5. k=4 has the highest score, which aligns with silhouette coefficient.

n_clusters	f1 score of GMM model on class labels
3	0.1166612995120018
4	0.26451855046180733
5	0.1598806987204774

Compared to the dataset class labels, this dataset has 3 class labels but the GMM model finds cluster number k=4 works the best. Such a discrepancy may be because some class labels are easier to identify than the others. GMM uses probability distribution to identify how likely a data point belongs to a certain class. A data point can be 50/50 likely belonging to two clusters or 33/33/33 likely belonging to three clusters. That means a data point can belong to many cluster labels if the features given cannot distinguish it confidently. As such, the optimal k found by GMM could differ from the dataset class labels.

Notice that the optimal k found by EM (k=4) is much closer to the dataset class labels than the optimal k found by k-means (k=10). It could be due to the soft clustering approach of EM which gives every data point a probability of belonging to which class labels than the hard clustering performed by k-means. Thus making the grouping better.

III. DIMENSIONALITY REDUCTION

3.1. Principal Component Analysis (PCA)

PCA reduces dimension by finding the orthogonal eigenvectors which can explain the maximum amount of variance in data points. The optimal number of components is found by observing cumulative explained variance and eigenvalues.

3.1.1. Divorce Dataset

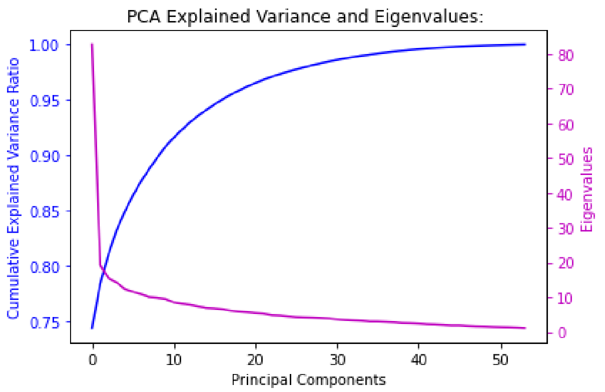


fig.9 Explained variance and eigenvalues vs #of PC (divorce dataset)

Fig,9 shows nearly 80% of the data variability can be explained by principal components (PC) 1 and 2. The eigenvalue drops significantly in the first 2 PCs and becomes flatten out after that. It means the latter PCs only explain a small fraction of the total variability.

The cumulative explained variance also shows over 80% variation in the dataset can be attributed by the first few PCs. PC=2 would be a good choice since it is the elbow point and over 80% of the dataset variability can be attributed to PC 1 and 2.

3.1.2. Contraceptive Dataset

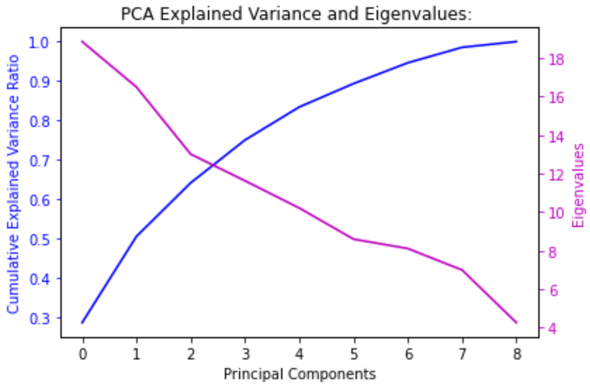
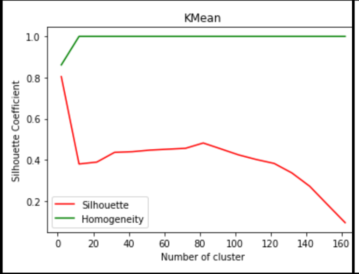
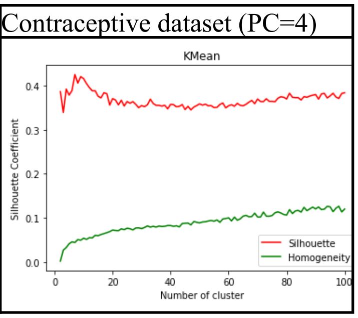
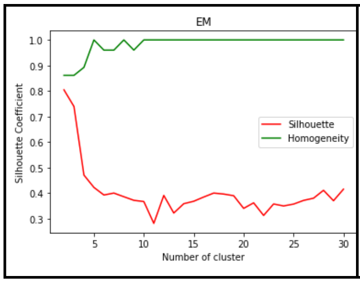
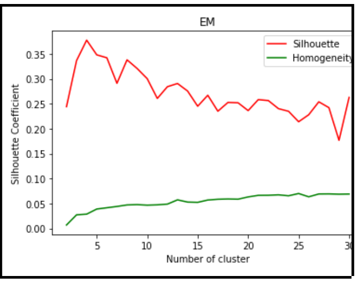


fig.10 Explained variance and eigenvalues vs #of PC (contraceptive dataset)

Fig.10 shows the eigenvalue drops moderately as PC increases. Unlike the previous dataset in which the eigenvalue drops significantly at first two PCs, this dataset shows a moderate amount of information can still be extracted even at PC4. The cumulative explained variance also shows at least 4 PCs is needed to describe over 80% of the dataset's total variance. Thus 4 PCs would be a better choice for this dataset.

3.1.3. Clustering Analysis after PCA

Divorce dataset (PC=2)	Contraceptive dataset (PC=4)
	
	

k-means and EM were re-run on both dimensionally reduced datasets. For both dimensionally reduced datasets, the shape of the silhouette and homogeneity curves looks similar to the original datasets but with smoother curves and higher silhouette scores. It means PCA removes unworthy information so that the clustering algorithms can sort the data points better. Thus giving better silhouette scores.

Divorce dataset:

highest silhouette score	before PCA	after PCA
k-means	~0.6	~0.8
EM	~0.6	~0.8

Contraceptive dataset:

highest silhouette score	before PCA	after PCA
k-means	~0.3	~0.4
EM	~0.3	~0.4

3.2. Independent Component Analysis (ICA)

ICA tries to find vectors which are independent of each other and maximize the difference between components. It attempts to separate mixed data points into subcomponents like separating sound signals from a mixture of sound sources. The optimal number of independent components (IC) is found by measuring the kurtosis (i.e. the non-gaussianity).

3.2.1. Divorce Dataset

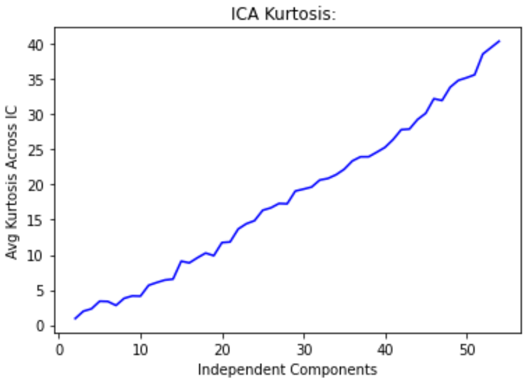


fig.11 Avg kurtosis across Independent Components (divorce dataset)

Fig.11 shows an almost linear increasing curve for the average kurtosis across IC. It makes sense that as more ICs are added, more data points can be separated from a mixture of points. Thus maximizing non-Gaussianity. But this curve has no sharp spikes which means ICA on this dataset might not be an effective dimension reduction method to extract information.

Small spikes appear at IC=15 and 45 which could be suitable choices for dimension reduction. IC=15 was picked for greater dimension reduction.

3.2.2. Contraceptive Dataset

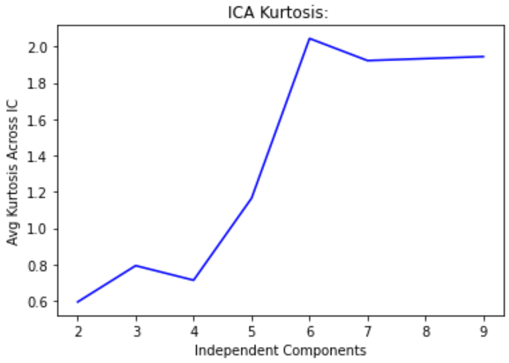
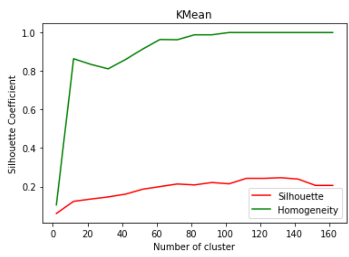
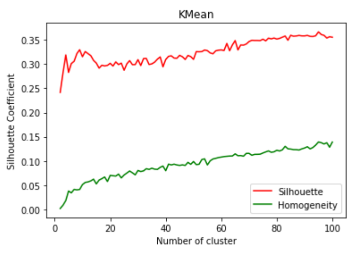
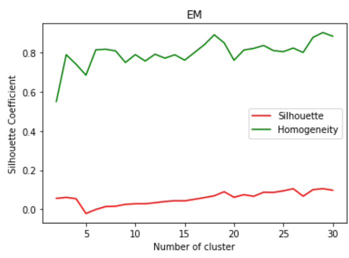
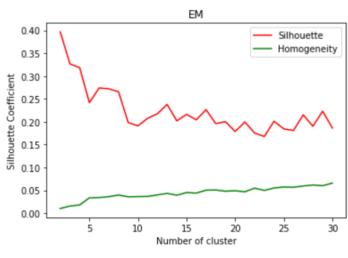


fig.12 Avg kurtosis across Independent Components (contraceptive dataset)

Unlike the previous dataset, this one (fig.12) shows two sharp spikes appearing at IC=3 and IC=6 which would be suitable choices for optimal values of IC. IC=6 was chosen since it has the highest average kurtosis which means more information can be extracted.

3.2.3. Clustering Analysis after ICA

Divorce dataset (IC=15)	Contraceptive dataset (IC=3)
	
	

Compared to the clustering results of the original dataset, ICA performs poorly on the Divorce Dataset as the silhouette scores for both k-means and EM are much lower than the original clustering results. The data points in this dataset may not be a linear combination of hidden variables thus ICA might not be a suitable method to reduce this dataset's dimension.

But ICA performs well on the Contraceptive Dataset for both k-means and EM algorithms since the silhouette coefficient is higher than the original clustering results. It could be because the data points in this dataset can easily be extracted by finding mutually independent vectors. This better clustering result is also supported by the kurtosis graph in previous discussion.

Divorce dataset:

highest silhouette score	before ICA	after ICA
k-means	~0.6	~0.2
EM	~0.6	~0.2

Contraceptive dataset:

highest silhouette score	before ICA	after ICA
k-means	~0.3	~0.35
EM	~0.3	~0.4

3.3. Randomized Projections (RP)

RP reduces dimensions by projecting the original high dimensional data onto a randomly generated matrix which is a lower dimensional space. It trades accuracy for faster processing times. Here reconstruction error was used as a measure to choose the optimal number of components.

3.3.1. Divorce Dataset

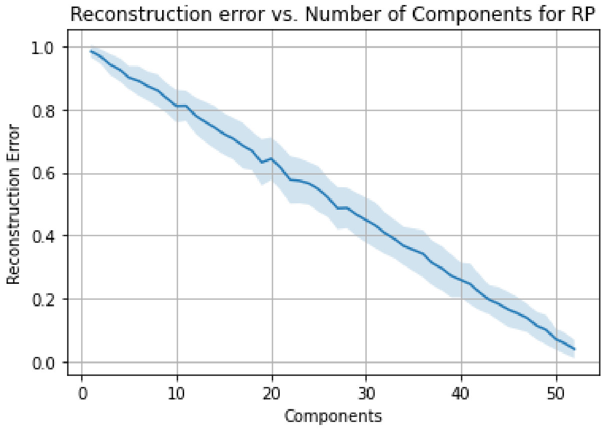


fig.13 Reconstruction error vs #of components in RP (Divorce dataset)

Fig.13 shows the reconstruction error is almost linearly decreasing with the number of components. The same curve was reproduced after several runs of the RP algorithm. Components=25 was chosen since it reduced the total reconstruction error by half.

3.3.2. Contraceptive Dataset

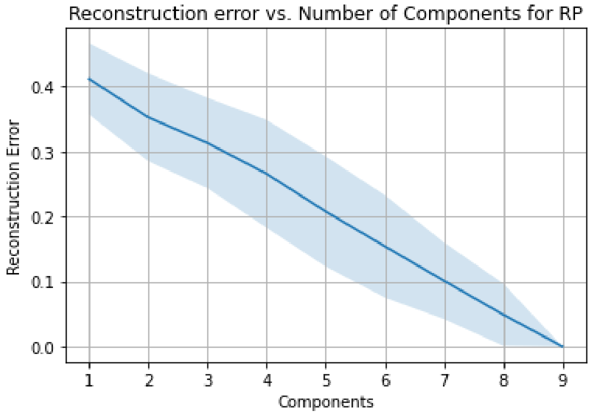
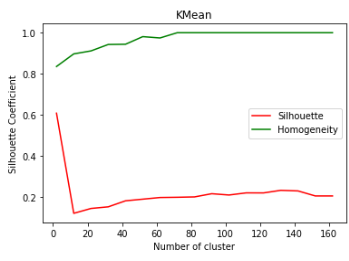
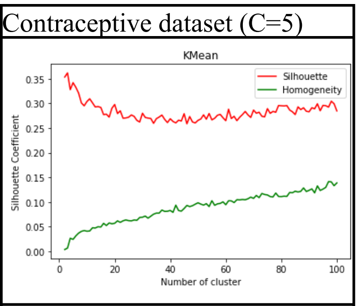
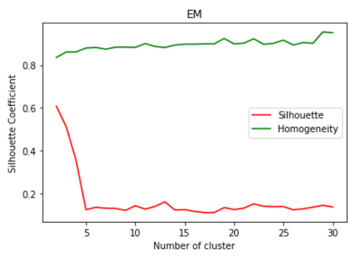
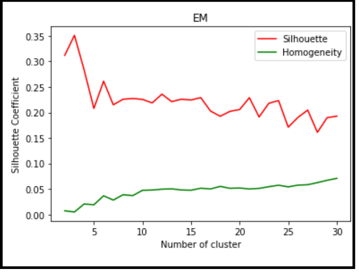


fig.14 Reconstruction error vs #of components in RP (contraceptive dataset)

Similar to the previous dataset, fig.14 shows this dataset also has a linearly decreasing reconstruction error with the number of components. Components=5 was chosen since it reduced the largest reconstruction error (0.4) by half.

3.3.3. Clustering Analysis after RP

Divorce dataset (C=25)	Contraceptive dataset (C=5)
	
	

Comparing to the original dataset, k-means and EM clustering results on Divroce Dataset are similar to the clustering results of its original dataset. They have the same highest silhouette and similar shape of curves. It means even after dimension reduction by RP, the data points are still sorted in a way that is similar to its original arrangement. It could be because RP reduced dimension by preserving the distance between two data points. Thus the data points grouping would be similar to its original.

The clustering results of the Contraceptive Dataset is slightly better than its original dataset since the dimensionally reduced dataset has higher silhouette score for both k-means and EM. Both curves also look smoother. It means RP helps this dataset rearrange the data points better such that both k-means and EM gives better clustering results.

Divorce dataset:

highest silhouette score	before RP	after RP
k-means	~0.6	~0.6
EM	~0.6	~0.6

Contraceptive dataset:

highest silhouette score	before RP	after RP
k-means	~0.3	~0.35
EM	~0.3	~0.35

3.4. Linear discriminant analysis (LDA)

LDA tries to find a linear combination of features which can maximize the separability of two or more classes of objects. It is a supervised algorithm since it has already known the labels and tries to find the best linear separators to divide the labeled data.

3.4.1. Divorce Dataset

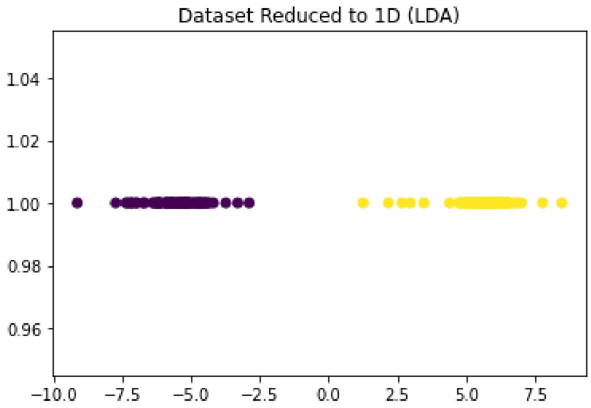


fig.15 LDA scatter plot (divorce dataset)

Fig.15 shows a one-dimension scatter plot of the data points of the Divorce Dataset. Since this dataset contains two classes labels, LDA reduces it to (classes-1) dimension i.e. one-dimension from 53 features by finding the best separator. Notice that the data points were separated clearly thus the homogeneity score should be very high when clustering algorithms run on this dimensionally reduced dataset.

3.4.2. Contraceptive Dataset

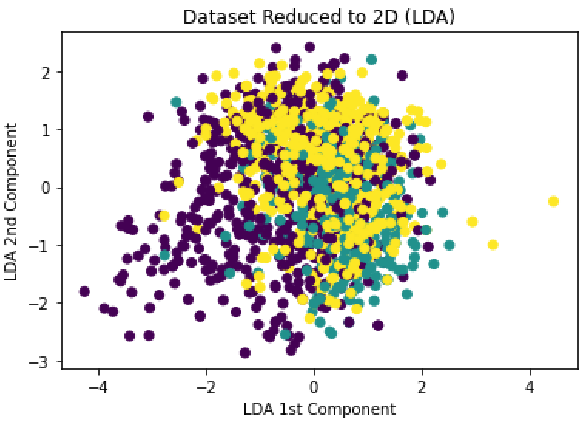
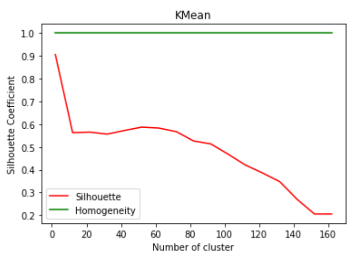
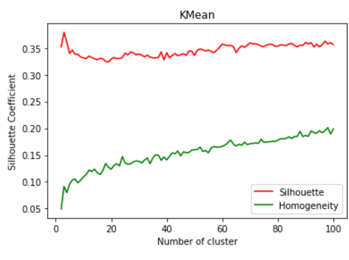
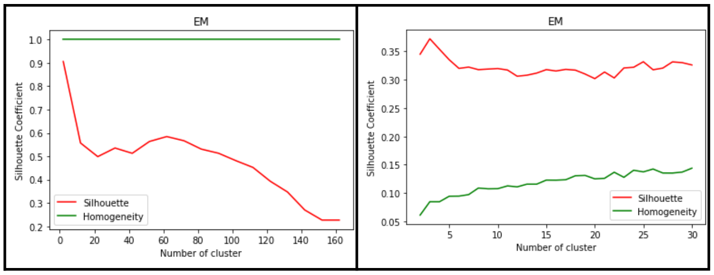


fig.16 LDA scatter plot (contraceptive dataset)

Since this dataset has 3 class labels (long-term, short-term and no-use), LDA reduces it to a 2D scatter plot. But the data points in this 2D plane still cannot be separated as cleanly as the previous dataset. The homogeneity score is expected to be low for the clustering results after LDA.

3.4.3. Clustering Analysis after LDA

Divorce dataset (1D)	Contraceptive dataset (2D)
	



As expected, the Divorce dataset has a perfect homogeneity score regardless of the number of clusters for both k-means and EM algorithms. It also has a very high silhouette coefficient at n_cluster=2 for both k-means and EM. It means LDA helps k-means and EM cluster data points better and finds the optimal k easier.

For the Contraceptive Dataset, the curves look similar to the original k-means and EM with a slightly better silhouette score and higher homogeneity score, though the improvement is not as significant as the Divorce Dataset. It makes sense that after LDA on the original data, the scatter plot does not give a clean and highly separable data points distribution as what the Divorce Dataset does. Thus the clustering result of this dataset is not as good as the Divorce dataset.

Divorce dataset:

highest silhouette score	before LDA	after LDA
k-means	~0.6	~0.9
EM	~0.6	~0.9

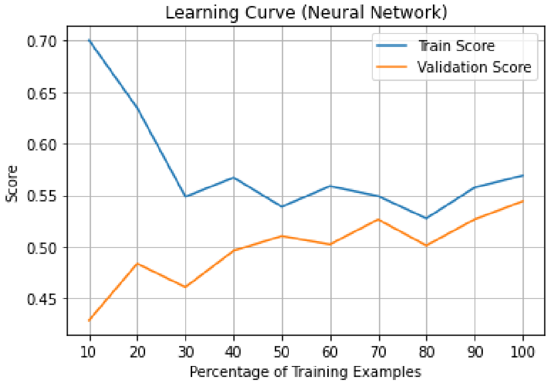
Contraceptive dataset:

highest silhouette score	before LDA	after LDA
k-means	~0.3	~0.35
EM	~0.3	~0.35

IV. NEURAL NETWORK AFTER DIMENSION REDUCTION

Contraceptive Dataset was chosen to re-run my original neural network learner on the dimensionally reduced dataset. Choosing this dataset because it has enough samples for train/test splits than the Divorce dataset. Divorce dataset only contains 170 entries which leaves a very small portion for the validation set and this could lead to biased validation results..

4.1. Original neural network (as benchmark for comparison)



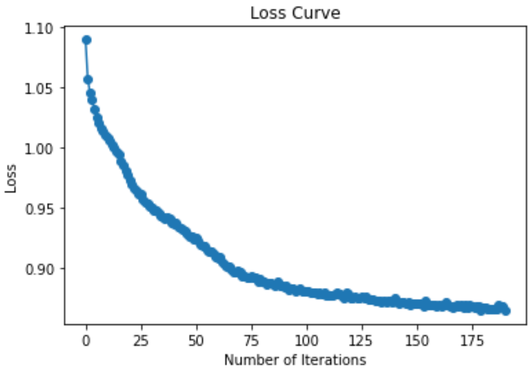
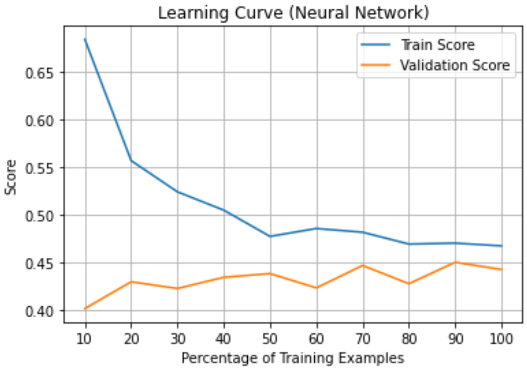


fig.17 Neural network based on original contraceptive dataset

In assignment 1, the best neural network was trained with two hidden layers (13, 13) and learning rate 0.01. The best f1 score for this neural network is 0.54. This serves as a benchmark to compare with the 4 neural networks trained after 4 dimension reduction algorithms.

original	Grid search best param
#of hidden layers	13
learning rate	0.01
f1 score	0.5466443435495659

4.2. PCA



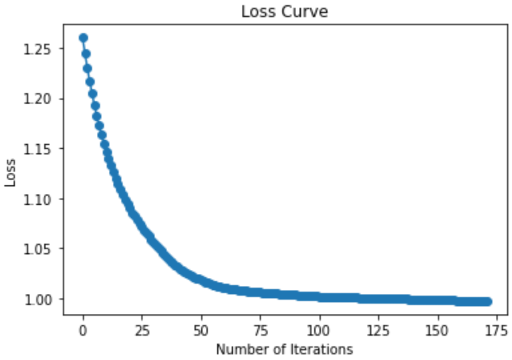


fig.18 Neural network on PCA dataset

Grid search was used to find the best f1 score that the neural network can reach in this PCA reduced dataset. PCA reduces the original dataset from 9 features to 4 features. The neural network trained can achieve a f1 score of 0.43 which is lower than 0.54 of the original neural network. This could be due to the information loss of dimension reduction. Since only 80% of the total dataset variance was explained by 4 PCs. Some information is yet to be extracted with a higher number of PCs in the PCA algorithms. Thus resulting in lower validation score of this neural network.

But the loss curve of this reduced dataset shows a faster loss rate to a stable low value within the first 50 iterations. It means PCA reduces the datasets in a way that helps the neural network converge faster.

PC=4	Grid search best param
#of hidden layers	17
learning rate	0.001
f1 score	0.4351261212763207

4.3. ICA

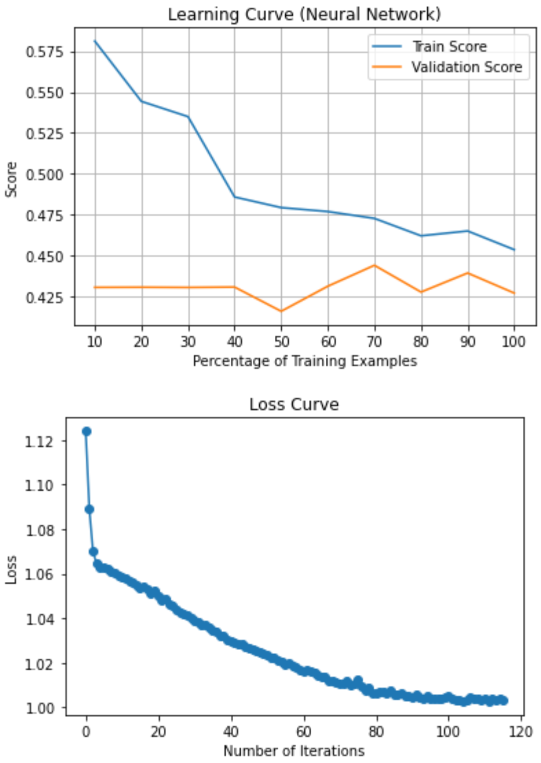


fig.19 Neural network on ICA dataset

This ICA algorithm reduces the dataset from 9 features to 6. Grid search best params for hidden layers and learning rate can reach a f1 score 0.47 which is better than PCA but still lower than the original neural network. It means some information loss in ICA causes the prediction accuracy not as high as the original neural network.

IC=6	Grid search best param
#of hidden layers	5
learning rate	0.01
f1 score	0.47231423275306256

But this neural network only uses 5 nodes in a hidden layer which is less than the original neural network to achieve its best f1 score. It means ICA sorts the data in a way that a less complex neural network model is enough to fit the dataset, though sacrificing some prediction accuracy.

The loss curve has an abrupt drop at the first few iterations and then linearly decreases with the number of iterations. It means the neural network learns most of the pattern during the first few iterations and then the learning rate decreases slowly.

4.3. RP

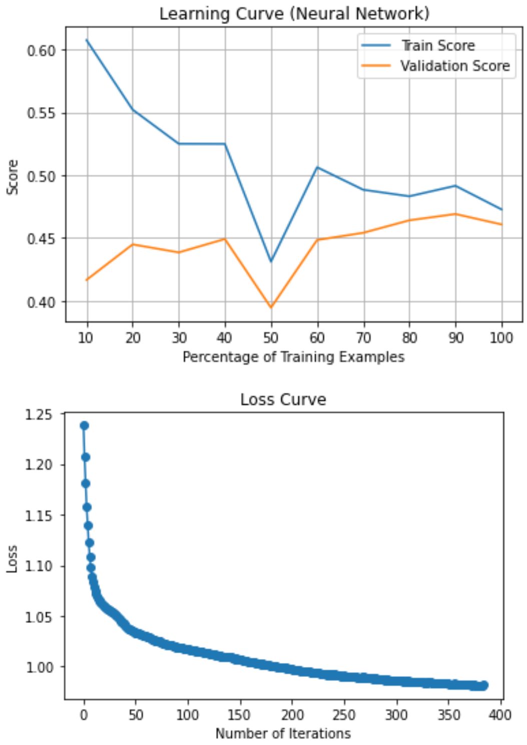


fig.20 Neural network on RP dataset (# of component=5)

Component=5 was chosen for RP dimension reduction. Grid search best params shows the highest f1 score is 0.43 which is the same as PCA but still lower than the original neural network. This loss curve is similar to the one in PCA but with more number of iterations to reach stable low values. It could be a sign of overfitting since the training loss continues to decrease with a number of iterations way more than PCA does.

Component=9 (i.e. total number of features of this dataset) was also tested to see if validation score get improved:
(best f1 score: 0.47031539888682744)

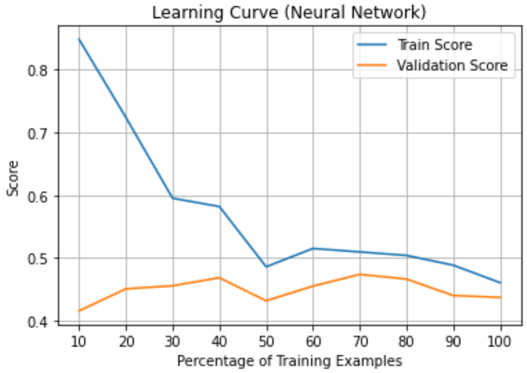


fig.21 Neural network on RP dataset (# of component=total features)

Even with all the components used to train the neural network, the best f1 score (0.47) is still lower than the original (0.54). It means RP cannot capture enough information from the original dataset.

component=5	Grid search best param
#of hidden layers	11
learning rate	0.001
f1 score	0.4366698511301947

component=9	Grid search best param
#of hidden layers	15
learning rate	0.01
f1 score	0.47031539888682744

4.4. LDA

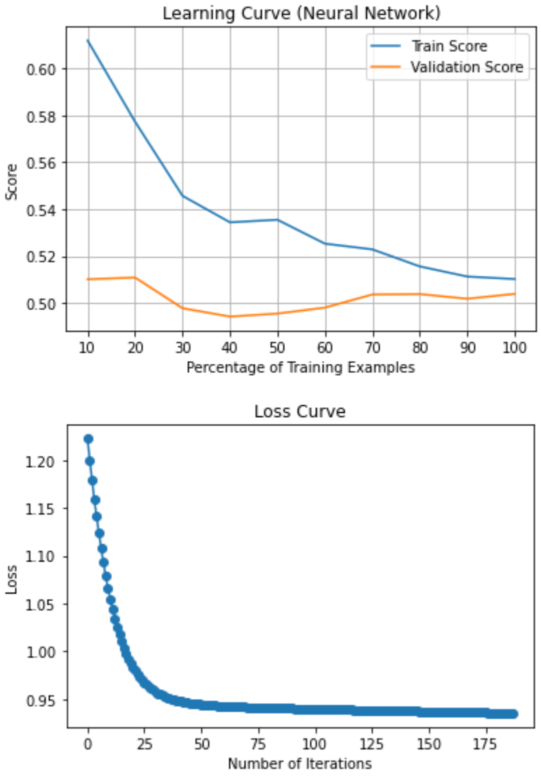


fig.22 Neural network on LDA dataset

LDA gives the highest f1 score among all the dimension reduction algorithms, though it is slightly lower than the f1 score in the original dataset (0.54). But considering the fact that LDA reduces dimension from 9 to 2 while still keeping its f1 score close to the benchmark, it is doing a very good job in preserving information from the original dataset.

component=2	Grid search best param
#of hidden layers	15
learning rate	0.001
f1 score	0.508973710819009

This loss curve looks similar to the PCA but with a faster loss rate than PCA. It converges at 25 iterations which means this dataset helps the neural networks converge in a way faster than PCA does.

Notice that LDA is a supervised learning algorithm which is different from the previous three dimension reduction algorithms. Thus it makes sense that LDA does a better job since it knows the class labels already.

V. NEURAL NETWORK AFTER CLUSTERING

K-means and EM clustering algorithms were run on the original Contraceptive Dataset. The clustering result is then added as a new feature to the original dataset, i.e. there are two datasets generated (one for k-means and one for EM). Then the neural network from assignment 1 was trained based on these two dataset for analysis.

5.1. K-means

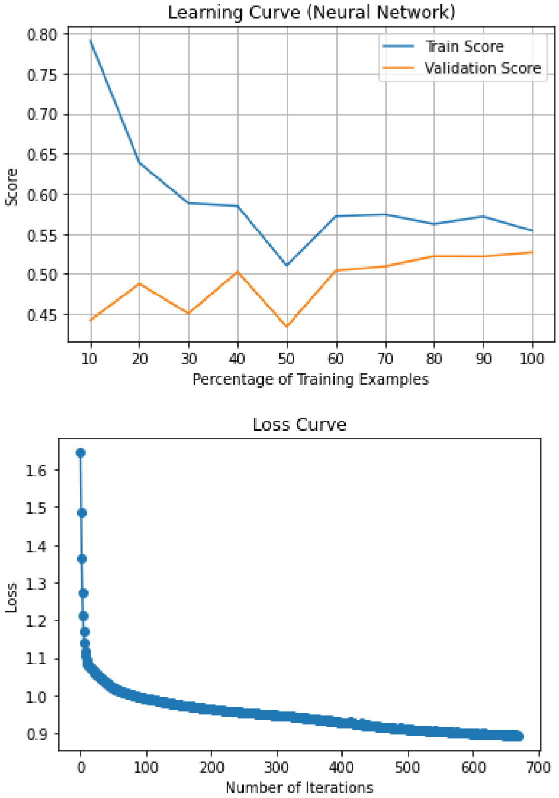


fig.23 Neural network on K-means clustering

The f1 score of this dataset is close to, yet still not better than the original f1 score. It means adding the clustering results (k=10) as one more feature to the original dataset does not help the neural network to predict the class label more accurately. The neural network trained is more complex than the original one with more nodes and slower learning rate.

Since this is a 3 class labels dataset, k=10 clustering labels are not lining up with the original class labels which might somehow mislead the neural network to learn the model. Thus making the prediction accuracy not as high as the original accuracy.

k=10	Grid search best param
#of hidden layers	14
learning rate	0.001
f1 score	0.517657261280495

This loss curve shows the neural networks need much more iterations to reach stable low values. But the first 50 iterations already contributed the most training loss. The linear decrease after iterations 50 could be a sign of overfitting thus an early stop may be helpful in training this neural network.

5.2 EM

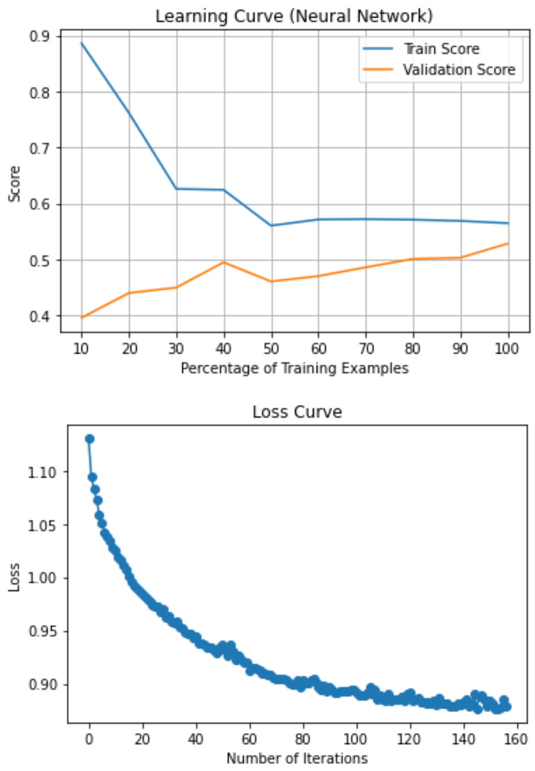


fig.24 Neural network on EM clustering

Similar to the k-means results, EM's f1 score is close to but not better than the original f1 score. But EM's prediction is better than k-means. It could be due to the soft-clustering nature of EM in which class labels are defined closer to the original class labels than k-means does. But still, EM's clustering labels are not helping the neural network to predict class labels better than the original neural network.

n_components=4	Grid search best param
#of hidden layers	9
learning rate	0.01
f1 score	0.5230782086860537

This loss curve looks similar to the one in the original neural network. It could be because EM's clustering results are similar to the original class labels thus giving a similar training loss to the original neural network.

VI. CONCLUSION

Among the six algorithms (4 dimension reduction, 2 clustering), none of them reaches a better f1 score than the original neural networks.

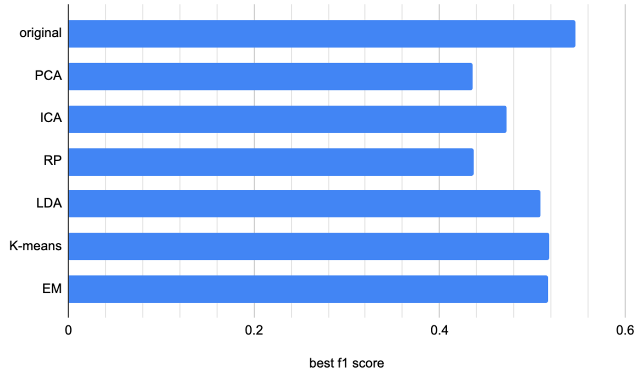


fig.25 F1 score of all neural networks tested

Among the 4 dimensionally reduced datasets, LDA performs the best due to its supervised nature in which class labels have been known already. RP and PCA perform the worst due to not enough information being extracted to explain the data's variance.

Among the 2 clustering-labels-added datasets, EM performs better than k-means. It could be because the EM soft-clustering labels are more similar to the original class labels than k-means clustering labels. But both clustering algorithms cannot achieve a better f1 score than the original neural networks. It could be because the clustered labels are not in line with the original class labels which mislead the neural networks.

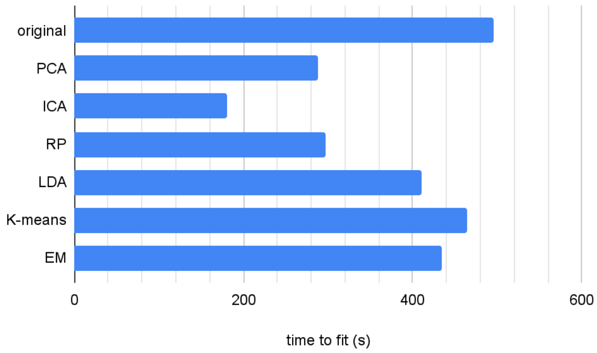


fig.26 Training time of all neural networks tested

In terms of training time, all six neural networks use less training time than the original networks. It makes sense that the 4 dimensionally reduced datasets use less training time due to less number of features being considered during neural network training.

But what's interesting is that the neural networks in k-means and EM also use less training than the original one. They are supposed to have longer training time since they added one more feature to the original dataset. One possible explanation to that would be that the clustered labels added as an additional feature summarize other features in the dataset which helps the neural networks learn the data pattern faster.

VII. REFERENCE

[1]	Introduction to Linear Discriminant Analysis in Supervised Learning. (2019, November 27). Analytics Steps. https://www.analyticssteps.com/blogs/introduction-linear-discriminant-analysis-supervised-learning
[2]	Gupta, M. (2020, August 9). Random Projection for Dimension Reduction | by Mehul Gupta | Data Science in your pocket. Medium. https://medium.com/data-science-in-your-pocket/random-projection-for-dimension-reduction-27d2ec7d40cd