CS7641 Assignment 3: Unsupervised Learning and Dimensionality Reduction

Xiaoming Su

Introduction

This assignment aims to analyze the performance of 2 clustering algorithms and four dimensionality reduction algorithms: 1. ***K-means*** 2. ***Expectation Maximization (EA)***

Four dimensionality reduction algorithms are explored by re-applying clustering algorithms and neural network algorithm using the data of reduced dimensionality: 1 ***PCA***; 2 ***ICA***; 3 ***Randomized Projections (RP)***; 4 ***LDA***

The strength and weakness of different algorithms are analyzed and the parameters for each problem are optimized as well. There are 5 parts in the report, which correspond to the 5 parts in the assignment requirements.

Dataset Introduction

Two datasets used are pima-indians-diabetes and spambase. Spam dataset is studied in assignment 1 and 2. The pima dataset is studied in assignment 2.

I think that pima dataset is complementary to the spam dataset, as one of the major differences is the number of attributes. Since the algorithms studied in this assignment focus on clustering and dimensionality reduction, the difference in dimension can provide better comparison between various algorithms. All the values of attributes are real numbers. All the data are normalized before the experiments.

Table 1. Datasets Information

Name of Database	Number of Attributes	Number of Instances	Number of Classes
Spam	57	4601	2 (binary)
Pima Indians diabetes	8	768	2 (binary)

1. Clustering algorithms on the datasets

The K-means and EM (Gaussian mixture model) algorithms are implemented using sklearn.

Four types of plots are generated to evaluate algorithm performance: 1 Information criterion plot; 2 Score summary plot; 3 Elbow plot; 4 Silhouette plot.

Determination of K: The two datasets are all binary which may indicate that K=2 is a good choice. The number of clusters for the experiments are explored ranges from 2 to 20. There are many experimental methods to determine K such as the elbow method, X-means clustering and information criterion approach [1]. The elbow method and information criterion approach are utilized in this part to determine the number of clusters.

1.1 K-means

L2 (Euclidean) is used as distance metric for similarity measurement, which fits the data sets' properties. The K-means algorithm clusters data by trying to separate samples in n groups of equal variance, minimizing a criterion called inertia or within-cluster sum-of-squares.

1.1.1 Experimental result

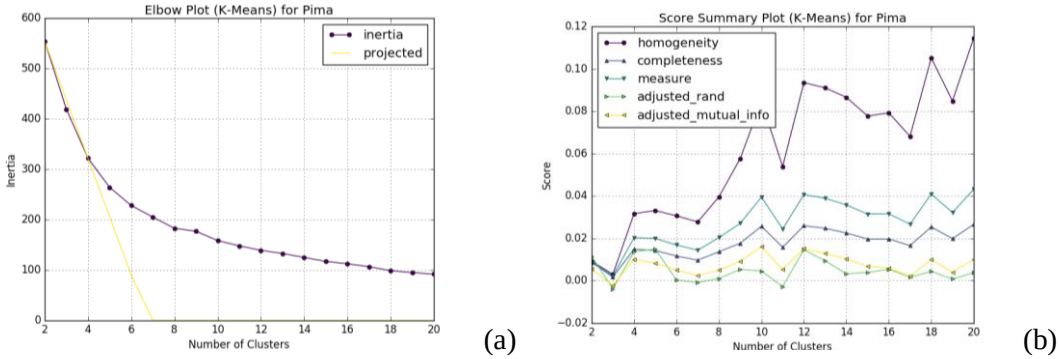


Figure 1. Pima dataset clustered by K-means (a) Elbow plot (b) Score summary plot

For the pima dataset, Figure 1 (a) shows that as the number of clusters increases, the inertia decreases. From the score plot, all the score are relatively low. The homogeneity increases as the number of clusters increases, all the

other scores increase very slightly. From the plot, based on the "elbow criterion", the best K for pima is 4.

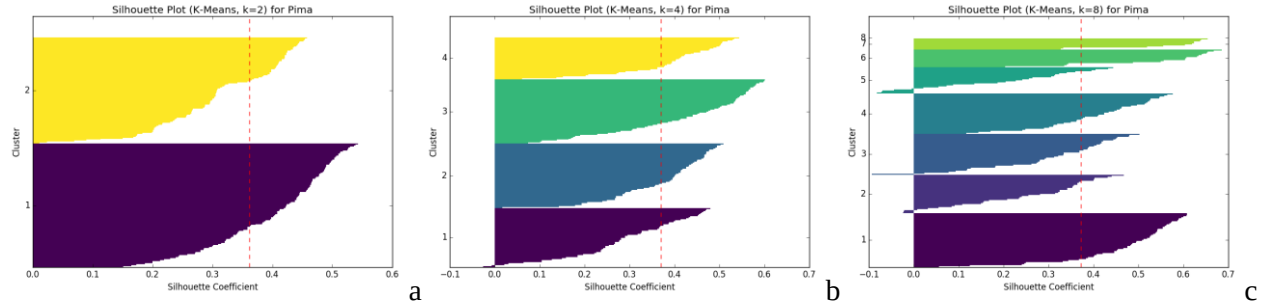


Figure 2. Silhouette plot for K-means of Pima (a) k = 2 (b) k =4 (c) k=8

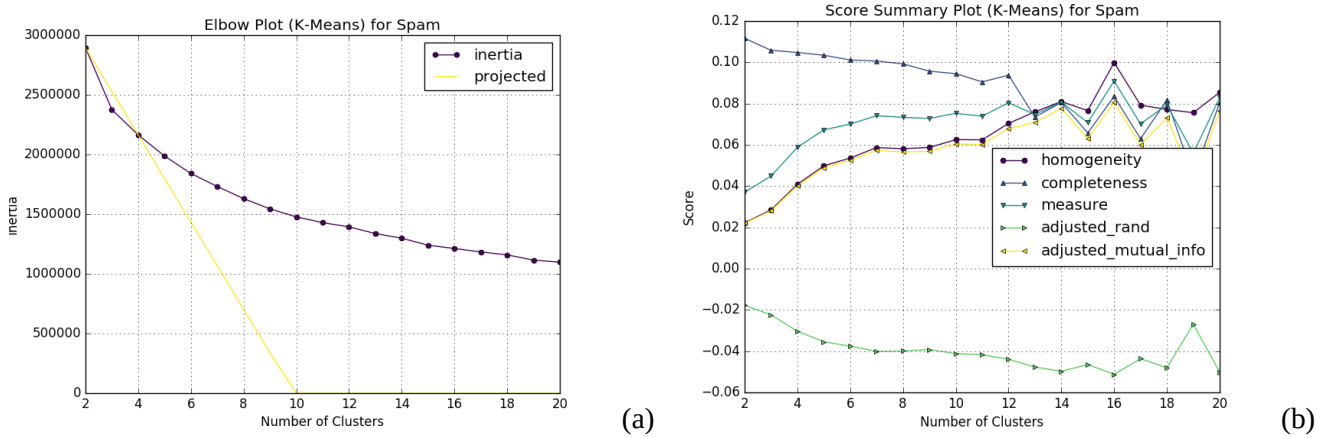


Figure 3. Spam dataset clustered by K-means (a) Elbow plot (b) Score summary plot

For the spam dataset, Figure 3 (a) shows that as the number of clusters increases, the inertia decreases. From the score plot, all the score are relatively low. Based on the "elbow criterion", the best K for spam is 2.

1.1.2 Discussion

The performance of K-means is evaluated by elbow plot which plots the inertia vs number of clusters, and score summary plot in which the measured scores are different measurement of the averaged “similarity” of all the data within each cluster. Inertia, or the within-cluster sum of squares criterion, can be recognized as a measure of how internally coherent clusters are, the smaller number, the better result. 0 is the optimal. For the scores, the higher score, the larger similarity within each cluster, thus the better results.

The correct choice of k is often ambiguous, with interpretations depending on the shape and scale of the distribution of points in a data set and the desired clustering resolution of the user. In addition, increasing k without penalty will always reduce the amount of error in the resulting clustering, to the extreme case of zero error if each data point is considered its own cluster. Thus the optimal choice of k will strike a balance between maximum compression of the data using a single cluster, and maximum accuracy by assigning each data point to its own cluster [1]. Based on the experiment and according to "elbow criterion", the best K for spam is 2, K for pima is 4. From the experimental results, the scores for both datasets are not very high, because the K-means minimizes inertia, which suffers from two drawbacks: **1.** Inertia makes the assumption that clusters are convex and isotropic, which is not always the case. It responds poorly to elongated clusters, or manifolds with irregular shapes. **2.** Inertia is not a normalized metric: in very high-dimensional spaces, Euclidean distances tend to become inflated [2]. The low scores of two datasets indicate that the properties of the datasets are not convex and isotropic clusters and the plots in part 3 further demonstrates this point.

1.2 EM

A Gaussian mixture model (EM) is a probabilistic model that assumes all the data points are generated from a mixture of a finite number of Gaussian distributions with unknown parameters. The mixture models can be considered as generalizing K-means clustering to incorporate information about the covariance structure of the

data as well as the centers of the latent Gaussians. Thus, by incorpartion of more parameters and less data assumption, EM usually generates better result than K-means. The experimental results confrim this point.

1.2.1 Experimental results

Based on the Pima information criterion plot Figure 4 (a), the lowest BIC score corresponds to the best number of clusters, which means that the best K for Pima dataset is 4. For Pima score plot, the scores are higher than that in the result of K-means method.

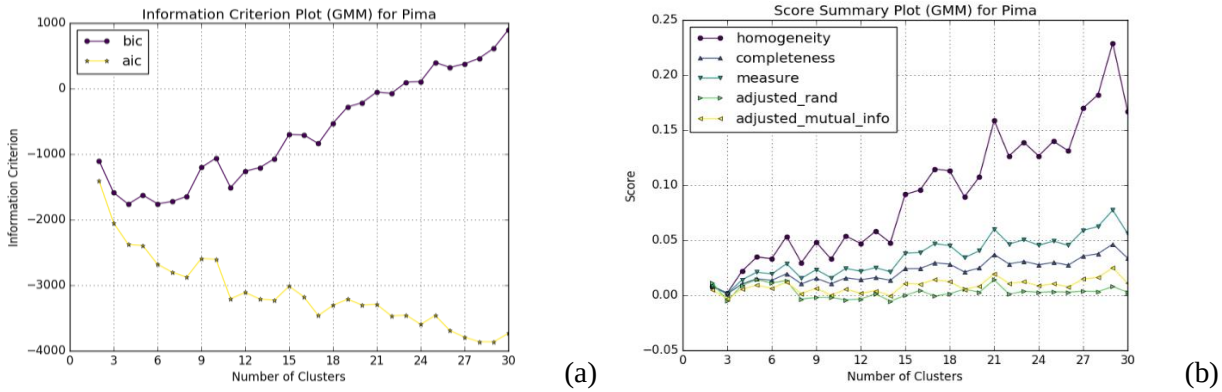


Figure 4. Pima dataset clustered by EM (a) Information criterion plot (b) Score summarty plot

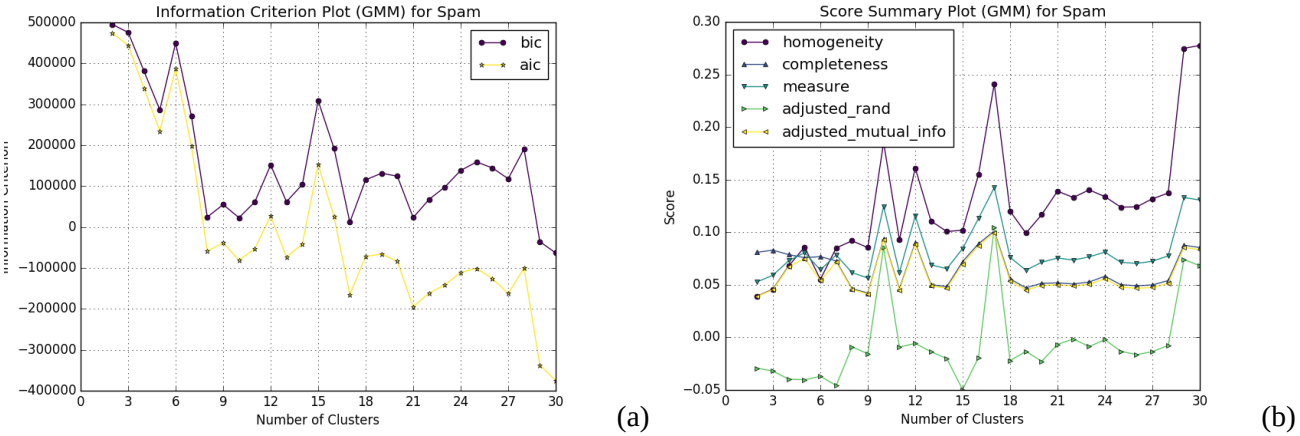


Figure 5. Spam dataset clustered by EM (a) Information criterion plot (b) Score summarty plot

Based on the spam information criterion plot Figure 5 (a), the lowest BIC score corresponds to the best number of clusters, which means that the best nubmer of clusters for spam dataset is 8. For Pima score plot, the scores are higher than the result of K-means method. Because the scores show no obivious change when the number of cluster ranges from 2 to 8 , based on the elbow method, the best number can be considered as 2. However, when the nubmer of clusters are larger (>8), there are some high scores.

1.2.2 Discussion

In the information criterion plot, Akaike information criterion (AIC) and the Bayes information criterion (BIC) can be used to estimate number of cluters. BIC is a criterion for model selection among a finite set of models; the model with the lowest BIC is preferred. It is based, in part, on the likelihood function and it is closely related to the AIC. In general, BIC generates better result than AIC. The score plot can also help to determine the best number of clusters based on “elbow criterion”.

1.3 Analysis and Comparison of K-means and EM

Difference: K-means scales well to large number of samples and has been used across a large range of application areas in many different fields. EM is not scalable. K-means ***hard*** assigns a data point to one particular cluster on convergence. It makes use of the L2 norm (a type of distance metric) when optimizing. EM ***soft*** assigns a point to clusters (it give a probability of any point belonging to any centroid). It doesn't depend on distances of points,

but is based on the Expectation, i.e., the probability of the point belonging to a particular cluster. This makes K-means biased towards spherical clusters.

Relationship: K-means is equivalent to the expectation-maximization algorithm with a small, all-equal, diagonal covariance matrix.

Performance: Given enough time, K-means will always converge, however this may be to a local minimum. This is highly dependent on the initialization of the centroids. Expectation-maximization is guaranteed to always converge to a local optimum [3]. For the two datasets studied, EM always performs K-means.

For the running time, the EM is much faster than K-means.

EM Algorithm: Good at handling Outliers; Better clustering results; Takes more iteration

K-Means Algorithm: Not too well with Outliers; Higher variations; Determined by initial means; Requires less iterations than EM [4]

Running time: EM is much faster than K-means. gmm_spam: 104.457 seconds gmm_pima: 6.381 seconds kmeans_spam: 396.640 seconds kmeans_pima: 153.400 seconds

Parameter tunning: For these two datasets, the results of EMS and K-means are mostly influenced by the number of clusters. Results can be improved by changing the number of clusters. Other parameters such as covariance type and convergence threshold do not have much influence for these two datasets.

Table 2. Comparison of K-means and GMM [2]

Method name	**Parameters**	**Scalability**	**Use case**	**Geometry (metric used)**
K-means	number of clusters	Very large n_samples, medium n_clusters	General-purpose, even cluster size, flat geometry, not too many clusters	Distances between points
EM (Gaussian mixtures)	many	Not scalable	Flat geometry, good for density estimation	Mahalanobis distances to centers

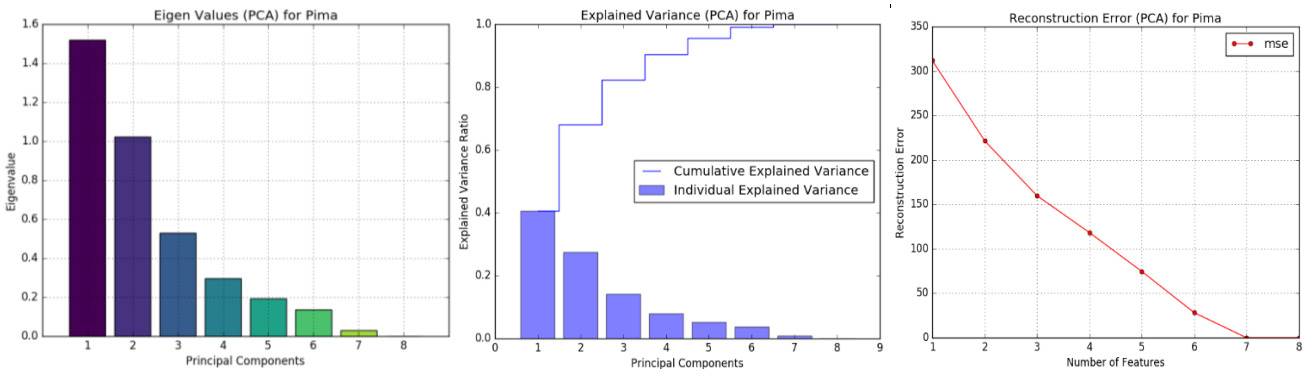
2. Dimensionality reduction algorithms

All the four algorithms studied are subspace analysis methods.

2.1 Principal Component Analysis (PCA).

PCA finds vectors such that projections on to the vectors capture maximum variance in the data.

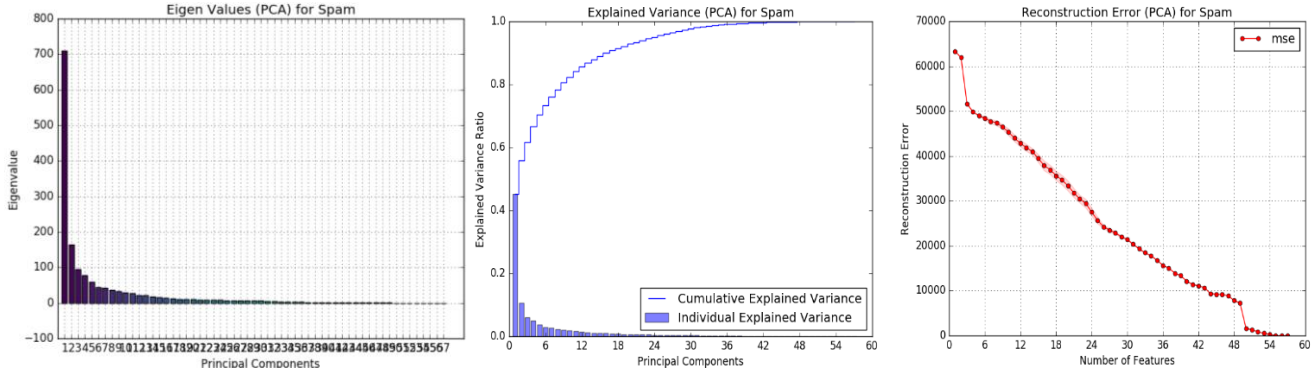
2.1.1 Experimental results



(a) (b) (c)

Figure 6. Pima dataset PCA (a) Eigen Values (b) Explained Variance (C) Reconstruction Error

For the pima dataset, Figure 6 (a) shows the Eigen values of principal components in a sorted order. There's steady decrease in the Eigen value and the Eigen value becomes 0 for the 8th component. Pima has 8 attributes. Plot (b) shows the variance for different components, it's very clear the PCA maximizes variance. Plot (c) shows that the construction error decreases almost linearly with the increase of number of components.



(a) (b) (c)

Figure 7. Spam dataset PCA (a) Egin Values (b) Explained Variance (C) Reconstruction Error

For the spam dataset, Figure 7 (a) shows the Eigen values of principal components in a sorted order. The steady Eigen value decreases really fast when the number of components are small. Spam has 57 attributes. Plot (b) shows the variance for different components, which looks very similar to the Eigen value plot. Plot (c) shows that the construction error decrease almost linearly with the increase of number of components.

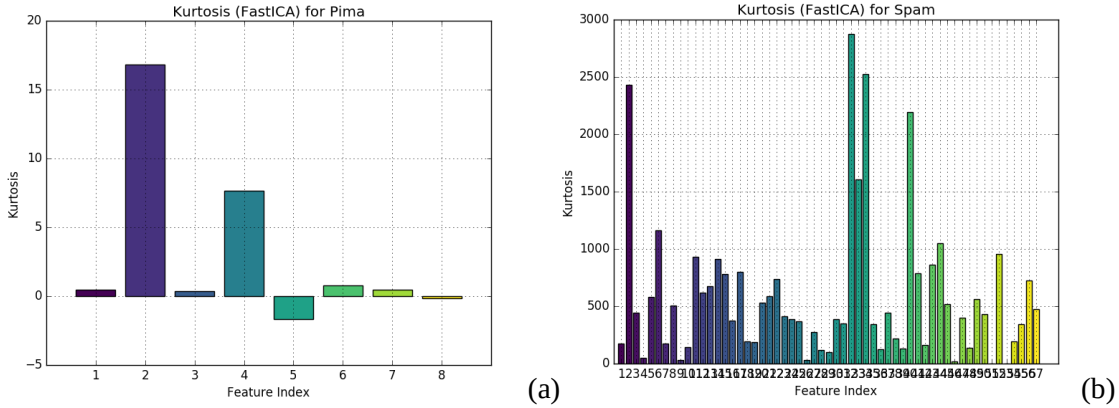
2.1.2 Discussion

PCA results for both datasets based on reconstruction error are good. PCA is based on the second-order statistics, and it works well for the data which fit Gaussian distribution. It's obvious the reconstruction error decrease with increase of number of components.

2.2 Independent Component Analysis (ICA)

PCA considered image elements as random variables with Gaussian distribution and minimized second-order statistics. For any non-Gaussian distribution, largest variances would not correspond to PCA basis vectors. ICA minimizes both second-order and higher order dependencies in the input data and attempts to find the basis along which the data (when projected onto them) are statistically independent [5].

2.2.1 Experimental results



(a) (b)

Figure 8. Kurtosis for (a) Pima with 8 features (b) Spam with 57 features

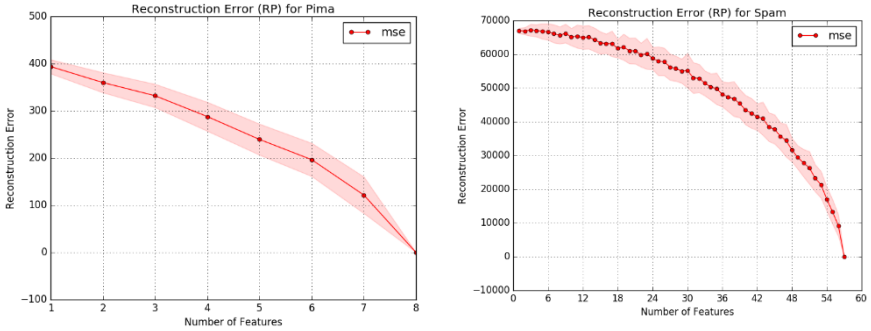


Figure 9. Reconstruction Error for
(a) Pima
(b) Spam

For pima, the kurtosis has large values for feature 2 and 4 and very small values for other features, which means that the distribution are spikier than Gaussian on features 2 and 4, and almost Gaussian distribution for other

features. For spam, the overall kurtosis are much larger than that of pima. Kurtosis are large positive values for most of the features which means that the distribution is supergaussian.

2.2.2 Discussion

The ICA method depends on certain measurement of the non-Gaussianity (kurtosis). Kurtosis measures the degree of peakedness of a distribution and it is zero only for Gaussian distribution. Any other distribution's kurtosis is either positive if it is supergaussian (spikier than Gaussian) or negative if it is subgaussian (flatter than Gaussian). Therefore the absolute value of the kurtosis or kurtosis squared can be used to measure the non-Gaussianity of a distribution [5]. However, kurtosis is very sensitive to outliers, and it is not a robust measurement of non-Gaussianity. The performance of ICA for spam is better for pima, and the projection axes for ICA are meaningful.

2.3 Random Projection (RP)

This part is implemented using Gaussian random projection. RP reduces the dimensionality by projecting the original input space on a randomly generated matrix where components are drawn based on certain probability.

2.3.1 Experimental results

RP has three parameters, n_components, eps and random state. The result is mainly influenced by the number of components. Eps and random state has little influence on the reconstruction error. Rerunning RP will result in different reconstruction error but the influence is not that much, there is not vary large variation for these two datasets. Figure 10 shows the kurtosis for pima and spam. All the features has similar kurtosis values for both datasets.

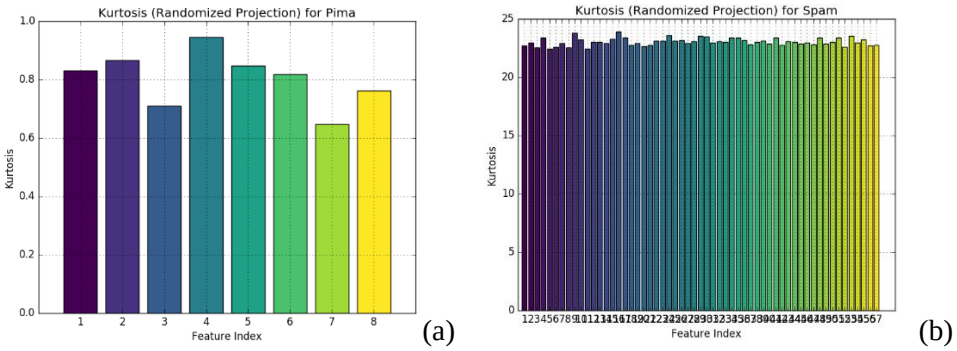


Figure 10. Kurtosis for (a) Pima (b) Spam

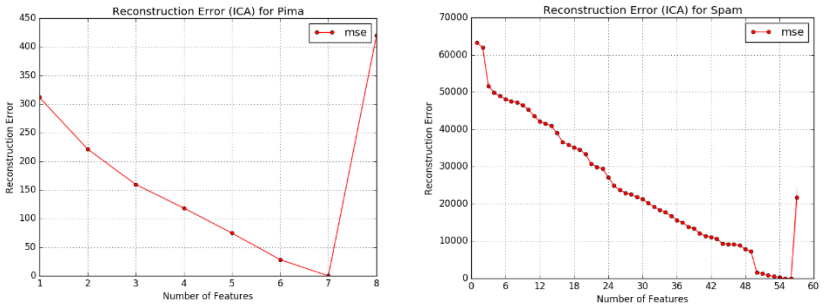


Figure 11. Reconstruction Error for
(a) Pima
(b) Spam

2.3.2 Discussion

RP reduces the dimensionality by projecting the original input space on a randomly generated matrix where components are drawn based on certain probability, thus the performance of each run may differ from each other because of randomness. In general, the performance of RP for these two datasets doesn't very that much, probably because the kurtosis calculated for different features are pretty similar.

2.4 Linear Discriminant Analysis (LDA)

The objective of LDA is to perform dimensionality reduction while preserving as much of the class discriminatory information as possible.

2.4.1 Experimental results

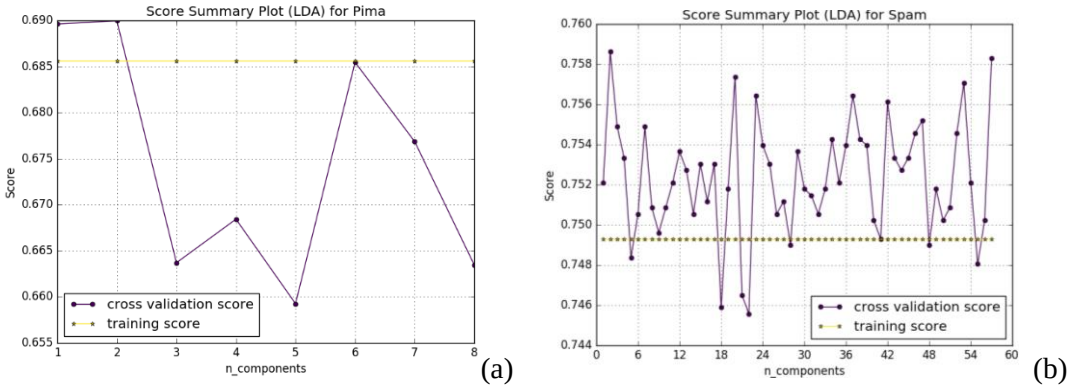


Figure 12. Score Summary Plot for (a) Pima 8 features (b) Spam 57 features

The score is the mean accuracy on the given test data and labels. For both of the datasets, the best score are when the number of components are 2. It's because both of the datasets are binary.

2.4.2 Discussion

LDA uses the class information and finds a set of vectors that maximize the between-class scatter while minimizing the within-class scatter.

2.5 Analysis and comparison of PCA, ICA, RP, LDA

RP: RP is very fast due to its random selection properties and thus it has relatively worse performance compared with PCA. RP is a simple and computationally efficient way to reduce the dimensionality of the data by trading a controlled amount of accuracy (as additional variance) for faster processing times and smaller model sizes. Based on reconstruction error, RP gets worse results for two datasets compared with ICA and PCA.

PCA, ICA, LDA comparison and discussion: The research results are often contradictory on comparing the performance of these three algorithms. For example, Liu and Wechsler (1999) claim that ICA outperforms PCA, while Baek et al. (2002) claim that PCA is better. Beveridge et al. (2001a) claim that in their tests LDA performed uniformly worse than PCA, and Belhumeur et al. (1996) and Navarrete and Ruiz-del-Solar (2002) claim that LDA outperforms PCA [5]. Thus, the performance of PCA, ICA and LDA are very dataset specific, and no particular projection–metric combination is the best across all tests [6].

PCA minimizes the covariance of the data; on the other hand ICA minimizes higher-order statistics such as fourth-order cummulant (or kurtosis), thus minimizing the mutual information of the output. Specifically, PCA yields orthogonal vectors of high energy contents in terms of the variance of the signals, whereas ICA identifies independent components for non-Gaussian signals. ICA thus possesses two ambiguities: First, the ICA model equation is underdetermined system; one cannot determine the variances of the independent components. Second, one cannot rank the order of dominant components.

LDA, ICA (not including Nonlinear ICA) and PCA are linear transformation techniques. LDA is a supervised whereas PCA and ICA are unsupervised (ignores class labels). LDA is a parametric method since it assumes unimodal Gaussian likelihoods. If the distributions are significantly non-Gaussian, the LDA projections will not be able to preserve any complex structure of the data that may be needed for classification. LDA will fail when the discriminatory information is not in the mean but rather in the variance of the data [7].

	PCA	**ICA**	**RP**	**LDA**
Spam	314.064	3.310	16.781	7.508
Pima	3.055	0.818	1.711	0.783

Table 3. Running time for different algorithms.

3. Reproduce your clustering experiment

Four dimension reduction algorithms were applied to two clustering algorithms for two datasets. Thus there were 4x2x2 = 16 combinations. The Silhouette plots were also generated, but there's no space to show all of them in

this report. Please refer to the results folder for all the plotted graphs.

3.1 Experimental results

8 score summary plots were generated for each of the datasets. Selected plots are shown in Figure 13. For both of the datasets, two clustering methods get worse results on dimensionally reduced data generated by PCA, ICA and RP compared with none-reduced one. LDA give comparable result which show that it works well for both datasets.
For pima dataset, for both EM and K-means, the performance ranking is the same LDA>RP>ICA>PCA.
For spam data set, for both EM and K-means, the performance ranking is the same: LDA>>PCA>RP>ICA.

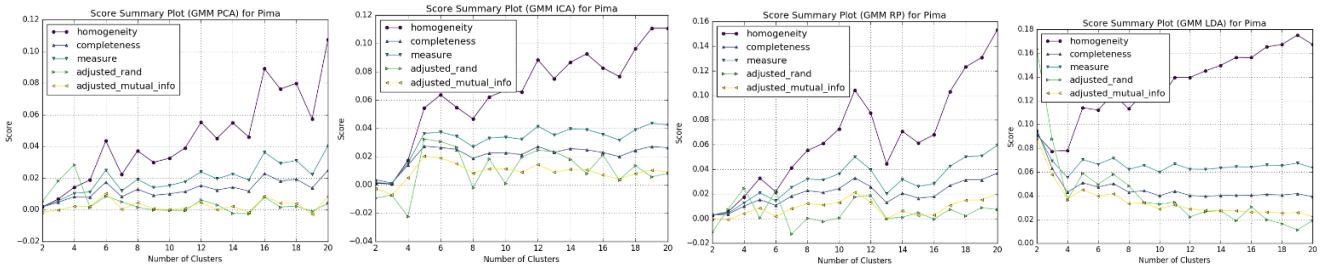
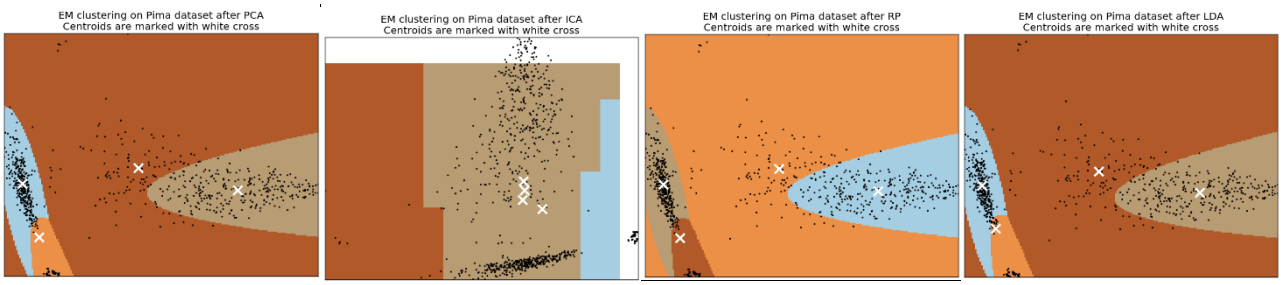
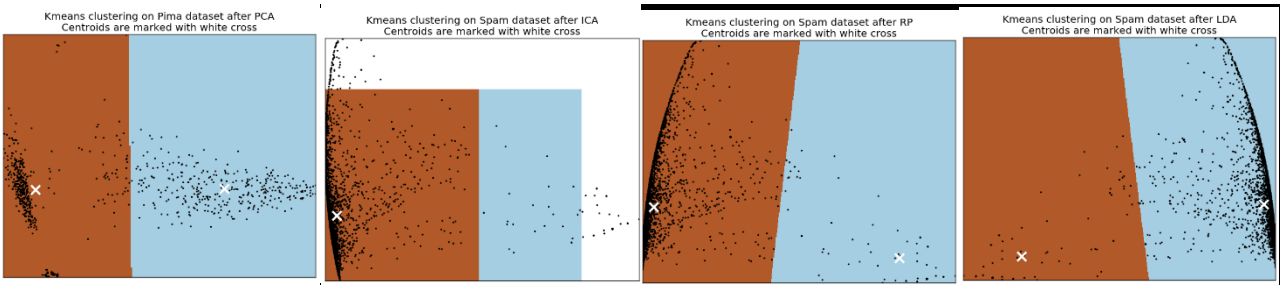


Figure 13. EM results for Pima (a) PCA (b) ICA (c) RP (d) LDA

I reproduced clustering experiments on the datasets projected onto the new spaces created by ICA, PCA, RP and LDA. The clusters are not the same as before. Besides the different clustering scores, Figure 14 can also demonstrate this point by visualizing the clusters.



EM _PCA _PIMA_4 (a) EM_ICA _PIMA_4 (b) EM_RP_PIMA_4 (c) EM_LDA_PIMA_4 (d)



Km_PCA _SPAM_2 (d) Km_ICA _SPAM_2 (f) Km_RP_SPAM_2 (g) Km_LDA_SPAM_2 (h)

Figure 14. The clustering method, dimension reduction method, name of dataset and number of clusters are shown in the plot name. For pima, the best number of cluster is 4. For spam, the best number of cluster is 2.

3.2 Analysis and comparison

LDA gives almost better result among the dimension reduction algorithms for spam and pima datasets. I think the exploitation of class information gives the LDA advantage over the studied datasets. To better view the clusters, the data are reduced to 2 dimensions and plotted in Fugure 13. The white mark represents centroid. Different algorithms have different clusters. Also different reduction algorithms have different assumptions, if the assumption matches the data property, then the algorithm can capture useful information which indicates good result. The running time is shown in Table 4 for K-means method.

	PCA	**ICA**	**RP**	**LDA**
Pima	145.271	193.360	100.707	102.566
Spam	397.686	354.628	237.836	611.214

Table 4. Running time (second) for different dimension reduction algorithms plus K-means clustering method.

4. Neural network learner on the newly projected data with PCA/ICA/RP/LDA

Due to the curse of dimensionality, the classification of highly dimensional data usually derives lower dimensional spaces to do the actual classification while retaining as much information (energy) from the original images as possible. The spam database was chosen for this experiment.

4.1 Experimental results

I re-ran the neural network for the reduced data using PCA, ICA, RP and LDA. The learning curve are shown in Figure 15. The result is similar to the performance of four algorithms for spam dataset in part 3.

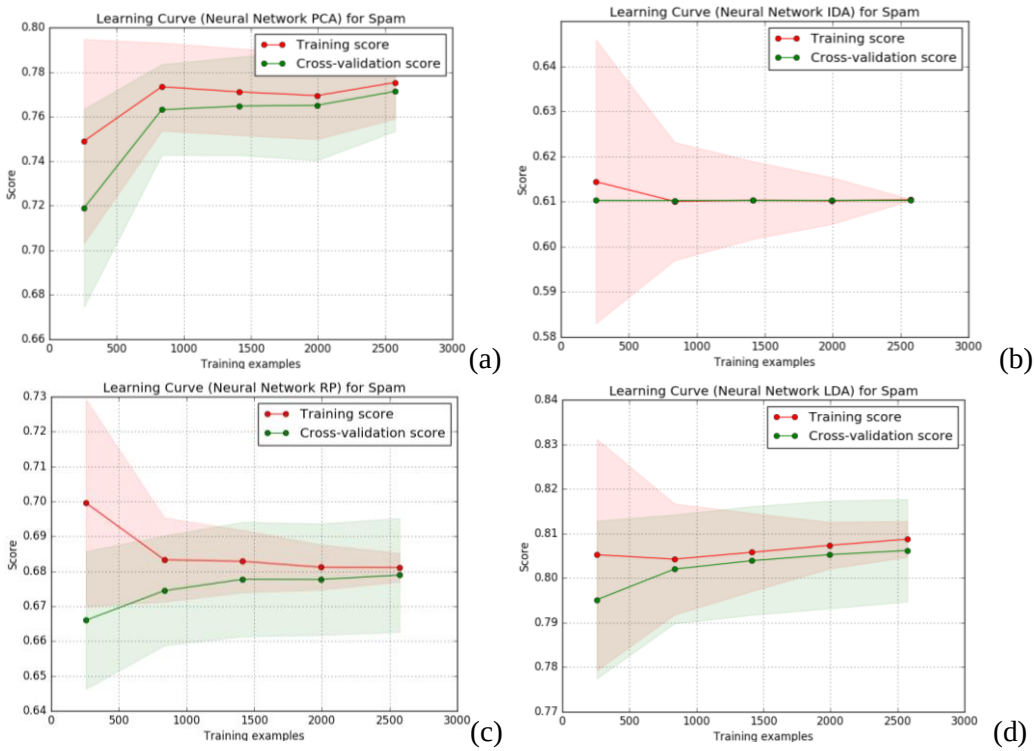


Figure 15. Learning curve of NN using reduced data generated by (a) PCA (b) ICA (c) RP (d) LDA

The performance of the four algorithms on pima dataset ranks as follows: LDA>PCA>RP>ICA. All the result if worse than the none reduced full dataset. The training time ranks as follows:

	PCA	**ICA**	**RP**	**LDA**
Pima	1.570 s	0.277 s	2.940 s	3.033

Table 5. Training time (second) for the reduced data using different dimension reduction algorithms

4.2 Discussion

The performance of fours dimension reduction algorithms is the same as the result found in part 3. It indicates that the reduction algorithm that can best capture the properties of the data and retain as much information as possible will give the best result in neural network. For the spam dataset, LDA has the best performance by utilizing class information. Compared with original data, the performance are all worse than the original data which is obvious because the original data contains all the information.

5. Neural network learner on the newly projected data with clustering

Two clustering algorithms ere applied to spam dataset to which I just applied the dimensionality reduction algorithms, clusters were treated as new features.

5.1 Experimental results

The results of K-means and EM clustering methods are very similar. The dimension reduction algorithm that generated the reduced data prior applying clustering plays important role in the learning accuracy. Generally, the clustering method applied to LDA-reduced data has the best performance based on the result of Figure 16.

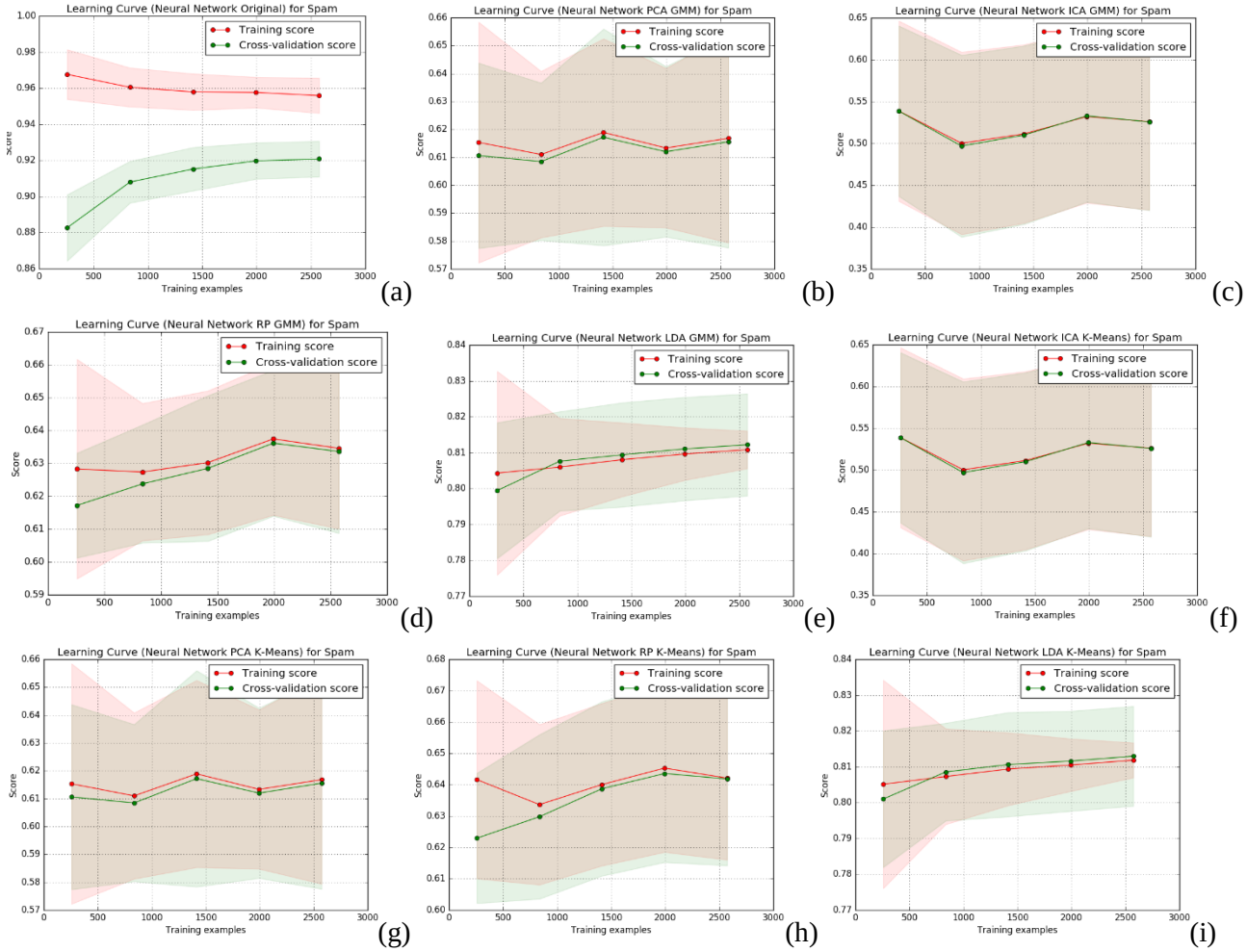


Figure 16. Learning curve for (a) original data (b) PCA EM (c) ICA EM (d) RP EM (d) LDA EM (f) PCA K-means (g) ICA K-means (h) RP K-means (i) LDA K-means

	PCA	**ICA**	**RP**	**LDA**
EM	0.249 s	0.202 s	1.192 s	0.552 s
K-means	0.317 s	0.254 s	0.825 s	0.972 s

Table 5. Training time including clustering as new features. Original data: training time is 1.814 s

5.2 Discussion

Clustering methods were treated as a dimension reduction algorithm. K-means and EM have similar results. Comparing the learning curve form part 4 and part 5, I found that the learning accuracy were very similar. However, the training time for the neural network is much less compared with the training time in part 4. Thus, clustering provides another method to reduce the data while maintaining comparable information. By utilizing clustering method, the training time can be reduced a lot. In order to get ideal performance, the dimension reduction algorithm that was used to generate the reduced data prior clustering need to be chosen carefully by considering data property.

Reference

[1] https://en.wikipedia.org/wiki/Determining_the_number_of_clusters_in_a_data_set

[2] http://scikit-learn.org/stable/modules/clustering.html

[3] http://scikit-learn.org/stable/modules/mixture.html

[4] http://athena.ecs.csus.edu/~vanggs/177_finalpres.pdf

[5] http://fourier.eng.hmc.edu/e161/lectures/ica/node4.html

[6] Delac K, Grgic M, Grgic S. Independent comparative study of PCA, ICA, and LDA on the FERET data set [J]. International Journal of Imaging Systems and Technology, 2005, 15(5): 252-260.

[7] http://courses.cs.tamu.edu/rgutier/cs790_w02/l6.pdf