COMP4641 Assignment #3
Liu Qinhan (qliu359@gatech.edu)

Unsupervised Learning Report

Dataset

Breast Cancer Data This is a set of data about the breast cancer diagnosis, that is, whether the tested cell is of malignant (bad) or benign (mild). 357 benign and 212 malignant samples with 30 features of each sample are included. The features are computed, as described on the website, based on the digitized images of a fine needle aspirate (FNA) of a breast mass. In the training process, the classification is indicated by "1" (malignant) and "0" (benign).

Credit Card Fraud Data This dataset contains totally 284,807 transactions made by European cardholders in Sep. 2013. 492 frauds are recorded inside taking up only 0.172% of all transactions. 29 features are included and the fraud transactions are classified by "1".

Clustering

The idea of clustering is attempting to somehow group the instances together so that the instances in the same group are rather more similar to each other than those of other groups. In this assignment, K-Means clustering and Expectation Maximization are implemented to explore the datasets.

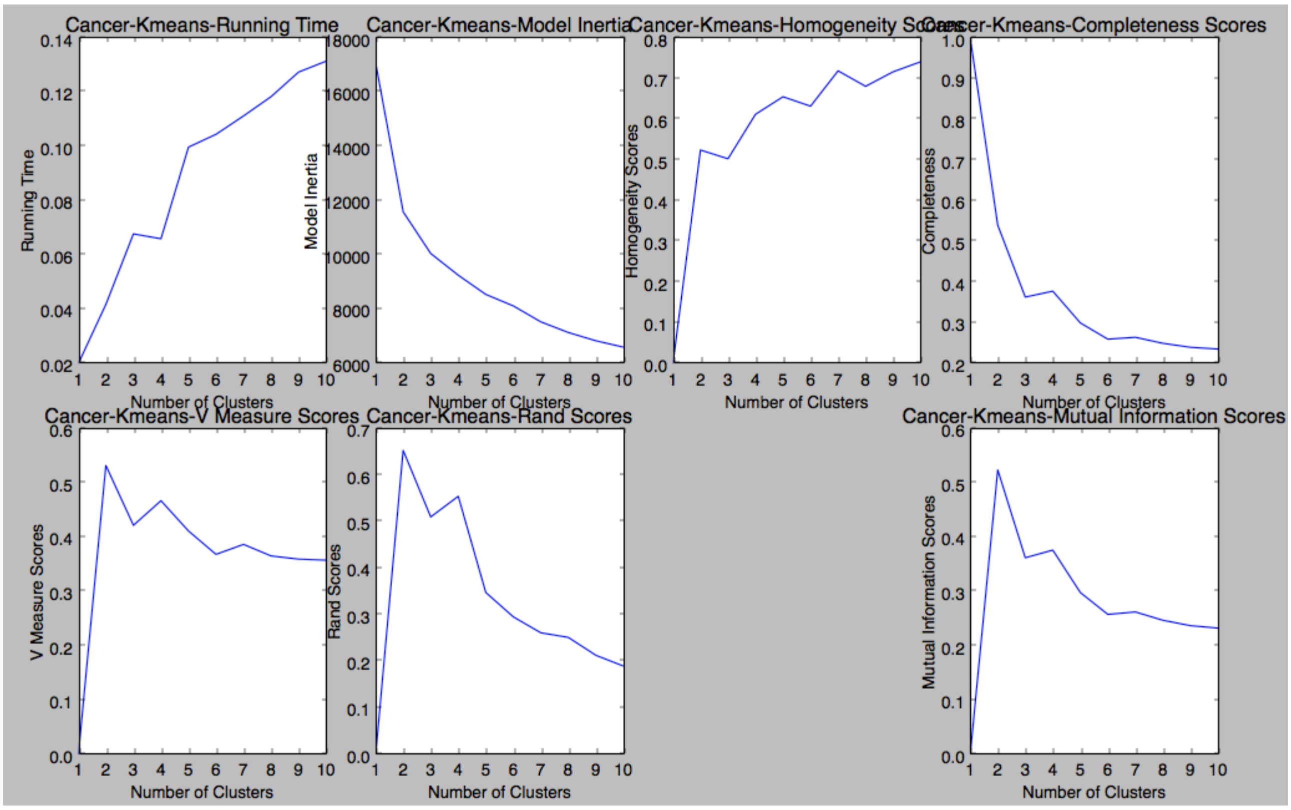
Particularly, for K-Means, Euclidean distance is used for determining similarities between instances. It is a native option but proves to be effective since other complex distances just may not converge on these two datasets.

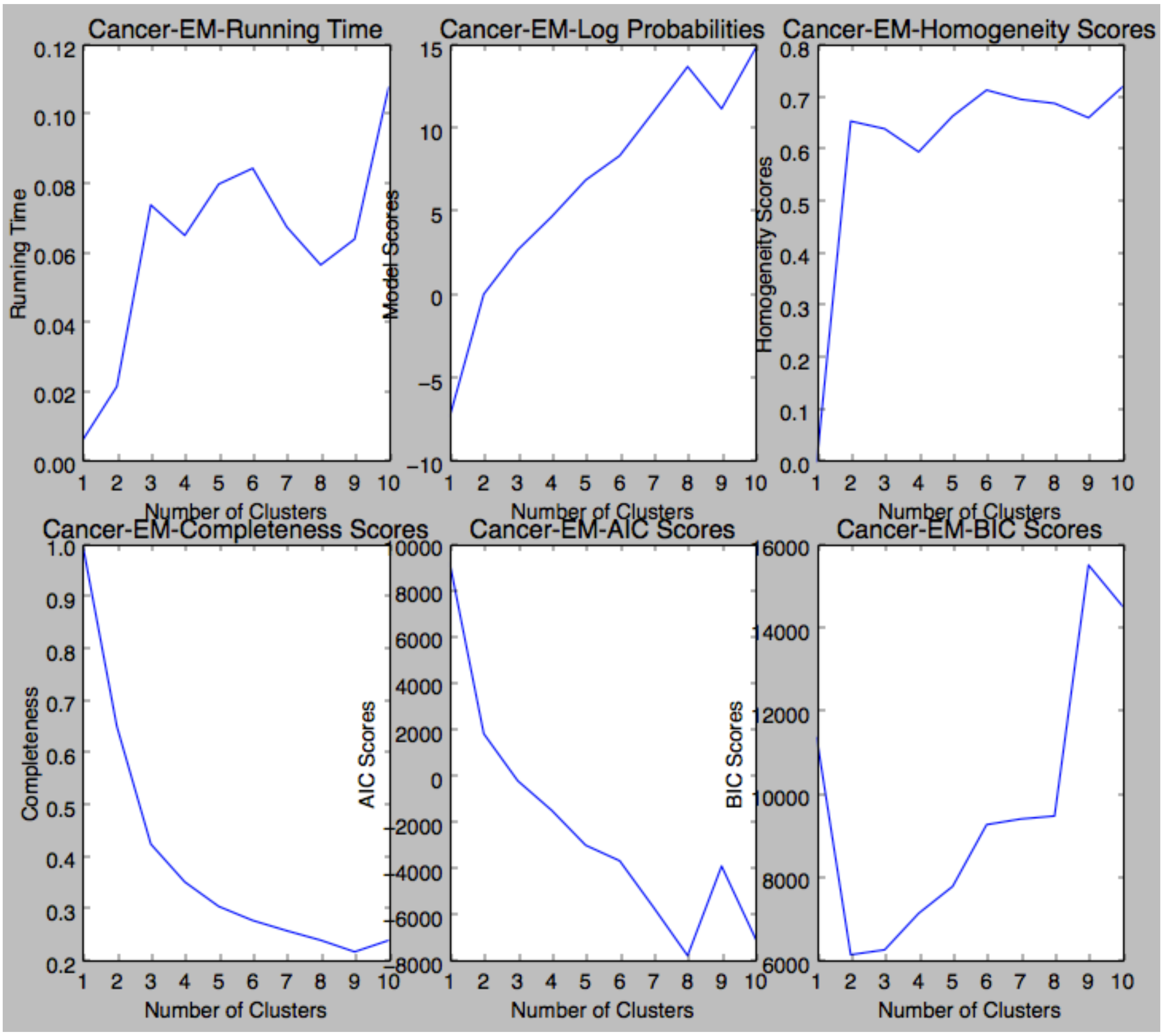
While for Expectation Maximization, probability distributions are considered in order to assign groups. Basically, a list of the possibilities that one particular instance belongs to each group is calculated based on Gaussian distributions. Then straight forwardly, the instance is included in the cluster with maximum probability.

The implementation of clustering algorithms are built on Scikit-learn. In order to evaluate the performance of K-Means, parameters of the final clustering are plotted including running time, model inertia, homogeneity, completeness, adjusted rand scores and adjusted mutual information scores. Also, for Expectation Maximization algorithm, running time, log probabilities, homogeneity, completeness, Akaike Information Criterion (AIC) and Bayesian Information Criterion (BIC) are recorded.

In this case, homogeneity describes how similar the instances in a cluster are and completeness measures the chance that all instances from same class indeed are assign to the same cluster in the end.

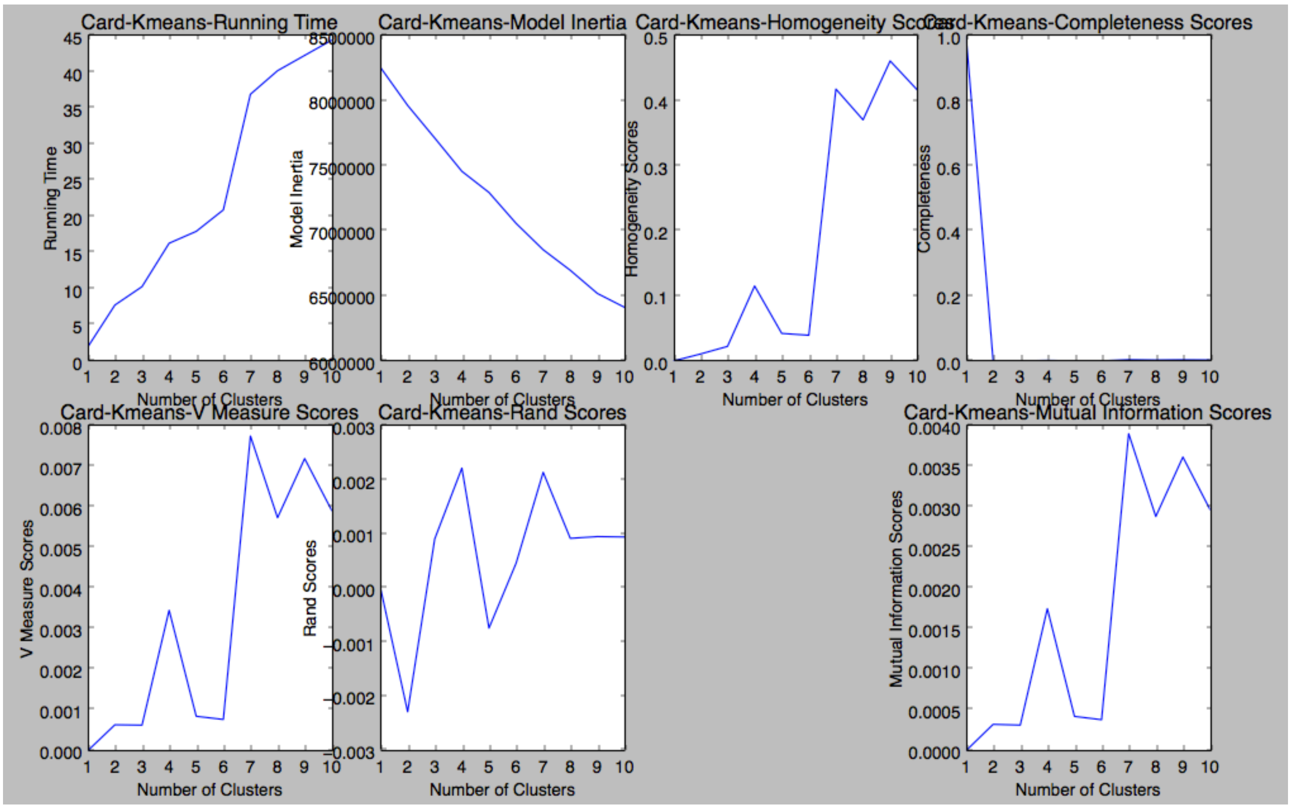
Breast Cancer

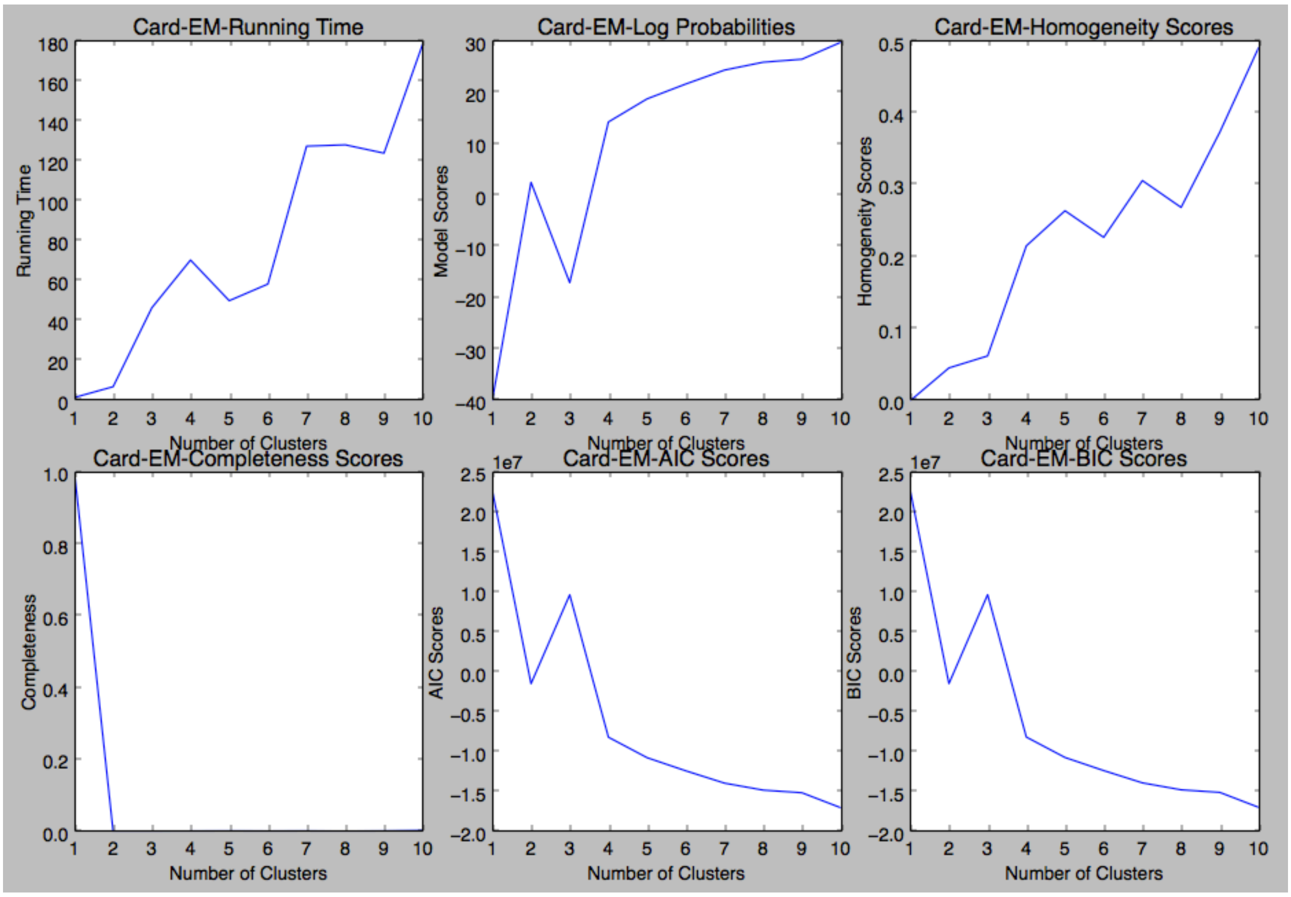




From the above plots, the elbow method indicates that cluster=2 seems to be the best choice. Because starting sharply from cluster number = 2, the curve for completeness tends to flatten. This indeed makes sense because there are only two classes in the breast cancer datasets, namely malignant and benign.

Credit Card Fraud Data





Applying elbow method to the completeness curve indicates that cluster=2 is the best number of clusters for this problem. And it does since there are only 2 classes of the data, either a credit card transaction is or not a fraud one.

Dimensionality Reduction and Clustering

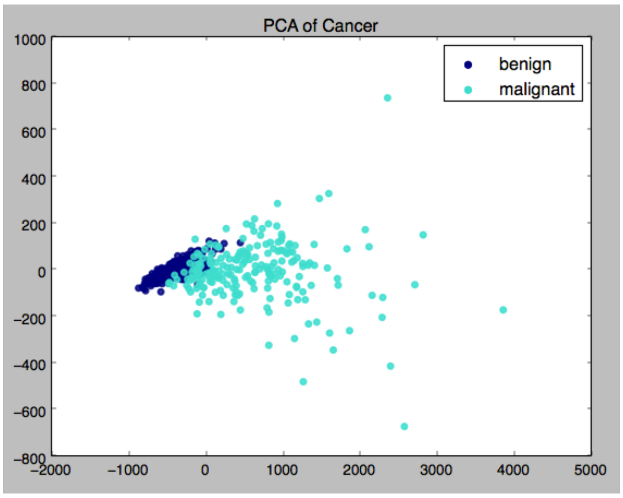
Dimensionality reduction is the operation performed on original data to get a simpler dataset that in various ways capturing original effective information. By simpler, it means less features in the new set of data particularly. When considering the "Curse of Dimensionality", this serves as a usually method to reduce the dimensionality of the data.

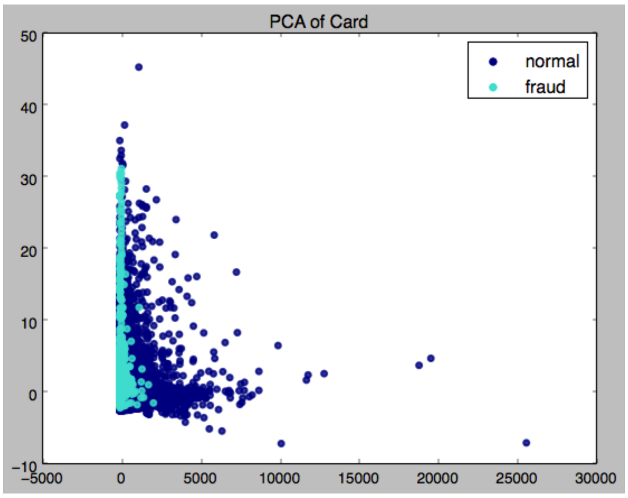
In this assignment, 4 dimensionality reduction algorithms are implemented, namely Principal Component Analysis (PCA), Independent Component Analysis (ICA), Random Projections (RP) and Linear Discriminant Analysis (LDA).

The following are the results of these dimensionality reduction algorithms performed on original dataset and further the clustering algorithms performed on the yielded dataset respectively.

Principal Component Analysis (PCA)

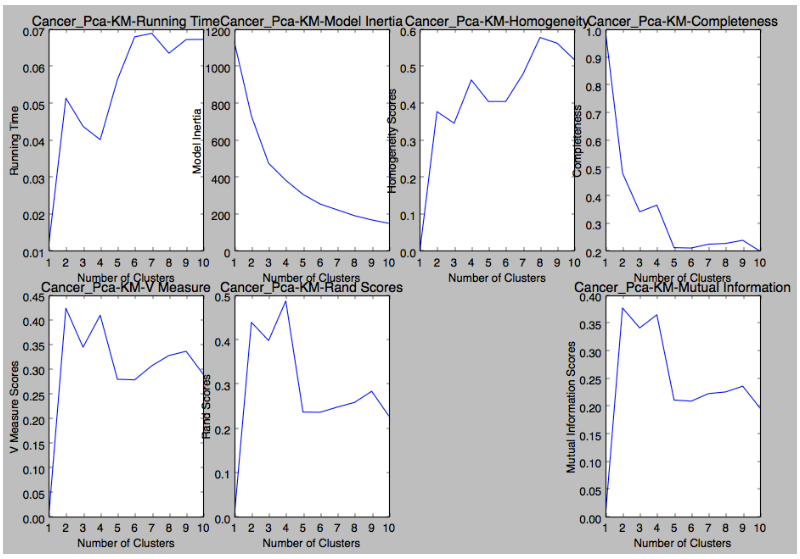
The following graph are plotted using the principal component and 2nd principal component computed from the original datasets.

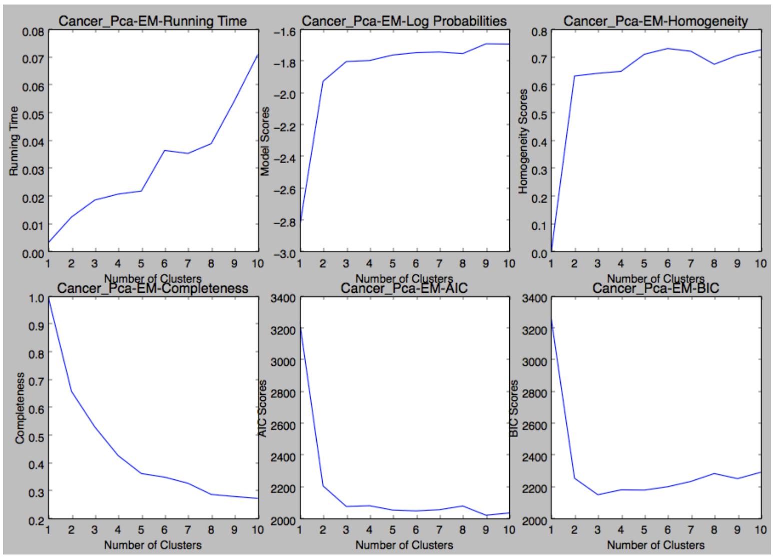




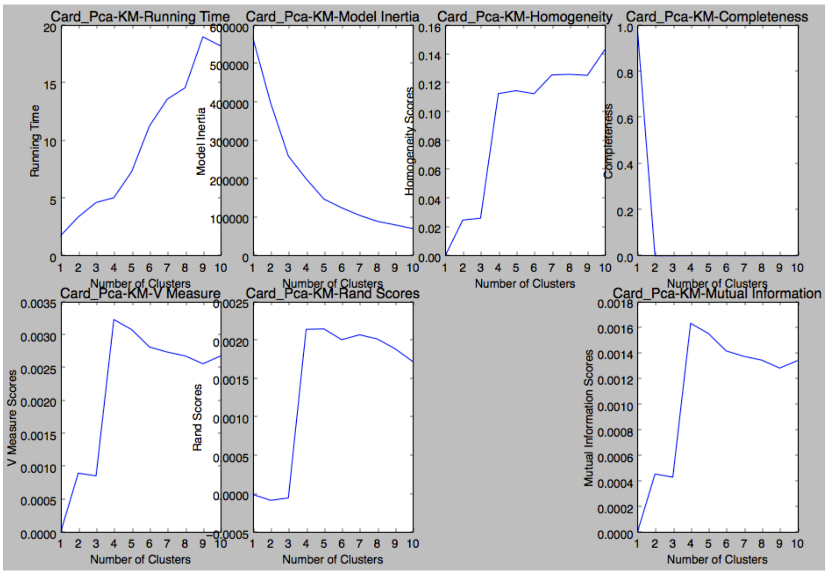
Explained Variance Ratio	1st	2nd
Cancer	0.98204467	0.01617649
Card	9.99538016E-01	5.98389763E-05

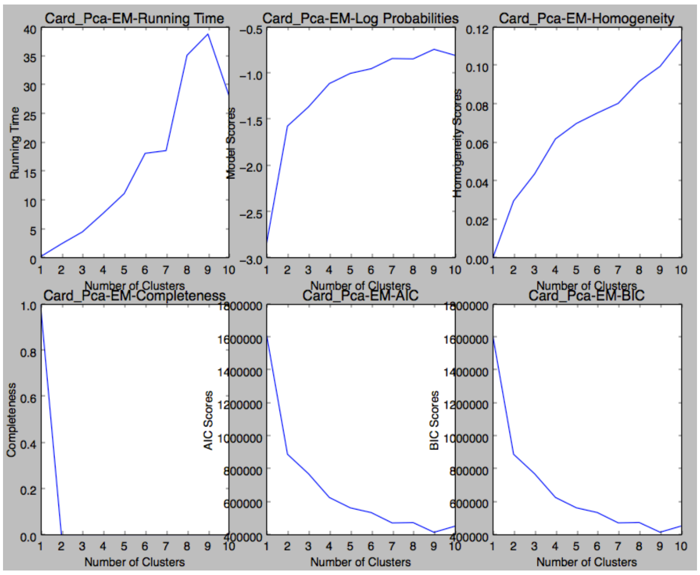
The explained variance ratio indicates the percentage of variance explained by each component. Basically, for the cancer dataset, the principal component accounts for 98% of its total variance, so this feature would be a very effective and informative one. Whereas for the card dataset, it gets only 9%.





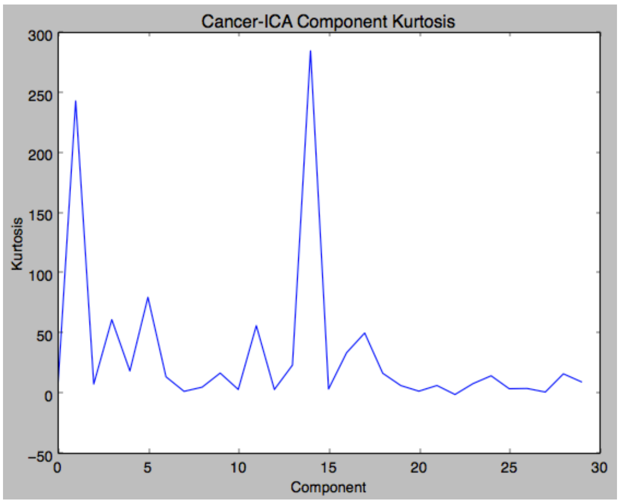
For cancer dataset, PCA transformed data gets shorter running time, since the number of features to be dealt with is less as expected. With much less features, both clustering algorithms achieve similar result as performed on original datasets. Actually, performing K Means on the transformed dataset yields clusters with less mutual information. And the EM gets better AIC and BIC scores.

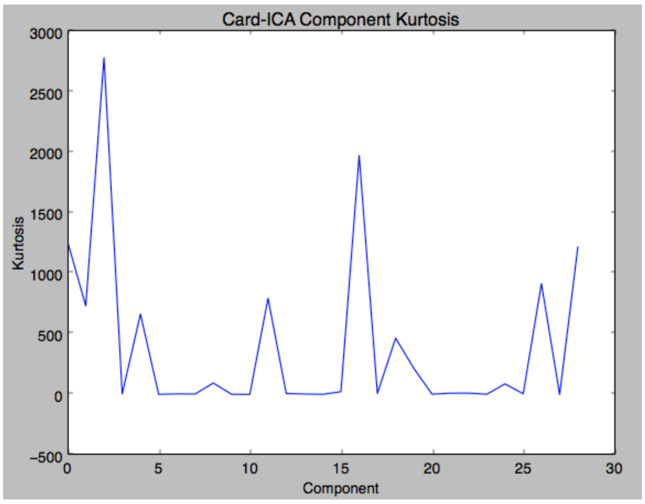




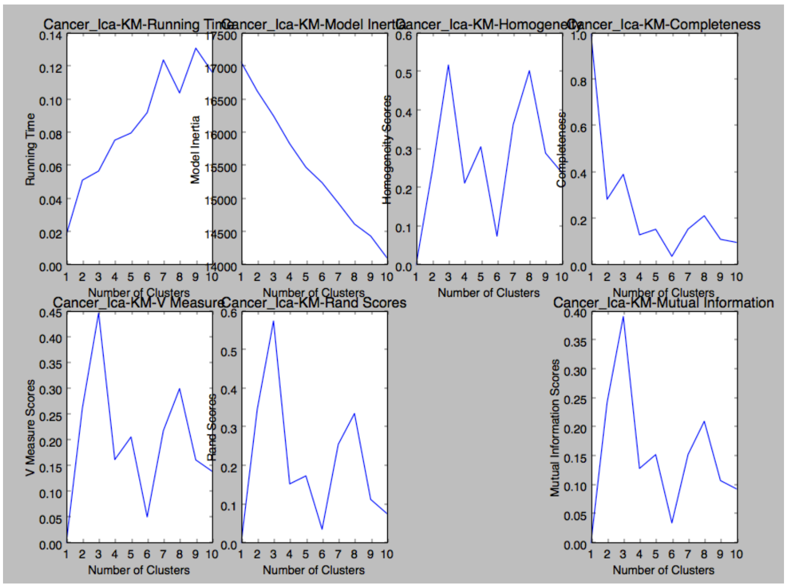
For card dataset, PCA transformed data does not preserve enough information to build clusters as good. This is actually foreseeable from the especially low explained variance ratio of its principal component. The reason may be that the differences of these two classes are very subtle and are not actually fitting in a Gaussian distribution.

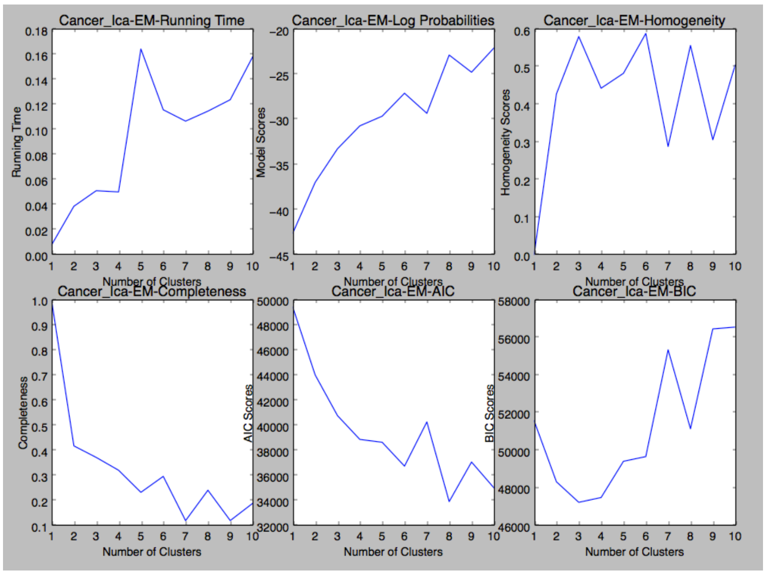
Independent Component Analysis (ICA)



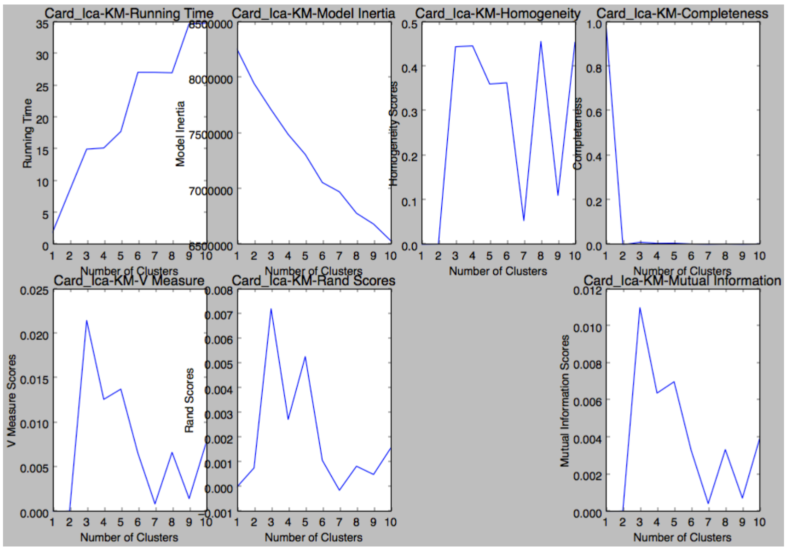


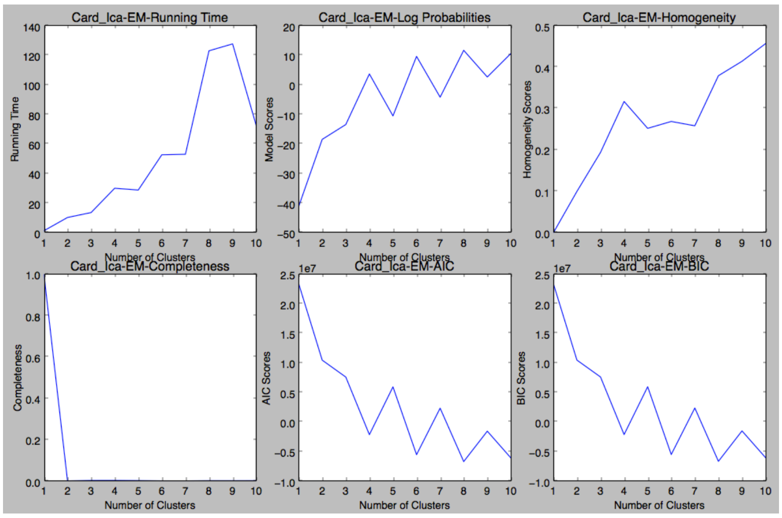
The tables above indicate the kurtosis of each component obtained from performing ICA. Since kurtosis measures the degree of non-Gaussianity, that is, the component has kurtosis farther to the value “3” is less similar to a Gaussian distribution. It actually supports the reason of bad performance of PCA considering the case of card data.





For cancer dataset, ICA helps reduce the mutual information of clusters and better log probabilities. Although other scores are not so well, the overall result is still acceptable. But since there is no particular way of filtering less informative components, ICA does not result in much less feature.

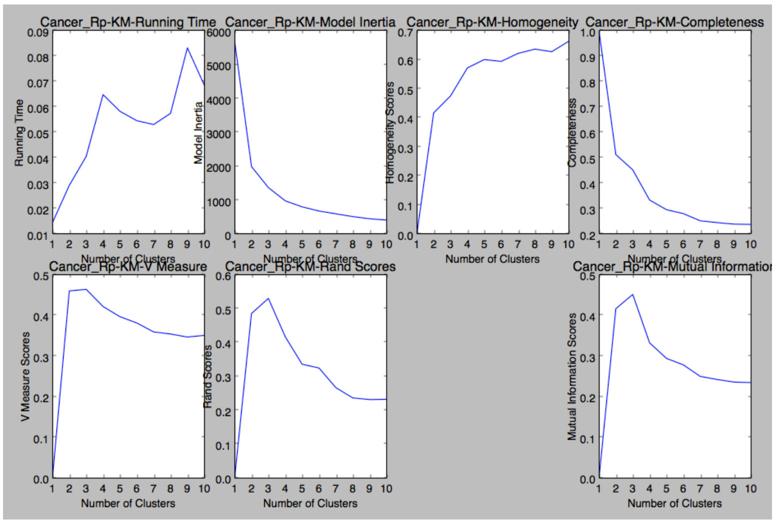


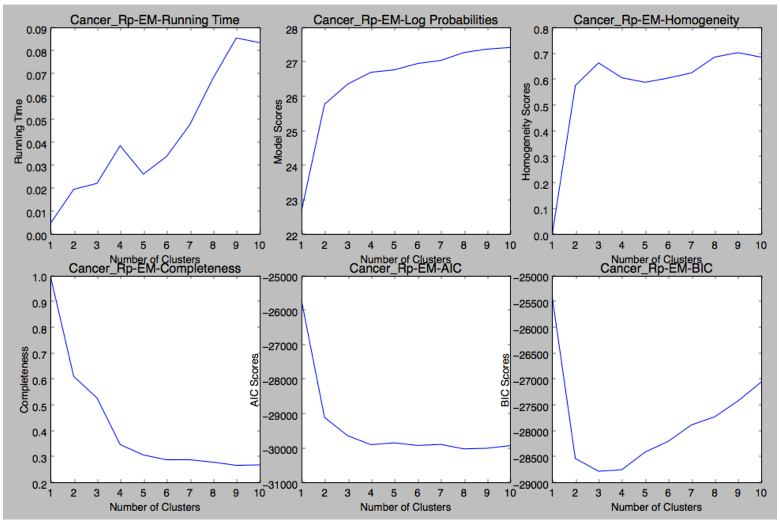


For card dataset, again, ICA does not provide less features. It achieves much better mutual information and rand scores. But the homogeneity and completeness scores are actually worse, there is no clear advantage over original dataset.

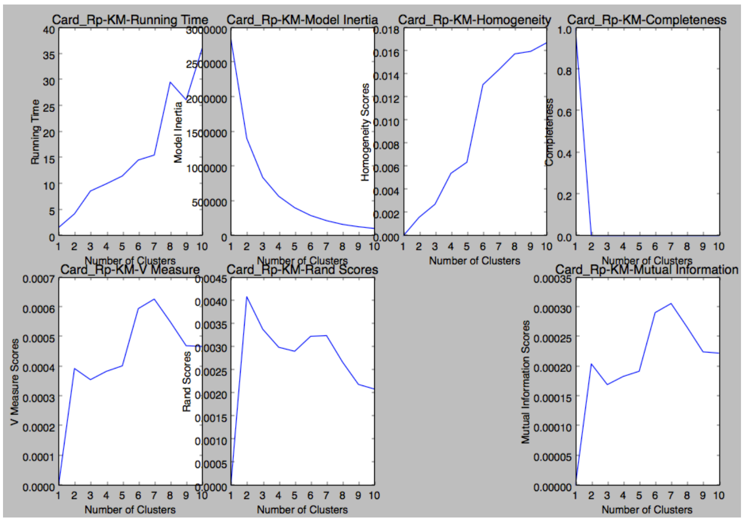
Random Projection (RP)

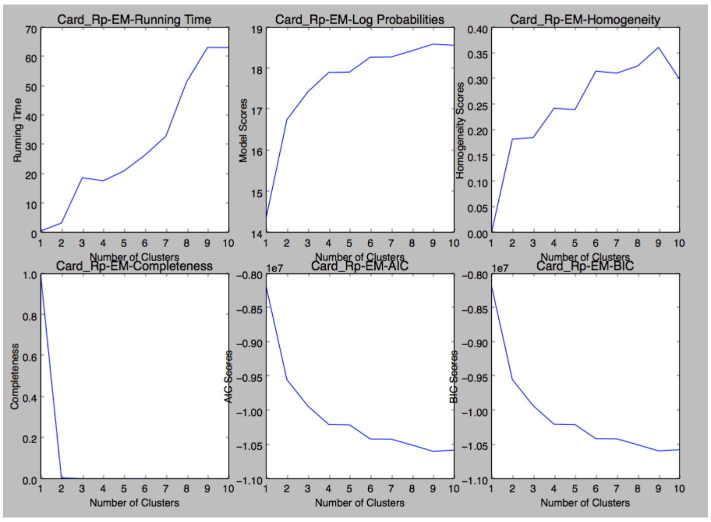
Random Projection is projecting all attributes to a lower dimensional space, by multiplying the original data with a randomly generated Gaussian matrix.





Although RP seems to be a pretty naive algorithm enough, from the graphs above, it works out impressively. The number of attributes is cut to 10 (originally 30 and 29), so the running time is reduced. While at the same time, all the measurement scores are perfectly reserved, meaning that the new dataset captured nearly all information for the clustering to work out for cancer data.



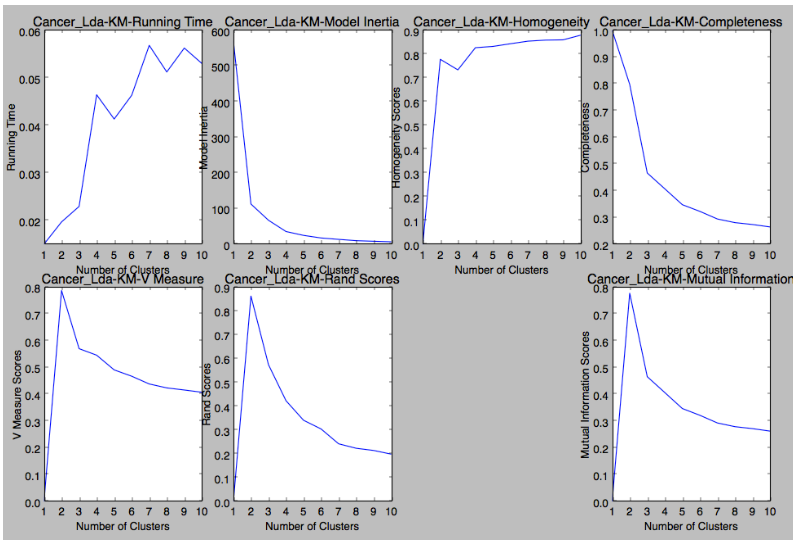


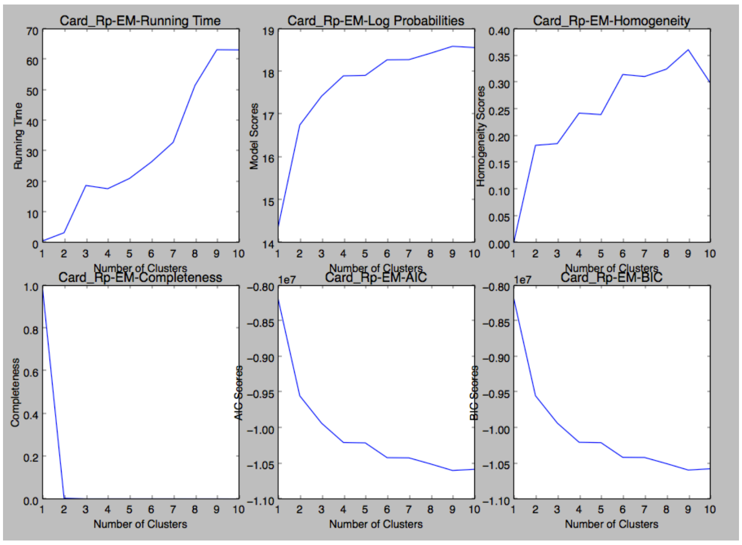
For card data, considering that 2 clusters are really needed in this case, the result of RP is great as well.

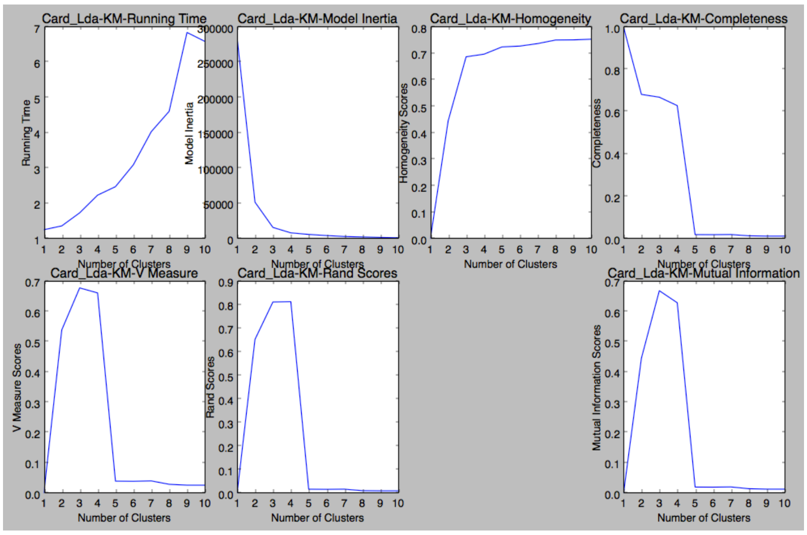
Linear Discriminant Analysis (LDA)

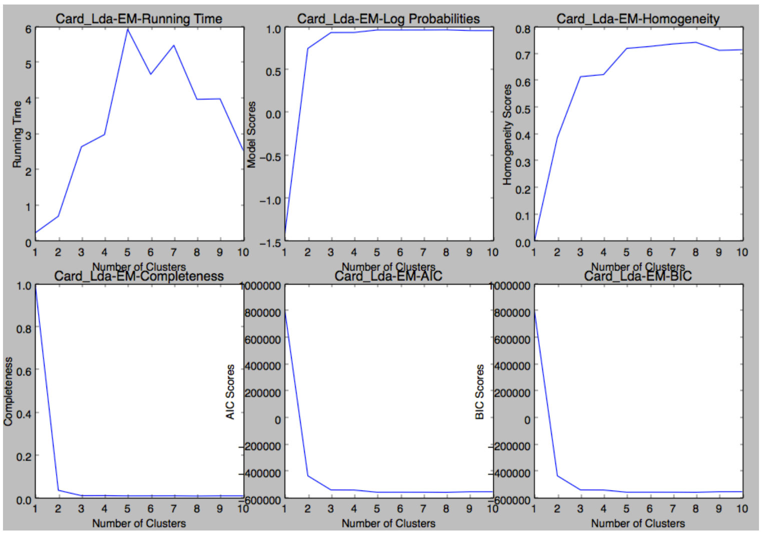
LDA tries to fit a Gaussian density to each class, assuming that all classes share the same covariance matrix, so that a linear decision boundary is generated for assigning clusters.

Since there are only two classes for both datasets, the resulting dataset contains only 1 parameter for clustering.



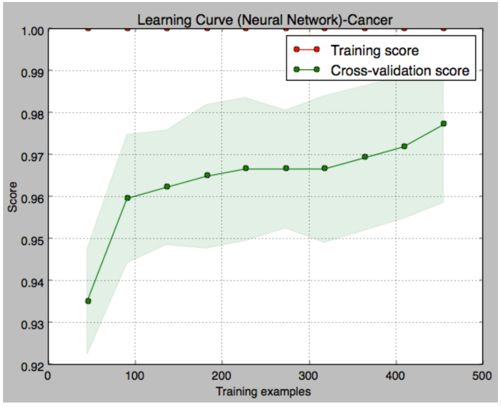


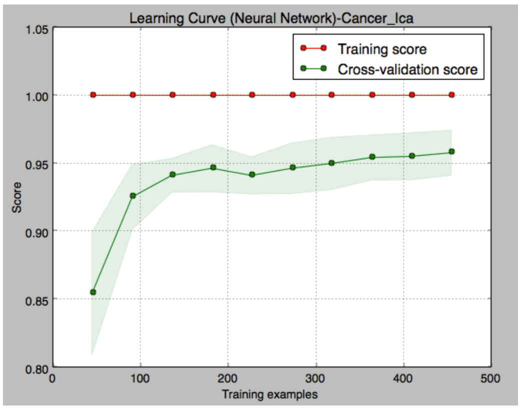


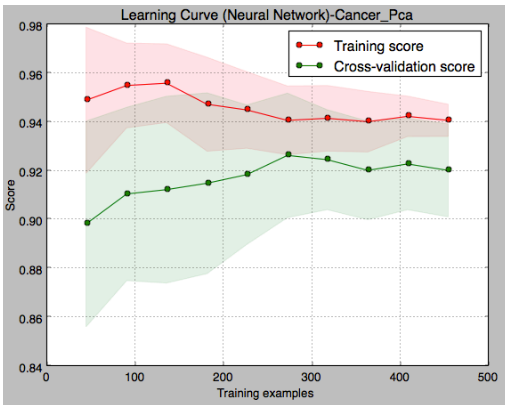


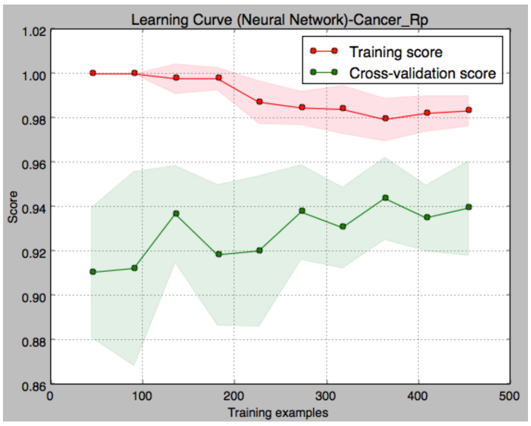
Comparing to other methods, the result from LDA not only contains all important information but actually helps the performance of clustering algorithms. With LDA, the clustering indeed achieves better score in almost every measurement. The advantages of LDA are huge against other ones.

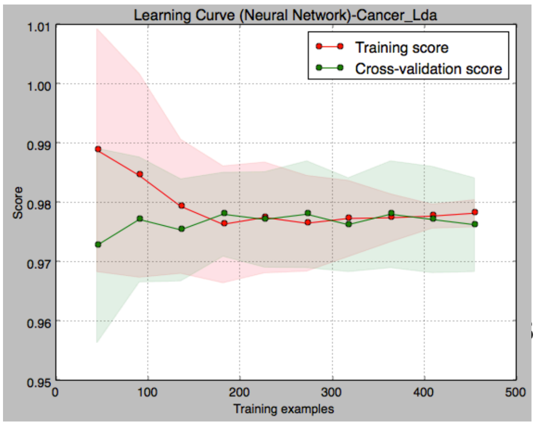
Dimensionality Reduction and Neural Network

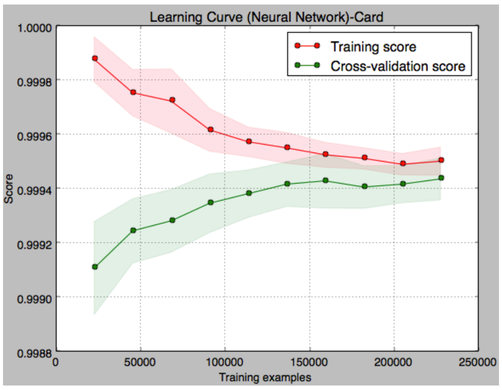


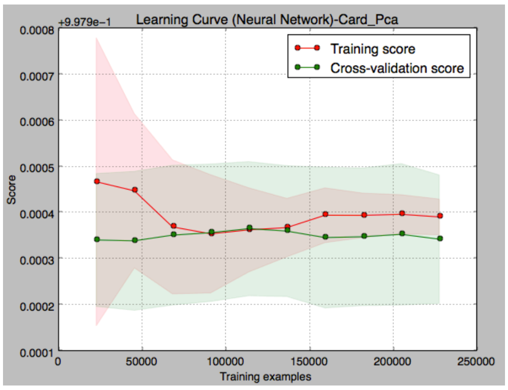


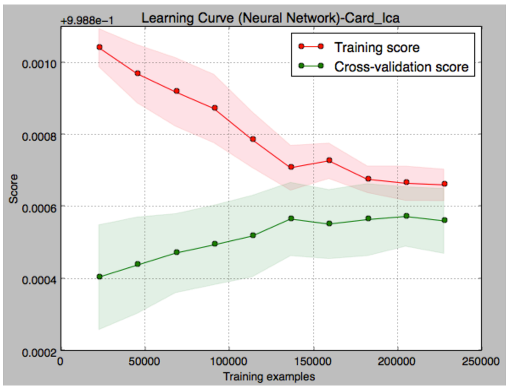


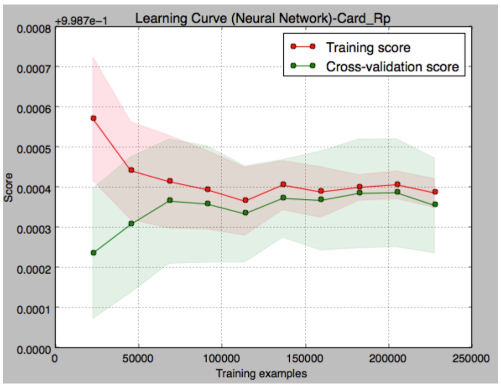


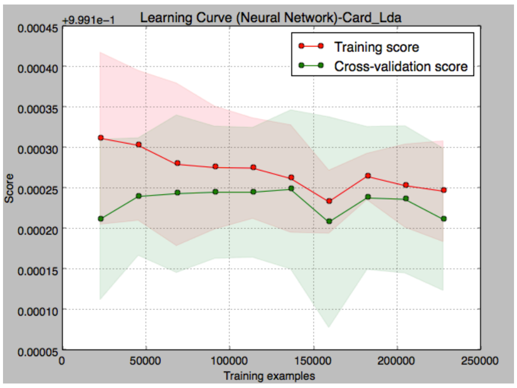






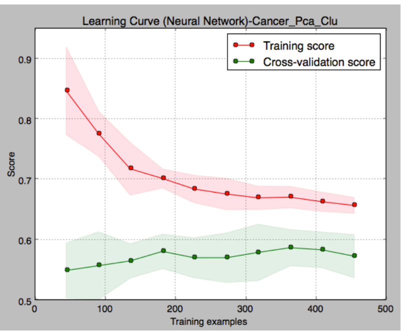


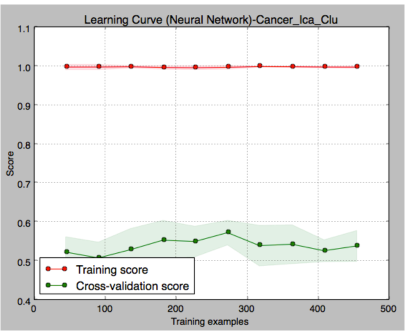


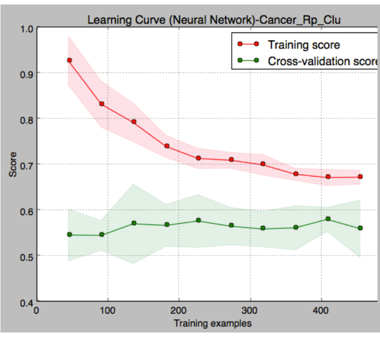


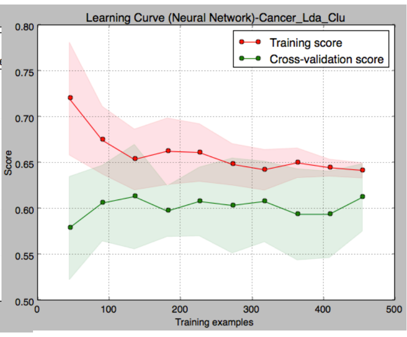
For cancer data, all 4 dimensionality reduction methodologies generally achieves the idea of representing the original dataset by a smaller dataset. Particularly, the LDA algorithm has a clear edge over other 3 methods in preserving most information by the least attributes kept, that is, only 1 attribute! As shown in the graphs above, the accuracy of Neural Nets built from all 4 new datasets are all quite close the the original result.

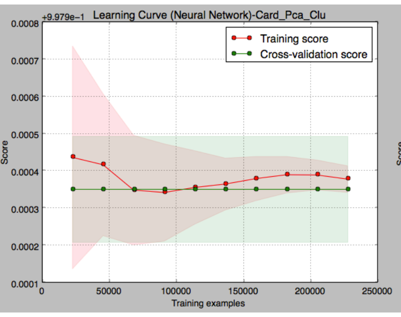
While for card data, the Neural Nets built from all these new datasets are much worse than the one built from original datasets. In fact, all 4 algorithms loses the important information in order to build an acceptable neural net. It may due to the relative rare occurrence of fraud examples, but there may be deeper reasons inside the data.

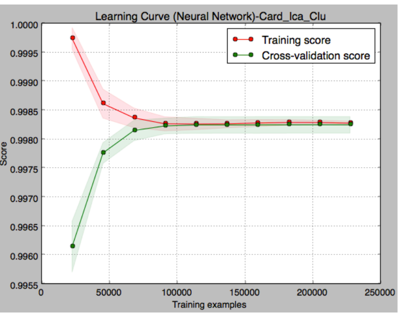


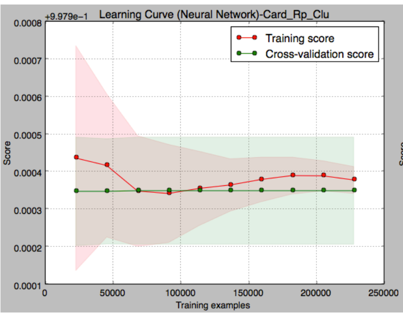


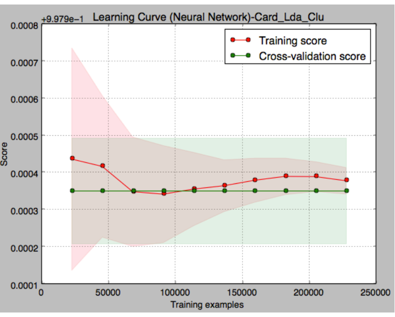












Adding the clustering results of every dimensionality reduction algorithm to the features of Neural Networks does not help improving the performance. The clustering method itself adds more error and interference against the neural network.

Conclusion

Linear Discriminant Analysis has been tested to achieve the best accuracy among all dimensionality reduction algorithms. Although it yields the least amount of feature for clustering and neural network, it does not require much more time to perform. Random Project, even it is a naive method, works out impressively good and even better than more complex models like PCA and ICA. Principal Component Analysis and Independent Component Analysis are not so well performed as expected.