CS7641 - Unsupervised Learning and Dimensionality Reduction

Shawn R. Mailo

November 5, 2018

Introduction

This Report explores clustering, feature transformation, and feature selection algorithms. More specifically, the two clustering algorithms used in this report are K-means and expectation maximization. The feature transformation/selection algorithms used to reduce dimensionality include PCA, RCA, ICA, and RandomForest. The three sections of this paper include: Part1 basic clustering of two data sets, Part2 applies dimension reduction to both data sets and running the clustering algorithms on the new data, Part3 applies dimension reduction and clustering methods to preprocess a data set before running it through a neural network.

Datasets

Wisconsin Breast Cancer

This data set is called the Wisconsin Breast Cancer data set and comes from SciKit Learn's Sample Data sets. This Data set is a classic binary classification data set in which the target attribute is a diagnosis if the patient has breast cancer or not. This data set contains 569 records of separate/individual patients. The dimension of this data set is 30 and all non-target attributes are real, positive, continuous numerical data types. The non-target attributes are computed from a digitized image of a fine needle aspirate (FNA) of a breast mass. This data set is slightly skewed toward malignant classifications at 63%. This data is interesting because all features are continuous numerical data types and contains a limited amount of training data both of which differentiate this data set from the other data set of this analysis.

Madelon

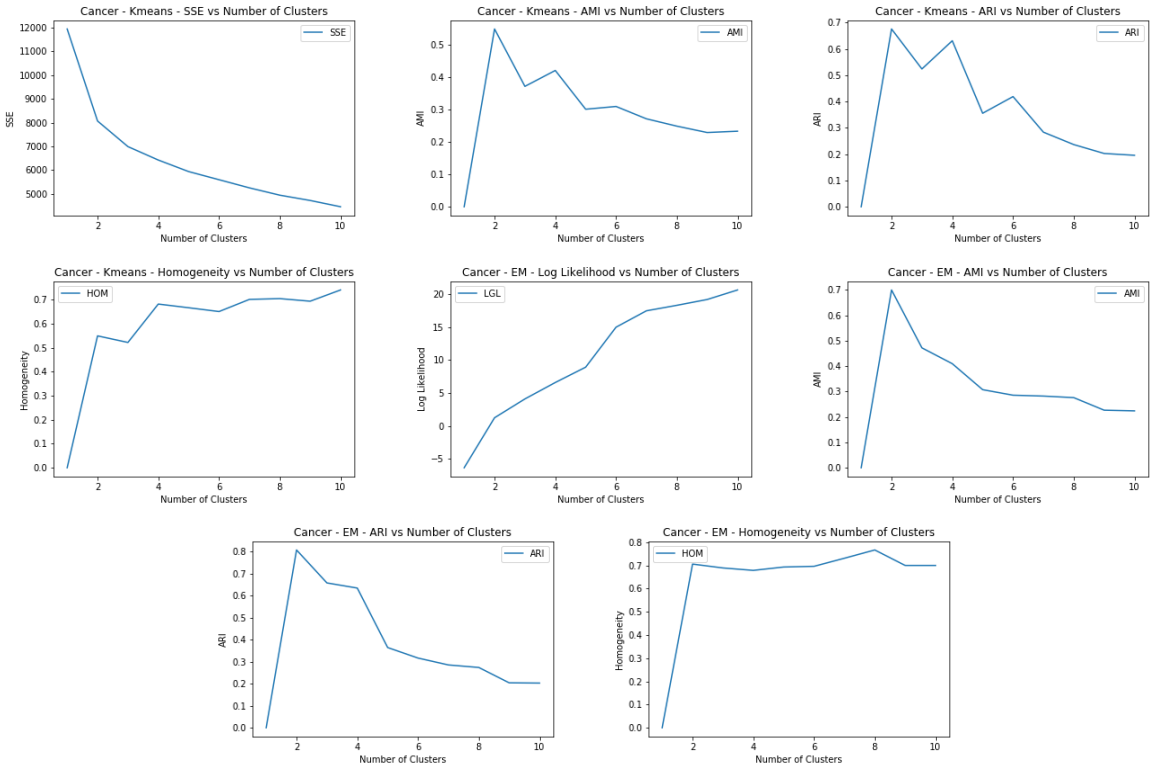
The Madelon dataset is an artificial dataset which contains 32 clusters of data points which are located on the vertices of a 5 dimensional hypercube. The dataset is a binary classification problem with labels -1,+1. There are 5 informative features based in the five dimensions and 15 linear combinations of those features which provide redundant information. In addition there are 480 "probe" features which contain no predictive power. There are 4400 samples in total, half of which are -1 and the other half +1. This dataset is very interesting because it is known to have informative data points, redundant data points, and distractor data points. It will allow for better understanding of our clustering and these feature transformation/reduction algorithms.

Part1 - Clustering

Clustering is an unsupervised machine learning group of algorithms that groups together data point instances into clusters based on similarity. The two clustering algorithms explored in this report are K-means and Expectation Maximization (EM). Although many distance functions can be used, the euclidean distance was used for the k-means algorithm. This allows us to use the commonly-used evaluation statistic "sum of squared errors" or SSE. Expectation Maximization is different from k-means in the fact that it is derived from probability distributions. The algorithm switches back and forth between estimating the log-likelihood of the current state (estimating step) and updating the state to maximize the likelihood (Maximizing step).

Due to being unsupervised, classification accuracy will not be used to evaluate the clusters. Instead other metrics will be used such as: intra-cluster sum of squared errors (k-means only), Log Likelihood (EM only), adjusted rand score, adjusted mutual information, homogeneity, and completeness.Adjusted rand score is a measure of similarity between clusters that is adjusted for chance.The Adjusted mutual information is a measure of the similarity between two labels of the same data adjusted for chance. Homogeneity is a score that looks for clusters that contain only data points which are members of a single class. These clustering methods and evaluation metrics were implemented with python's scikit-learn library.

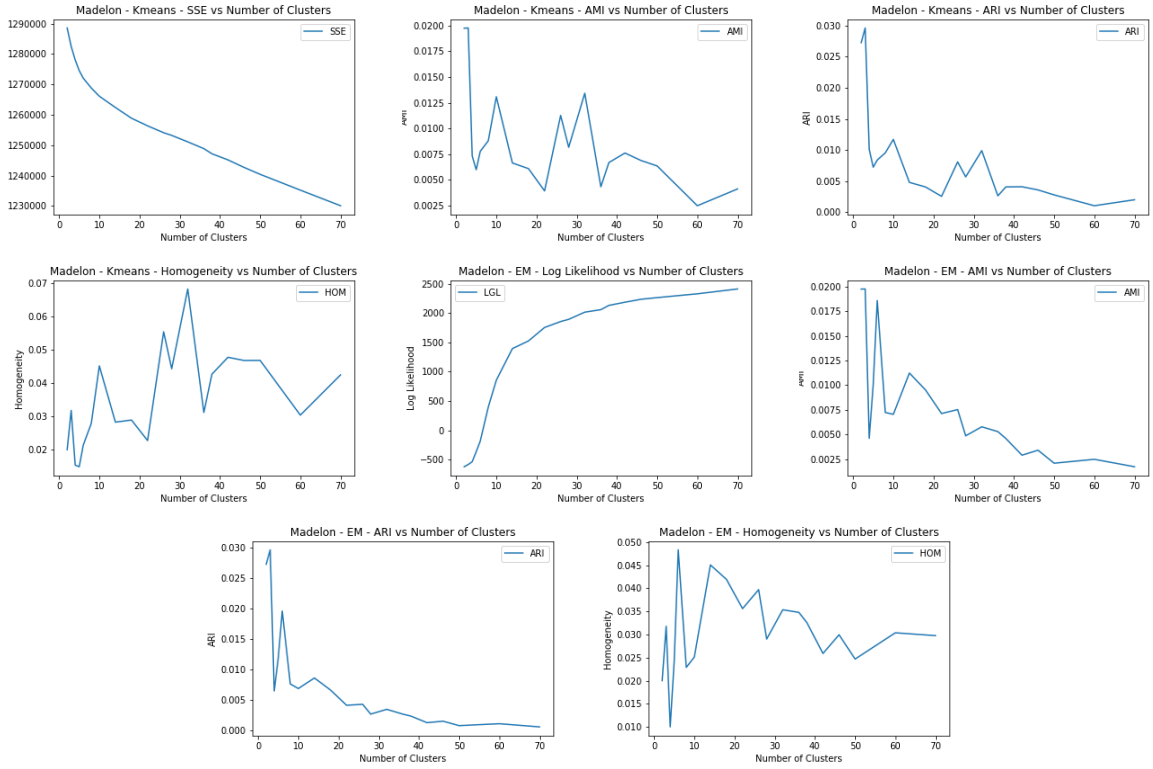
Wisconsin Breast Cancer



Usually to evaluate these graphs, one can use the elbow method to find the number of clusters. An example of using the elbow method can be seen on the K-means SSE vs Number of clusters graph. The elbow method would find that 2 is the optimal number of clusters. This is where the absolute value of the derivative of the curve is the greatest. For all graphs, except for the EM

Log-likelihood graph, the elbow method finds the optimal number of clusters to be 2. The number of clusters equals the number of classes. Each cluster most likely represents a different class.

Madelon



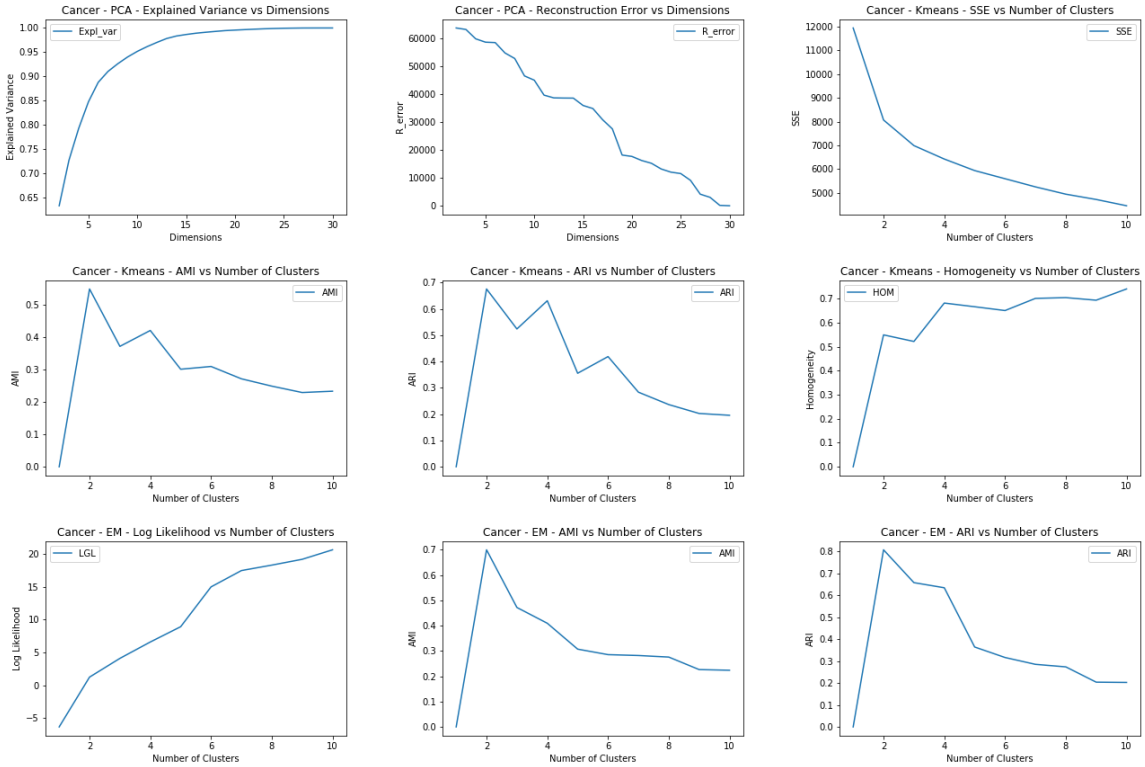
Many of the graphs here have an elbow around 8 clusters. This might be because each of the 8 clusters may have 4 clusters inside of it. 8 clusters of 4 clusters equals 32 clusters which is the actual number of clusters given by the data set. It can also be seen on kmeans-AMI, kmeans-ARI, and kmeans homogeneity that there are peaks at 32 clusters. This is most likely due to the fact that 32 is the real number of clusters. Overall EM performed more poorly than k-means. This is probably because there are clusters of subclusters. A cluster maybe real close to 3 clusters but these 4 clusters can be significantly farther away from other subcluster groups.EM tends to blur classification of what cluster a data point belongs too.

Part2 - Dimensionality Reduction and Clustering

Principal Component Analysis - PCA

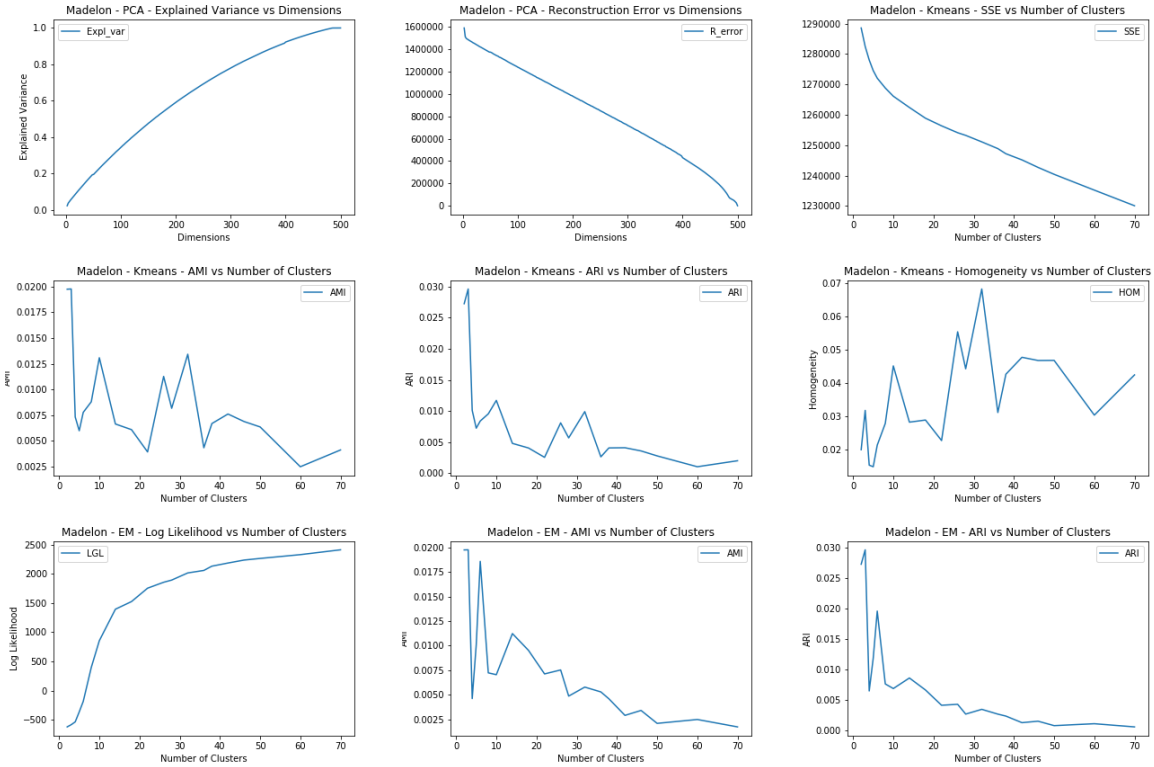
Principal Component Analysis is a feature transformation and dimensional reduction algorithm. This algorithm finds orthogonal eigenvectors and its eigenvalues that best explains the variance among the data points. These eigenvectors are then used to transform the features into smaller dimensions.

PCA - Wisconsin Breast Cancer



To evaluate PCA, one can examine an explained variance graph and/or a reconstruction error graph. It is obvious that the explained variance plateaus after about 15 dimensions of PCA. 15 dimensions cuts down the dimensionality by more than half! With the transformed 15 dimension data, the clustering experiment was ran again. When compared to the graphs of the original data it is obvious that the new data still keeps most of the information from the old data (a goal of PCA). Looking at SSE, AMI, Homogeneity and using the elbow method, these graphs suggest the best number of clusters is 2, which agrees with the true data.

PCA - Madelon

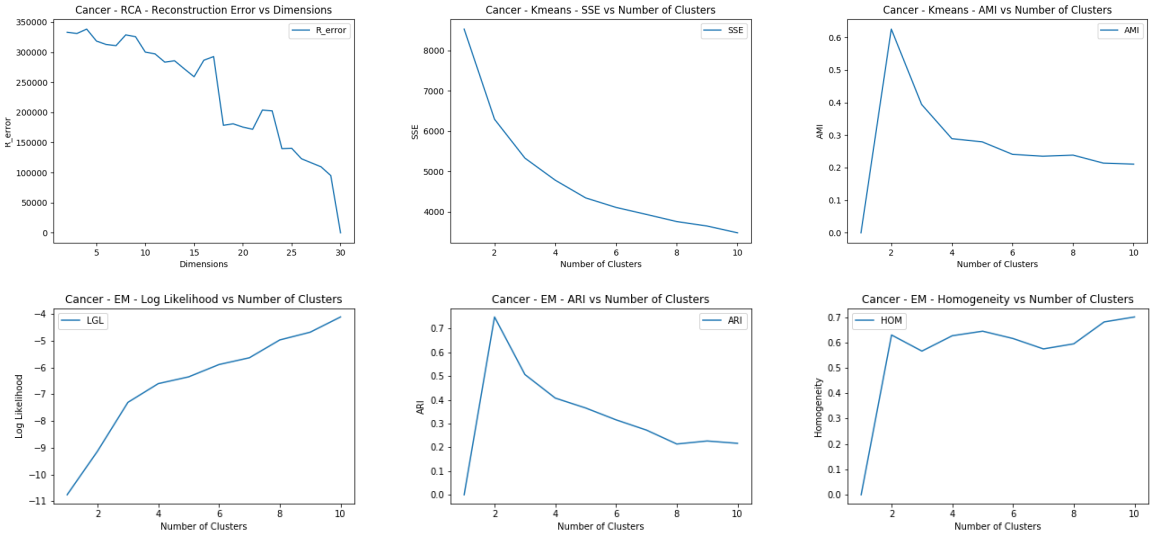


PCA, when applied to the Madelon dataset did help much. Both the explained variance and reconstruction error graphs are smooth curves that continue to increase(exp.var)/decrease(rec error) as dimensionality increase. This probably signifies that there is a lot of noise in the data. Although not being significant, a PCA with 300 dimensions was chosen to apply to the data and to be used in the clustering experiment. Again the PCA transformation seemed to keep most of the information of the data as the graphs look very similar to the original clustering experiment. We still see the same peaks around 32 clusters for the kmeans graphs.

Random Component Analysis - RCA

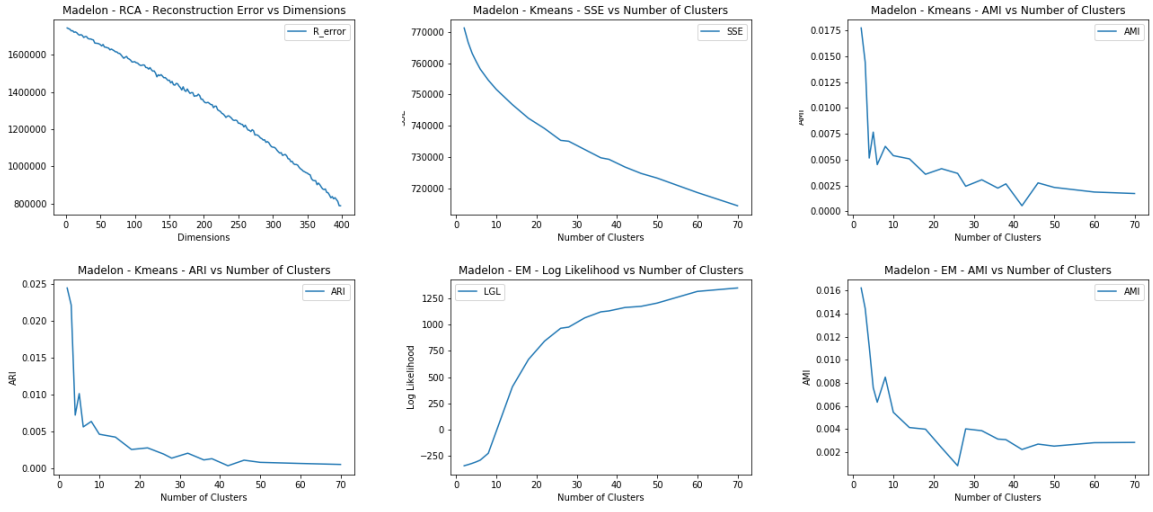
Random Component Analysis also known as Random Projection performs a similar transformation as PCA except it chooses eigenvectors at random to project the original data on a randomly generated Gaussian matrix. Although counter intuitive, RCA actually performs decently well and is significantly faster at computations than PCA.

RCA - Wisconsin Breast Cancer



Reconstruction error was used to evaluate the RCA output across different dimensions. As seen by the graph, there wasn't a very clear elbow in the curve. 15 dimensions was chosen since it had a sharp downward turn. Taking the 15 dimension transformed data, the clustering experiment was ran again. The graphs relatively look the same which means that RCA was able to keep most of the information of the data. One thing to notice is that the SSE curve of the kmeans algorithm has become smoother while the Log likelihood has a more visible elbow. AMI, ARI, and Homogeneity graphs all point to 2 clusters which represents the true data.

RCA - Madelon



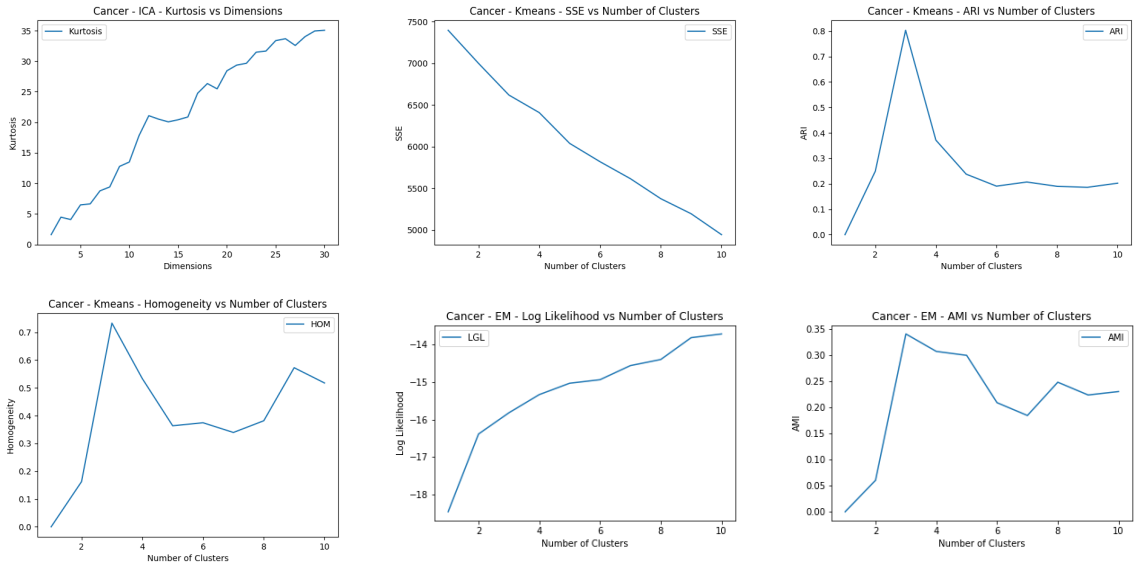
The reconstruction error is similar to what was seen in the PCA data. A continue decrease in error as the number of dimensions increase. Again there is no clear dimension to pick. 300 was chosen to decrease dimensionality while not allowing too much reconstruction error. The transformed data was then used in the clustering experiment. Spikes at 8 and 28 clusters can be

seen in many of the graphs. Although not the correct number of clusters, they are a multiple of 4 which maybe due to the cube orientation of the clusters.

Independent Component Analysis - ICA

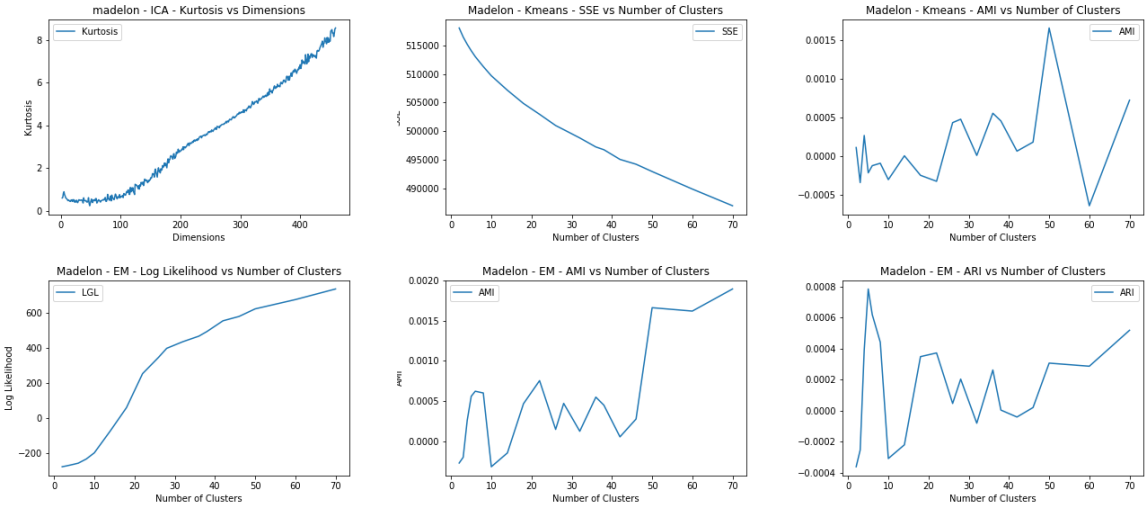
Independent Component Analysis or ICA is another feature transformation and dimensionality reduction algorithm. The goal of ICA is to explain non-Gaussian data by finding linear combinations of features that prove to be statistically independent of each other.

ICA - Wisconsin Breast Cancer



To find a good number of components to use for ICA, one can look at the kurtosis of each dimension. The goal of ICA is to create independent components which implies that a high kurtosis is desired. According to the the graph, the kurtosis continues to increase as dimensionality increases with a little peak around 13 dimensions. Using the 13 dimension ICA-transformed data we implemented the clustering experiment again. It was found that the ICA data performed poorly compared to the original dataset. AMI, ARI, and Homogeneity all suggest 3 clusters. Furthermore, the SSE graph for kmeans is almost a straight line. The slight elbow on the EM log likelihood graph suggests 2 clusters, however its not very clear.

ICA - Madelon

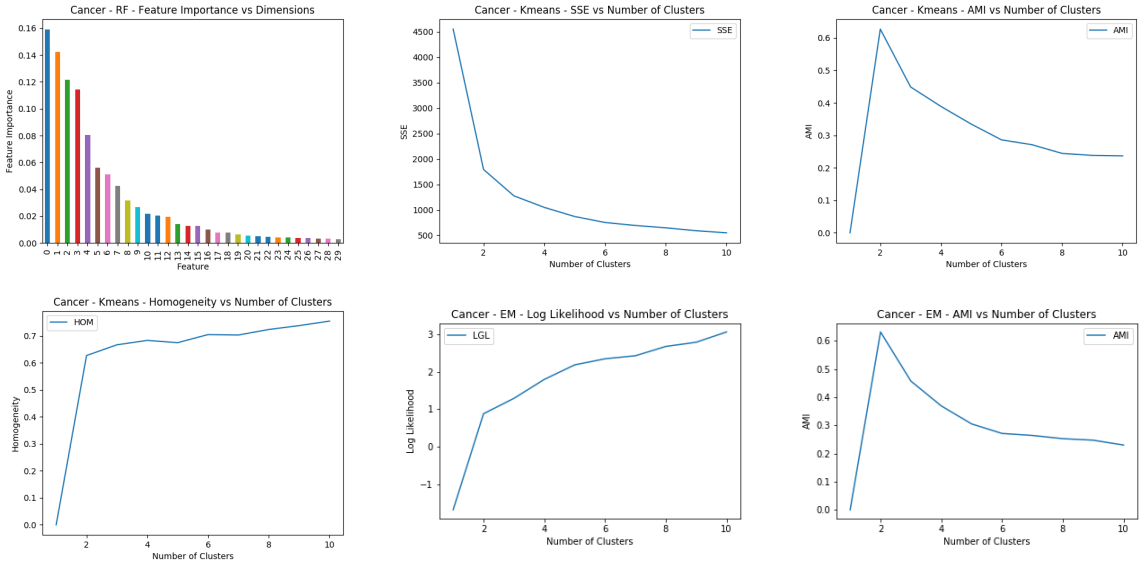


ICA performed even worse for the Madelon data set. The kurtosis grew as the dimensions increased with a change in slope around a dimension of 200. Although not a very high value for kurtosis, the change of slope could have some significance. The data was transformed with ICA to a 200 dimension data set. When subjected to the clustering experiment, results were found to be very unclear. AMI, ARI, and Homogeniety graphs all had very noisy data. The kmeans SSE graph was very smooth which doesn't suggest a optimal number of clusters. The best graph was the EM log likelihood which started to plateau around 28-32 clusters.This suggestion would agree with the real data set.

Random Forest - RF

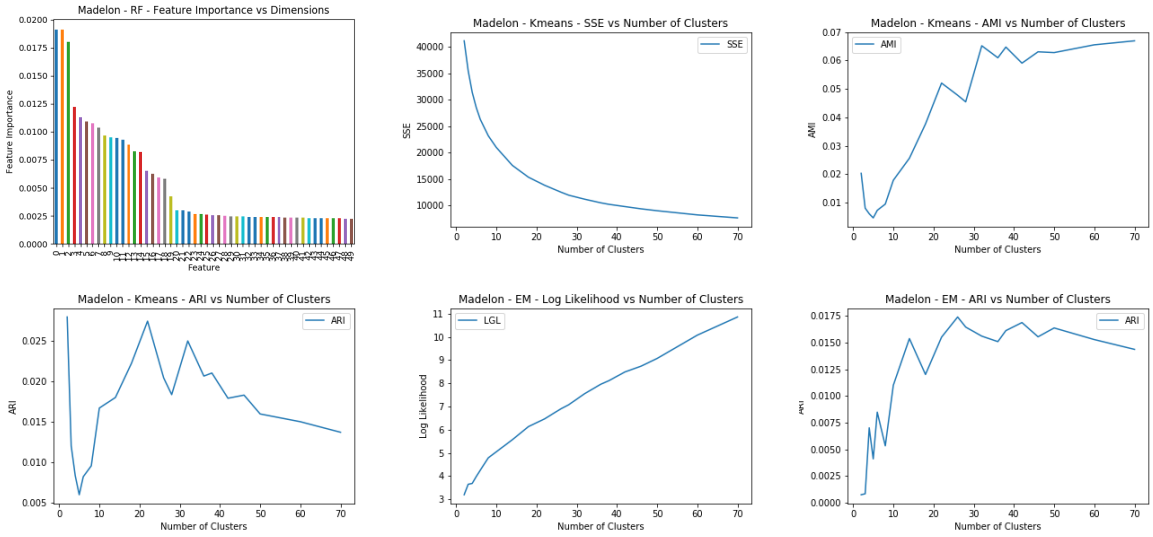
Random forests can be used as a supervised method to dimensionaly reduce the feature set. Unlike the other methods, random forest dimensionality reduction does not transform the data. Here a random forest of 100 trees was used to rank the features in order of importance. Based on the importance magnitude of the feature, selected features were chosen to be kept or dropped.

RF - Wisconsin Breast Cancer



Looking at the Feature Importance graph, the elbow method can be used here as well. It was decided that after 10 features the importance of each feature was minimal. Putting these features through the clustering experiment gave good results. None of the graphs were noisy. All graphs (SSE, log likelihood, AMI, ARI, Homogeneity) had elbows/spikes at a cluster number of 2. Overall this matches the real data set and gave the best results.

RF - Madelon



The Random forest showed a sharp drop off in feature importance after dimension 20. All the dimensions above 20 were dropped from the original data set. After running the new data set through the clustering algorithms, it was clear that significant information was lost from the data. Both the SSE and log likelihood graphs were very smooth, showing no suggestions for optimal cluster size. Although ARI, AMI, and Homogeneity showed peaks around 32, they also had significant peaks elsewhere. Overall Random forest feature selection did not work well with the Madelon data set.

Part3 - Dimensionality Reduction, Clustering, and Neural Networks

To be continued .....