Keith Mertan
CS 7641 - Machine Learning
Assignment 3

Unsupervised Learning and Dimensionality Reduction

Introduction

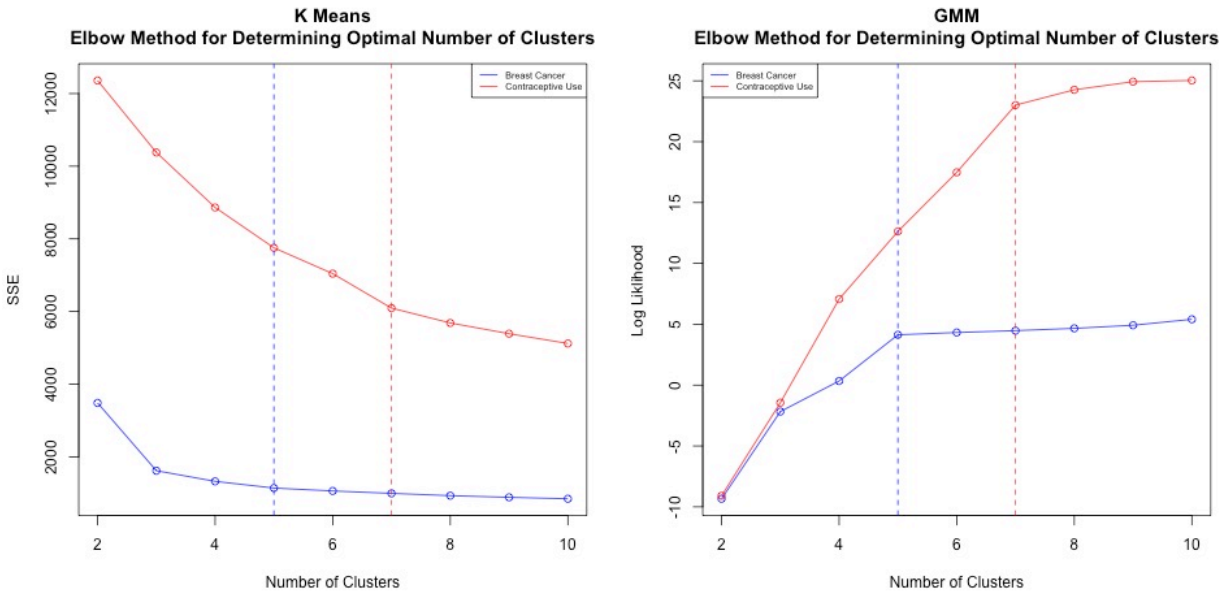
This analysis explores the effectiveness of two unsupervised learning algorithms (K-means clustering and Expectation-maximization clustering) in regard to creating meaningful clusters given two datasets. In addition, four dimensionality reduction techniques (Principal Components Analysis, Independent Components Analysis, Random Projections and selection using feature importance given a trained Random Forest model) are employed on the same two datasets and the results are compared and contrasted in an effort to understand which methods are appropriate for each dataset based on the relationships among variables in the data.

These unsupervised learning and dimensionality reduction algorithms are also tested in tandem to understand how clusters change when using data abstracted into a lower dimensional space, hopefully with redundant or irrelevant information removed in the transformed dataset. Finally these transformed datasets are used as inputs to a neural network and compared to one another along with the results found on the original data from assignment one.

The two datasets used are from assignment one. The breast cancer dataset contains 569 rows which each represent a patient, and 10 independent variables. All 10 variables are integers which measure attributes of cell nuclei, such as shape and size. The independent variable is the diagnosis: either “benign” or “malignant”. The Indonesian contraception choices data contains 1472 rows, each corresponding to attributes of an Indonesian woman’s life and her contraceptive choices. There are nine independent variables, one of which is one-hot encoded. This creates a total of 12 dependent variables. The independent variable, contraceptive choice, takes three values: “no-use”, “short-term” and “long-term”. The independent variables are ignored when applying the unsupervised learning and dimensionality reduction techniques. Any involvement of the independent variable for validation of algorithmic outputs is explicitly mentioned.

Code is adapted from Jonathan Tay’s assignment 3 Github repository which utilizes Python’s scikit-learn library.

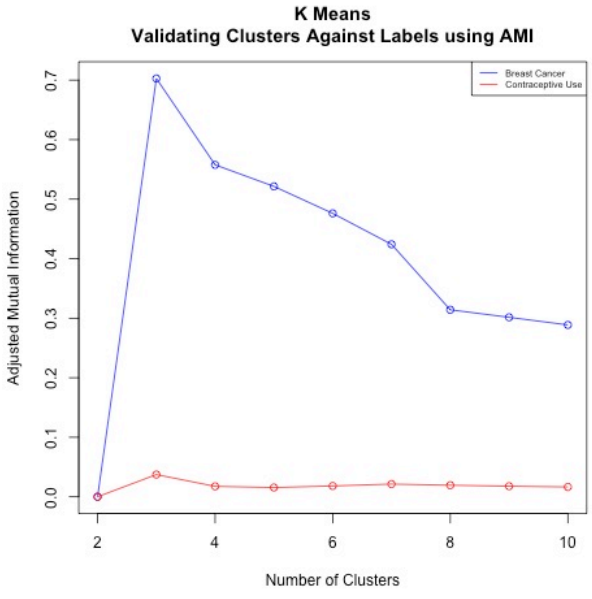
Clustering on Original Data

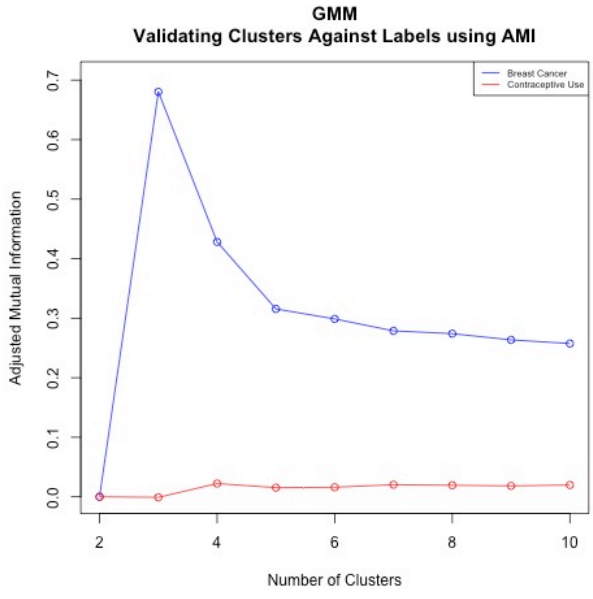


First clustering is performed on the original datasets. K-Means, which aims to minimize the distances of each point to the cluster center, is a hard assignment clustering algorithm. At the end of its iterations, each point in the data is assigned a class. GMM, which is the implementation of the EM algorithm used in this analysis, assigns probabilities of each data point belonging to each cluster by creating Gaussian distributions around cluster centers. This allows more flexibility in terms of measuring error since cluster confidence is taken into account.

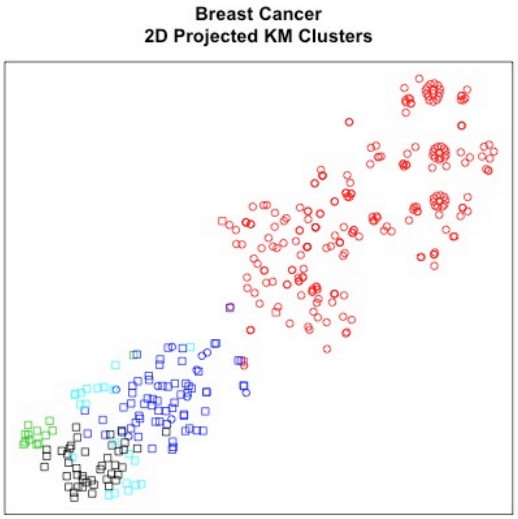
The elbow method is used to select the optimal number of clusters for each dataset. The silhouette method was also considered but ultimately abandoned because it does not perform well with binary inputs. Given that the datasets truly have two and three classifications, respectively, the values two through ten were chosen as cluster sizes to search over. This allowed the inclusion of each of the pre-defined cluster sizes along with larger cluster sizes to potentially capture subcategories within the pre-defined groups. The elbow method approximates the best cluster size by finding a point where the metric to optimize (SSE for K-Means and Log-Likelihood for GMM) sees a “significant” decline in the rate of change. This is a subjective measurement but is clear in certain cases.

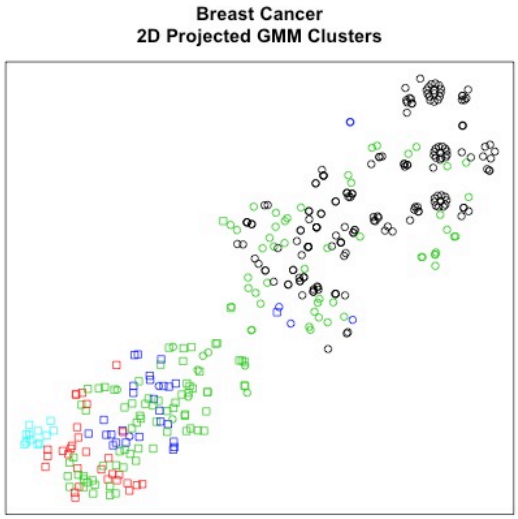
The charts above show that both algorithms agree on the ideal number of clusters according to the elbow method. The breast cancer data saw its elbow at five clusters for both algorithms (although an argument can be made for three clusters in the case of K-Means) and the contraceptive use data agree that seven clusters are ideal. This leads to questions about the structure of the data and the level of granularity the ground truth labels capture. Perhaps, in the case of the breast cancer data, there are clusters within the ‘malignant’ group that could indicate the level of severity of the cancer. For the contraceptive choice data, there may be subgroups within the categories ‘short-term use’ and ‘long-term use’ that could represent a data point that is absent from the dataset, such as frequency of use during those periods. The process of clustering could shed light on these additional levels of granularity and potentially influence the data collection process in the future.

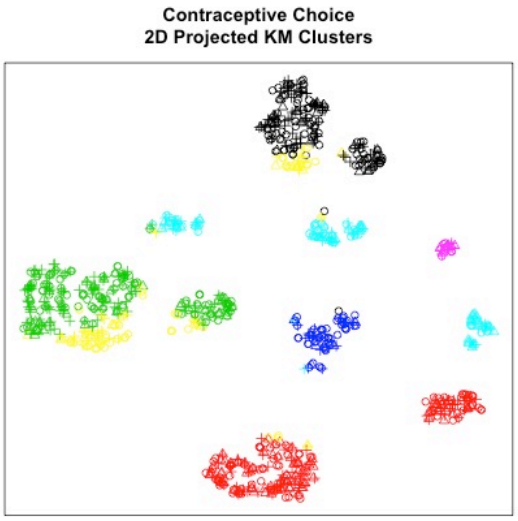


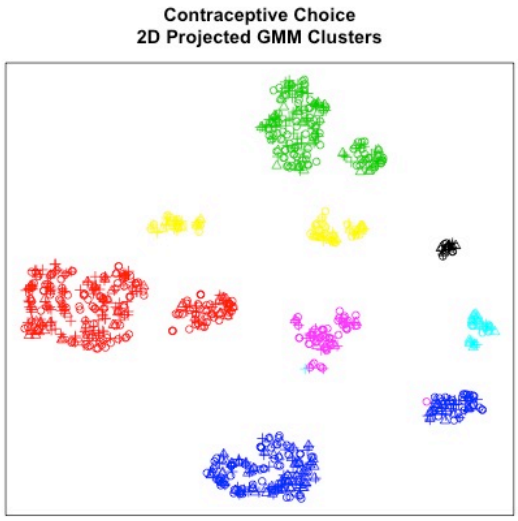


Given that this is labeled data, this analysis enjoys the luxury of being able to compare cluster assignments to the ground truth. The adjusted mutual information, which much like entropy measures the similarity of clusters based on their true labels, is used to validate the cluster size choices of the two algorithms. The plot above gives more credibility to the argument for three clusters when it comes to the breast cancer data. The mutual information is significantly higher for the case of three clusters compared to the other clusters tested. Five clusters still do relatively well for K-Means, but fall short in the GMM validation. The contraceptive use data performs poorly overall when attempting to recreate the true class labels, never exceeding 0.1 AMI for either clustering algorithm. Since this data reflects human activity, it makes intuitive sense that clustering is difficult as human activity is highly nondeterministic.







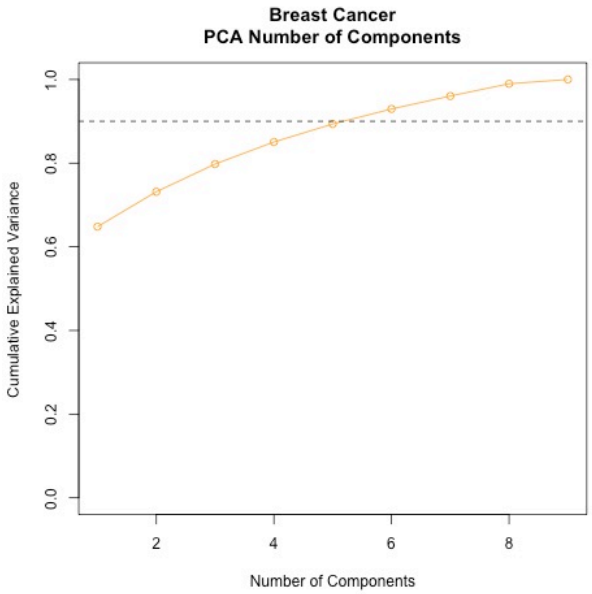


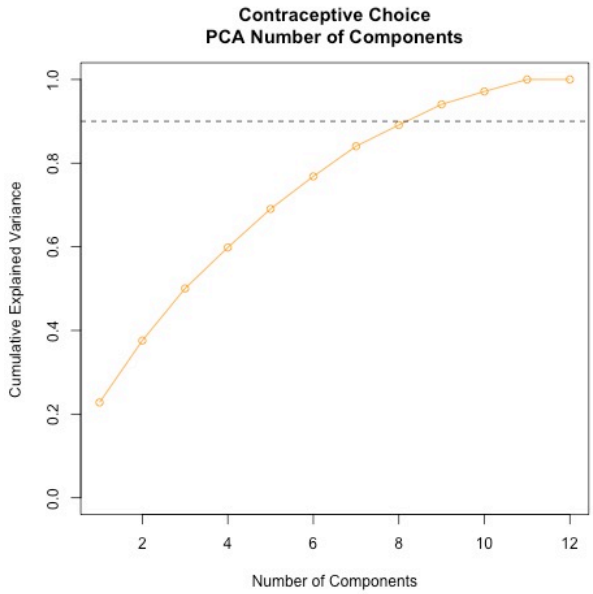
Using t-SNE to visualize the cluster assignments in two-dimensional space, interesting patterns emerge from each dataset. The colors in the plot above represent the cluster assignments given the optimal cluster numbers found using the elbow method and the shapes denote the true labels. The breast cancer data seems fairly well separated in this projected space and K-Means seems to capture this better than GMM as the validation plot suggests. The upper right corner of the plots show three tightly clustered groupings of benign tumors (marked as circles) which may be of interest to clinical researchers. On the contraceptive choice data, GMM seems to capture the natural, human-recognizable groupings in the projected space better than K-Means. Despite this, each cluster clearly has many of each label which again supports the low AMI values seen in the validation chart. This may suggest that additional variables about women's behavior may be necessary to create accurate contraceptive choice clusters, or that this type of human activity data is not well suited for clustering due to its inherent noise.

Dimensionality Reduction

Dimensionality reduction can be used to extract relevant information from a dataset while reducing the number of features needed to train a model. A major issue this course points out is the curse of dimensionality, which the following techniques attempt to mitigate. This is beneficial from the standpoint of computational and time complexity along with the confidence one has in a model, but can also unveil hidden variables in the transformed space. This Google Problem presented in the lectures is a great example of this, as dimensionality reduction helps address polysemy and synonymy.

Principal Components Analysis, or PCA, reduces dimensionality by finding directions of maximal variance in the data. This can be thought of as capturing correlations, then using linear combinations of original variables to represent these correlations as new variables.



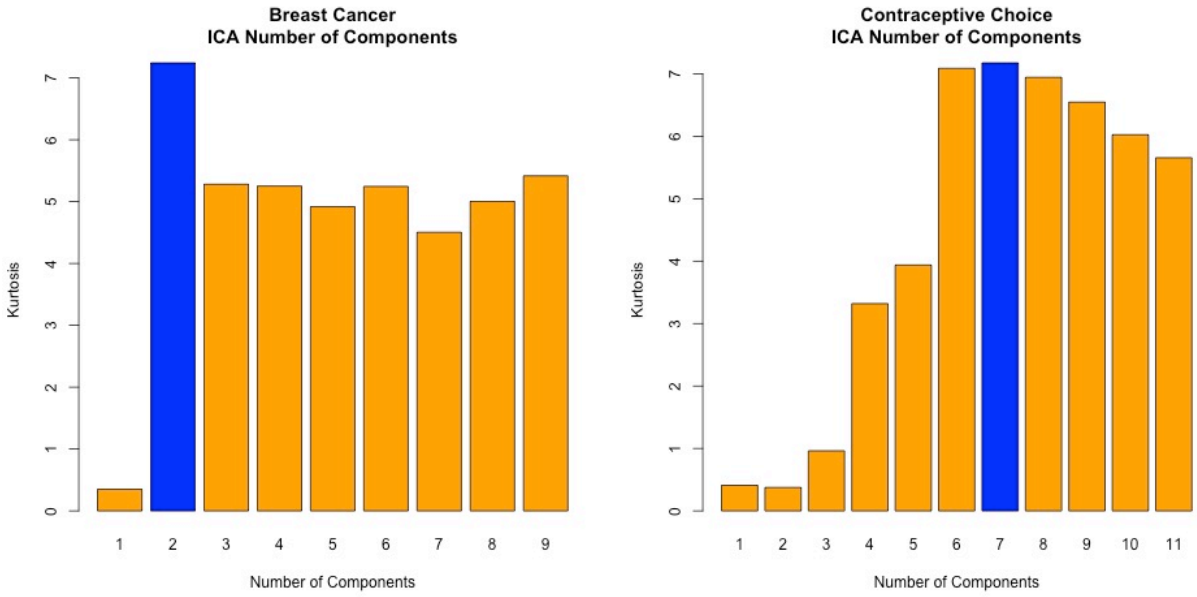


Each dataset had the number of components, k, tested from one through n, the number of variables present in the data. The plots above show the cumulative variance in the data that is explained with each additional principal component used. For this analysis, 90% cumulative variance was arbitrarily chosen as a cutoff after which principal components are excluded. Explaining 90% of the variance seems fairly high, while still eliminating a few principal

components to allow an analysis of how the reduced space affects the clustering and neural network outputs further down the line.

The breast cancer data can explain over 60% of the overall variance with only its first principal component. This shows that linear relationships between variables in this data exist, especially when compared to the contraceptive choice data. Only ~20% of variance is explained by the first principal component in that case. For both datasets, there are steady incremental increases as k approaches n. Given the 90% cumulative variance cutoff, the breast cancer data is reduced to six variables from ten and the contraceptive use data is reduced to nine variables from twelve.

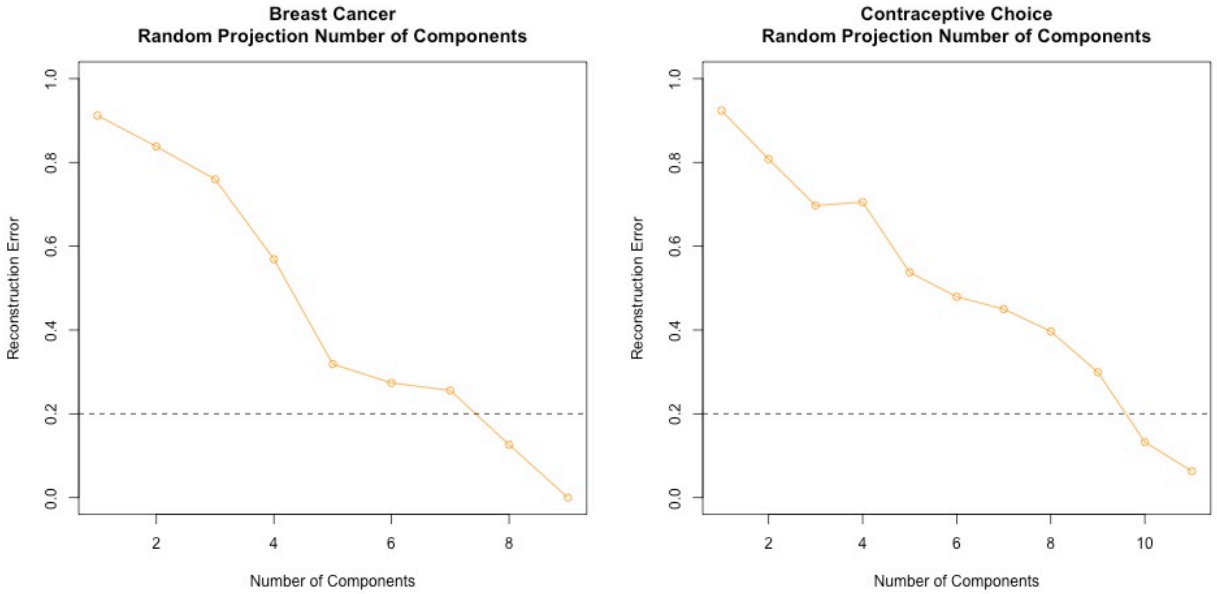
Independent Components Analysis, or ICA, sharply contrasts to PCA. ICA assumes that the data can be separated into a new set of non-Gaussian, statistically independent variables. Unlike PCA, which will produce the same first components as larger values of k are tested, ICA can produce completely unique components for each value of k. For this analysis, kurtosis, which measures non-Gaussianity, is used as a selection criteria for the value of k.



The plots above show the breast cancer data has a strong preference for two independent components while the contraceptive choice data favors seven, albeit less convincingly. The kurtosis of the breast cancer data quickly climbs to its maximal value then stagnates around five for the remaining values of k. On the other hand, the kurtosis for the contraceptive data incrementally climbs to its maximum. This complements the results of PCA well. The contraceptive data shows low variance explained by its first few principal components and climbs slowly in kurtosis vs. k because it truly has underlying phenomena that are independent and require additional components to express. The opposite can be seen when comparing the breast cancer data's PCA vs. ICA results.

Random projections are used to project data into a reduced (or sometimes even *increased*) dimensional space along random vectors. This is often empirically effective, particularly when it is run multiple times to survey the landscape of potential projections, because it can capture relationships between variables without the constraints of PCA. If, for example, many of your variables are Gaussian noise with high variance and the one (or few) truly predictive variables in your data have small variance, PCA would throw most of the information about your relevant and useful variables away! Random projections give an opportunity for these relationships to be expressed. For this analysis, random projections were

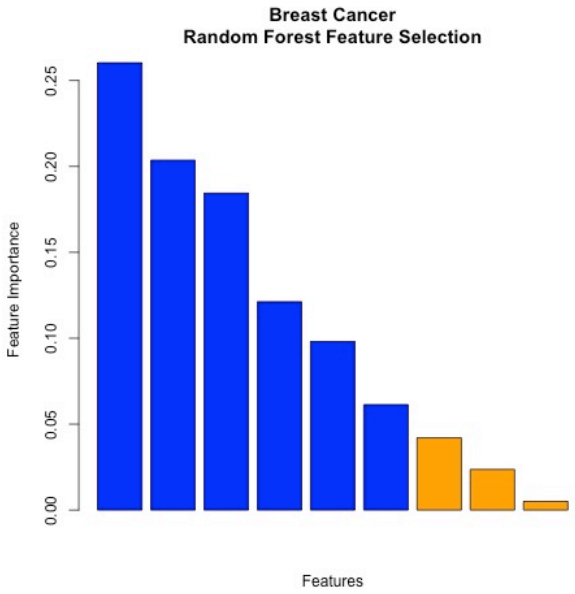
run ten times and the projection with the overall lowest reconstruction error was chosen to be displayed.

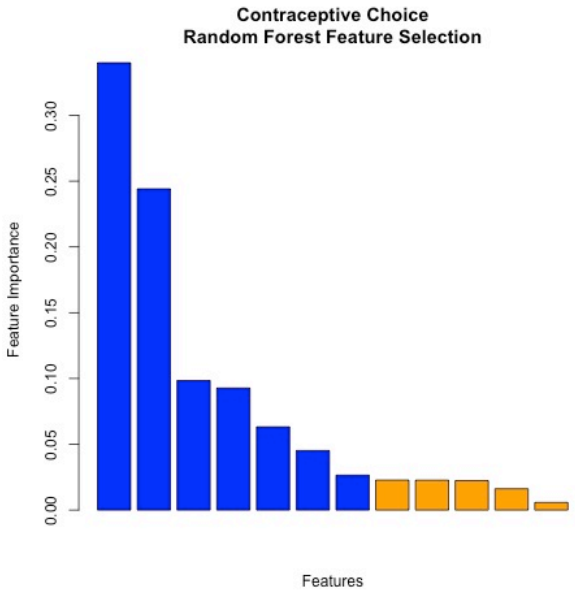


The charts above show the generally declining reconstruction error as larger values of component sizes are tested. With similar logic to that of the arbitrary threshold for PCA's cumulative explained variance, 20% is used here as a threshold for reconstruction error. The reconstruction error measures the similarity between the original data and the reduced-dimension data when it is projected back into the original space. To achieve an error lower than this, the breast cancer data uses eight random directions of projection while the contraceptive data uses ten.

Finally, random forests are used as a form of feature selection. This method will utilize the labels in the data to gain insight into each variable's significance in the model. Since random forests take an aggregate of feature importance from all of their decision trees, these scores can be used to subsample the most predictive variables and reap the benefits of fewer dimensions. One benefit that is unique to this method relative to the others employed throughout this analysis is that the features are not transformed. If one wishes to perform supervised learning and interpretability, not solely predictive power, is important, this method makes understanding the inputs much more accessible when compared to explaining, for example, the linear combinations produced in PCA.

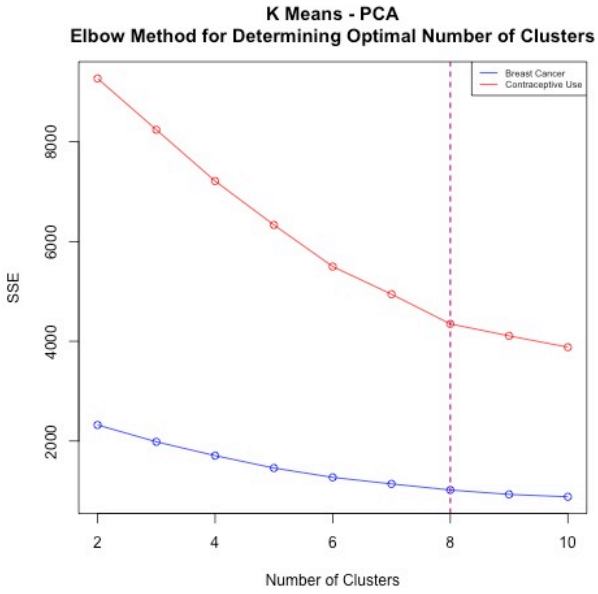
Below are the number of features selected for each data set, sorted in decreasing order by feature importance. A 90% cumulative importance cutoff established which variables to remove. The breast cancer data retained six variables with feature importance falling nearly linearly. The contraceptive data retained seven, but most of the importance lies within the first two features. These results compare favorably with the other feature selection techniques we've seen. It seems that the breast cancer data "shares" importance amongst its features, which is why PCA favors it compared to the contraceptive data. The contraceptive data holds most of its predictive power in a pair of powerful variables and ICA detects those strong signals much better.

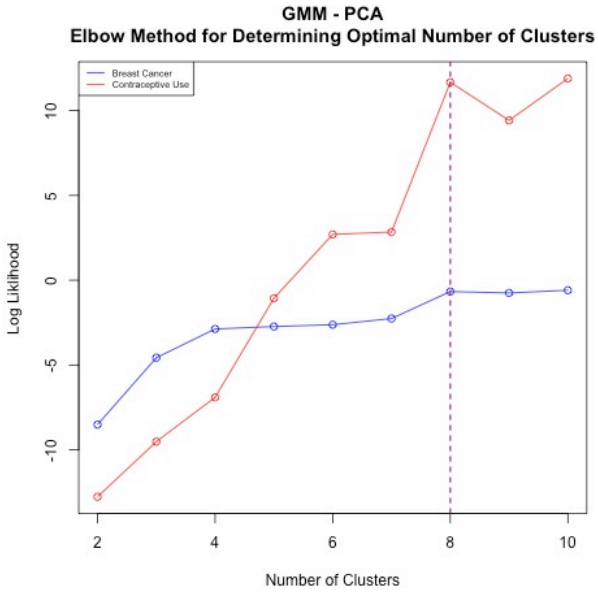


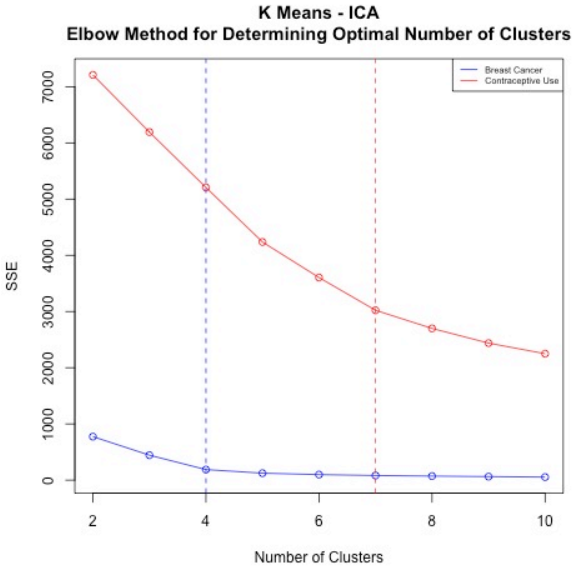


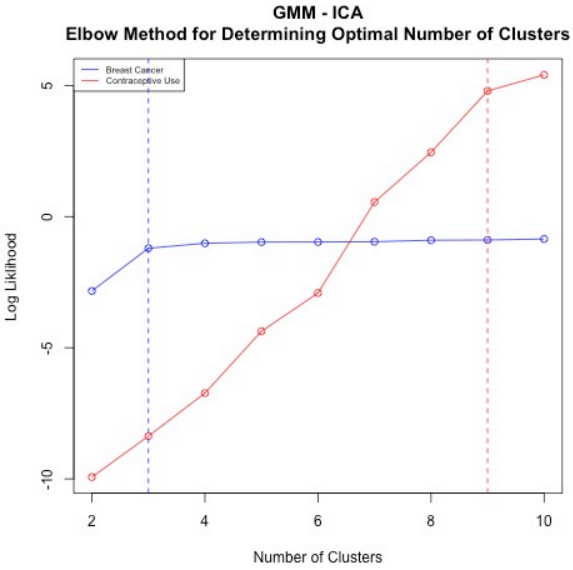
Clustering on Dimensionally Reduced Data

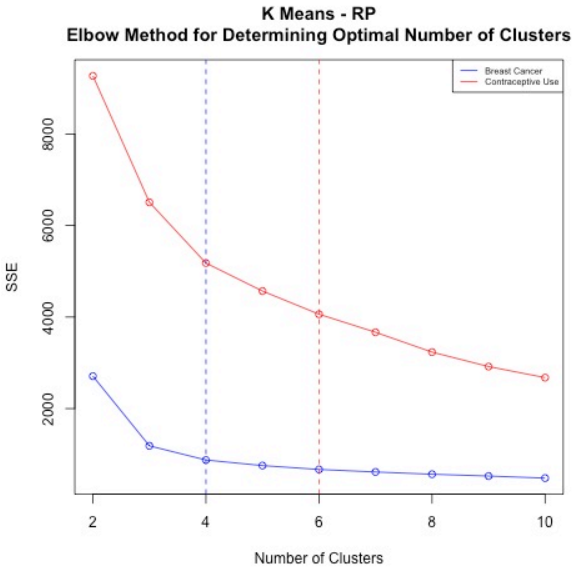
To get an idea of how these dimensionality reduction transformations affect our data, the clustering algorithms are run on each of the four reduced datasets using the optimal number of features for each dimensionality reduction algorithm. The plots below detail how the clusters would be chosen using the elbow method on unlabeled data. It should be noted that, although the results are not pictured, the AMI validation was run for each of these experiments and had largely the same results as clustering on the original dataset. The contraceptive data had a near-zero AMI and the cancer data's three clusters outperformed all other across all algorithm combinations, except RF + GMM where four clusters had a marginally higher AMI. This suggests that, although clustering the reduced data may help summarize and understand it in a different way, the clusters may have been discoverable given the original data.

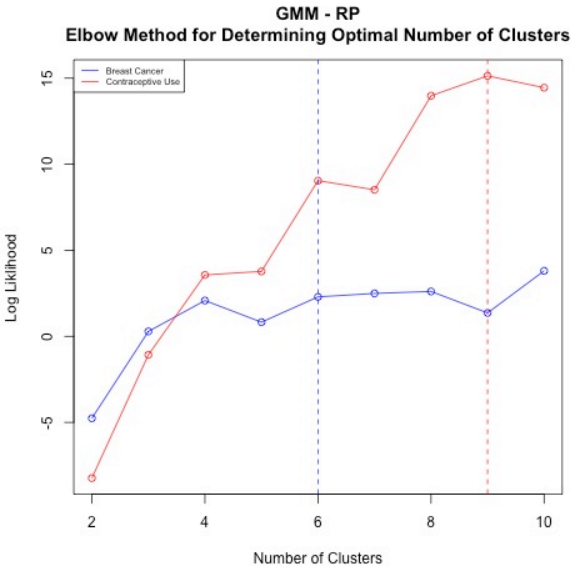


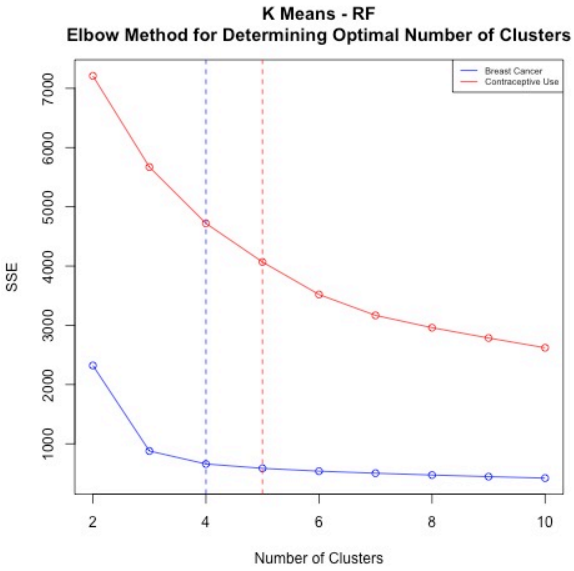


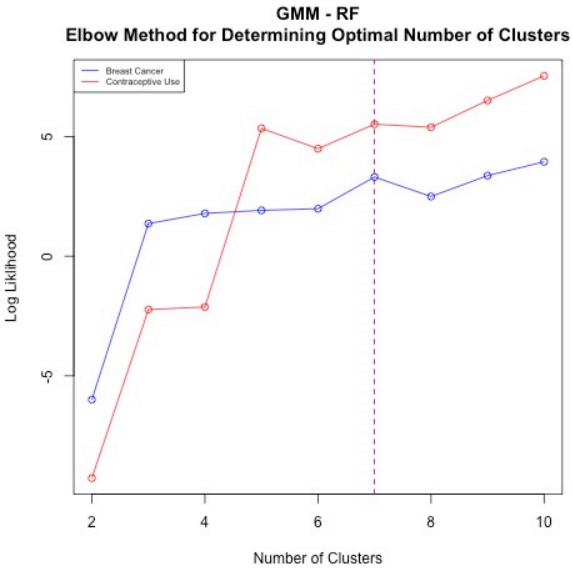

K Means - ICA
Elbow Method for Determining Optimal Number of Clusters
Breast Cancer
Contraceptive Use
SSE
Number of Clusters


GMM - ICA
Elbow Method for Determining Optimal Number of Clusters
Breast Cancer
Contraceptive Use
Log Likilihood
Number of Clusters


K Means - RP
Elbow Method for Determining Optimal Number of Clusters
Breast Cancer
Contraceptive Use
SSE
Number of Clusters

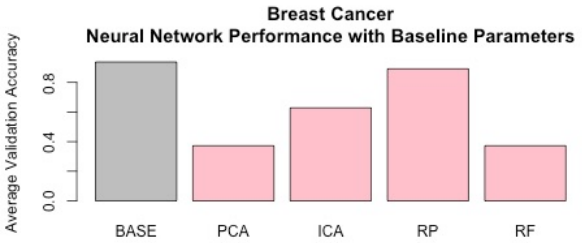

GMM - RP
Elbow Method for Determining Optimal Number of Clusters
Breast Cancer
Contraceptive Use
Log Likilihood
Number of Clusters

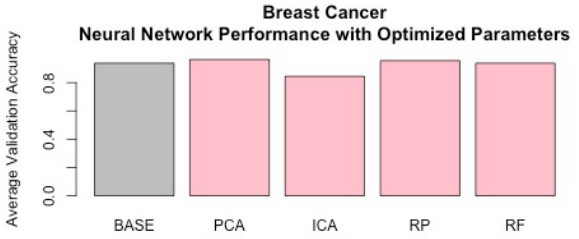

K Means - RF
Elbow Method for Determining Optimal Number of Clusters
Breast Cancer
Contraceptive Use
SSE
Number of Clusters

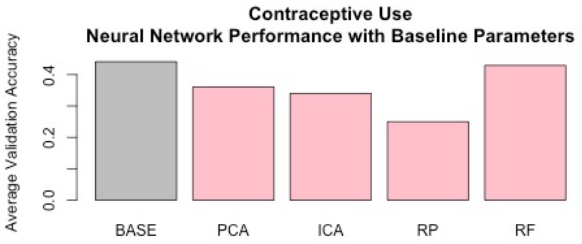

GMM - RF
Elbow Method for Determining Optimal Number of Clusters
Breast Cancer
Contraceptive Use
Log Likilihood
Number of Clusters

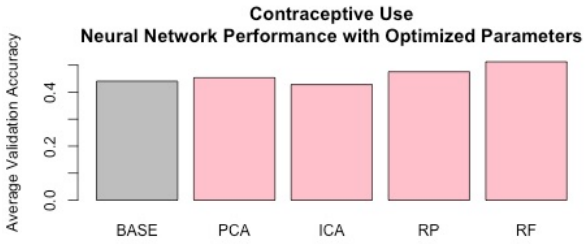
K-Means favors lower values of k for both datasets compared to GMM. This may be due to GMM's soft clustering which allows probabilities for each class and could encourage more classes if many points have flatter distributions. Another distinct feature separating the clustering algorithms is that, when directly comparing the cluster numbers that the original data found optimal, K-Means saw all SSE scores improve using reduced data while GMM suffered a loss in log likelihood across the board. The, in some cases stark, decrease in SSE could be due to the different scaling in the transformed space for K-Means. GMM's log-likelihood takes a per sample average, so it can be said that it doesn't benefit from dimensionality reduction on these two datasets.

Neural Networks

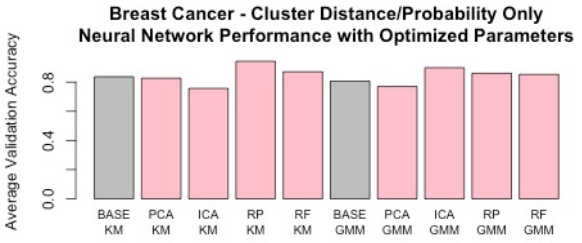


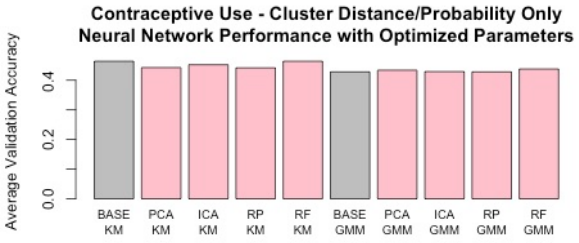




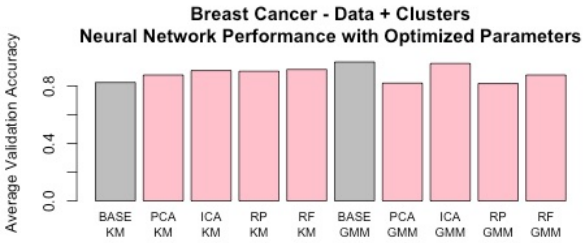


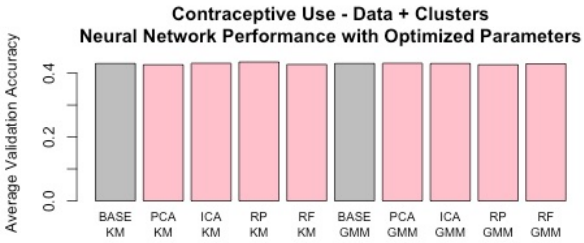
The neural network created in assignment one was recreated and its optimal architecture used to test the performance of the dimensionally reduced data, using each algorithms respective optimal value of k. Using the architecture from assignment one is a requirement mentioned in office hours, but the charts above (on the left) show relatively poor performance across the board using the parameters optimized to the baseline data. On the right, however, are the average validation scores of the networks when the same grid search for parameters used on the baseline data is performed. There is a considerable difference in validation scores when each network's weights are optimized to the reduced dimension data. For both breast cancer and contraceptive choice data, all of the reduced datasets except ICA perform better than the network ran on the original data alone with an optimized network.





The charts above show the affect of using the distance to each cluster for K-Means and probability of belonging to each cluster for GMM as predictors in the neural network. All clustering/dimensionality reduction combinations use optimal values of cluster and dimension number but networks have their hyperparameters selected through grid search. Although these don't quite reach the potential of the dimensionally reduced data in most cases, these models don't perform all that poorly for being trained on such abstractions of the data. GMM tends to perform slightly worse on average than K-Means on both datasets.





Finally each network was trained on each data set along with the cluster assignments as an additional predictor. For the breast cancer data, GMM again performs worse on average across all dimensionality reduction algorithms despite. It is interesting to note that despite this, the best scoring combination in this category is the baseline data along with GMM clusters. The contraceptive data has fairly consistent performance on all combinations, although these numbers fall below those given by training the network on the dimensionally reduced datasets alone. Although the addition of the clusters as features may seem like relevant information, it appears that its redundancy actually hurt the models for both datasets on average. As should be learned when studying dimensionality reduction: fewer features are (generally) preferred.

Overall, the best validation performance for the breast cancer data belongs to the network trained on all of the original data with the GMM features attached, scoring 96.64%. The contraceptive choice data saw its best validation accuracy from the random forest feature importance filter, scoring 51.26%. Dimensionality reduction has shown to improve the performance of neural networks in the cases of these two datasets. Clustering has revealed interesting groupings within the data and suggests that smaller subcategories may exist within or even between the labels assigned. With domain expertise in the areas of these datasets, one may even be able to understand phenomena unrelated to the labels that the independent variables represent by studying the clusters into which they fall.

Improvements

This analysis would benefit from looking at additional dimensionality reduction techniques such as forward and backward stepwise selection. The K-Means portion could also be improved by specifying a different distance measurement other than Euclidean, such as the Gower distance, to properly handle the binary variables in both datasets. Cluster sizes and average statistics of the columns within each cluster would allow a deeper understanding of the groupings and each dataset's structure. Lastly, the elbow method used could be improved by setting a threshold for the percentage decrease at each step to make it less subjective.