CS 7641 – Unsupervised Learning and Dimensionality Reduction (Assignment 3)

Introduction

This paper explores unsupervised learning and dimensionality reduction. Six different algorithms are implemented; the first two are clustering – k-means clustering and Expectation Maximization and the last four are dimensionality reduction algorithms – PCA, ICA, Randomized Projections, and Random Forest. The experiments are split into four main parts. Part 1 applies the clustering algorithms on both datasets. Part 2 applies the dimensionality reduction algorithms and reproduces clustering experiments with dimensionality reduction on both datasets. Part 3 applies dimensionality reduction algorithms to the Faulty Plates dataset and runs a neural network on the data. Part 4 applies clustering algorithms and uses the clusters as new features for the neural network learner. All algorithms are evaluated using Scikit-learn in Python.

Description of Datasets

Both datasets can be found on the UCI Machine Learning Repository.

Breast Cancer:

The Breast Cancer Wisconsin dataset contains 569 instances 31 attributes that classifies whether a breast mass is malignant or benign. The features are measurements calculated from a digitized image of a breast mass, describing characteristics of the cell nuclei in the image. All features are real-valued. This is an interesting problem because breast cancer affects about 200,000 people a year in the U.S. and about 12% of U.S. women develop breast cancer in their lifetime. Early identification of malignant breast mass cells using machine learning algorithms is very beneficial for prevention, treatment, and potentially save lives. Only 30 attributes are used here since one feature is the ID.

Faulty Steel Plates:

The Steel Plates Faults dataset contains 1,941 instances with 27 attributes that classifies faulty steel plates into 7 different categories – Pastry, Z_Scratch, K_Scratch, Stains, Dirtiness, Bumps, and Other_Faults. Other_Faults comprises about 35% of the classes, while Bumps and K_Scratch make up about 20% each. All of the features are numerical. Analyzing faulty steel plates is important in improving safety and reducing costs. Automatic pattern recognition of faulty steel plates can reduce the amount of defective plates that are in circulation and possibly used. Defective plates may be dangerous and pose safety concerns or not be aesthetically pleasing. Classifying defective steel plates before they leave the factory can save high return shipping fees and lead to better customer satisfaction. This technique can also be widely applicable to evaluate other types of defective metals. It's also good because it's challenging to classify the dataset into seven different classes.

Part 1: Clustering

Clustering is an unsupervised learning algorithm that groups a set of observations together that are similar to each other compared to those in other groups.

The K-means algorithms takes in a parameter k, for the amount of clusters, and randomly generates k means. K-clusters are formed based on associating each observation to the closest mean. The least squared Euclidean distance is used for measurement. The center of each of the clusters becomes the new mean. These steps are iterated until convergence is reach. K-means is one of the faster clustering algorithms, but it can fall into local minima.

Expectation-Maximization is an iterative method using maximum likelihood to find the clusters means. This algorithm alternates between a soft clustering (Expectation) and computing the means of a soft cluster (Maximization). Expectation calculates the likelihood that the observation is in a certain cluster based on the mean. Maximization computes the means from likelihoods, using weighted averages of the data points. At each step it maximizes the

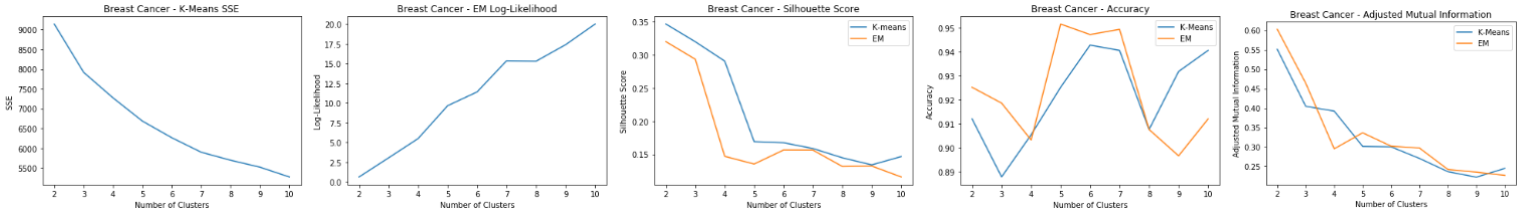
likelihood of the distribution until convergence. Sci-kit learn uses the Gaussian Mixture Model to calculate EM. It uses the mean and the covariance of the gaussian distribution to fit the clusters.

There are several methods to choose a good k. For k-means, the sum of squared distances within clusters (between each member of the cluster and its centroid) can be plotted against the number of clusters. The elbow point, where SSE decreases sharply, can be used to determine k. For EM, log-likelihood calculates the likelihood that the data is to be generated by the parameters estimated. Higher likelihood means that the data is more likely to be generated by the estimated parameters. Higher values aren't always good because of overfitting, so there's a tradeoff.

The Silhouette score can also be used for both k-means and EM. This measure takes into account both intra and inter cluster distances, which explains how similar an observation is to its own cluster compared to other clusters. The range is from -1 to 1. The best value is 1, meaning it matches well with its own cluster and is far from other clusters. Values near 0 indicate overlapping clusters.

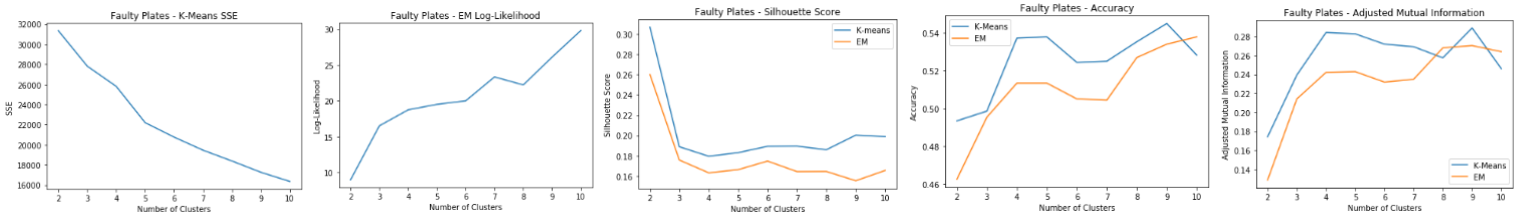
To evaluate the clusters, we can measure accuracy and adjusted mutual information score. Accuracy measures the percentage of the predicted label matching with the true label since we have the classification. Mutual information score measures the similarity between two labels of the same data between two clusters. The adjusted score just takes into account chance. The range is from 0 to 1, and 1 means that the labels agree.

<u>Breast Cancer</u>



The SSE plot shows an elbow point at k = 7 where it levels out a bit and there is also a spike in the EM log-likelihood at k = 7. The silhouette score also shows a spike at k=7 for both k-means and EM. The clusters do not always have to match up with the amount of labels. In this case, there are two classifications, malignant and benign. A cluster of k=7 makes sense since there could possibly be seven different types of breast mass cells that could be identified as either malignant or benign. To evaluate the clusters, the accuracy is highest around k=7 and the adjusted mutual information has leveled out somewhat and shows a balance at k=7. This indicates good prediction and clustering.

<u>Faulty Plates</u>



Using the elbow method, the k-means SSE starts leveling out around k=9. The log-likelihood method shows a spike at k=7 and k=9 also produces a good result. A good cluster may be k=7 and k=9. The silhouette score for k-means shows a spike at k=9, but a dip when evaluated by EM. K=6 experiences a small spike for silhouette score for both clustering methods, so it could also be a candidate. There are seven different labels for the faulty plates dataset, so k=7 matches the labels. K=9 also makes sense because one of the labels is categorized as "other" and there could be two distinct

faulty plates within the "other" category. In evaluating the k's, accuracy is highest and adjusted mutual information peaks at k=9. K=6 and k=7 do not perform as well in accuracy.

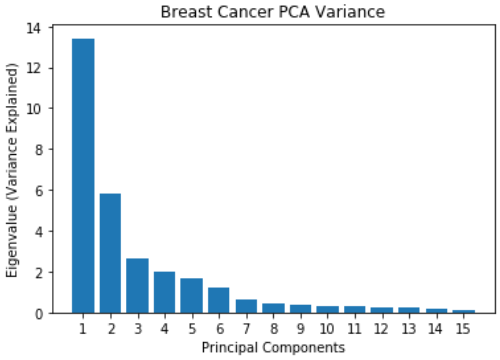
Part 2: Dimensional Reduction and Clustering

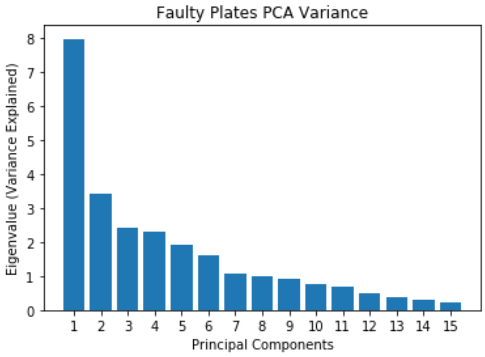
Dimensionality reduction selects a subset of important relevant features to simplify models, decrease training times, reduce overfitting, and avoid the curse of dimensionality. PCA, ICA, Randomized Projection, and Random Forests are experimented for dimensionality reduction. After finding the amount of dimensions to reduce, the data with reduced features are used to run a clustering algorithm.

Principal Component Analysis (PCA):

Principal component analysis uses orthogonal transformation and linear combination to identify important components that maximizes variance. PCA is used to reduce a large set of features into a subset that still contains most of the information.

To determine dimensional reduction for PCA, the variance explained by the components or distribution of eigenvalues in PCA is examined. The elbow method can be used to evaluate the number of principal components to choose.



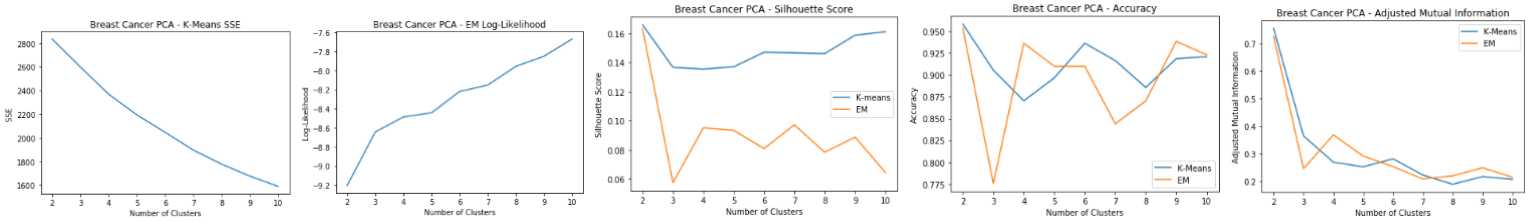


For the Breast Cancer dataset, most of the variance is explained by the first 7 principal components. The original dataset has 30 features, and now the dimensions have been reduced to 7 principal components.

For the Faulty Plates dataset, most of the variance can be explained by the first 12 principal components. The original dataset has 27 features, and now has been reduced to 12 principal components.

PCA Clustering:

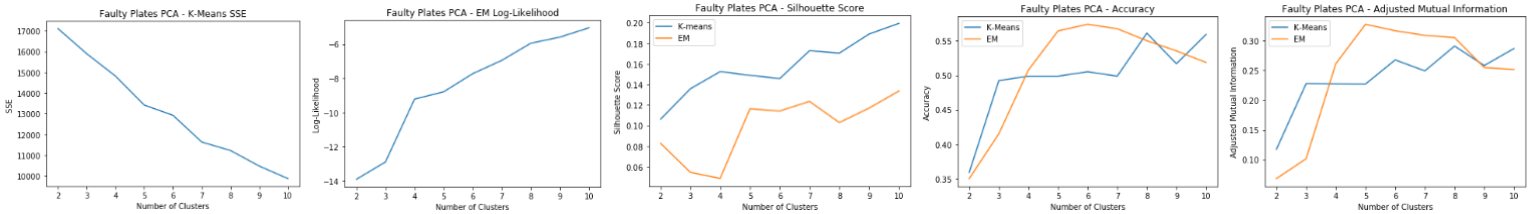
<u>Breast Cancer</u>



Using 7 principal components, the clustering algorithm shows a possible good cluster at k=6. At k=6, the k-means elbow method doesn't show much, but the log-likelihood for EM and Silhouette score for k-means show a slight jump. The low silhouette score for EM shows that it doesn't do as well as k-means in separating the clusters. The low

silhouette score may mean that I have lost too much information in dimensionality reduction. The accuracy experiences a sharp dip for EM at k=3. Compared to the original clustering algorithm, PCA with clustering does not do a better job accuracy wise. The lower silhouette score for both k-means and EM compared to the clustering algorithm shows that the PC's don't necessarily improve cluster quality or capture the structure well. The SSE from k-means shows a reduction of 1/3 from the original clustering, which means that PCA reduces within cluster error, but does worse intra-cluster. This is also backed by the lower log-likelihood.

Faulty Plates



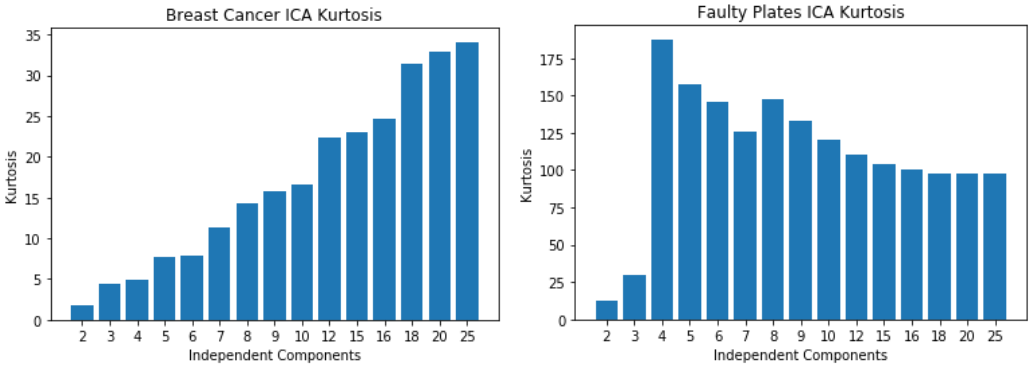
The elbow method from k-means, log-likelihood from EM, and peak in silhouette score for both, show that k=7 may be good. The silhouette score for EM is much lower than k-means, which indicates possibly overlapping clusters. However, the accuracy and adjMI shows that EM actually has relatively decent prediction accuracy.

PCA does do a good job of reducing the SSE in k-means from ~24,000 to ~14,000. This means that PCA does a good job reducing inter cluster error, but the lower silhouette score shows that intra cluster distance gets worse. EM clustering with PCA performs better in prediction accuracy and k-means with PCA performs worse.

Independent Component Analysis (ICA):

Independent component analysis tries to decompose data into independent non-Gaussian components. It does this by maximizing mutual information between the original data and the independent components. The sub-components are assumed to be non-Gaussian and independent from each other. FastICA in Sci-kit learn is used.

The number of independent components to choose can be evaluated by their kurtosis values since kurtosis measures gaussianity and ICA tries to maximize non-gaussianity. A kurtosis near 3 is gaussian, so it's best to find a kurtosis that has the highest absolute value of the mean of the kurtosis.



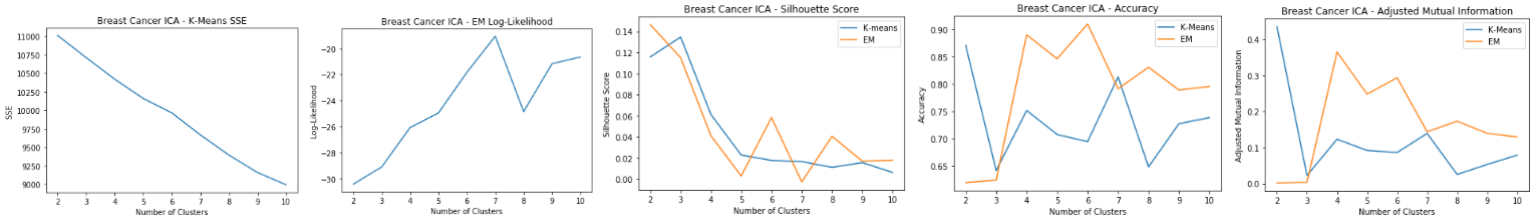
For the Breast Cancer dataset, the Kurtosis keeps increasing, meaning non-gaussianity increases as the number of independent components increases. The original dataset has 30 features, but 25 IC's is enough since it starts leveling off.

For the Faulty Plates dataset, the Kurtosis peaks at 4. The original dataset has 27 features and now 4 independent components are selected. This is a huge reduction in dimensionality.

CS 7641 – Unsupervised Learning and Dimensionality Reduction (Assignment 3)

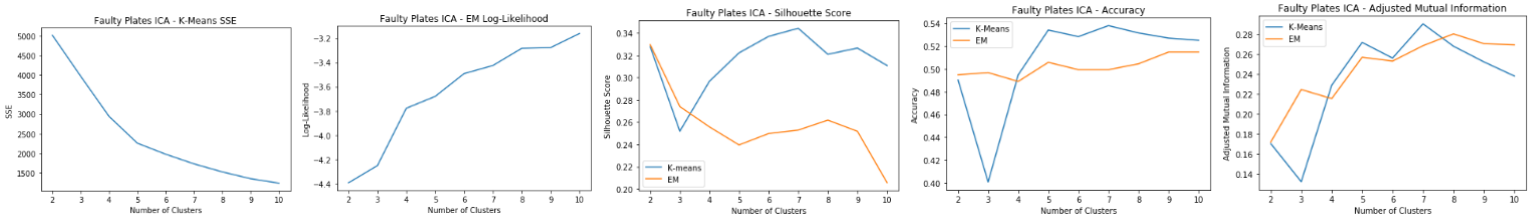
ICA Clustering:

Breast Cancer



The SSE plot from ICA doesn't show any elbow point, but it does decrease the SSE within clusters from the original dataset. From the EM log-likelihood plot, there's a clear peak where k = 7. The silhouette scores for k-means and EM are more similar to each other, meaning they produce similar cluster distances. K-means shows a slight peak at k=7, but EM has visible peaks at k = 6 and 8. In evaluating the plots, EM has highest accuracy at k=7 and k-means has a peak at k=7. ICA overall has less accuracy than PCA and the original dataset. The AdjMI for EM is consistently higher than k-means. A good amount of clusters are k = 6 and 7.

Faulty Plates

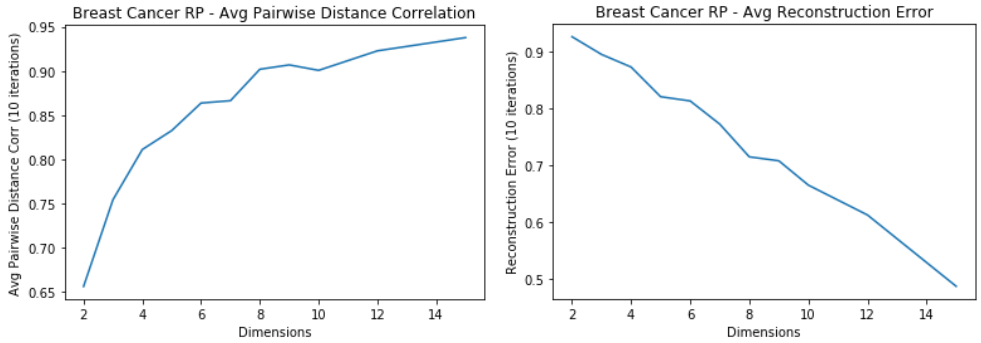


For faulty plates, the within cluster SSE decreases significantly compared to the original dataset. An elbow exists at k=5. The Silhouette score shows that EM produces clusters that are much closer together than k-means and that a good cluster exists at k=7 for both k-means and EM. This is backed by the accuracy from EM and k-means. K-means accuracy is highest at k=7 and levels off around there for EM. AdjMI is also high for both at k=7.

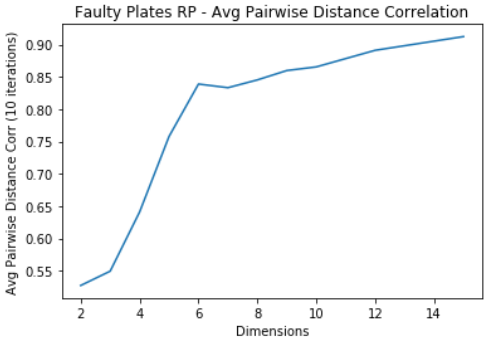
Random Projection (RP):

Random projection also reduces dimensions, but does so randomly using a Gaussian distribution. The benefits of random projection is that it's computationally efficient and works well on low dimensions. SparseRandomProjection is used in Sci-kit learn.

Random projections may perform pretty poorly based on one random generation, so 10 iterations are run and averaged for evaluation. Random Projection is evaluated based on Average Pairwise Distance Correlation and Average Reconstruction Error. The goal of Random Projection is to preserve the pairwise distances between any two samples of the dataset, so we want to maximize the variance and average pairwise distance correlation. We also want to minimize the reconstruction error, which is the squared distance between the original data and the estimate.



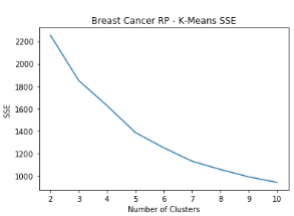
The Breast Cancer dataset shows that Avg Pairwise Distance Correlation starts level off around 7 dimensions. The average reconstruction error keeps decreasing for higher dimensions. The higher the dimensions, the lower reconstruction error, but there is a tradeoff between overfitting and computational time.

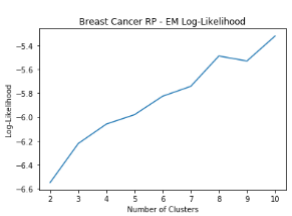


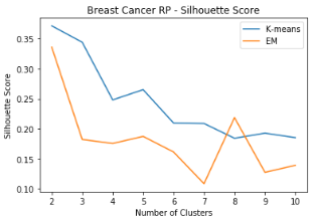
The Faulty Plates dataset also shows that Avg Pairwise Distance Correlation starts level off around 7 dimensions. The average reconstruction error keeps decreasing for higher dimensions. Again, there is a tradeoff. For clustering, the number of dimensions produced by Random Projection is 7.

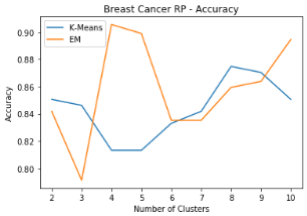
RP Clustering:

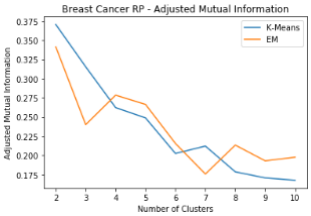
<u>Breast Cancer</u>





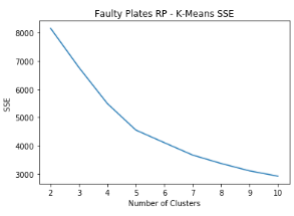


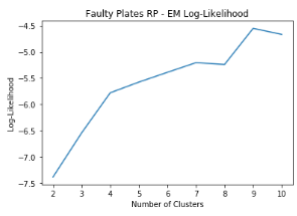


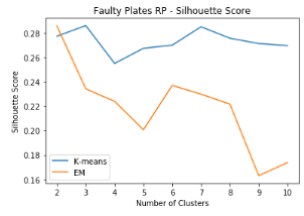


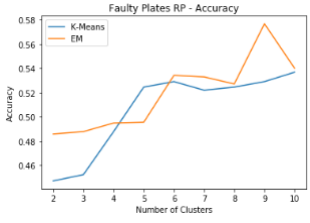
The SSE has an elbow point around k=5 and shows a much lower SSE within clusters due to dimension reduction. The log-likelihood has a peak around k=8. The silhouette score on average has imporved for k-means from the original clustering. There's a peak at k=8 for EM and k=5 for k-means. The accuracy is much lower than the original dataset and EM has better accuracy than k-means that peaks at k=5. A good cluster seems to be k=5 for both.

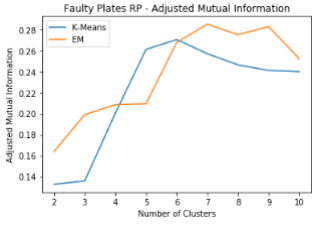
<u>Faulty Plates</u>









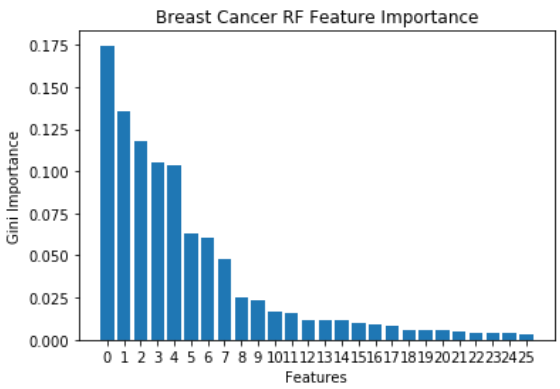


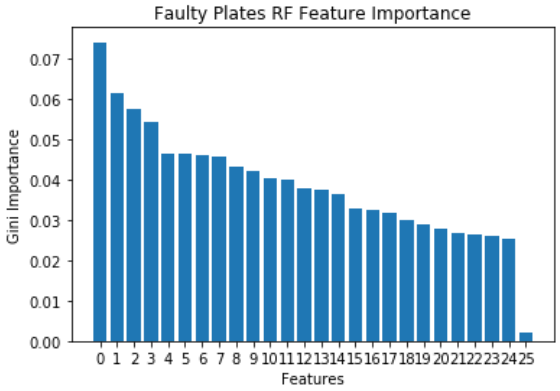
The faulty plates dataset shows a slight dip at k=5 and EM shows a good log-likelihood of k=9. Based on the silhouette score, k=7 is a good since both EM and k-means sort of peak around there. The accuracy for EM backs that k=9 is a good cluster for EM. The accuracy for RP EM is the highest so far for all the dimensionality algorithms at k=9. The accuracy for k-means shows that k=7 is good because it starts level off around there.

Random Forest (RF):

Random Forest is a strong learner that is an ensemble of weak learner decision trees. Feature selection for Random Forest is based on feature importance. Feature importance is measured by Gini importance, which is the total

decrease in node impurity reaching that node averaged over all trees of the ensemble. The higher the Gini Importance value, the more important the feature. The elbow method is used to evaluate the number of features to use. RandomForestClassifier in Sci-kit Learn is used.



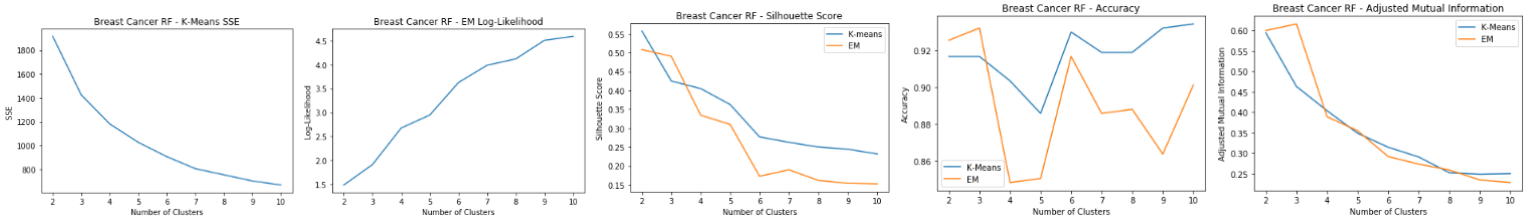


For the Breast Cancer dataset, there seems to be 10 important features, because the Gini Importance starts to level off after that. The features have reduced from 30 to 10 using Random Forest.

The Faulty Plates dataset has a lot of important features, and has a sharp decline at 25 features. Random Forest evaluates 24 important features. This is just a decrease in dimensionality from 27 to 24.

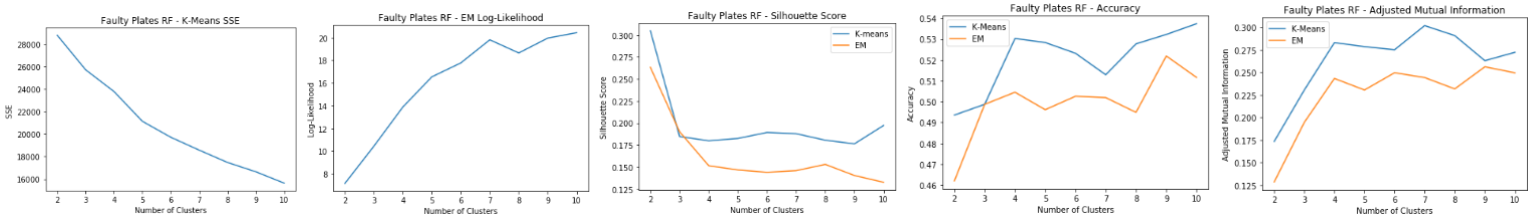
RF Clustering:

Breast Cancer



The Breast Cancer dataset shows an elbow point around k=7. The EM log-likelihood keeps increasing and begins to level out around k=9. The silhouette score for both show a consistent decrease. As the number of clusters get larger, the cluster distances for both start decreasing. It's hard to tell what a good k is from the graphs. However, based on the accuracy it shows that k=3, 6, and 10 are good clusters.

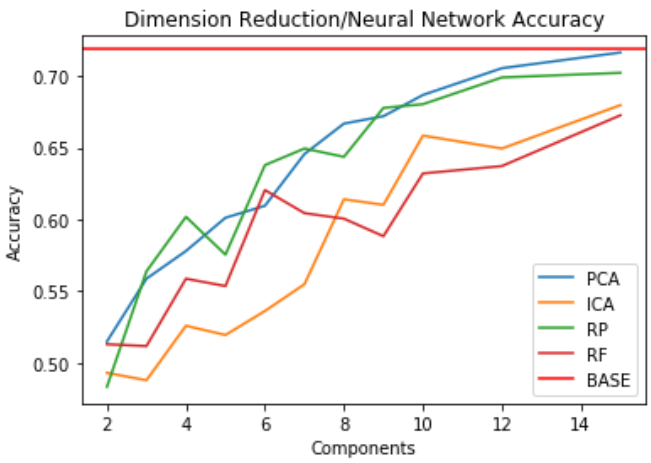
Faulty Plates

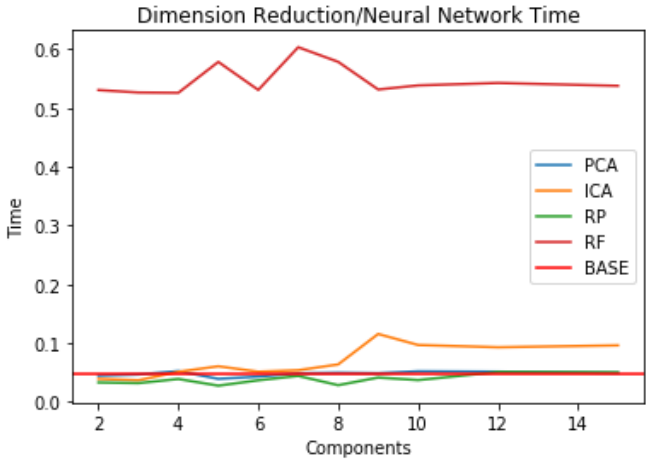


Random forest doesn't decrease within cluster error in k-means as much as the other algorithms. There isn't really an elbow point either. The EM log likelihood shows a peak at k=7. The silhouette score is the highest for k-means around 7 and 8 for EM. The accuracy shows that k=9 and k=4 are good clusters for both.

Part 3: Dimensional Reduction Algorithms with Neural Network

The faulty plates dataset from assignment 1 is used to evaluate the dimensional reduction algorithms with the Neural Network. The baseline for the Neural Network is a 71.9% accuracy with hyperparameters of learning rate = 0.1, momentum = 0.3, and one hidden layer with 14 neurons.

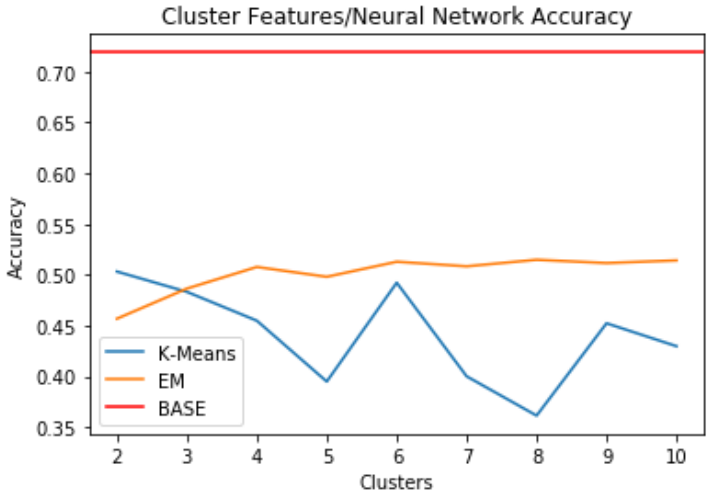




This experiment compares accuracy of the different reduction algorithms on the Neural Network to the benchmark Neural Network. The accuracy shows that PCA and Random Projection perform well on the Neural Network at higher components. With around 15 components, the accuracy is close to the benchmark. The dimensions are much lower than the original 27 features, so it simplifies the model complexity of the model. The compute times are also very similar between the benchmark, random projection, and PCA. Random Forest has a much higher computational time than the rest of the algorithms. The ICA algorithm increases in time as there are more components.

Part 4: Cluster Features with Neural Network

The clusters are used as features for the Neural Network. Again, the baseline for the Neural Network is a 71.9% accuracy with hyperparameters of learning rate = 0.1, momentum = 0.3, and one hidden layer with 14 neurons.



When clusters are used as an additional feature along with the original data, it doesn't do a good job of predicting the labels. It performs far below the benchmark accuracy of 71.9%. Although it performs poorly, it seems like accuracy is best for EM and k-means when the number of clusters is 6.

Conclusion

The dimensions for each of the components in the clustering algorithms in Part 2 were as follows: PCA = 12, ICA = 4, RP = 7, RF = 24.

Dim Red	Time (ms)
PCA = 12	0.0510
ICA = 4	0.0510
RP = 7	0.0440
RF = 24	0.5800

The computational times are similar for PCA and ICA. Random Projection did best in computational time and in lower dimensions. Random Forest used as feature selection takes too much computational time compared to the other algorithms. ICA on average also has higher computing times than PCA and Random Projection.

The silhouette score for clustering is on average lower for EM, which makes sense because it's a soft classifier compared to K-means, a hard classifier. Most of the dimensionality reductions do help in reducing SSE within clusters. PCA and ICA do not increase the intra-cluster distance, but Random Projection and Random Forest does.