Unsupervised Learning Report

# Introduction

Clustering algorithms and dimensionality reduction algorithms were run on data sets explored in a classification setting previously in assignment one to consider the unlabeled data alone and to compare clusters with the ground truth. The two data sets considered, as in assignment one, were the adult census data set and the banknote data set.

*Adult Census Income Data Set:*

This data set contains one thousand instances of adults, classifying them into two groups – those making above $50,000 and those making below. The data set contains a thousand instances and fourteen attributes of information (not including the >50k/<50k classifier). The attributes are mixed between nominal and continuous integer. Gender, race, native country, relationship status, occupation, education-level, marital-status, and workclass are all recorded as nominal attributes. Age, capital gain, capital loss, hours-per-week, number of years in education, and final sampling weight, are all among the integer attributes. The dataset is interesting to consider in an unsupervised setting for the same reasons it was interesting in a classification setting – a mix in attribution type and difficulty mapping nominal attributes – plus the fact that this data set was never too successful in the classification setting, with error always above 20%, and the inclusion of attributes that have high variance but limited correlation with the labels. The high variance but low information gain attributes, mainly the final sampling weight attribute, really pose difficulty for unsupervised learning algorithms.

*Banknote Authentication Data Set:*

This data set contains information from 1372 instances of images of genuine and forged banknote-like specimens. The information is across four continuous value attributes – the variance, skewness, kurtosis, and entropy of the wavelet transformed image. The labels of this data set designate if the banknote is forged or genuine.

This data set also complements my other data set quite nicely; the adult data set had validation error consistently over 20% in the classification setting, but the banknote data set was able to achieve 0% validation error quite frequently. Moreover, the challenges present in the adult data set, mixed attribute type and uncorrelated attributes, are not present in the banknote data set.

*Methodology:*

For all the following algorithms and exercises, the k used for clustering is 2. Because both of these datasets have binary labels clustering into two groups makes the most sense. Evaluating each algorithm posed its own unique challenges, because the algorithm could produce output that “describes” the data set usefully without lining up with the ground truth labels, and so each algorithm was evaluated uniquely while also being compared against the ground truth. For every algorithm in this paper, a number of different hyperparameters are manipulated, and the algorithm is run with every possible configuration to consider the effect of each hyperparameter.

For the adult data set, an additional data set, named “adult\_lite” was created to overcome the challenges presented by the nominal attributes. This data set has far higher dimensionality because each nominal attribute is turned into multiple binary attributes, which poses its own unique challenges.

# K-Means without Dimensionality Reduction

K-means clustering algorithms were run on both data sets using scikit-learn’s clustering package. The number of clusters was 2 for both data sets as explained earlier, meanwhile the number of initializations and the maximum number of iterations were manipulated alongside the method of initialization and use of normalization.

*Adult Data Set:*

 Over the several dozen different configurations run, only four different values were outputted for “error”, measured as the difference between the clusters and ground truth. The values were .250, .368, .369, and .388. Given the similarity of the latter three values, it is likely that those clusters differed very slightly, and so for our purposes we can consider the algorithm to have clustered in two different ways total, with one of them being far closer to the ground truth.

The table on the right shows the effect of the number of initializations and maximum number of iterations on the “error”. “Error” is used to understand these algorithms more so than evaluate them as successful of not in this setting – from considering the hyperparameters’ effect on “error”, we see that maximal iterations and number of initializations barely affect the adult data set.

 From the table below, we see the effect of the normalization (data was normalized for true values, and was not for false values) and initialization method. Clearly, normalizing the data had the biggest effect on the clusters, and brought the clusters to more closely resemble the ground truth. Normalizing the data makes a lot of sense because an attribute with a greater range of values is not more important or less important, and normalizing removes that bias. The k-means++ initialization method is also clearly useful. For future use, the optimal k-means algorithm is determined to be one with normalization on, k-means++ initialization, and 10 initializations and 200 maximal iterations. The latter two were chosen to conserve computation time while also being conservative with respect to getting stuck at local minima or not converging.

The graph below on the left shows the optimal k-means’ clusters, and the one below on the right shows the ground truth, across attributes age, education-number, and hours per week. These three attributes yield the most understandable image, but also do not tell us a lot because there’s 11 other attributes, and therefore dimensions, that cannot be seen. Plotting for the adult data set does not truly help understand the clusters.

# 

*Adult Data K-Means (left is output, right is ground truth)*

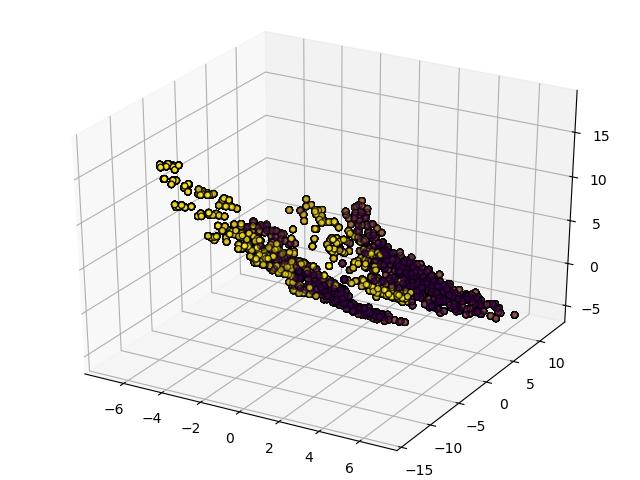
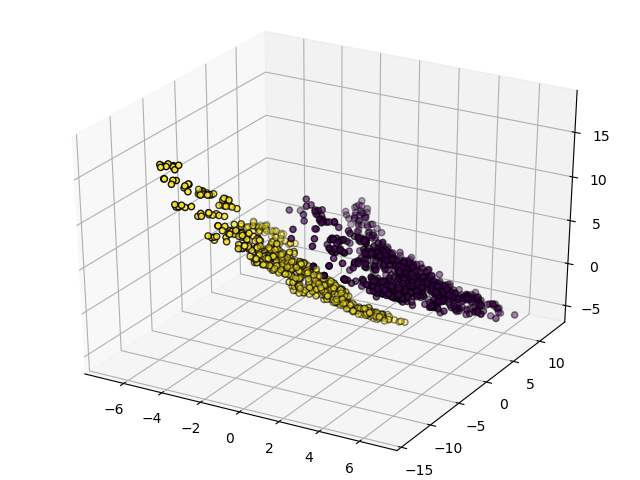
The two cluster centers for the model closer to the ground truth did not differ at all for any of the categorical attributes, which essentially means that the information provided in those attributes is not being used. Therefore, I ran the K-means algorithm on the adult\_lite data set, which represented every nominal attribute as multiple binary attributes, to see how the cluster centers worked.

The adult\_lite plots are limited in use for the same reasons the above plots are unhelpful. There were three significantly unique values for the “error”, with the smallest “error” being .249, essentially the same as the adult data set. The difference between the two cluster centers for the .249 output algorithm was zero was for three attributes, currently married, white-collar, and white. It makes sense that these three attributes held no import on the cluster participation, because other attributes made those redundant. Capital gain, capital loss, and hours worked a week had the most difference in terms of cluster center difference in that order, which signifies that those are three most significant attributes when partitioning the data set into two.

*Banknote Data Set:*

Somewhat similarly to the adult data set, very few distinct “error” values were produced – only two in this case, .42 and .38. In this case, the .38 “error” clusters were created for both initialization methods, all number of initializations, and all maximal iterations, but only when there was no nominalization. When there was normalization, the clusters yielded an “error” of .42. In this situation, the .42 clusters are the more objectively successful clusters because normalization fundamentally agrees with unsupervised methods – it just so happens that the attributes with greater variance were in fact more highly correlated with the labels, and therefore the algorithm was closer to the ground truth without normalization. The optimal clustering algorithm configuration for the banknote dataset would therefore be the same as the adult data set.

The cluster origins differed most consistently across the skewness and kurtosis attributes, signifying that in an unlabeled setting the dataset’s most defining features are skewness and kurtosis, both when normalized and when not normalized. The plots below (k-means clusters on the left, ground truth on the right) show the difference between the clusters when considering variance, skewness, and kurtosis, omitting entropy, the attribute consistently with the least difference between cluster centers.



*Banknote Data K-Means (left is output, right is ground truth)*

# Expectation Maximization without Dimensionality Reduction

The scikit learn GaussianMixture package was used to run expectation maximization on both datasets, where the difference against the ground truth as well as the log-likelihood, the value EM tries to optimize, were considered when evaluating the configurations of the algorithm.

*Banknote Dataset:*

The table to the right shows the effect of the maximal iterations and the number of iterations – as can be seen, these two hyperparameters have no real effect on the performance of the EM algorithm for the banknote data set.

The table below shows that normalization works best, attaining far higher log-likelihoods, while spherical is the best covariance parameter (most closely resembling the ground truth), which is consistent with the findings of assignment one, where radial outliers are typically the forged banknotes.

As was with the k-means algorithm, the greatest difference between the two cluster centers lied in the skewness and kurtosis attributes consistently across all configurations. K-means and EM were similar in terms of clustering, both clustering in ways that were logical, being unimpacted by manipulation of maximal iterations and number of restarts, and finding the same attributes important to result in clusters that were similarly different from the ground truth for the banknote dataset.

*Adult Data Set:*

 Across the several dozen configurations tested of Expectation Maximization on the adult data set, several different “error” values were produced, all of them very similar to the ones seen with the K-means algorithm. The lowest “error”, .238, was produced alongside the highest log\_likelihood, 59.72. As was the case with the banknote data set for EM, and both datasets for k-means, the number of initializations and the maximal iterations held no effect on the output. This optimal result was produced with normalization, as was with k-means, and with the full covariance parameter. Under the full covariance parameter each component has its own general covariance matrix. As can be seen from the table on the right, the covariance parameters were similarly successful, except for spherical which worked terribly with the adult data set, while also being the most successful with the banknote dataset. Spherical probably works poorly with the adult dataset because the data is not radially distributed, while the banknote data is, with outliers being the forged notes.

 When the adult\_lite data set was run, similar results were achieved, with normalized data and full covariance parameters attaining the highest log\_likelihood, of 55.53, but in this case the “error”, is significantly higher, meaning that although this data is logically separated in this case, it does not match as well with the ground truth. The regular data set emphasized the capital-gain, capital-loss, and hours per week attributes more, whereas the converted data set emphasized the workclass binary attributes greatly in addition to the earlier three.

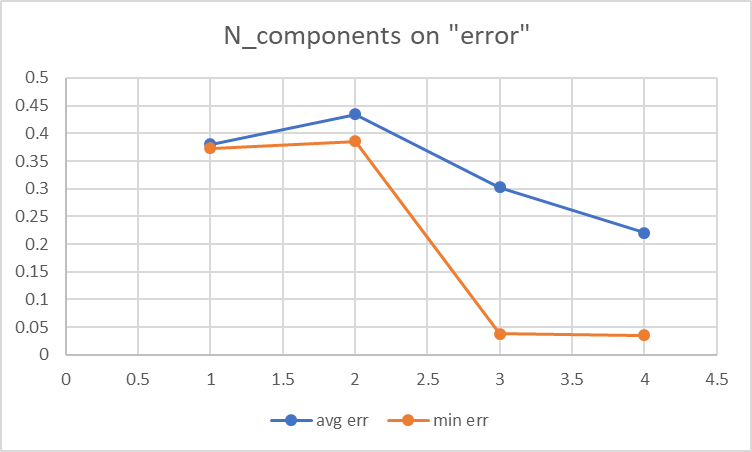
As it was in the k-means output, the same attributes, currently married, white-collar, and white, had no difference on the cluster centers because they were made redundant by their counterpart binary attributes. My hypothesis is that the adult\_lite data set clusters will be able to very closely mirror the ground truth after dimensionality reduction. Without dimensionality reduction, the adult data set is achieving higher log\_likelihood values and more successfully mirroring the ground truth because of the greater dimensionality of the adult\_lite dataset – the adult data set is essentially currently eliminating the categorical attributes as evidenced by the lack of difference in cluster centers for those attributes.

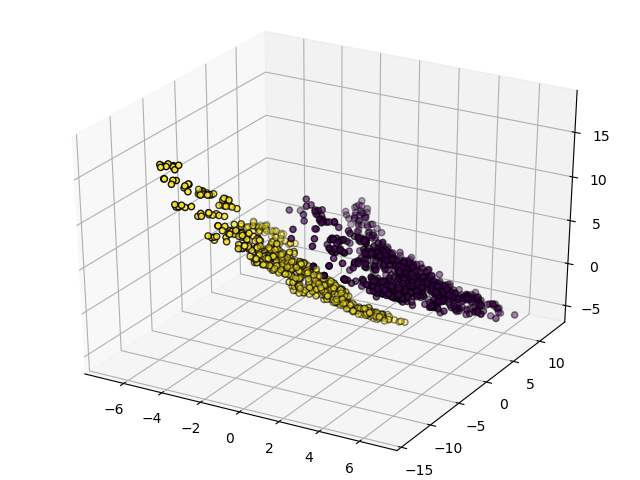
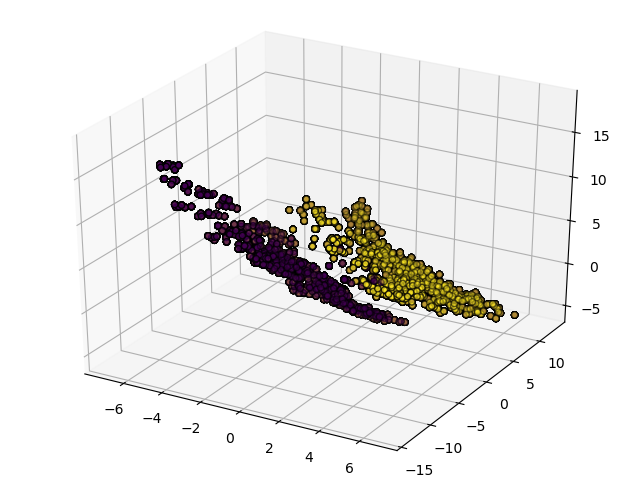
# PCA and Clustering

The PCA algorithm was run on both data sets, with the number of components to use being manipulated from 1 to the number of attributes in the dataset. When maximal components are being used, the full distribution of the eigenvalues is clear – for the sake of clarity, the method called expressed the distribution of eigenvectors as percentage values.

*Banknote Data Set:*

The distribution of the eigenvalue variance percentages goes 76.1%, 14.2%, 6.7%, and 2.9% across the four components. The first component has the greatest magnitude of change for the skewness attribute, signifying that skewness is the most variant attribute. The next component relied most on the variance attribute, followed by the third component relying most on the kurtosis attribute, and the last component relying on the entropy attribute the most.

As can be hypothesized from the distribution, the first three components carry the most weight, and therefore the addition of the fourth component is marginal, as seen in the graph to the right. It is also clear from the graph to the right, (the result of k-means) that the clustering algorithm was able to attain similarity to the ground truth far greater than before without the dimensionality reduction. Whitening was necessary to achieve these results, but given use of all four components, normalization becomes arbitrary. The images below on the left shows the result of the optimal k-means algorithm after pca dimensionality reduction and the image on the right shows the ground truth, displayed across the three attributes with the greatest variance.



The EM algorithm performed almost identically to itself without pca dimensionality reduction, attaining similar log\_likelihood and “error”, between .3 and .4.

*Adult Data Set:*

The distribution of the components goes 32.9%, 24.3%, 21.3%, 10.6%, 9.1%, 1.75%, followed by values to the power of -34 and smaller. After the sixth component, the values truly fall of in terms of importance. The biggest pull on the first component, and therefore essentially the most vital attribute, is final sampling weight, which actually has no correlation with the ground truth.

The adult\_lite data set was run as well to see how it would compare. The adult\_lite dataset has the final sampling weight attribute removed and stands to benefit far more from dimensionality reduction. The distribution of the components gives the first component 99.3%, and of course every latter component holds very small eigenvalues. The first component relied on capital gain the most, followed by capital loss, and then hours worked per week.

Both data sets still clustered very similarly, both performing their best when whitening their data, for greatest similarity with “errors” of .22, which is lower than either data set had been able to achieve without dimensionality reduction applied.

The dimensionality reduction had limited effect on the EM algorithm, producing very similar results to earlier without the pca reduction, with error rates between .25 and .4 as well. For both datasets, pca had limited benefit for the EM algorithm.

# ICA and Clustering

*Banknote Data Set:*

With K-means, the number of components to use was iterated from 1 to 4 in combination with altering use of the parallel and deflationary algorithm to yield the exact same clusters every time, for an “error” value of .42. This is the same clustering configuration seen with both EM and k-means without dimensionality reduction. While PCA was able to hugely impact the clustering, ICA did not achieve the same results. In terms of the clusters, there was no difference between the two algorithms, but in terms of change in kurtosis, the parallel algorithm was more successful.

Meanwhile, ICA in conjunction with EM was far more successful in significantly impacting the clusters. EM had not been able to get more similar to the ground truth than .38 in “error” with PCA and without dimensionality reduction, but with EM reached the same levels of similarity KMeans reached with PCA, .03. The error did not change between using three and four components, showing the redundancy of the fourth component for the clustering. The log likelihood was also greatest when the “error” was the lowest, which shows that the EM algorithm considers the clusters that line up with the ground truth to be the best, and is able to achieve those with data reduced by ICA and whitening post-transformation. The most effective models increased the kurtosis significantly, while the lesser effective models decreased the kurtosis of the data set.

*Adult Data Set:*

ICA seemed to have held no effect on the adult data set when used in conjunction with K-means, as was with the banknote data. The same clustering configurations and error values came up as when k-means was run without dimensionality reduction. Even with one component being used, the algorithm worked as successfully as a dozen. The PCA algorithm on the adult\_lite dataset had a single component reaching 99% variance, and so it makes sense that the first ICA component here could be so valuable as it is intrinsically the same information in both datasets.

While EM worked well in conjunction with ICA for the banknote data set, EM was not greatly successful either for the adult data set. EM worked well with the banknote data set when the transformed data was whitened, but the scikit ICA method would not work successfully with whitened data, probably due to the converted nominal attributes, for both the adult data set and the adult\_lite data set.

# Randomized Projections

*Banknote:*

Randomized projections in conjunction with k-means had more success with the banknote dataset than ICA had, able to bring the error down. Because the RP algorithm is not deterministic, I ran the model multiple times with each configuration. Normalizing the data worked better as has been consistently true with the banknote data set. Every iteration of the algorithm with three components and normalization worked exceptionally well, but not as quite as EM and ICA. Meanwhile the EM algorithm combined with RP worked pretty poorly, creating the same clusters as when EM was run without dimensionality reduction.

*Adult Data Set:*

No configuration performed consistently better with the RP transformation than without for k-means clustering on the adult data set. Normalizing got the clusters to resemble the ground truth marginally more on average, but greatly increased the variance of the output clusters. As was with ICA, a single component performed as well as two or more components. The same error values occurred for the adult\_lite data set however, which means the randomized projections transformation was successful in mitigating the difficulties of nominal attributes. EM error values ranged between .25 and .4, the same range as the ICA-EM combination, for the adult data set, doing no better than k-means. It did yield exceptionally consistent results however, although not deterministic, and randomized projections took less computational time than the other feature transformation methods.

# Variance Threshold

*Banknote:*

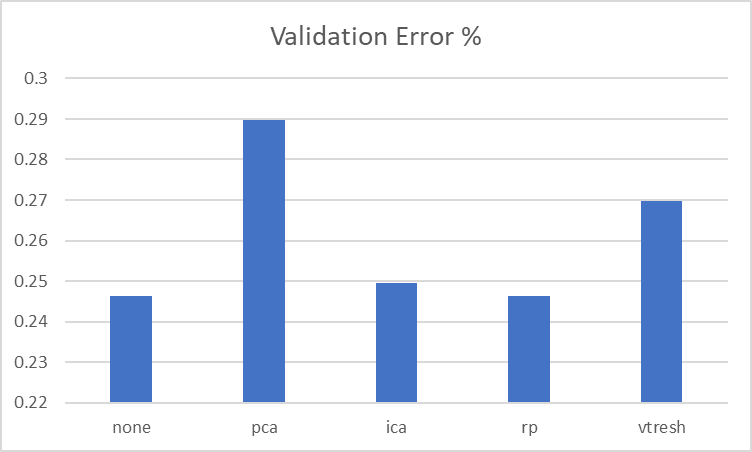
The feature selection method used is exceptionally simple – the algorithm holds on to the features with a variance above a certain threshold. This is a fairly similar method to PCA, so the expectation is similar results. The banknote data set normalized had variances between .034 and .048, so variance thresholds of 0, .035, .04, and .045 were tested. The global minima occurred with two features in play, decreasing to one brought error back up.

With EM, the largest log likelihoods occurred with the 0 threshold which can be expected because all features are in play, as that is an intrinsic bias of the EM algorithm. Also predictably, the error was minimized with the .04 threshold just as with k-means, going down to .13 as did k-means. All covariance measures worked similarly for EM – manipulation of the threshold held the biggest effect on the clusters. The results did not mirror PCA quite as much as expected; PCA attained clusters with far greater similarity to the ground truth.

*Adult Data Set:*

With k-means, error shot up from .25 to .38 once the variance threshold came up from 0 to nonzero values. EM’s results were similar to that off the banknote data set – the clusters were the same as seen previously and the log likelihood was greatest without any feature elimination, however error was smallest for the adult data set without any feature elimination as well. Error values ranged from .25 to .4, being the smallest without feature elimination signifying that feature selection is entirely unsuccessful with the adult data set.

# Dimensionality Reduction and Neural Nets

 The adult data set was chosen to run neural nets on alongside dimensionality reduction algorithms because optimal neural nets were already able to attain zero validation error on the banknote data set, but not the adult data set where improvement could be used. Without dimensionality reduction, the optimal neural net configuration achieved 24.62% validation error and a testing error of 21.8%. Optimal models of the PCA, ICA, RP, and Variance Threshold algorithms were all used to transform the data.

From the chart on the right, it is clear that PCA reduction did the poorest with the adult data, at 28.9% while randomized projections and ICA did comparably well to when the model was run without reduction.

The PCA algorithm was run with 6 components, because after 6 the eigenvalue percentages drop off really sharply as discussed earlier. Whitening was also used as that was more successful with the adult data earlier. However, the PCA reduction hurt the model because it would keep the attributes with high variance, such as final sampling weight, while removing other attributes with low variance that held high correlation. With the adult\_lite data, which has the final sampling weight feature removed, PCA still worsened the result but was not the worst reduction algorithm.

ICA performed well compared to the other algorithms with 3 components in use. With the adult data set, ICA is a fundamentally better approach then PCA – instead of maximizing variance, ICA maximizes mutually exclusive information. The adult data set has important data with low variance and unimportant data with high variance. This is also why the variance threshold algorithm did poorly but not as poorly as PCA. PCA gives importance to the more variant features, while VT simply keeps them, but VT still removes the lesser variant features.

Of all the reduction algorithms, RP worked the best. This is because randomized projection does not attempt to transform or filter the information, but simply express the same information in fewer columns – in this case, 1. This is a fairly similar approach to that of ICA.

However, this backfired in some ways because the computational time increased by 50% as can be seen in the chart on the right. When considering computational time, clearly ICA was the best dimensionality reduction algorithm for the adult data set – keeping very similar error rates but halving the computational time. The others do not contend for being viable options to be considered with this data set.

# Dimensionality Reduction, Clustering, and Neural Nets

After running neural nets on the adult data set in conjunction with dimensionality reduction, I ran neural nets in conjunction with clustering after reduction – concatenating the cluster value with the rest of the input data. My initial expectation is that this method will work phenomenally with the banknote data set, where clusters were as close as 3% different from the labels, and so the neural nets could take far fewer iterations and save significant computational time. But the clustering algorithms alone most frequently got as low as 25% different from the ground truth, and rarely dipped below 20%.

 From the table on the right is clear that the best configuration involves no dimensionality reduction. Using EM and K-Means was more effective than without clustering, and for not too much greater computational time. Given the findings above with the other dimensionality reduction algorithms, it is to be expected that the validation error was worsened with the reductions. In terms of time, PCA and ICA were better than no reduction, while RP and VT were obviously thrice as expensive. Of all the ways tested thus far, the best classification approach in terms of time and error would be k-means clustering before neural net classification. The optimal ‘k-means’ clustering algorithm included 10 initializations and 200 maximal iterations.

# Conclusion

When combining unsupervised and supervised algorithms, the adult data set obviously presents lots of challenges to these dimensionality reduction algorithms, which is what makes it an interesting data set to consider in this setting. The ideal dimensionality reduction algorithm that would help classification of this data set would be a filtering algorithm using information gain against the labels (similar to decision trees), which would of course be a supervised method rather than unsupervised. Another efficient transformation of the data would be turning the nominal data into logically continuous data by replacing each value with the average of the label for that category – essentially ranking each nominal value. The banknote data set would have worked very well with both reduction and clustering, but the adult data set is simply more interesting due to its difficulties.

The most unexpected finding for me was how certain combinations of reduction and clustering algorithms were far more successful than other similar combinations. PCA and K-Means worked incredibly well in combination with each other, as did ICA and EM. My understanding is that PCA, in maximizing variance, would have produced data was more easily clustered with certainty while ICA would allow for reduction that works better probabilistically.