Jeffrey Minowa
CS 4641
Unsupervised Learning and Dimension Reduction

Datasets

The two datasets I am using for this write up are the Titanic dataset from Kaggle and the Adult dataset. They are interesting to contrast because of their differences in their data. The Titanic dataset hosts a smaller number of instances (~900) with all discrete values and small number of attributes. The adult dataset hosts a larger number of instances (~32000), has a mix of discrete and continuous attributes, and certain values have unknown categories.

Clustering

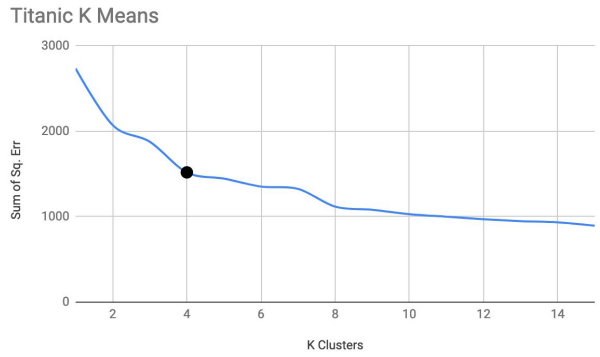
The two algorithms used to cluster together the instances were K means and Expectation Maximization. K means is an algorithm that chooses K points and attempts to cluster instances based on distance calculator. Then it reevaluates based on the mean values in each cluster. Expectation Maximization does something similar to K means in that it clusters values and re-evaluates based on the mean in its specific cluster, but the difference between K means and EM is that EM provides probabilities of being in each cluster, forming a Gaussian, where all instances can potentially be in any of the K clusters. This allows for values to be in a sort of grey area where the instances may belong to multiple clusters.

In regards to K means, the method chosen to identify the best K clusters for K means was to look for the elbow in the graph trend. By looking at the sum of squared values per K clusters, one can visualize the point where the error dips while minimizing overfitting. Minimizing the number of clusters also allows for greater generalization for future values to be used for categorization. Euclidean distance was used to determine the sum of squared error value because centroid is implicitly calculated by euclidean distance. As for expectation maximization, the leveling of the line was used to choose the smallest K that had a maximized slope. This was chosen because once the slope leveled off, the log likelihood did not changed much afterwards. And as mentioned before, we want to minimize the number of clusters to keep generalization while getting the maximal information possible.

K means

Titanic:

The best K for Titanic was found to be 4. It had the smallest error in the elbow while not overfitting.



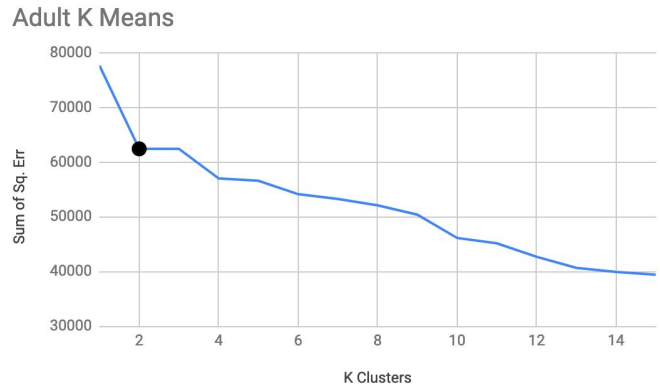
The four clusters hint to the fact that the ship was prevalently male, middle aged, and were mostly alone. From the figure below, it can be seen that there were some general trends between surviving the titanic and the generalizations made from the clusters. This leads me to believe that the people in group 2 were less privileged and were not seen as valuable to the decision makers on the titanic.

```
Class attribute:  Survived
Classes to Clusters:

  0   1   2   3  <-- assigned to cluster
202 121 179  47 | 0
103  91  48 100 | 1
```

Adult:

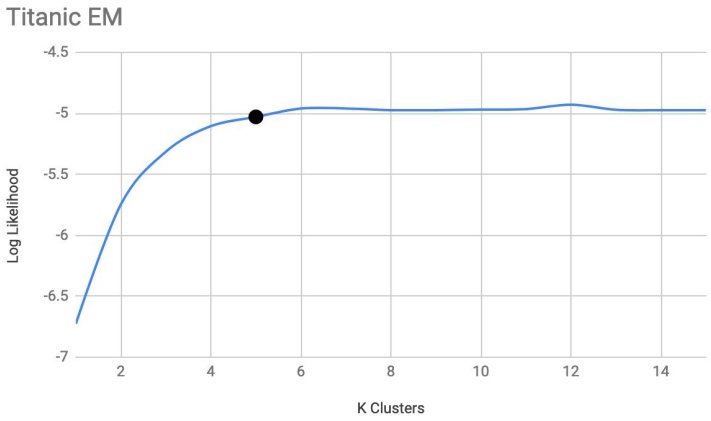
The best K for Adult was also found to be 2. This had the largest change in slope. The result of the two clusters are intuitive as can be seen from the centroid clusters. The clusters considered can be generalized as married, white, educated males in cluster 0, and single, educated, white females in cluster 1. These two stand out because the survey used in the dataset was taken in the 1980s and was heavily biased toward surveying white males. It can also be seen as a distribution in power and money given the wage gap amongst the two categories that the algorithm decided to split instances into.



Expectation Maximization

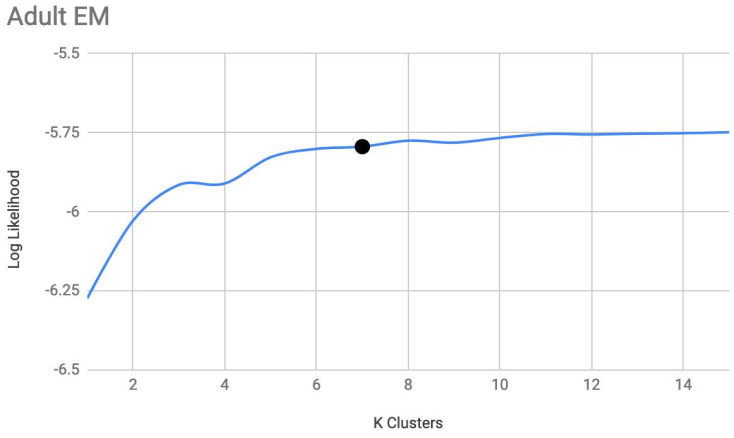
Titanic:

K = 5 was chosen for its largest log likelihood while still being a part of sloping.



Adult:

Because Expectation Maximization finds probabilities of attribute belonging to a cluster, there can be grey areas between what should be in what cluster. Because adult has continuous data, specifically age, I believe EM struggled to find definite clusters to put instances into. K = 7 was the number of clusters chosen from the visual smoothing of the graph.



In general the clusters make sense, they are intuitive in the way that they are categorized. The adult clusters in particular are intuitive because of the knowledge that there was historical significance in the bias of the dataset. The clusters did not line up with the labels for both datasets. I believe this is the result of the high dependence between each attribute which makes it hard for clusters to know what instance to put into each cluster.

Ideally, the two clustering algorithms will find the same number of clusters if they both find the optimal clusters. This is an important observation because from project 2, the Titanic dataset had few local optima, based on the results of simulated annealing while the adult dataset had many local optima. Because the cluster algorithms are trying to maximize the distances between each other, they will fall into local optima as well. This is relevant because the number of clusters for K means and EM are very close for Titanic, and adult dataset has very different K values for its clusters for K means and EM. Therefore, I believe the clustering algorithms for Titanic have found the global optima while the algorithms for adult have found only local optima. Therefore, clustering does poorly when there are multiple local optima for the problem while performing much better for those with fewer local optima. To improve performance of the algorithms, maximal euclidean distances can be used as starting points to provide the most room for clusters to grow as well as random restarts or a sort of simulated annealing to avoid falling into local optima.

Principal Component Analysis

PCA tries to reduce the dimensionality by changing the frame of reference and maximizing the variances between the dimensions by finding the orthogonal eigenvectors. This is similar to how knn tries to maximize the distances between its margins.

PCA as a whole will attempt to help clustering along because it identifies maximal disparaging features amongst the components that it finds which would inherently help clustering algorithms decide what to make as clusters. Another important note is that PCA weights components differently and gives preference towards those with higher variance as those are the components with more isolation.

Titanic:

The best K value for K means for variance of .75 was 6 and the best K value for EM was 3. As before, a range of K cluster values were used to find the optimal K clusters, and as well as a range of variances (.75, .95, and 1). The variances shown were deemed the best the charts for providing insight into the finding the elbow and the smoothest curve for EM. The data was reconstructed and changed from 7 attributes to 11 attributes. This is likely due to the dataset being non binary and the attributes increase because there are more hidden variables generated from the PCA.

Based on the eigenvalue distribution, one of the components created by the analysis had a large variance value which provided a good descriptor and a lot of smaller variances which were not as good in finding commonalities as a component.

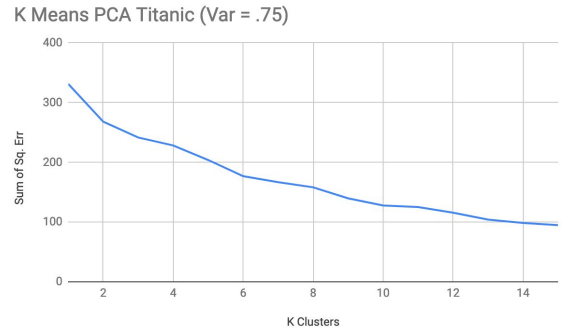
The K values were similar each other in this algorithm but were not the same as the pre PCA datasets. The data looks very different and the smoothness of the graph changed dramatically to the point where the K means had no identifiable elbow, and the EM had many peaks to choose from.

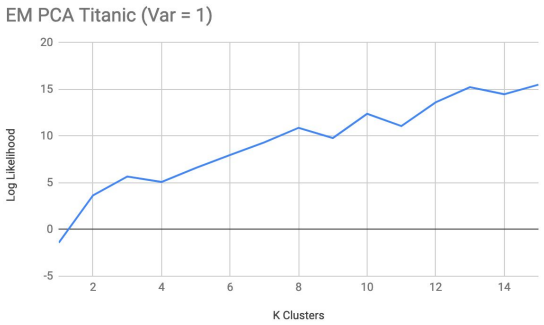
Eigenvalues

4.15872	2.61955	1.95223	1.60973	1.45577	1.29384	1.12922	1.11478	1.06826	1.02708	1.01272

K means: K value = 2

EM : K value = 3





Adult:

The adult dataset changed from 8 to 13. Like in the Titanic example, I believe the number of components is greater than the number of attributes because of the fact that the values are non binary. In regards to the eigenvalues, they are smaller than the Titanic eigenvalues. This means their variances on the first principal component do not have much spacing; therefore, the components is not optimal for this dataset as the attributes could not be well separated into different identifying similarities.

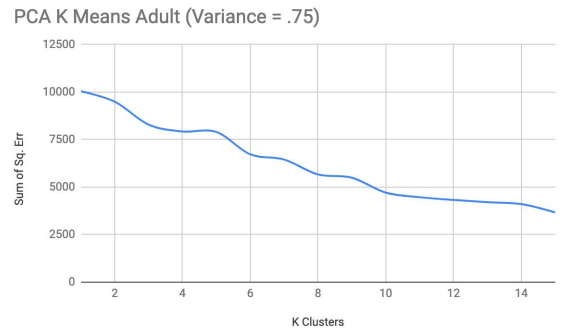
The K values found for the clustering algorithms after PCA were chosen based on the previous methods. Though the K values were the same for both algorithms, there were multiple hills to choose from, and the optima was a blurry choice.

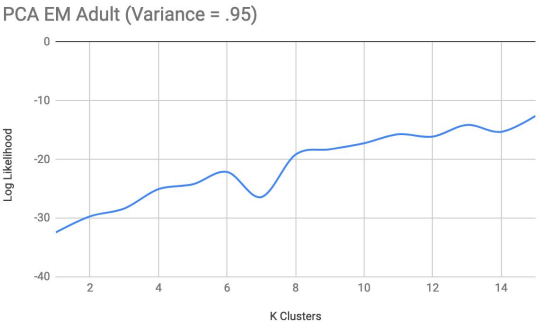
Eigenvalues

2.614	2.504	2.032	1.981	1.67	1.504	1.435	1.351	1.25	1.187	1.118	1.1	1.092	1.073	1.057

K means: K value = 3

EM : K value = 3





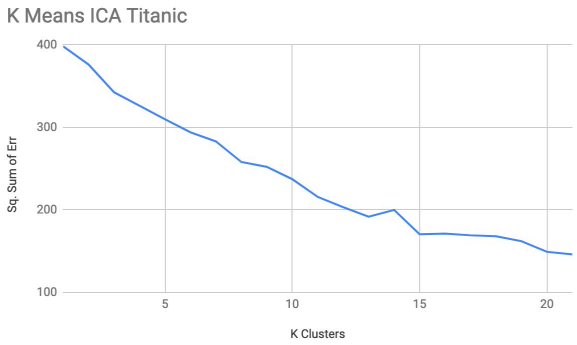
ICA

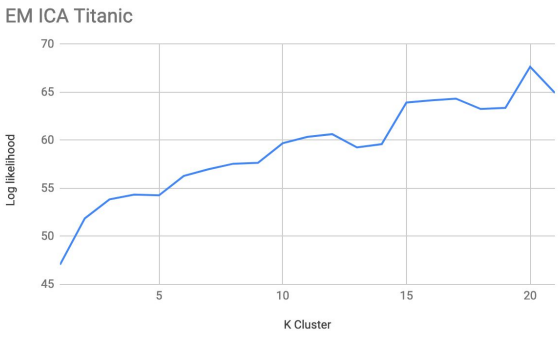
Independent component analysis has a goal to linearly separate independent components taken from several sources of information. Put simply, the idea behind ICA is that is takes in two

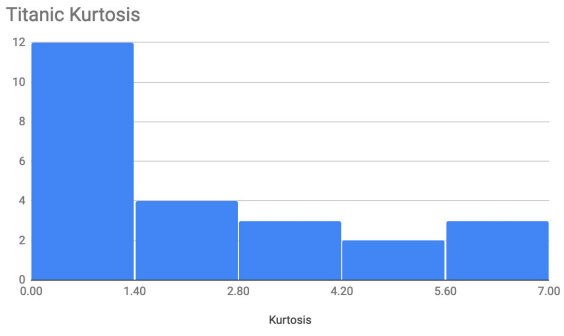
signals, and attempts to find trends that split them into components by finding maximal non-gaussianity.
Like PCA, ICA will ideally help clustering algorithms because it finds components that are independent of each other which is essentially the goal of clustering algorithms.
The choice of K values were consistent for all clustering algorithms.
An important note is that ICA weighs all inputs equally, which contrasts with the PCA counterpart. Consequently, this allows for noise is be handled poorly.

Titanic:

Titanic's distributions were distributed in a way that most of the kurtosis were below 3 or below gaussian. This is a good split because it shows that the values can be split up independently.
Most of the kurtosis values are non-gaussian, providing proof that the components do not rely on each other.
The clustering algorithms have varying values with a smoothness for the K means and many peaks for the EM. The K values are different from each other, alluding to the fact that they may be caught in local optima due to the equal weighing of attributes as mentioned before.
K means: K value = 7
EM : K value = 2



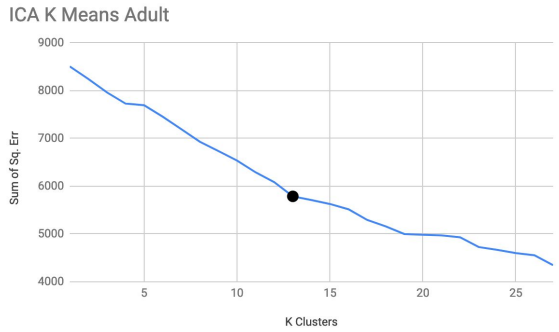


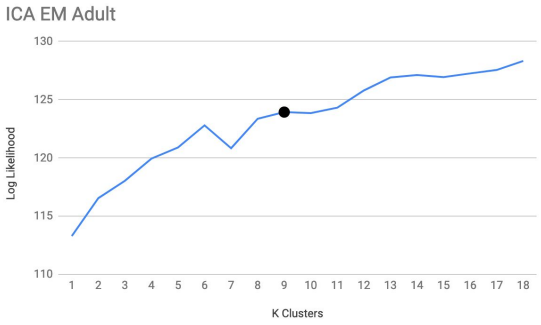


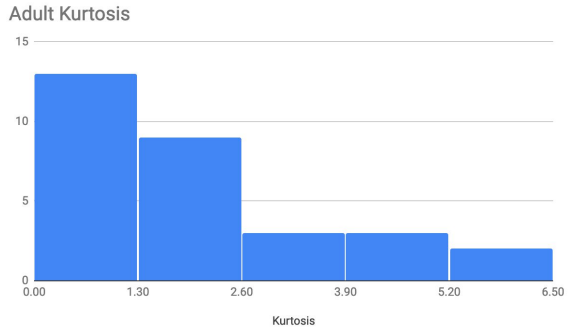
Adult:

The adult dataset was also pretty well split up. Many of the kurtosis were again below 3, which shows that the independent components were non-gaussian.

Similar to Titanic, the k values are different which may mean that the values are trapped in local optima as well.
K means: K value = 13
EM : K value = 9







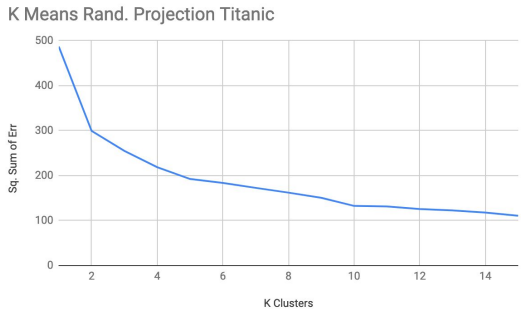
Random Projections

As the name suggests, random projections makes a random matrix that is transformed and projected to estimate different components. If run multiple times, it can do very well because it does not fall into the trap of overfitting or follow specific trends. In the experiments that were ran, random projections were run 5 times. Between averaging and choosing the best trend, the best trend was chosen for both Titanic and Adult. As for variation amongst the samples, if the random seed is changed, the random projection for k means sum of sq. errors can range within 100 for Titanic and within 1000 for the adult dataset. For the EM algorithm, the values vary by margins of Log likelihood of 2 for both Titanic and Adult. SSE difference makes sense because the larger the dataset, the more error the algorithm will have.
Surprisingly, both K values for both datasets are very similar, meaning they may have found global optima. Random projections may have done very well because it was given the chance to redo from the restarts given to it, having a bias towards its best charts.

Titanic:

K means: K value = 2

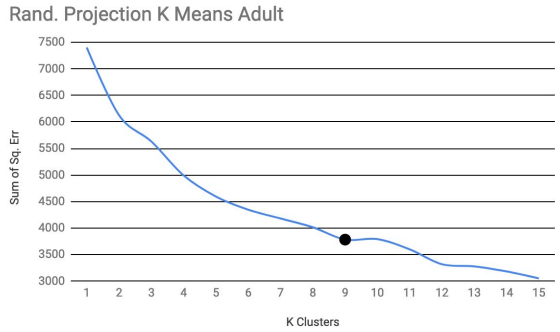
EM : K value = 2

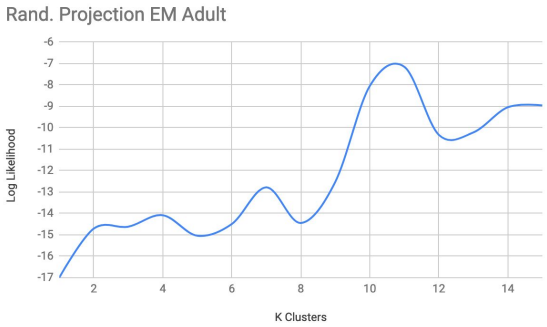


Adult:

K means: K value = 9

EM: K value = 10





Information Gain

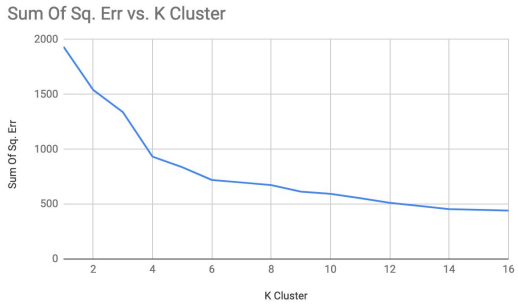
Information gain is similar to the idea used in decision trees in that it shows the provides the maximum information obtained per attribute. This can then be used as a feature engineering system. Information values were given after running this on the original test data and then the attributes that were equal to or below the lower quartile were removed from the dataset and then rerun.

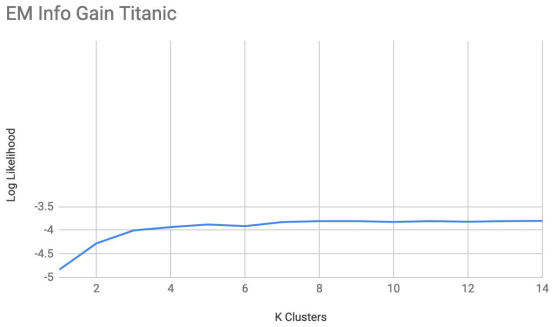
Information gain also had similar K values to each other for the algorithms for both datasets. The Titanic K values for information gain were also similar to the ones from the original K values. This makes sense because attributes are being taken away instead of changing their classifications to become new components.

Titanic:

K means: K value = 4

EM : K value = 3

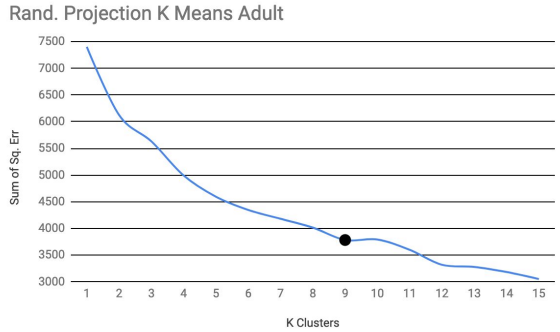


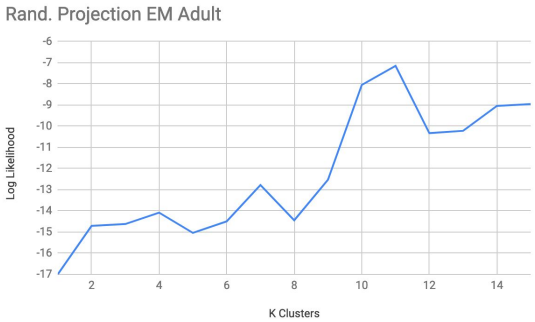


Adult:

K means: K value = 9

EM : K value = 10





Neural Net Evaluation

The performances were similar when comparing pre clustering and post clustering when looking at all the attributes. When comparing only clusters and the results based on those who survived, all of the cluster algorithms did worse than their counterparts.

In terms of performance, there was no difference in the speed at which the neural net ran, which is unsurprising given that the dataset is small.

Before Clustering for Neural Net Results

PCA	ICA	Rand Projection	Info Gain	Normal
716	715	703	715	717

After Clustering for Neural Net Results

	PCA	ICA	Rand Projection	Info Gain
EM cluster only	623	643	614	606
K Means cluster only	529	640	562	607
EM all	711	717	706	715
K means all	710	714	708	714