David Li

Machine Learning, C. Isbell

March 23, 2019

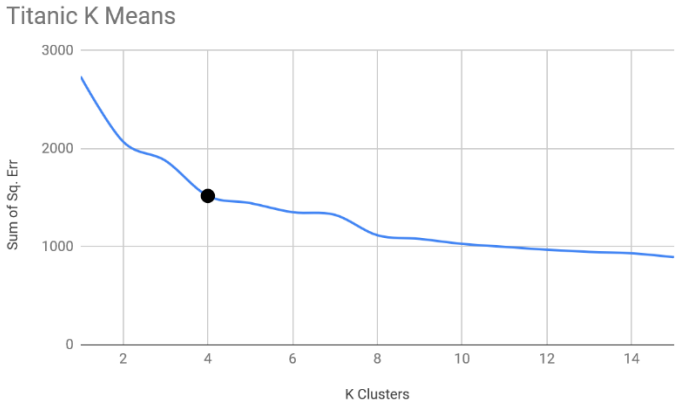
Unsupervised Learning & Dimensionality Reduction

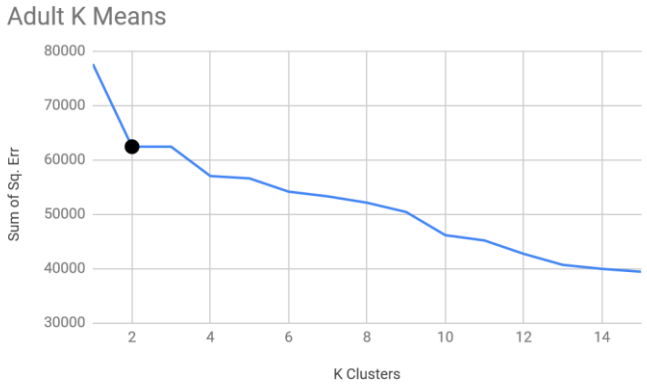
Overview: In this project, we shall look at two unsupervised learning algorithms and four dimensionality reduction algorithms and their effects on our past two datasets: Titanic survivors and Adult Income. We will compare their results ran on a neural from the first assignment with no modification against the data after their dimensions have been reduced, and lastly against data that is reduced and clustered. This will allow us to analyze the strengths and weaknesses of each algorithm in relation to the types of data we give it and the different combinations of preprocessing we subject it through.

Data: The Titanic dataset has 891 instances and 7 attributes while the Adult dataset has 32561 instances and 7 attributes after discretization and preprocessing from assignment 1. This will be interesting to analyze how extreme grouping and discretization affects clustering, as some values become nominal and loses its distance value in dimensions or groups certain clusters that were close separately. And we can see how reduction affects a data set like adult, where each of the attributes are very dependent upon each other.

Unsupervised Learning

First, we ran iterations of K-means and Expectation Maximization clustering with varying number of clusters, K.





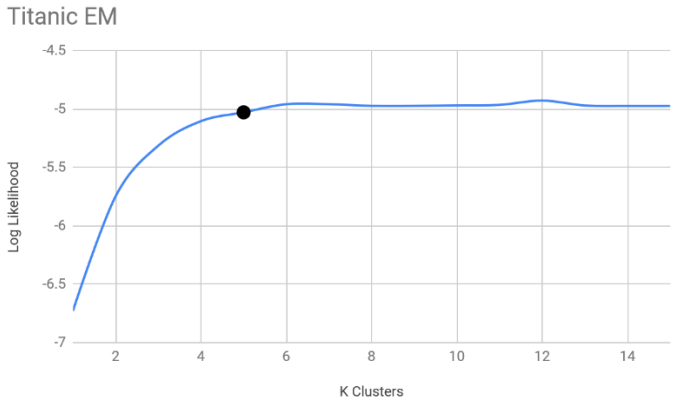
For K means, we looked at the sum of squared error as we increased the clusters, and picked the biggest decrease in first derivative, or the biggest elbow, of the graph. Since SSE is a measure of distance of points from the center of clusters, the smaller the SSE the more accurately placed towards the centers are for each cluster. After a large decrease in rate of decrease, the amount of rate decrease per increase in K is diminished. We want to catch this as early as possible, because we can have virtually 0 SSE as we approach a K that is close to how many instances there are, which would make the clustering pointless and represent each individual point, and we can find larger trends with larger and fewer clusters. In a sense, if we use too many clusters, the clusters overfit the individual points as opposed to capturing an entire trend. For Titanic we picked 4 clusters, as the elbow fit it quite well.

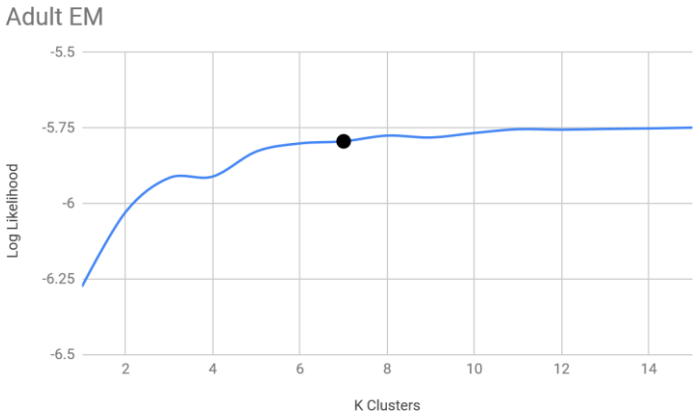
For adults, the elbow appears at 2 clusters, not because of the change in SSE, but rather the lack of, as going from 2 to 3 clusters, the algorithm could not find a separable third cluster and thus removed its attempt at a third cluster altogether. This means that the data strongly biases towards two groups of people.

Final cluster centroids:

Attribute	Full Data (32561.0)	Cluster# 0 (23277.0)	1 (9284.0)
ï»¿age	'(43.5-54.5]'	'(43.5-54.5]'	'(-inf-21.5]'
workclass	Private	Private	Private
education	College	College	College
marital status	Married	Married	Never-married
race	White	White	White
sex	Male	Male	Female
Native Region	US and Canada	US and Canada	US and Canada

Based on the results of the cluster, and the time period and location this survey was taken, which was in the U.S., clusters were heavily focused on those who are from North America, white, worked in private class and went to college, and distinguished between middle aged male and younger females. This instance of response bias made it difficult to find a third group during the hyperparameter testing.



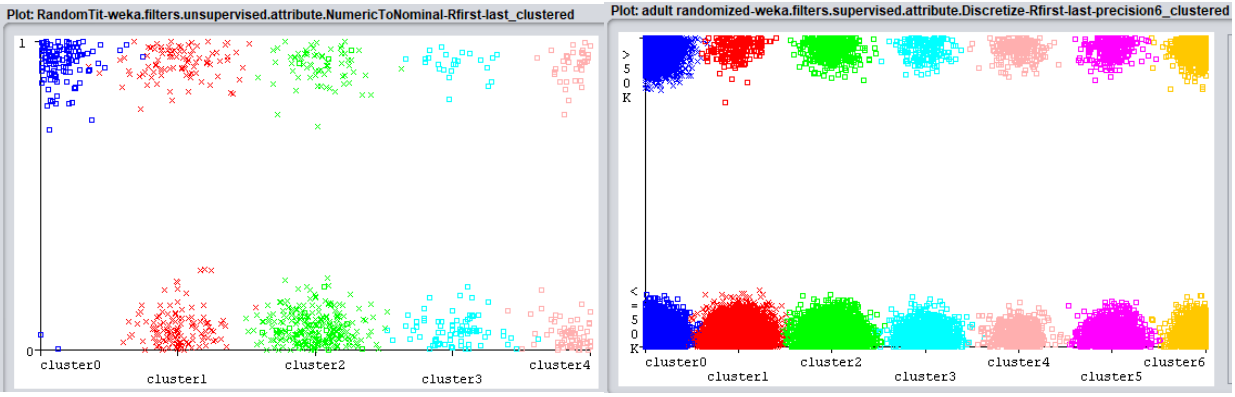


For Expectation Maximization, a likelihood of a cluster is calculated based on the summation of the probability that each point exists within its given cluster based on distance, and thus a higher log likelihood is preferred. But as mentioned before, to prevent overfitting and maintain some trend and generalization capture, a lower K is preferred, so we want to find the first value where the likelihood tapers off, which is also where returns start to diminish in this logarithmic scale. This gives us 5 clusters for titanic and 7 for adults.

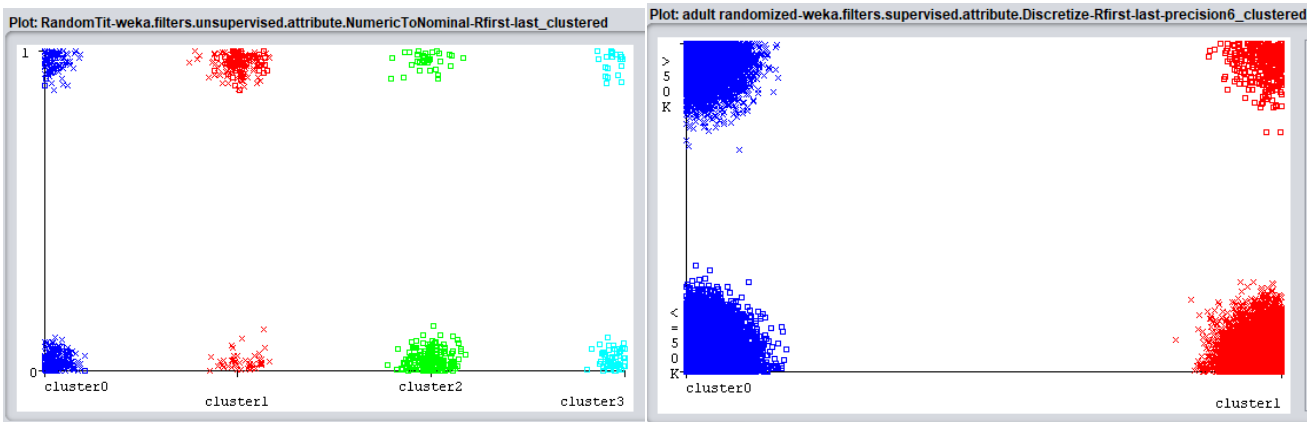
One key point is why does EM and K means give us different number of optimal clusters? Two things come to mind, one is that K means forces points that are close to each cluster into said cluster, while EM might take a proportion of the likelihood into each cluster. This means that K means might get stuck in a local optima or extreme point when one point barely fits into one cluster but gets placed in there, while EM will take a weight proportional to how likely it would be there versus another cluster, which allows the existence of more clusters in between areas during the formation process. This is furthered because of the aggressive discretization where someone on the edge of two zones are distinguished when it shouldn't be, where two groups (1 to 5) and (5 to 10) exists and a 5 or 6 might be incorrectly placed for K means, a prevalent case when looking at age. And the missing attributes might group or bridge unrelated clusters together in general, something to keep an eye out for.

On the other hand, the reason why the two cluster for Titanic, 4 and 5, are so close while Adults 2 and 7 are so far apart might be because of local optimas. As we analyzed in the previous project through optimization algorithms, we found that Titanic was a relatively simple algorithm in terms of patterns towards the label, while adult in general had a lot of local optimas and odd clusters. This might be one during K means, adult might have gotten stuck at a lower number of clusters to repeatedly gotten variable cluster amounts despite running it so many times, while Titanic's was relatively consistent besides the one group that was separated during EM.

Looking at the K means cluster graph of cluster vs classification, both groups are very evenly distributed, and in conjunction with the relatively lower SSE without overfitting, shows that trends were found, and it correlated little with the actual results we were trying to predict, highlighting a key feature of unsupervised learning.



When looking at the EM graphs, adult gave relatively even clusters as well, separating each cluster proportionally to the overall data based on income. With Titanic, however, cluster 0 seemed to have picked up a trend matching that of survivors, with a predominately young adult, rich group with little to no family, the cluster was 96% survivors. The difference between K means and EM in this case might be due a higher cluster that allowed a separation in one of the preexisting clusters when considering given probabilities of some borderline points. But it is possible that it was just a small population that was overfitted by the algorithm.

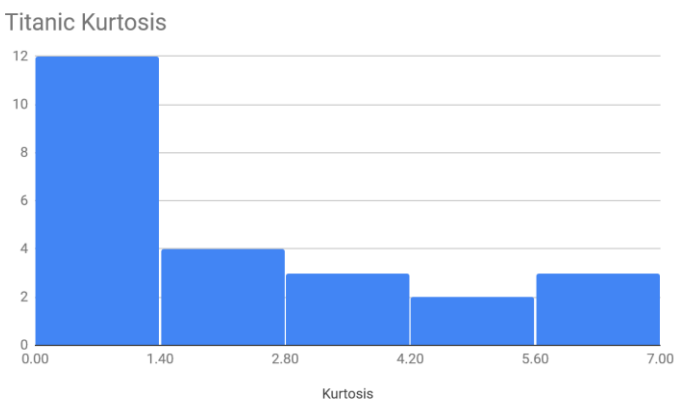


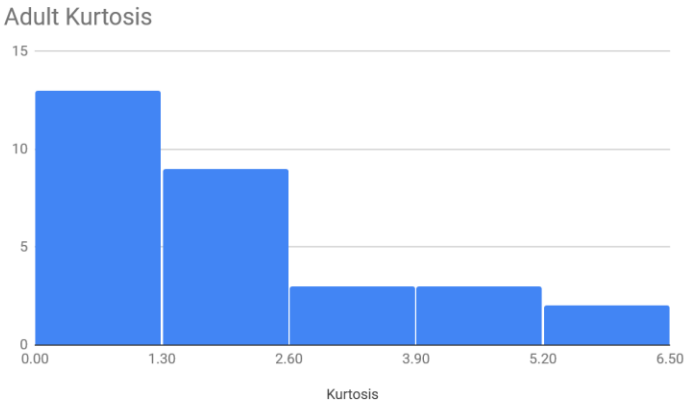
Unsupervised with Reduction

From this step, we ran four dimensionality reduction algorithms: PCA, ICA, Random Projection, and Information Gain.

ICA

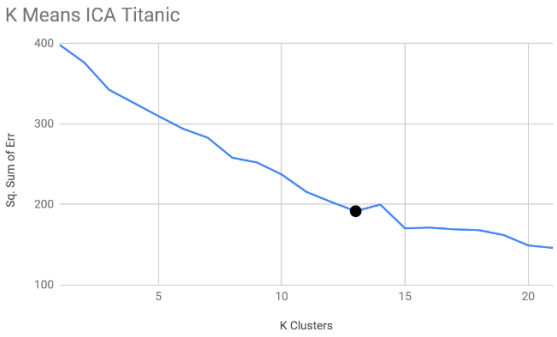
During ICA, despite it being a dimensionality reduction, both data set ended up with more attributes than started with, which unfortunately actually makes the curse of dimensionality worse. I am led to believe this is because, despite the discretization, many attributes in both data sets, especially age, are worked out to be ranges or continuous, making each section be held responsible for many sets of input. This in a sense created an artificially complicated attributes that had to hold more information and brought a lot of overlap over each range of age, ticket price, or education level in their respective data sets.

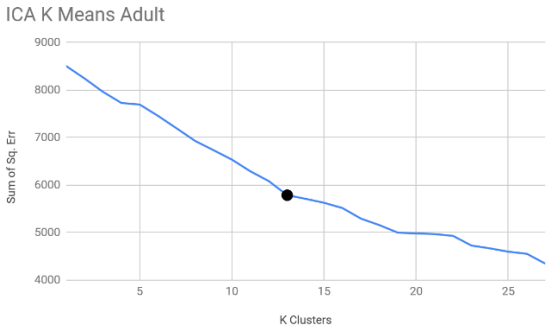


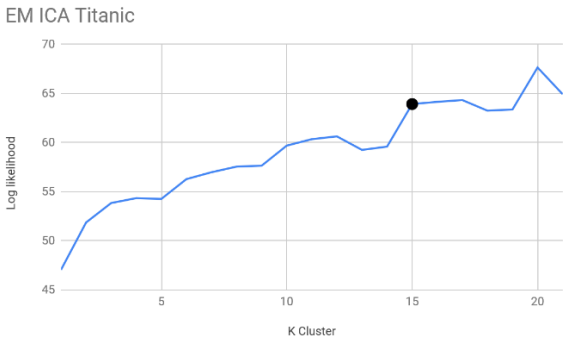


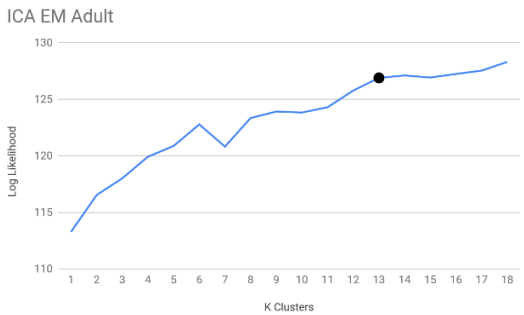
When looking at the kurtosis of attributes as a result of ICA, I made histograms to see what proportion of data had less than 3, and what proportion of data had greater than. In both data sets, the data had majority attributes less than 3, meaning they are very peaked distributions with very little spread or outliers. Looking at the data, ICA fit well for it as most of the data is not overly gaussian, as ICA tries unmix the attributes to independent no gaussian components.

I believe the low kurtosis is as a result of overfitting, as ICA saw that some attributes or linear combinations were independent, or that it was very center, and took the few points to make that combination of attributes, while it had little spread and was just a few very strong instances in the center. While they are very independent, it is not necessarily a result of the reduction, but just a forced separation of instances. As said before, it could be because of the ranges of discretization that placed too much overlap and responsibility into different ranges while creating distance between values that should have been close such as ages 20 or 22 when ranges cutoff are at 21 in adults.







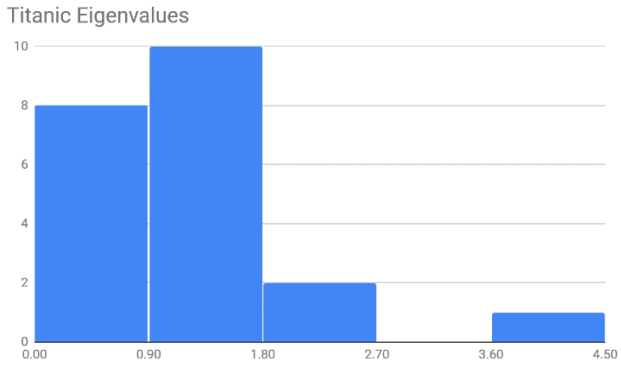


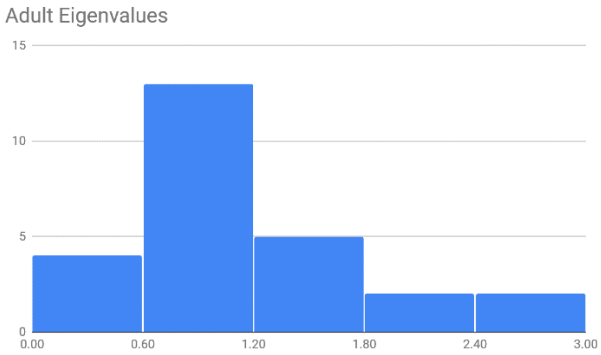
With the ICA reduction, both k means and EM have similar optima clusters, at all around 13, and EM for Titanic at 15. This could be explained by the reduction by ICA that highlighted combinations that became important features which better separated the clusters and accounted and fixed the ranges of discretization or gaps made by missing values and overlaps. The increase in cluster is not necessarily a good thing with over fitting, but the consistency between k

means and EM speaks well ICA. In the results for adult, it ended up putting 33% of the data in one cluster for EM while K means was more evenly distributed. This result maybe be a result of the local optima in the adult data set, or that EM ICA picked up on some attributes that was too broad or generalized, picking up a lot of different data points into this one cluster, while Titanic's simpler dataset allowed for separation into independent and non-gaussian components.

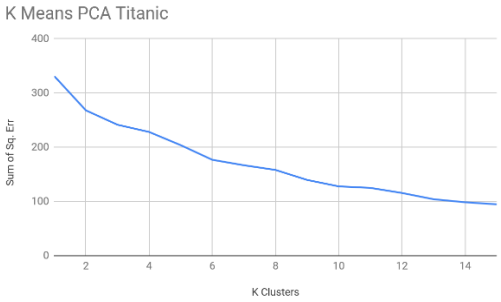
PCA

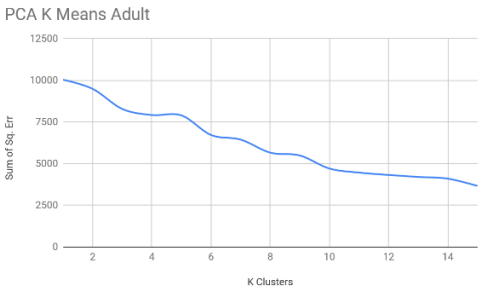
For PCA, we ran a range of cluster values at varying variance cover to determine the best hyperparameter, which is .75 for adults K means and all of Titanic, and 1 for EM for adults. Like ICA's reason for discretization ranges and missing values, PCA also made more attributes than the base attributes, at 21 for adults at 1 variance covered and 11 for Titanic at .75 variance and 21 at 1 variance covered. We can see that as we increase the variance covered, more components are needed to explain the variance and distances in data from over components, sometimes this may lead to overfitting.

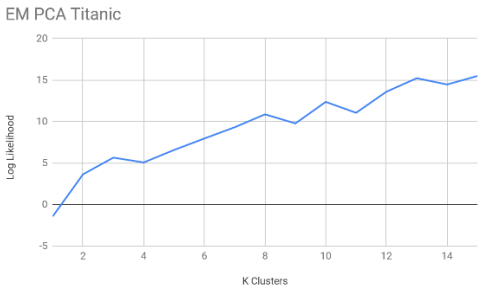


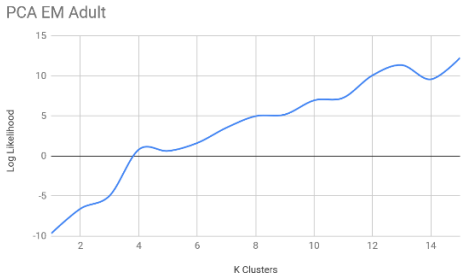


In Titanic, one of the Principal Component was at an eigen value of 4.158, contributing greatly to the algorithm, and 2 were above 2, the rest were relatively low and 2 were at 0, and were completely useless, and were not included. Less extreme were the cases in adult, where there still existed two values or 0, but the most impactful attributes were not complete outliers. This shows that in Titanic, there existed a single component that was able to maximize variance and cover many instances, and that this component can describe the data extremely well. It is possible that with a lower variance covered, the data will still be able to be generalized well enough with a lot less components created.





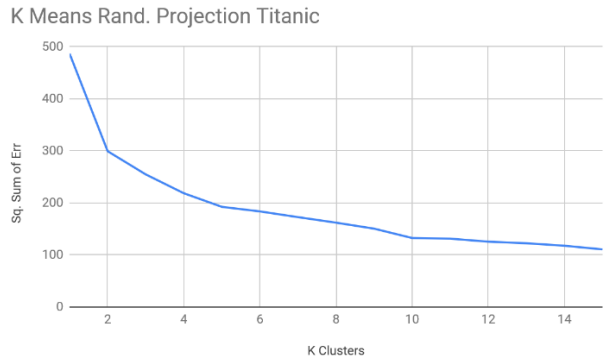


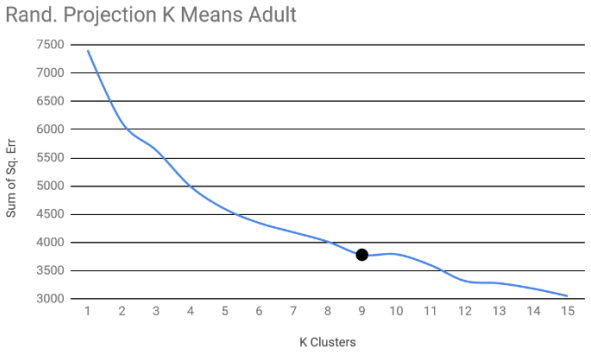


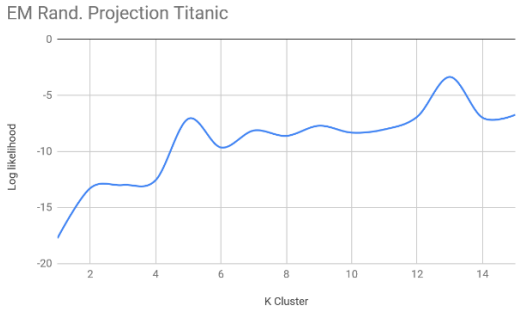
Based on the graphs, Titanic its optimal at 6 clusters for K means and 8 for EM, 4 for K means adult and 10 for EM. One key point is that none of these graphs had defining elbows or leveling off points, which indicates that PCA was not very compatible with the data. Because of the large gaps in data for adults, PCA might have a difficult time finding principal components amongst the missing values, and the new attributes drawn won't carry the variance needed along the vectors if everything is bridged together. Meanwhile Titanic's data is single peaked and may prove to be too difficult to separate into new frames of references by PCA.

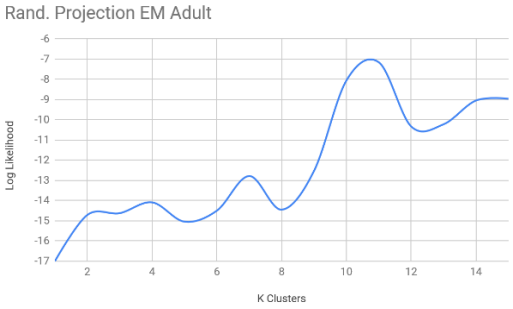
Random Projection

Random projection basically creates components randomly, and so we ran the algorithm multiple times. We are given the option of either averaging and combining the components or choosing the best one. Since the average or summation of random distributions will approach 0, we decided to use our best result for this experiment. The key concept is that despite it being completely random, it will still maintain some trends from the previous attributes and demonstrate enough information to form clusters.









The elbows in K mean for both were relatively obvious, at 2 for Titanic and 9 for adults, while the log likelihood increased at random points or spiked, leading to Titanic with 7 and Adult with 10 for EM. This can be explained that because K means does not consider the fact that some points might be on the borderline or edge, so these random components and axis drawn right through a cluster will not affect it as heavily as it would with EM, where some clusters are completely cut through. In the end the random projection generally made around the same number of attributes as given, but a few times more, which in the case of these two it gave 8 for Titanic sand 11 for Adults.

As we looked at random seeding, the variance for Titanic at the optimal SSE ranged from 299 (the one we used) to nearly 400 in K means, and adult ranged from 3781 to 4121. It becomes obvious that as the sample grows larger, as does the variance and error in clusters. For EM Titanic ranged from -7 to -3, while adult ranged from -13 to -8. There may be biases for this experiment since normally we will not be able to hand pick the lowest values, and even if we had the computer pick the minimum value, the medium might be more representative, but less optimal, leaving room for what type of data and experiment to be ran as the determining factor.

Information Gain

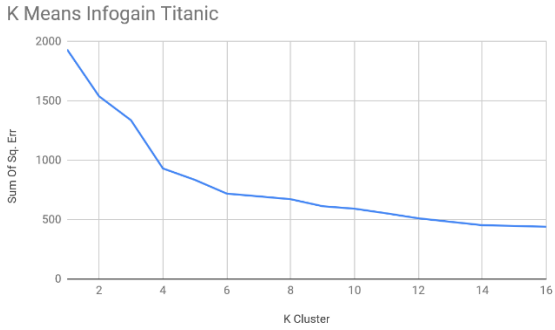
As opposed to the other three, information gain will try to find the best attributes that contribute the most amount of information, based on entropy and information theory. With this in mind, we ran the info gain attribute evaluation and removed the bottom quartile of attributes that contributed the least, based on these results.

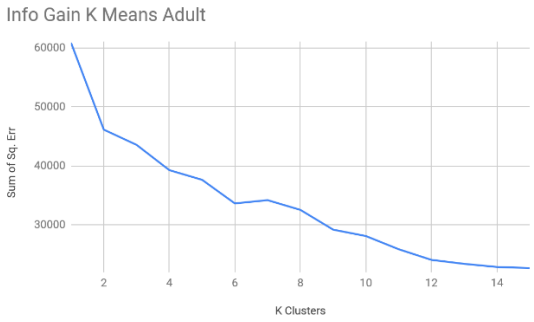
average merit	average rank	attribute
0.218 +- 0.009	1 +- 0	3 Sex
0.084 +- 0.006	2 +- 0	2 Pclass
0.071 +- 0.004	3.4 +- 0.49	6 Has_Cabin
0.069 +- 0.003	4 +- 0.77	7 FamilySize
0.065 +- 0.005	4.6 +- 0.66	5 Fare
0.03 +- 0.003	6 +- 0	8 IsAlone
0.015 +- 0.002	7 +- 0	4 Age

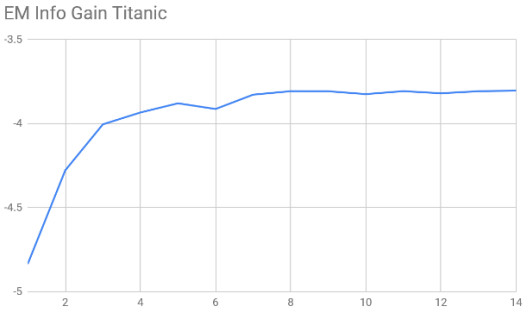
average merit	average rank	attribute
0.15 +- 0.001	1 +- 0	4 marital status
0.097 +- 0.001	2 +- 0	1 ï»¿age
0.063 +- 0.001	3 +- 0	3 education
0.037 +- 0	4 +- 0	6 sex
0.008 +- 0	5 +- 0	5 race
0.008 +- 0	6 +- 0	2 workclass
0.006 +- 0	7 +- 0	7 Native Region

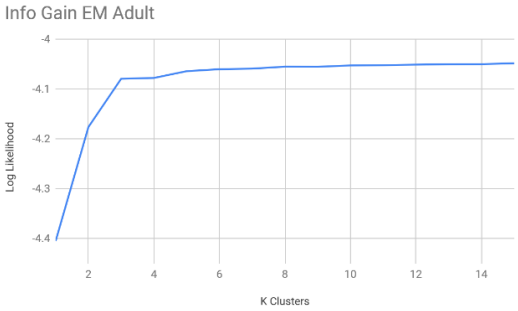
For Titanic we removed isAlone and age, while for adult we removed race, work class, and native region. This can be explained as isAlone overlaps with information provided with family size in titanic, and is unnecessary, while age has been discretized heavily and it became hard to distinguish between who survived based on ranges of ages, as well as the fact that only certain age of people would be on the ship anyways.

For adults we took out race, work class, and native region. Due to response and sampling bias, a large portion of people were already from North America, rendering race and native region relatively useless, and work class was vague to start and was missing a lot of values, thus contributing little.





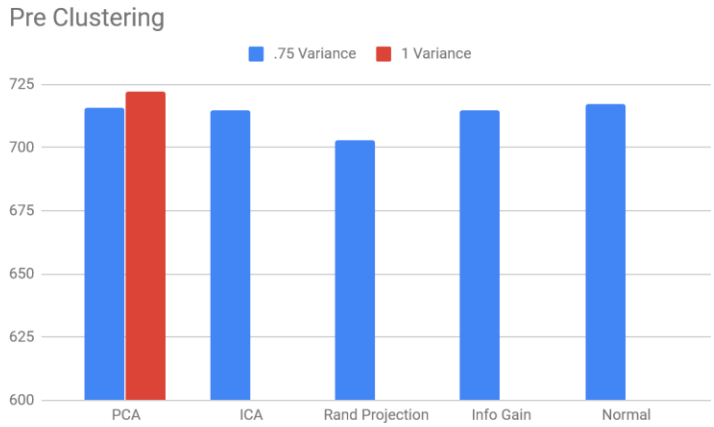




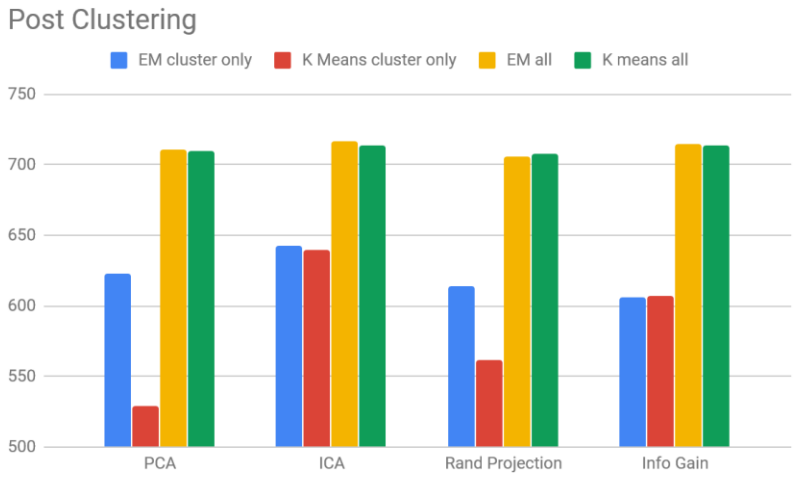
In terms of clusters, Titanic used 4 for both, and adult uses 2 and 3 for k means and EM respectively. The main comparison to made here is to the data before reduction, where k means used the same amounts of clusters, EM created less clusters this time, most likely because there were several really obvious groups and that the removed attributes were there more as a hindrance to bring in noise, requiring more clusters where it is not necessary, such as an entire cluster representing missing values etc.

Neural Nets

In the final part, using Titanic data set, we ran each of the reduced data through unsupervised learning algorithms to create clustered data, and then compared that to how well it predicts against pre clustered data and only clusters and see how the neural net predicts its labels.



Without any clustering, there is very miniscule amount of differences between the neural net accuracies; each of them ranged between 710 to 720 out of 831 instances correct. PCA did do better with 1 variance covered, but it is not significant enough to be commented out. This is probably because the data was simple to start with, and neural net already does its best to place weight on important and less weight or on less important attributes. Due to the small data set speed was about the same, but info gain ran the fastest due to less attributes while ICA ran the longest because of the 31 attributes, but the difference was less than .005 seconds.



Using all the attributes, a similar trend appeared where they performed similarly, but predicting labels only based on clusters did horribly, especially using K means. As mentioned before, there exists a lot of overlap between different attributes, that even with dimension reduction it is hard to remove all grey areas. More importantly, there might be bigger trends to be captured, such as the types of people on the ship based on class and economy, rather than whether they survived or not, and that survival was just not as obvious in terms of K means and EM under unsupervised learning. They did not do below 50% in the end, because some correlation exists between these trends, such as richer people or upper-class people with better cabins and location will survive easier, but it is just not the strongest correlation amongst attributes.

Conclusion

Starting from early on it became apparent that running the neural net with and without clustering would make very little difference. The biggest thing was that as stated above, survival was not the biggest trend in Titanic, as those who boarded the ship did not entire the situation purely for survival, and it just happened that certain traits helped them based on location, money, or family. Between the different dimensional reducing algorithms, the optimal cluster that appeared the most common was around 2 to 4 for the majority, and around 12 for ICA and PCA; random projection's trend was harder to predict due to the nature of the algorithm.

Next time, to improve consistency and improve the results of the reduction algorithms, especially ICA and PCA, it will be good to fill out some of the missing values, as well as be less aggressive with the discretization. That way two items on the ends of a range will not be viewed as similar while two items at two sides of the cut off will not be viewed as so different, help making differences clearer, and allowing components to be added more clearly. This will also allow clusters to draw better inference on what is truly near it and what is far away. This is important as ICA and PCA made later algorithms suffer as they could not reduce amount of components and caused curse of dimensionality.

References:

Ronny Kohavi and Barry Becker (1996). UCI Machine Learning Repository [http://archive.ics.uci.edu/ml]. Data Mining and Visualization ,Silicon Graphics.