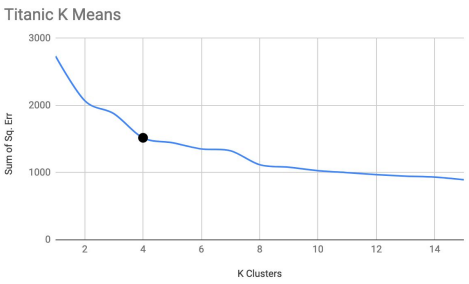
Arthur Shim
ashim6
CS 4641

Analysis

For this assignment, I used the Titanic and Adult datasets. Both of these were used in my previous assignments. The Titanic dataset holds information about passengers of the Titanic, including their age, sex, and family size, which can be used to classify whether they survived. It has 7 features and 891 instances. This dataset is interesting because it is composed solely of discrete features, and is not an incredibly enormous dataset, so it will be intriguing to see how different algorithms perform under these two conditions. The Adult Dataset contains information about adults around the world including age, education, residence, marriage-status, etc, and can be used to classify whether an adult has an income of above $50k. The Adult Dataset has 14 features and 48,842 instances, so it is a significantly larger dataset than the Titanic Dataset. It also contains a mixture of discrete and continuous features, so it'll be interesting to see how the learning algorithms perform on this dataset compared to the Titanic dataset.

Clustering

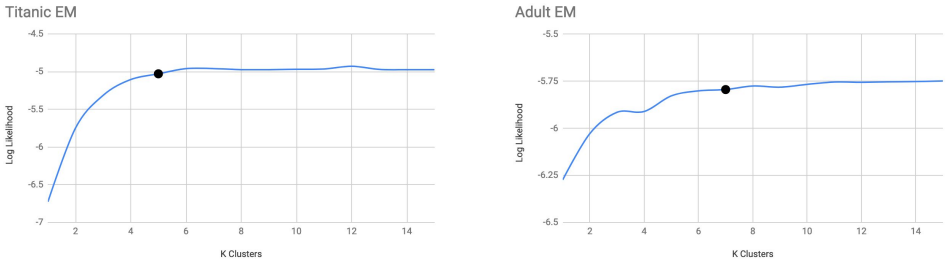
In this assignment, I examined the performance of K Means and Expectation Maximization on the two datasets. K Means works by selecting k random centers, not necessarily points in the data. The centers claim the nearest points in the data and then each is recomputed using the average of the points of the clusters. I used K Means using euclidean distance to cluster the data. To find the value for k, I tried different numbers of clusters and graphed the sum of squared errors that each produced.



I then used the elbow method to determine the optimal value of k. When selecting k, we're looking for the smallest value of k that minimizes the sum of squared errors. However, the value of k with the minimum sum of squared errors will be when k equals the number of instances, since each instance will be in its own cluster. This is clearly overfitting, so the elbow method looks for the value of k where there is a "bend" in the curve, like an elbow. This method assumes that values of k past the "elbow" have diminishing returns. Using this method, I

identified that the optimal value for k is 4 for the Titanic dataset and 2 for the Adult dataset. Both graphs contain multiple elbows, so I selected the points with the greatest changes in slope.

The second clustering algorithm used was Expectation Maximization. This is a soft clustering algorithm that creates k probabilities for each point that represents the likelihood of the point being in each of the k clusters. This effectively generates a Gaussian distribution for each cluster with the probability at a point representing the probability that the point could've been generated by the distribution. We then try to maximize these likelihoods.



To find the optimal k value, I found the log likelihood of a range of k values for both datasets and then selected the smallest k value where the log likelihood begins to converge. It's better to pick smaller k values because more clusters will lead to higher chances of overfitting. Therefore for Titanic I selected k = 5 and for Adult k = 7 as the optimal k values.

Cluster Results

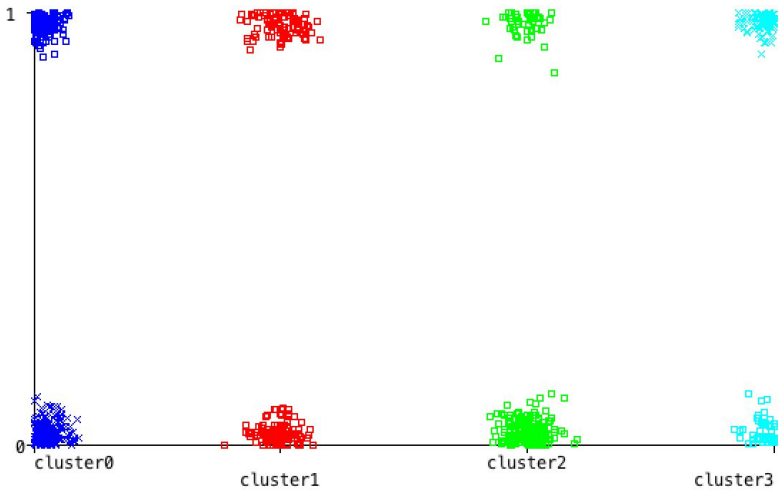
For the Titanic dataset, K Means with 4 clusters, the algorithm generated clusters with data points composed of certain combinations of features. The four clusters appear as follows:

Cluster 1: Contains 34% of the instances. Composed of young, lone, middle-class men without cabins.

Cluster 2: Contains 24% of the instances. Composed of young women with families.

Cluster 3: Contains 25% of the instances. Composed of young, lone lower-class men without cabins.

Cluster 4: Contains 16% of the instances. Composed of older, upper-class men with families and cabins.



This graph is a visualization of the clusters grouped with their instances that survived and didn't (1 - survived, 0 - didn't survive). Each cluster is an almost even split between survivors and non-survivors. This indicates that the clusters produced by K Means did not line up with the labels. Instead, K Means found groupings of instances with similar attributes, which in this case are groupings of similar passengers.

The clusters are organized in this way because most passengers can be put into these groups. There are no clusters with lone women because most of the passengers that were female were with their families. There are more clusters with men because there are more male passengers than female and therefore K Means produced clusters with different variations of male passengers.

For the Adult dataset, K Means with 2 clusters produced the following clusters:

Cluster 1: Contains 79% of the instances. Composed of 43-54 year old, college educated, married white men from the USA and Canada.

Cluster 2: Contains 21% of the instances. Composed of college educated, single white women of all ages from the USA and Canada.

This graph shows the two clusters and how their instances are labeled in terms of income. Like the Titanic dataset, the clusters for this dataset also contain an almost even split between the labels, indicating that the clusters produced by K Means are not in line with the labels. K Means found groupings of instances with similar attributes, but not groupings of instances with similar labels.

The clustering produced by K Means displays the bias in this dataset. Most of the instances are from the USA and Canada and most are white and college-educated which is why the best clustering produced are two clusters, one essentially composed of married white men and the other composed of single white women. The majority of the instances are in the first cluster, indicating that the dataset is mostly composed of data about educated, white men from the USA and Canada.

For the Titanic dataset, EM with 5 clusters produced the following clusters:

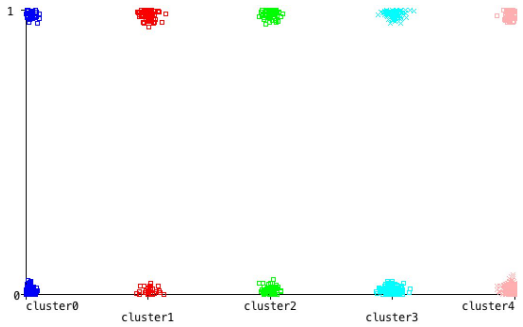
Cluster 1: Contains 11% of the instances. Composed primarily of middle-class, young, lone people without cabins and paid cheap fares. ⅓ women and ⅔ men.

Cluster 2: Contains 12% of the instances. Composed primarily of lower-class, middle-aged people with cabins and paid expensive fares. Most had families and were not alone. 61% women and 49% men.

Cluster 3: Contains 12% of the instances. Composed primarily of young, lone people with cabins and paid expensive fares. 30% women and 70% men.

Cluster 4: Contains 28% of the instances. Composed primarily of middle and upper-class, young families and couples without cabins and paid expensive fares. ½ women and ½ men.

Cluster 5: Contains 37% of the instances. Composed of upper-class, young, lone men, without cabins and paid cheap fares.



As with K Means, EM also produced clusters that, as seen, did not line up with the labels. Each cluster is an almost even split between survivors and non-survivors. EM created clusters that focused less on gender like K Means and more on wealth, class, and family.

For the Adult dataset, EM with 7 clusters produced the following clusters:

Cluster 1 (33% of instances): Primarily middle-aged, private and government sector, college educated, married white men from the USA and Canada.

Cluster 2 (25% of instances): Primarily young (less than 20), private sector, college educated, single white men and women from the USA and Canada.

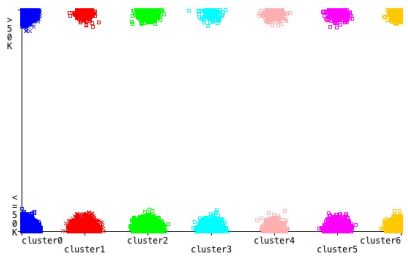
Cluster 3 (18% of instances): Primarily older, government sector, college educated, separated/widowed white women from the USA and Canada.

Cluster 4 (6% of instances): Primarily middle-aged, private sector, pre-high school educated, married white men from North and South America.

Cluster 5 (2% of instances): Primarily middle-aged, private sector, college educated, married Asian men and women from Asia.

Cluster 6 (9% of instances): Primarily middle-aged, private sector, college educated, married white men from the USA and Canada.

Cluster 7 (7% of instances): Primarily elderly, private sector, college educated, married white men from the USA and Canada.



EM produced clusters that are more fine tuned towards the attributes that K Means for the Adult dataset. While K Means found clusters that simply distinguished between men and women, EM found clusters that went into greater detail with the attributes.

While K Means and EM produced a similar number of optimal clusters for the Titanic dataset, they produced very different optimal values of k for the Adult dataset. EM produced many more clusters for the Adult dataset than K Means. This could be due to the fact that the Adult dataset contains many non-binary discrete features (ex. Race: white, black, asian, etc.).

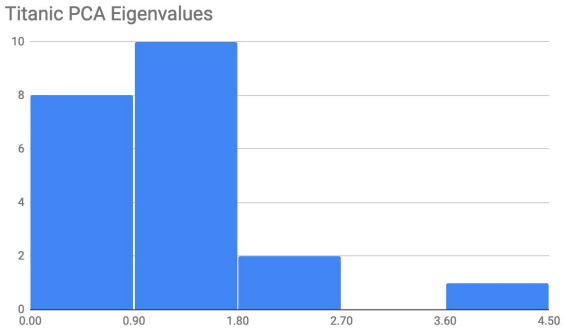
Overall, I observed that EM produced a larger number of optimal clusters than K Means. For these two datasets, K Means created fewer clusters with a more general look at the features of the instances while EM created more clusters with more detailed features. This can especially be seen in the Adult dataset, where K Means created 2 clusters that essentially just separated men and women and EM created 7 clusters that distinguished instances by age, job, and marriage status as well.

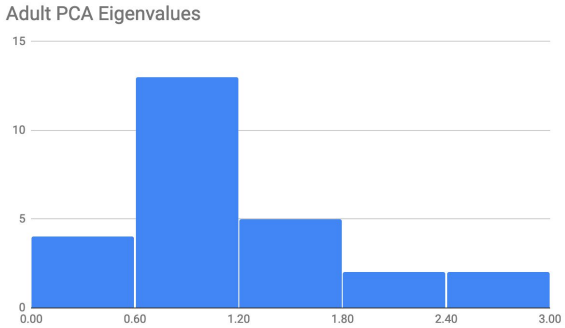
Dimensionality Reduction

In the next portion of the assignment, I performed four dimensionality reduction algorithms on the datasets. The four algorithms were PCA, ICA, random projection, and information gain.

PCA

These are the eigenvalue distributions for the Titanic dataset and Adult dataset using variance coverages of 0.75.





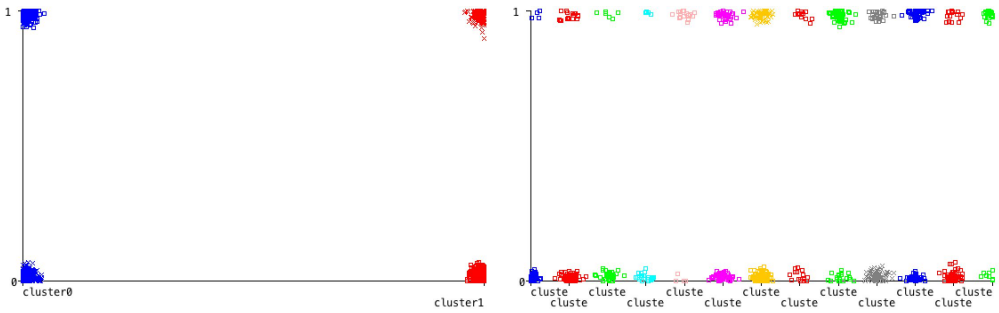
Based on these values, it can be seen that PCA was able to find one very strong component for the Titanic dataset. This component has a high eigenvalue (between 3.6 and 4.5) which means that there is high variance along this component. Essentially, PCA found a component that is able to effectively describe a lot of the data, which is why its eigenvalue is high.

For the Adult dataset, PCA was not able to create as strong of a component, but was able to create a couple fairly strong components. This can be seen by the Adult PCA generating a couple eigenvalues above 2. While for the Titanic dataset, PCA created one very strong

component and many weak components, for the Adult dataset PCA created a few somewhat strong components and several weak components. This could be attributed to the fact that the Titanic dataset is a smaller, more straightforward dataset with less features and complexity while the Adult dataset is a larger, more complex dataset.

I ran PCA with K Means on the Titanic dataset with variance coverages of 0.75, 0.95, and 1. PCA with var = 0.75 had the least sum of squares error for the Titanic dataset so I used this version of PCA to determine the new value for k, which for Titanic was 2. The new clusters each contained about 50% of the instances.

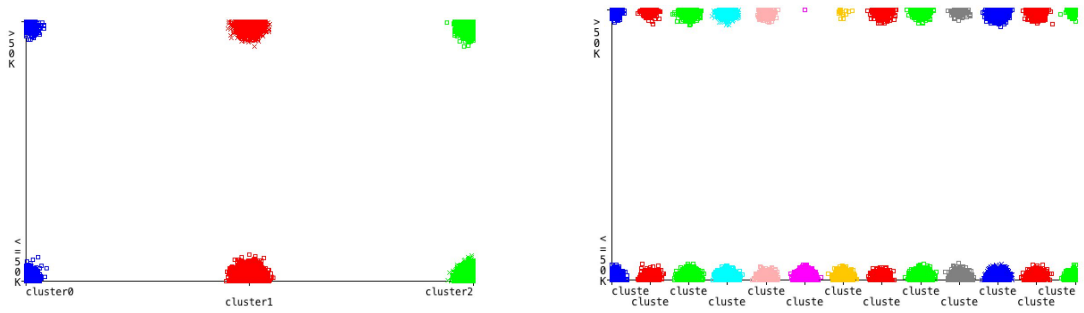
I repeated the same process for EM and PCA. I selected a variance coverage value of 1 because it had the greatest log likelihoods. I then found the optimal value of k to be 13. The new clusters had a fairly even distribution of the instances, with the biggest cluster containing 18% of the attributes and the smallest containing 3% of the attributes.



For both of these new clusterings, the trend in labels not lining up with the clusters maintained.

I then ran PCA with K Means on the Adult dataset with the same variance coverages. PCA with var = 0.75 had the least sum of squares error so I used this value to determine the new value for k. Using the elbow method, I found that value to be 3. The new clusters created were composed of two large clusters containing 50% and 48% of the data, and one smaller cluster containing 2% of the data.

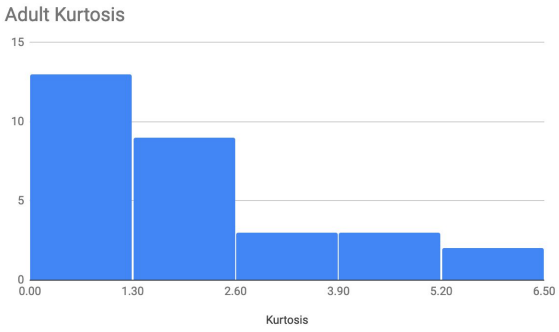
I repeated this with PCA and EM and found that EM with var = 1 had the highest log likelihood. I found the optimal value of k to be 13. The new clusters ranged from containing 27% of the instances to 2% of the instances.

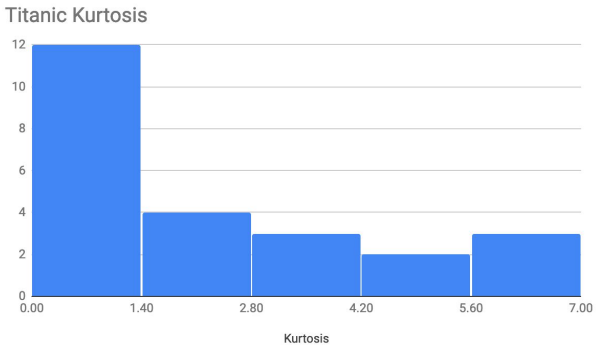


The trend in the clusters was the same as before, with the clusters not following the labels.

ICA

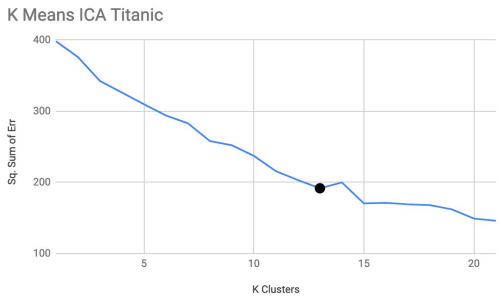
Below are the kurtosis values for the two datasets.

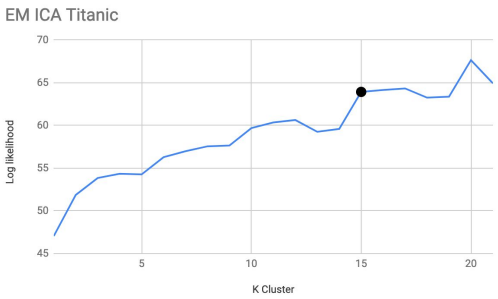




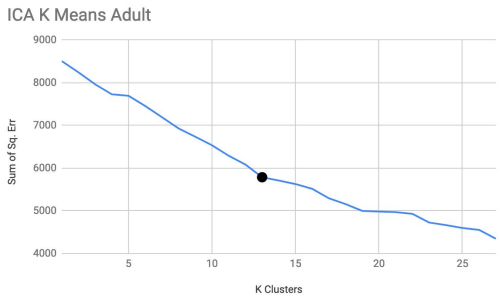
Both datasets generated high kurtosis values after ICA was run, indicating that most of the distributions generated by ICA are narrow with high peaks. Since a kurtosis value of 3 indicates a perfect gaussian, most of the independent components are not perfect gaussians. Both datasets have mostly small kurtosis values. This is good because ICA is looking to create independent, non-gaussian components.

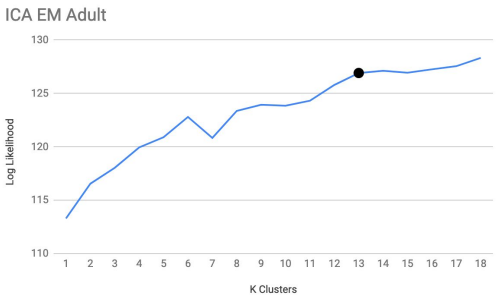
After applying ICA, I re-ran K Means and EM on both datasets. The Titanic dataset produced the following curves:





I then found the optimal k values for both, which are 13 and 15 respectively. The new clusters for K Means and EM both had similar, even spreads of instances, and the trend among the clusters with respect to the labels remained unchanged.

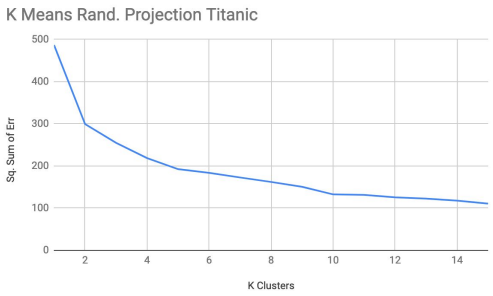


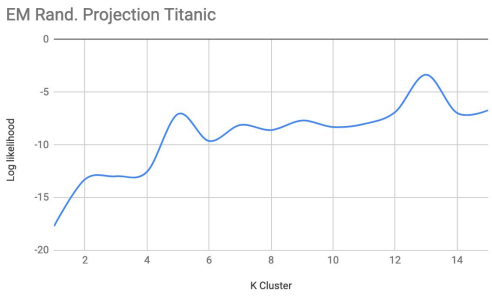


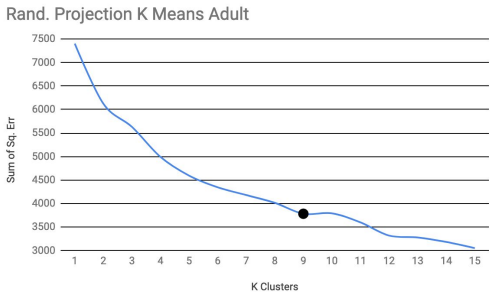
Repeating this process for the Adult dataset led to k values of 13 for both K Means and EM. K Means produced clusters including one large cluster containing 30% of the data and the rest containing much smaller amounts. The new clusters for EM contained two large clusters containing 18% and 22% of the instances and the rest containing much smaller amounts.

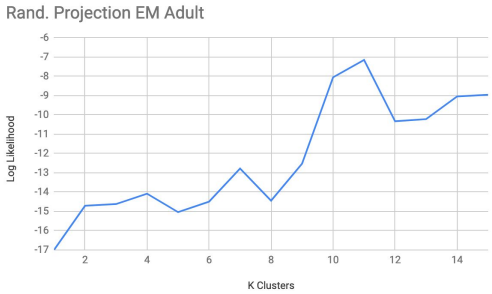
Random Projection

Random projection works by essentially creating random components. I ran this several times and selected the best results to use.









Running K Means on the Titanic dataset led to an optimal value of k = 2. EM on the Titanic dataset after random projection led to an optimal value of k = 9. For the Adult dataset, K means led to an optimal k value of k = 9, and EM led to an optimal value of k = 14.

For the Titanic dataset, the optimal sum of squared errors ranged from 299 to 400 for K means. For the Adult dataset, the optimal sum of squared errors ranged from 3781 to 4121. As the sample size increased, the variance and error in the clusters increased as well. In terms of EM, the Titanic dataset ranged from -7 to -3 and the Adult dataset ranged from -13 to -8.

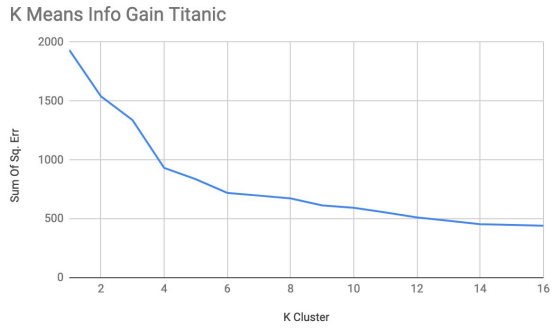
Info Gain

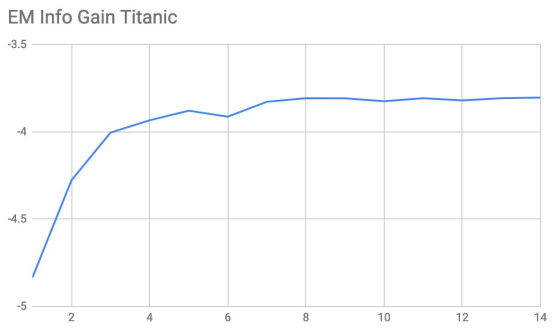
Information gain works by attempting to pick the features that provide the most amount of information using information theory and entropy. Using the following results, I removed the bottom quartile of features with the least information gain:

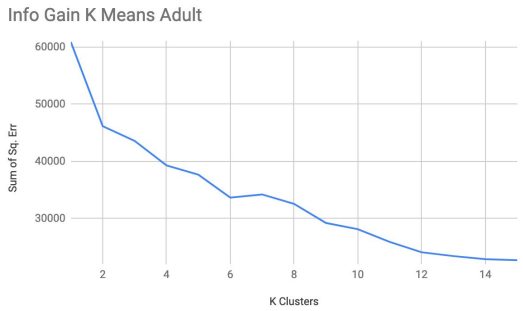
average merit	average rank	attribute
0.218 +- 0.009	1 +- 0	3 Sex
0.084 +- 0.006	2 +- 0	2 Pclass
0.071 +- 0.004	3.4 +- 0.49	6 Has_Cabin
0.069 +- 0.003	4 +- 0.77	7 FamilySize
0.065 +- 0.005	4.6 +- 0.66	5 Fare
0.03 +- 0.003	6 +- 0	8 IsAlone
0.015 +- 0.002	7 +- 0	4 Age

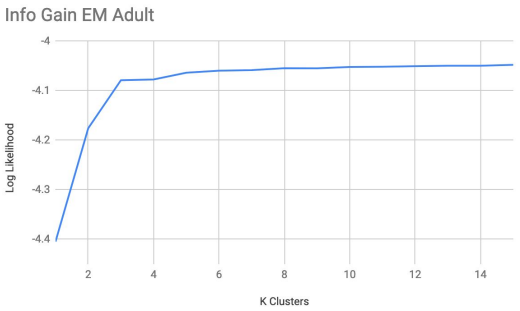
average merit	average rank	attribute
0.15 +- 0.001	1 +- 0	4 marital status
0.097 +- 0.001	2 +- 0	1 ï»¿age
0.063 +- 0.001	3 +- 0	3 education
0.037 +- 0	4 +- 0	6 sex
0.008 +- 0	5 +- 0	5 race
0.008 +- 0	6 +- 0	2 workclass
0.006 +- 0	7 +- 0	7 Native Region

I removed isAlone and Age from the Titanic dataset and Native Region, Workclass, Race from the Adult dataset. This can be explained by the fact that isAlone and FamilySize overlap. Age is a heavily discretized feature making it difficult to determine if a passenger survived based on their age. Native Region provided little information gain since the vast majority of instances were people from the USA or Canada. The majority of people in the Adult dataset also worked in the private sector and were white, so there is little information gain from these features as well.





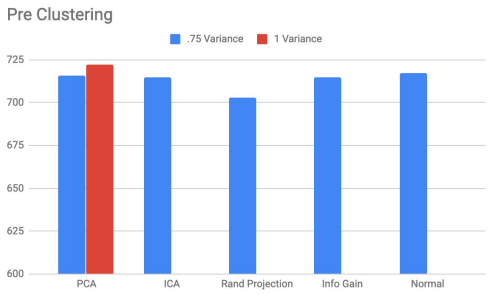




K Means produced 4 clusters for Titanic and 2 clusters for the Adult dataset. EM produced 5 clusters for Titanic and 3 clusters for Adult. While the number of clusters before info gain for Titanic K Means, Titanic EM, and Adult K Means didn't change after info gain, there was a significant reduction in clusters after info gain for Adult EM. This could indicate that the removal of unnecessary features like Native Region and Workclass was significant enough to eliminate some previously created clusters.

Neural Nets

The final part of this assignment involved running the clustered datasets through a neural network using 10-fold cross validation and comparing its performance to a neural network before clustering. For this portion I only used the Titanic dataset.



I first compared the performance of the neural net on the original dataset with the performance of neural nets on the datasets after running them through each of the dimension reduction algorithms. I performed two different tests with PCA, one with a variance coverage value of 0.75 and another with a variance coverage value of 1. This was due to the fact that for K Means, PCA with var = 0.75 performed best while for EM, PCA with var = 1 performed best. PCA with var = 1 was the only algorithm to demonstrate an increase in the number of correctly classified instances for this dataset compared to the original dataset. This indicates that most of the features in the original dataset are better to have when solving this classification problem instead of the features created by each of the dimension reduction algorithms. This is most likely due to the original simplicity of the Titanic dataset.

The next step was clustering the data and running a neural net on the dataset while including the clusters as features. Each reduced dataset was taken and clustered using K Means and EM. Then the original features and the newly formed clusters were used as features in a neural network (K Means all and EM all) and the performance was tested. Then, all features except the clusters were removed and the dataset was run through the neural network a second time (K Means cluster only and EM cluster only). For all reduced datasets, using just the clusters as features in the neural net demonstrated markedly decreased performance, especially for K Means. When the clusters and original features were included, both EM and K Means demonstrated similar levels of performance. This indicates that the clusters did little to improve the neural net's performance, and likely the neural net relied mostly on the original features instead of the clusters to perform its classification.