Joao Matheus Nascimento Francolin

CS 4641

Unsupervised Learning and Dimensionality Reduction

Sunday, April 7, 2019

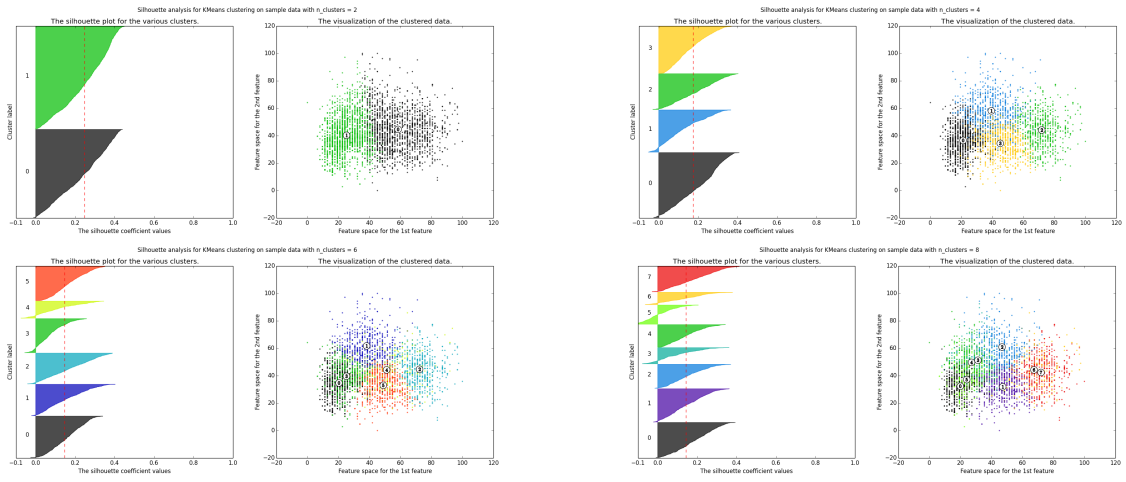
Introduction

In this assignment, I explore unsupervised learning algorithms by applying k-means clustering and Expectation Maximization, as well as performing four dimensionality reduction algorithms (PCA, ICA, Randomized Projections, and Chi-Square Feature Selection) on two distinct data-sets. The first data-set is wine quality data-set with 8 classes used mainly as a proof of concept for the correct implementation of the algorithms. The second data-set is a vectorized dictionary of the language used by politicians extracted from twitter on assignment 1 (Supervised Learning) with two classes, Democrats are Republicans. In the paper, I will first present the results from the wine data-set to demonstrate the expected grouping of the wine classes, and finally, apply the same algorithms on the dictionary in order to group individuals into their respective political affiliation.

Clustering Algorithms

Wine Data Set

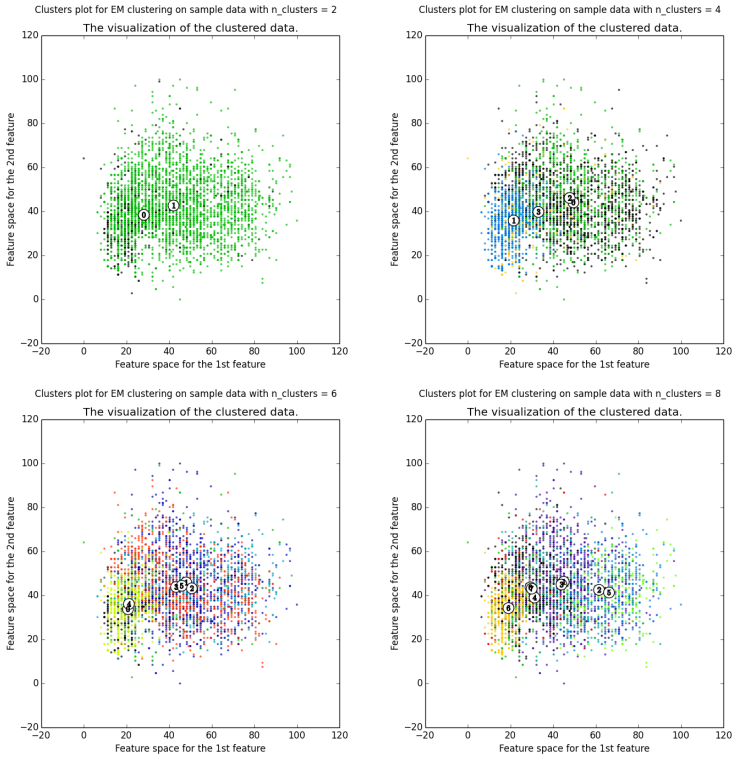
The wine data-set has 11 features. The data were clustered varying K for K = [2, 4, 6, 8]. The silhouette coefficients were plotted for every run for the most relevant features of the data-set. The algorithm clustered well for iterations up to k = 4, but it seems to start doing poorly when k is greater than 4.



For k = 6 and k = 8, the data is only subdivided into groups already created on the previous iterations. Moreover, the Silhouette Coefficient and Normalized Mutual Information were both plotted on every iteration. By inspection, we see that NMI is greater when k = 2. This is probably

because as of this implementation the algorithm assumes that all eleven features of the data-set as being equally relevant, which is probably not the case.

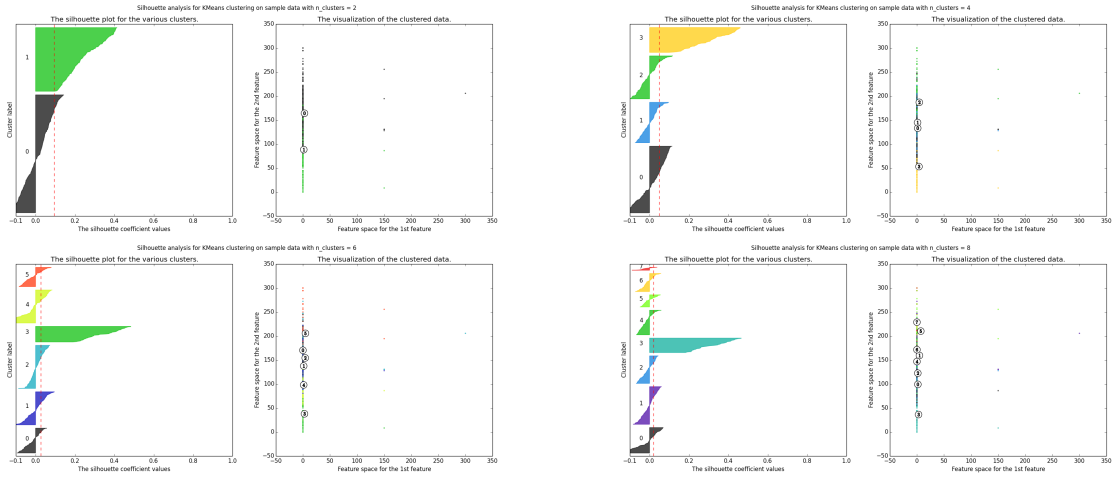
The dataset was also clustered using Expectation Maximization. The feature values were normalized, and the clusters were plotted for K = [2, 4, 6, 8].



The scores for n_clusters = [2, 4, 6, 8] were NMI = [0.049031, 0.056479, 0.048976, 0.078178]. The scores are not particularly exciting. While I performed the experiments various times, it is my current hypothesis that the NMI scores were slow due to a poor selection of clustering features.

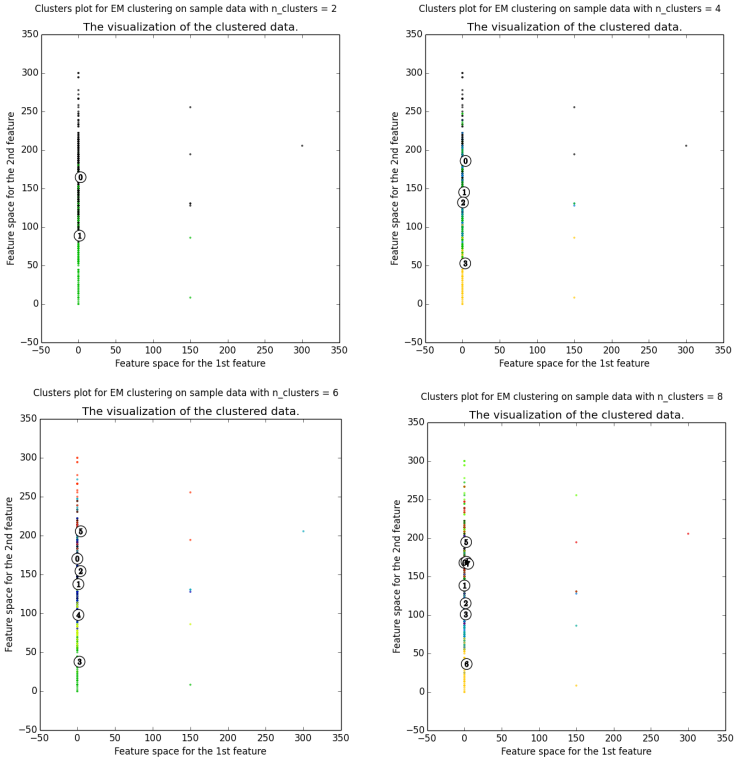
Political Affiliation Data Set

The dictionary of the political affiliation data-set has 300 features and two classes, republican and democrat. The data were clustered varying K for K = [2, 4, 6, 8]. The silhouette coefficients were plotted for every run for the most relevant features of the data-set.



As expected, the algorithm clustered the best for k = 2. For values of k greater than 2, the data is only subdivided into groups already created on the previous iterations. The Silhouette Coefficient and Normalized Mutual Information were both plotted on every iteration. Similarly, to the wine data-set, we see that NMI is greater when k = 2. The hypothesis of poor deprecated NMI scores due to a naive weight of features holds. Now given 300 features, we observe much lower NMIs.

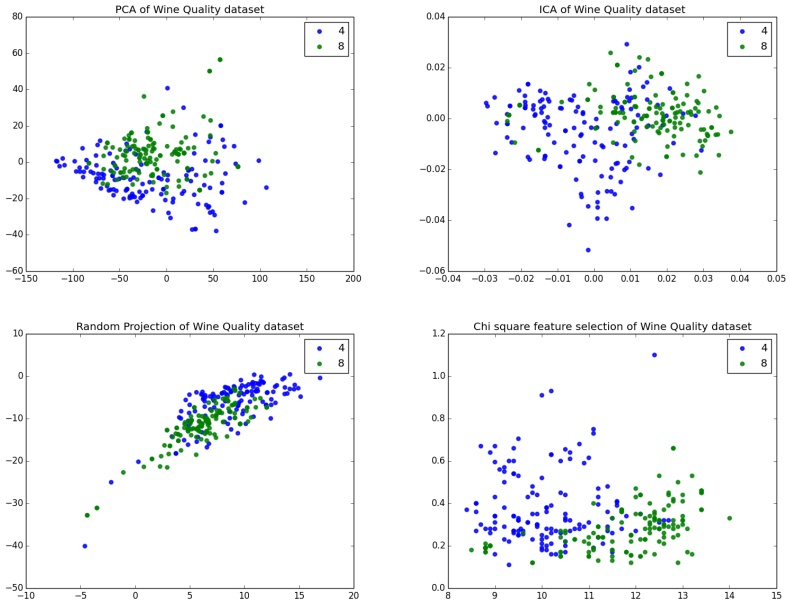
The dataset was also clustered using Expectation Maximization. The feature values were normalized, and the clusters were plotted for K = [2, 4, 6, 8]. The scores for n_clusters = [2, 4, 6, 8] were NMI = [0.104596, 0.084505, 0.081843, 0.090088].



Dimensionality Reduction Algorithms

Wine Data Set

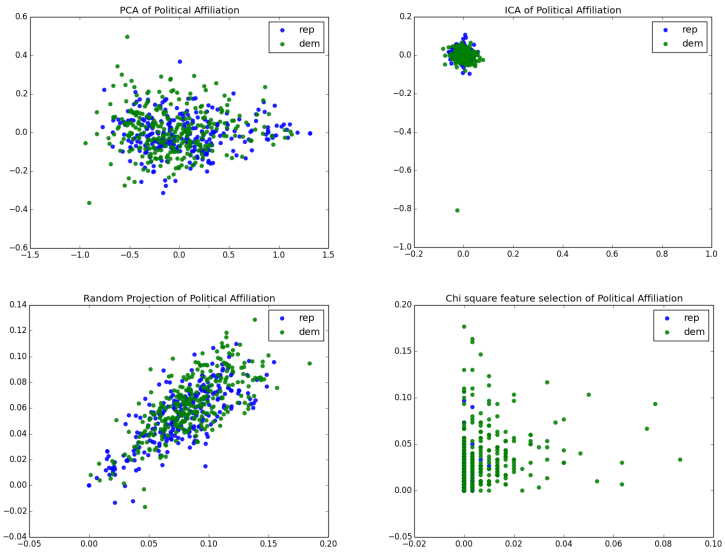
The four dimensionality reduction algorithms (PCA, ICA, Randomized Projections, and Chi-Square Feature Selection) were applied in the wine data-set. The results were plotted for separating classes 4 and 8.



Classes 4 and 8 were chosen because they clustered the best. ICA and Random Projection, even having very different clusterings, ended up performing similarly. PCA was slightly better, and Chi-Square Feature Selection performed the best showing a very clear separation of the samples. It is interesting to point out that on average the feature selection and transformation methods performed equally when clustering the data, a rather surprising outcome since the feature selection should take advantage of the most significant features to perform such task.

Political Affiliation Data Set

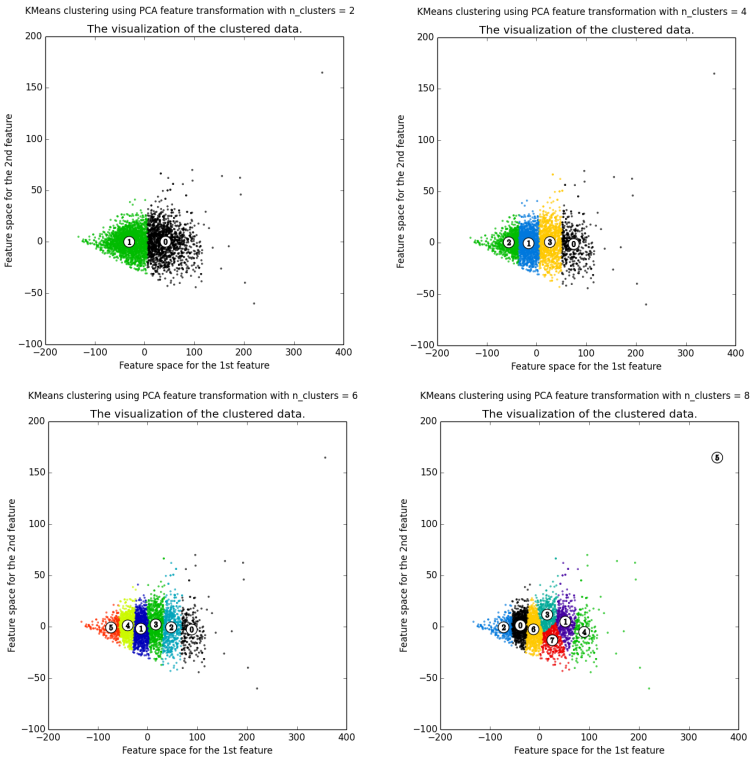
The same four dimensionality reduction algorithms were now applied in the political affiliation data-set. The only two classes, republicans, and Democrats were plotted for all algorithms. ICA and Chi-Square Feature Selection performed the worst showing overlapping points for republicans and democrats, and therefore no meaningful separation in the data. PCA and Random Projection performed better, and still showcasing a lot of overlapping data points among the classes. There was not clearly better method between feature selection and transformation algorithms. Results shown below.



Cluster After Dimensionality Reduction

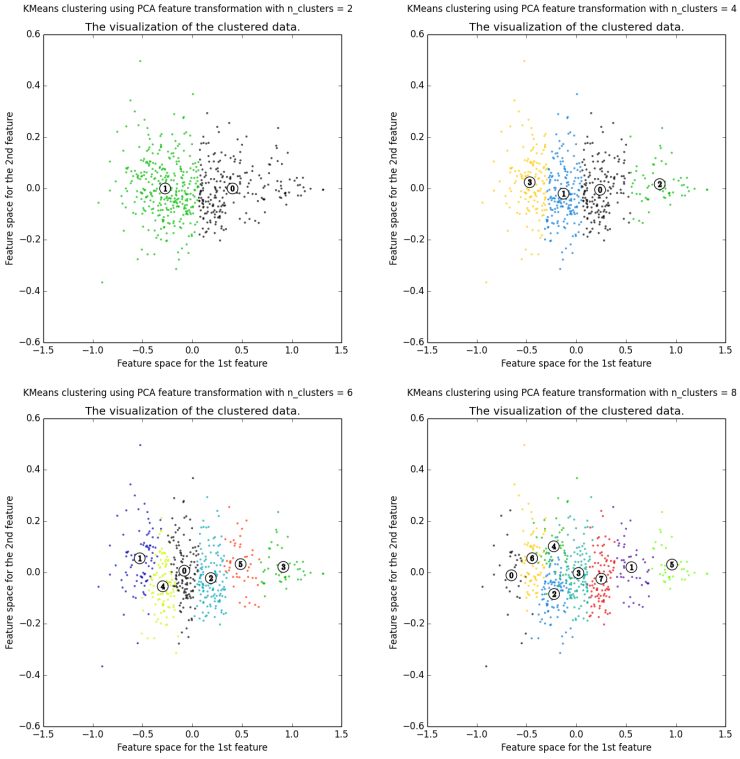
Wine Data Set

The data were re-clustered after dimensionality reduction for K = [2, 4, 6, 8]. The results were plotted for every run for the most relevant features of the data-set. The algorithm clustered the best for k = 6. The new scores for n_clusters = [2, 4, 6, 8] were NMI = [0.024817, 0.025186, 0.030188, 0.029908].



Political Affiliation Data Set

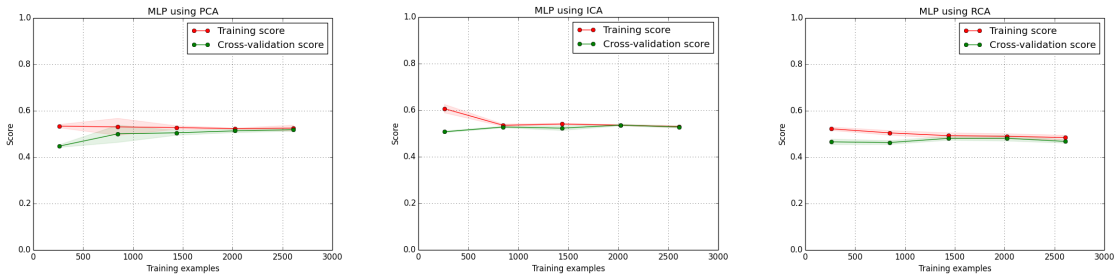
The data were re-clustered after dimensionality reduction for K = [2, 4, 6, 8]. The results were plotted for every run for the most relevant features of the data-set. The algorithm clustered the best for k = 6. The new scores for n_clusters = [2, 4, 6, 8] were NMI = [0.007235, 0. 007404, 0. 0.014631, 0.011630].



Neural Network

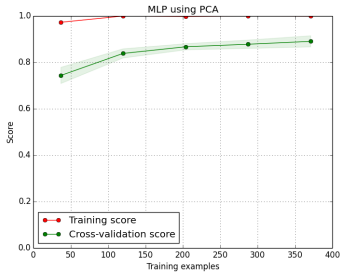
Wine Data Set

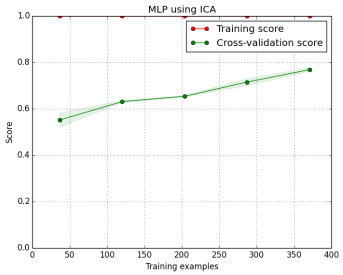
The data-set was trained and tested on an ANN model with 20 nodes and 5 hidden layers. PCA, ICA, and RCA all had very similar scores around 0.5, which is not very encouraging. Normally the close gap between training score and cross-validation score would indicate that the amount of data used for training was sufficient. However, with scores so close to a representation of pure chance, the close distance between the lines lose their significance.

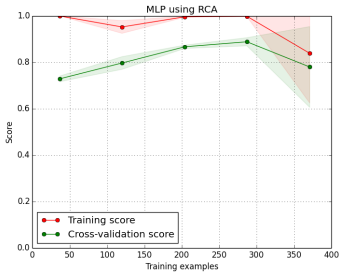


Political Affiliation Data Set

Once again, the data-set was trained as tested using the same ANN model with 20 nodes and 5 hidden layers. This time we observe different behavior for PCA, ICA, and RCA. PCA arguably performed the best with very high training scores, and relatively low gap the training scores and cross-validation scores. ICA performance was the worst among the three, but still relatively good; given the trend in the lines, probably better given more training data. MPL had interesting results, showing good scoring data, but signs of overfitting.



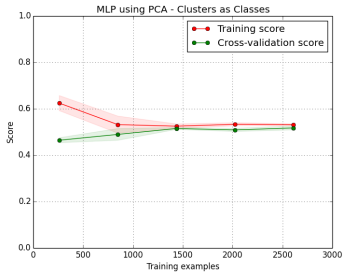


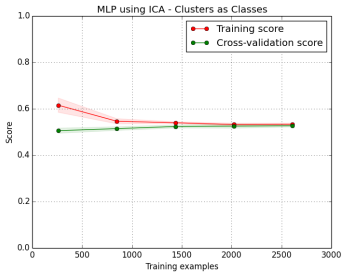


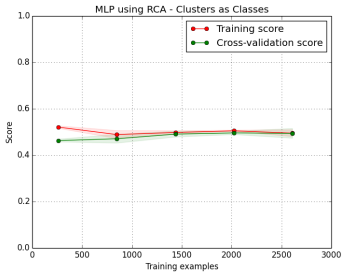
Neural Network with Clusters as Classes

Wine Data Set

The same ANN model with 20 nodes and 5 hidden layers were used to train and test different models. Comparing the results with the ones found on the Neural Network section we see little improvement on the the algorithms. The scoring average scoring is still around 0.5 and the training score and cross validation score do not vary much.







Political Affiliation Data Set

The same ANN model with 20 nodes and 5 hidden layers were used to train and test different models. Comparing the results with the ones found on the Neural Network section we see improvements on both ICA and RCA, while PCAs behavior remains virtually the same. RCA performs best now with training scores and cross validation score both on their 0.90s and lines very close to converge.