Georgia Institute of Technology CS 7641 – Machine Learning

Unsupervised Learning and Dimensionality Reduction

Santhanu Venugopal Sunitha (ssunitha3)

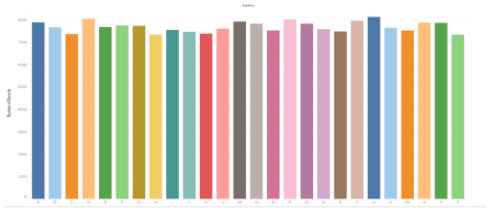
The purpose of this paper is to explore unsupervised learning algorithms, namely clustering and dimensionality reduction algorithms.

Datasets

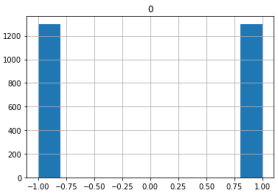
The datasets chosen for the analysis are the same from Assignment 1. They are interesting as Madelon is a high dimensional dataset with 500 dimensions and artificially created using clustered data points from the summit of a 5-D hypercube, while Letter Recognition dataset is not highly dimensional, and most features seems to be non-correlated. Hence, it would be interesting to compare and contrast these two datasets by applying clustering and dimensionality reduction algorithms. Below is the brief description of these two datasets.

- **Letter Recognition Dataset** from UCI ML repository (20000 instances, 16 attributes, 26 classes)
  This dataset contains information about character image features. The objective is to identify each of a large number of black-and-white rectangular pixel displays as one of the 26 capital letters in the English alphabet. The character images were based on 20 different fonts and each letter within these 20 fonts was randomly distorted to produce a file of 20,000 unique stimuli. Each stimulus was converted into 16 primitive numerical attributes (statistical moments and edge counts) which were then scaled to fit into a range of integer values from 0 through 15.

- **Madelon Dataset** from UCI ML repository (4400 instances, 500 attributes, 2 classes)
  MADELON is an artificial dataset containing data points grouped in 32 clusters placed on the vertices of a five-dimensional hypercube and randomly labeled +1 or -1. The five dimensions constitute 5 informative features. 15 linear combinations of those features were added to form a set of 20 (redundant) informative features. Based on those 20 features one must separate the examples into the 2 classes (corresponding to the +-1 labels). It has a number of distractor feature called 'probes' having no predictive power. The order of the features and patterns were randomized. The difficulty is that the problem is multivariate and highly non-linear. This is also a balanced dataset as depicted below.



Madelon Class Distribution



Letter Recognition Class Distribution

Part 1.

Clustering

Clustering is an unsupervised machine learning technique that involves classifying data into specific groups or clusters.

K-means clustering is a type of unsupervised learning, which is used to find groups in the unlabeled data, with the no. of groups represented by the parameter K. This algorithm works iteratively to assign each data point to one of K groups based on the features that are provided.

The EM algorithm is an efficient iterative procedure to compute the Maximum Likelihood (ML) estimate in the presence of missing or hidden data. It is a type of soft clustering technique where we wish to estimate the gaussian model parameters for which the observed data are the most likely.

In this section, we explore K-means and Expectation Maximization clustering techniques.

Methodology

Using scikitlearn library, both algorithms were applied to Letter Recognition and Madelon datasets. K means uses Euclidean distance metric to form clusters. Both algorithms were applied over a range of cluster sizes and the resultant metrics like Sum of Square Error (SSE), Accuracy, Loglikelihood, Homogeneity score, Completeness score, adjusted mutual information (AMI) were plotted and further analyzed. Silhouette plots were analyzed to understand how the inter cluster and intra cluster distances varied across different cluster sizes.

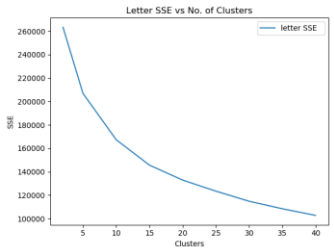
Kmeans Algorithm

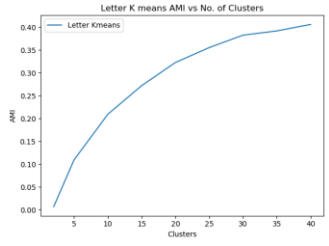
Letter Recognition Analysis

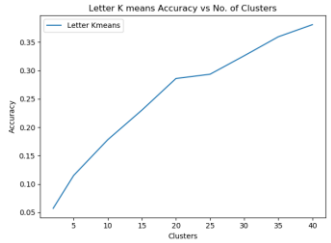
Using the elbow method on the SSE plot, we could find a sharp change in slope around 15 clusters. However, as we look at the silhouette plot for 15 clusters, we can find that though the cluster thickness seems uniform, few of them have high silhouette coefficient values. This can be attributed to the fact that there are different letters in the dataset sharing similar feature values. For instance, 'I' and 'J' share similar width and height features in the dataset. We could also find few samples being close to the decision boundaries between two neighboring clusters

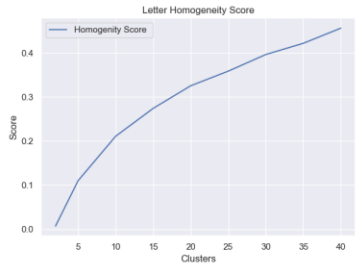
Upon further observation of the silhouette plots, with the increase in cluster size, we can find more samples being closer to the decision boundary of adjacent clusters. Correlating with Accuracy and Completeness score plots, a spike can be observed around ~25 clusters, which can be attributed to the number of label classes in the dataset, which is 26 letters and the silhouette plots shows equal width clustering between 25 and 30 clusters.

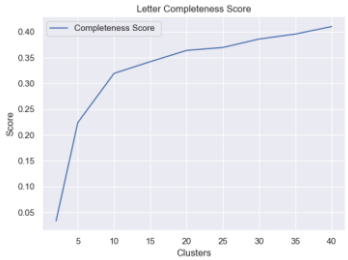
The AMI, Accuracy, Homogeneity and Completeness plots shows increase in score linearly, with 40 clusters having the maximum value. This maybe because there are 20 different fonts for the same letter and the algorithm is trying to capture the same by creating mini clusters of the same letter based on varying font attributes.

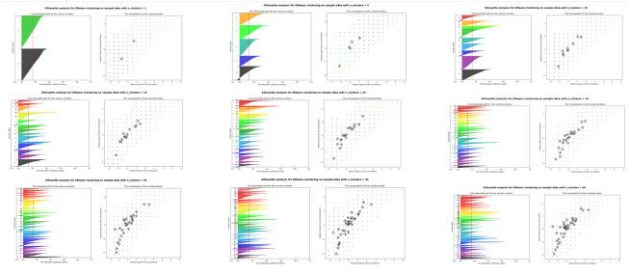












Silhouette Plots

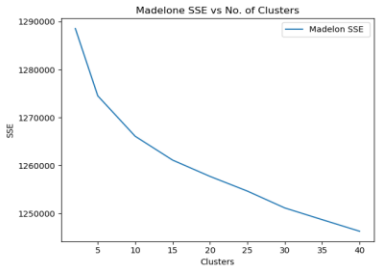
Madelon Analysis

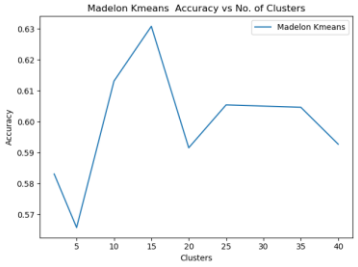
Based on the SSE plot, using an elbow method would propose 5 clusters as an optimal no. of clusters. Madelon dataset is artificially created using a 5-dimensional hypercube and hence K mean algorithm could classify the samples created using the linear combination of these 5 dimensions. If we observe the silhouette plot for 5 clusters, we could find the clusters have equal width, with none of them having any samples being closer to adjacent cluster decision boundaries.

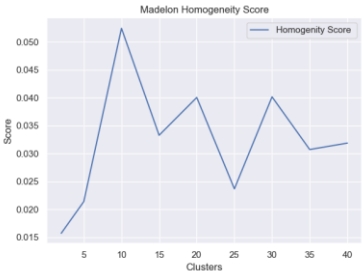
Based on the AMI, Accuracy, Completeness and homogeneity score plots, we could find least scores when numbers of clusters equal 5. This is because the clusters will not have the same labels in them as they are randomly assigned to each of the 32 clusters at the summit of the hypercube, while creating this dataset.

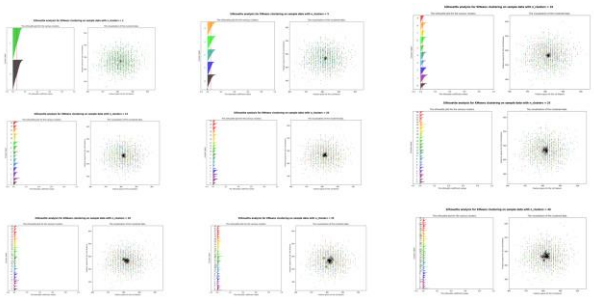
High Accuracy score and AMI score at k = ~16 indicates that the algorithm was able to recreate the initial 16 clusters of same labels (-1 or 1), with which the dataset was artificially generated. This shows that the dataset has a structure with 16 groups of observation having some similarity, which is none other than the linear combination of the 5-dimensional coordinates. A spike can also be

observed at around k=32, which correlates with the 32 clusters with which the dataset is generated but as they don't have uniform labels, the scores are not high as k= ~16.









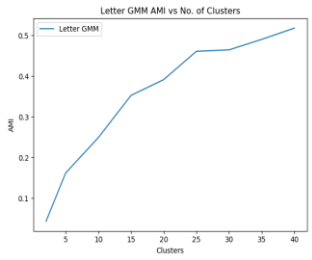
Silhouette Plots

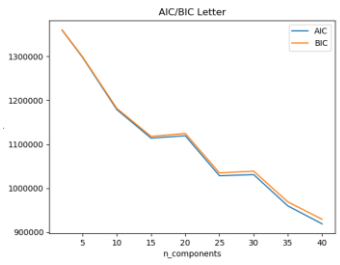
Expectation Maximization Algorithm

Letter Recognition Analysis

Based on the loglikelihood and AMI plot, there appears to be spike at 15, which can be interpreted as different letters falling into the same gaussian distribution with high probability as they share some similar feature values like box measurements, pixel numbers etc.

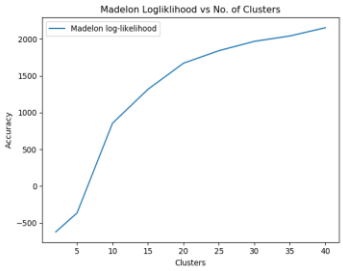
The spike at ~26 can be attributed to the clustering of data based on the labels, which contain 26 letters. Higher accuracy at higher cluster # shows further refinement of the clusters based on font characteristics as well. As lower BIC values are preferred for cluster selection, so, in conjunction with other plots, cluster no. between 25 and 30 would be an ideal choice here. So, we can pick cluster = 26.

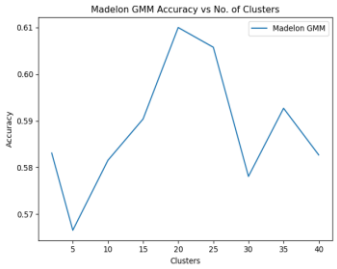


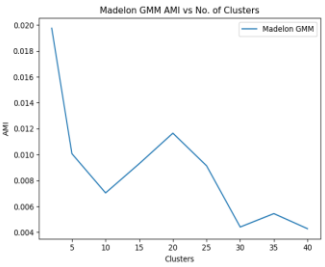


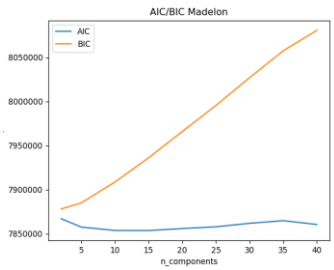
Madelon Analysis

In the case of Madelon, we can observer BIC values start to increase rapidly after 5 clusters, which indicates that the model complexity is very high at higher no. of clusters. AIC seems to be stabilized after 5 clusters, with a slight peak at 35. In terms of accuracy, we can notice that at 5 clusters the accuracy is the lowest and then peaking up at 25 clusters. We already interpreted the importance of cluster size of 5 during K means and the same reason applies here as well as there is a nice gaussian distribution among the 5-Dimensional coordinates of the hypercube summit. Hence, we will choose the no. of clusters as 5 here as well.









Part 2.

Dimensionality Reduction

Dimensionality Reduction is the technique of converting a set of data having vast dimensions into data of lesser dimensions by obtaining a set of principal variables. This part focus on apply the following dimensionality techniques to Madelon dataset – Principal Component Analysis (PCA), Independent Component Analysis (ICA), Random Projection (RP) and Random Forests (RF).

Methodology

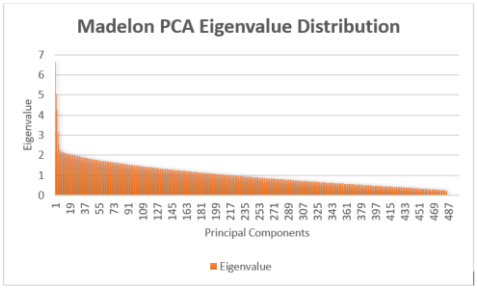
The four dimensionality reduction algorithms were applied using scikitlearn on Madelon and Letter Recognition datasets, after scaling its features. The algorithms were applied over a range of dimensions and for each dimension, the resulting data with lesser dimensions were fed to a neural network learner to observe the effect of the reduced dimensions.

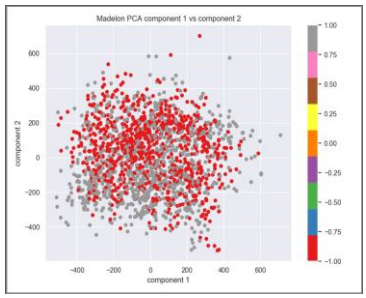
Principal Component Analysis

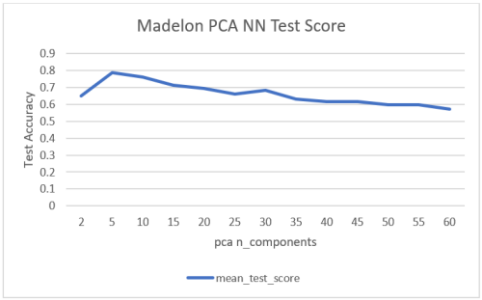
Principal component analysis is a technique used to transform a number of possibly correlated variables into a smaller number of uncorrelated variables called principal components. The first principal component will preserve the most variance in the data and the succeeding components accounts for as much of the remaining variance as possible.

Madelon

Based on the below distribution of eigenvalues over the 500 principal components, the most variance in the data is captured in the first 5 principal components. As higher variability indicates some sort of signal and little variability indicates noise, we can drop the unimportant components and hence reduce the dimensionality of the dataset to 5.







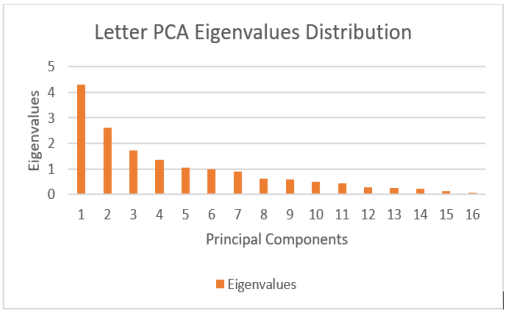
Below plot shows the performace of a Neural Network , trained using cross validation , on the Madelon dataset after reducing it's dimension from 2 to 60 principal components. As we can see, the Neural Network achieves the most accuracy when the dimension is reducted to 5.

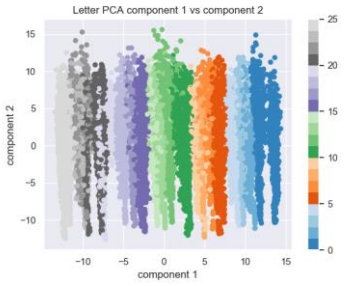
PCA algorithm was able to eliminate the noise in the dataset by projecting the samples to lower dimensions using their eigenvectors and also capture the maximum variance in the dataset at such lower dimensionality.

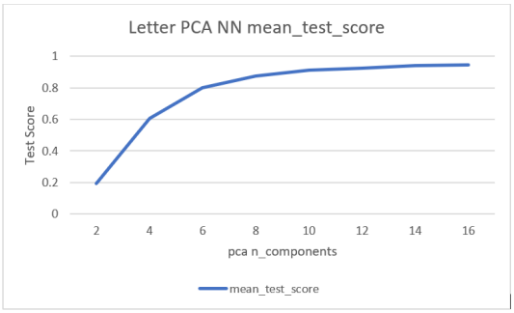
Hence , no. of principal components for Madelon dataset is chosen as 5.

Letter Recognition

Below plot on eigenvalue distribution shows maximum value of ~4 for 1st principal component with decreasing variance for the rest.



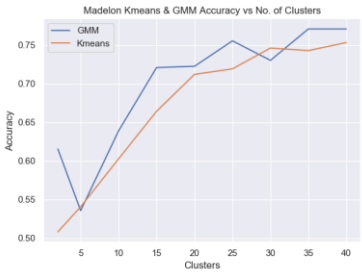


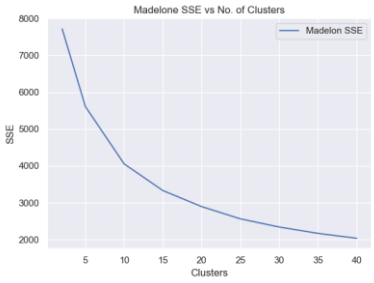


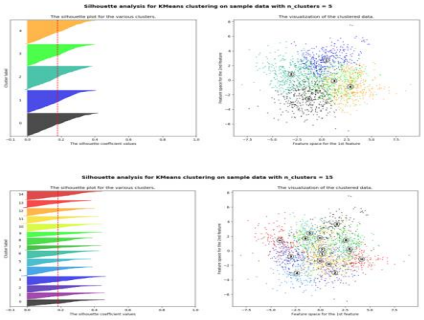
As per the cross validated Neural Network performance on varying the dimensionality of the dataset from 2 to 16, it can be observed that the maximum accuracy of ~94 % is achieved at principal components equal to 16, which is basically the same no. of dimensions that the dataset actually has. However, it can also be noted that reducing the dimension to 14 also achieves accuracy closer to ~94. Hence, the no. of principal components is chosen as 14 here.

Clustering Analysis

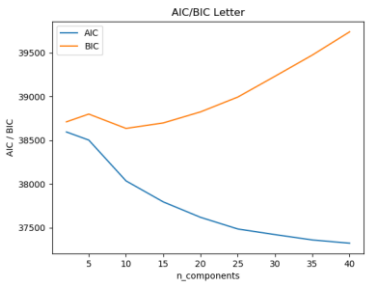
After reducing the dimension of madelon dataset to 5, clustering algorithms were applied again. We can note a significant drop in the SSE values, which goes to show the effect of removing noisy and unwanted features from the dataset.





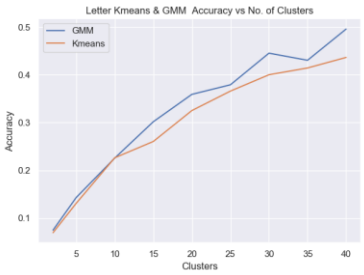


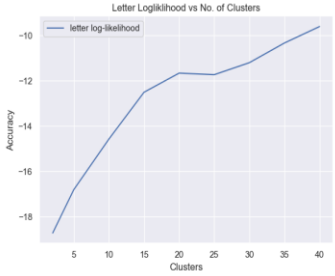
We could also find the AIC/BIC value have also decreased, which shows overall reduction in model complexity and now cluster = 10 looks promising in comparison with clustering done before dimensionality reduction, where we choose 5.

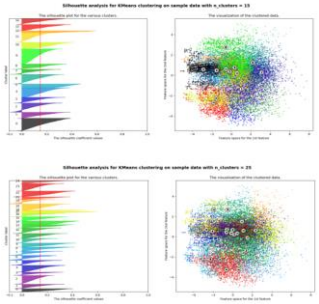


Based on the silhouette plots for clusters 5 and 15, which were chosen earlier as consistent with the dataset, shows better visualization and separation of the data points into groups and shows very less overlap between adjacent clusters. Hence, dimensionality reduction has not only removed the noise but also improved the spread of the data in the 2 principal component space.

When it comes to Letter Recognition, with the dimensionality reduced to 14 from 16, k means clustering algorithm did not show much difference in terms of SSE and Accuracy. However, after re-clustering, silhouette plots show better refinement of the data points for k=15 and 25. There is no notable change in AIC/BIC plots.







Independent Component Analysis

Independent component analysis attempts to decompose a multivariate signal into independent non-Gaussian signals. Here we use the FastICA algorithm for our analysis.

Madelon

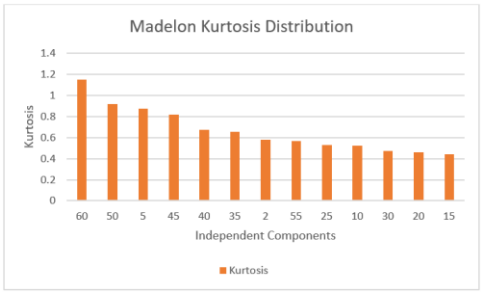
Kurtosis is defined as the normalized form of the fourth central moment of a distribution. Kurtosis measures the degree of peakedness (spikiness) of a distribution and it is zero only for Gaussian distribution. Kurtosis is calculated using pandas kurt api, which considers normal distribution as having 0 kurtosis.

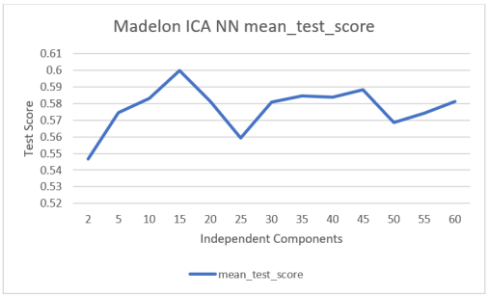
To maximize the statistical independence among the components, the maximum kurtosis should be considered, which is observed for ICA components = 60.

For each Independent component (ranging from 2 to 60), the ICA output was fed to a cross validated Neural Network learner and based on its accuracy, we could find peak performance for ICA components at 15 and 45.

As 15 has lower kurtosis than 45, and since 60 has lower accuracy than 45, the ICA components for Madelon was chosen as 45.

ICA was able to remove the noise present in the dataset and reduce the total no. of dimensions from 500 to 60 statistically independent components.

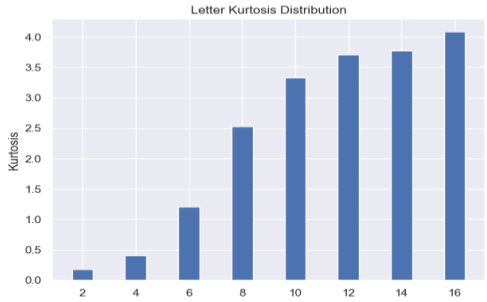


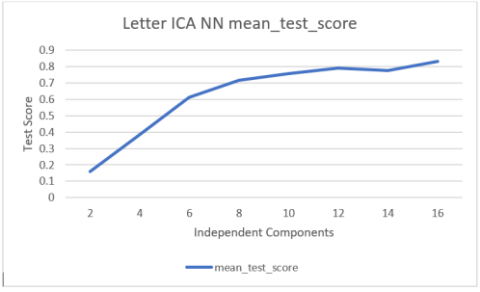


Letter Recognition

When it comes to Letter Recognition dataset, we could observe maximum kurtosis at IC=16 and based on the cross validated neural network performance, IC=16 provides the best performance.

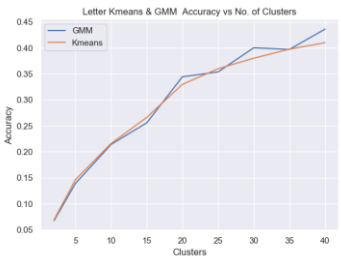
So, ICA does not seem to help to reduce the dimensionality of this dataset, mostly because many of the features of this dataset are independent already and is not a mixture of multiple signals. This can be attributed to the dataset having features which were statistically generated like measurements of the box, no. of pixels , mean variance etc.

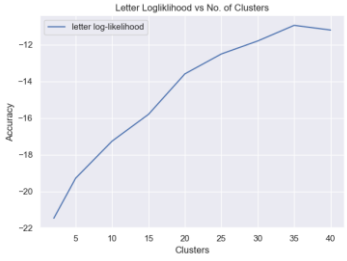


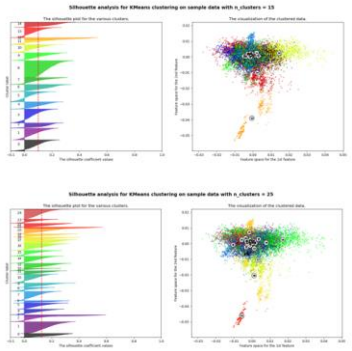


Clustering Analysis

Based on the below plots, Letter recognition did worse after the dimensionality reduction as the SSE error seems to have gone up.

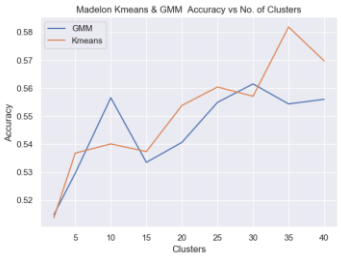


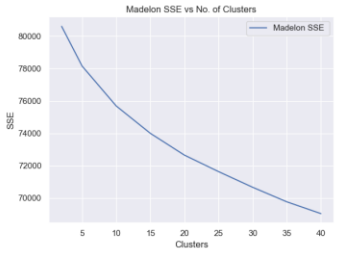


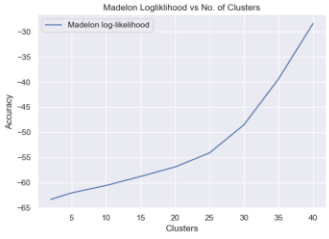


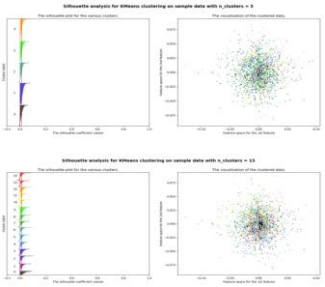
The silhouette plots shows the effect of using Independent components of 16, which seems to show that many samples were assigned to wrong clusters as they have silhouette coefficient of less than 0. The lowered accuracy may be because of further linear transformation of already independent features causing distortion in the data.

For Madelon, we can again notice a significant reduction in SSE due to the reduction in dimensions, hence eliminating noise from the dataset.









Based on the silhouette plots for k = 5 and 15, due to higher dimensionality compared to PCA analysis, the clusters doesnot look as visually separated and thick as PCA but does have better spread of cluster centres compared to clustering done on original dataset.

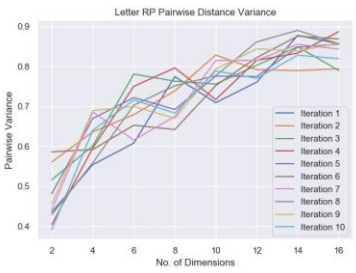
Random Projection

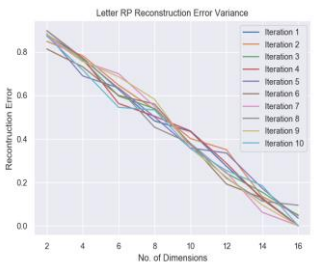
Random Projection is a dimensionality reduction technique which reduces the dimensionality by projecting the input space randomly to lower dimensional subspace. Here, we use the Sparse Random projection in scikitlearn to implement the algorithm. The Random projection is done using a sparse random matrix which maintains the distance between points in lower dimensions. To do the projection, the algorithm does not need to see the whole data and hence this is computationally faster than other dimensionality reduction algorithms like PCA.

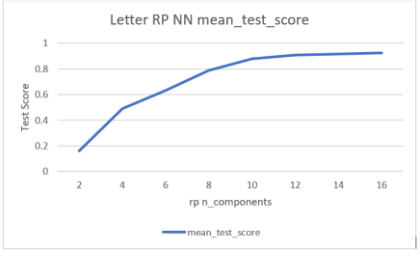
Letter Recognition

For all the 10 iterations of the algorithm implementation, the pairwise distance variance increases with increase in dimensions. We can observe high fluctuation in the curve which is questionable. For couple of iterations, lower dimension like 10 or 12 has higher value compared to higher dimension 16, which shows the consequence of randomized embedding.

Based on the cross validated Neural network accuracy, we can reduce the dimension of this dataset to 12, as the curve seems to flatten from that point onwards.

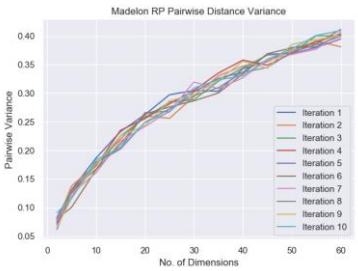


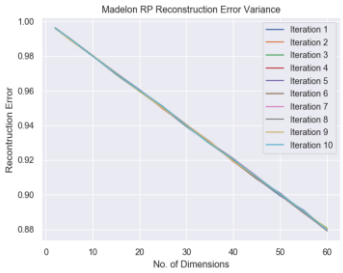


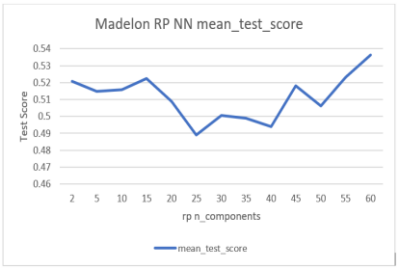


Madelon

In contrast to Letter Recognition dataset, we do not observe much fluctuation in the pairwise distance between the iterations. The cross validated Neural network has high accuracy at 60 dimensions and hence the same is chosen for further analysis. As the Neural network accuracy seems low, it seems like randomly projecting the data did not preserve the informative features and seems to have just distorted this dataset to lower no. of features.

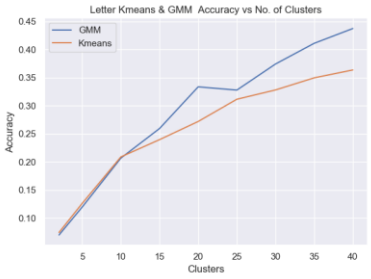


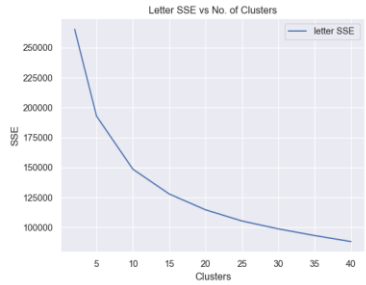


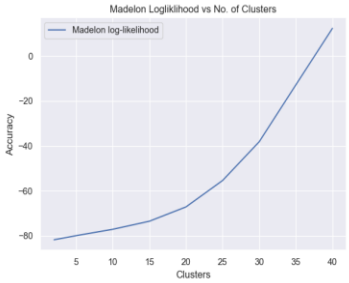


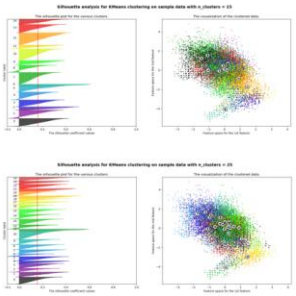
Clustering Analysis

Though we don't observe any decrease in SSE for Letter Recognition dataset, the silhouette plots show better cluster dispersion and based on SSE and Loglikelihood plots, an elbow method would suggest cluster size of ~26, which corelates to the no. of labels in the dataset.

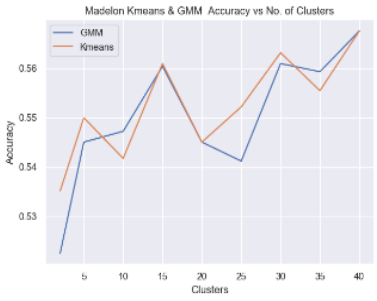


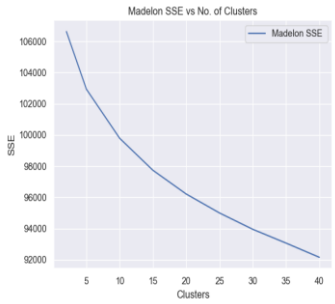


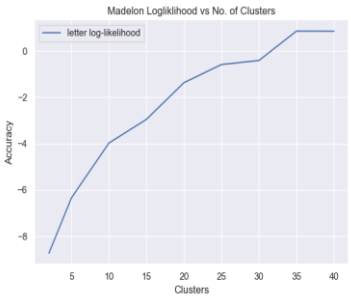


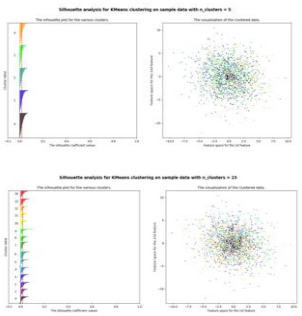


Madelon showed decrease in SSE scores due to the removal of 490 dimensions and silhouette plots shows bit better representation of the clusters as well. Using elbow method, we can assume cluster size of 10 as the optimal value here.









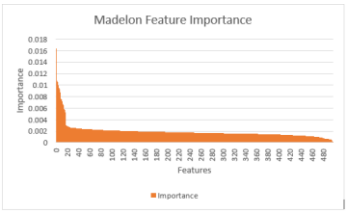
Random Forest

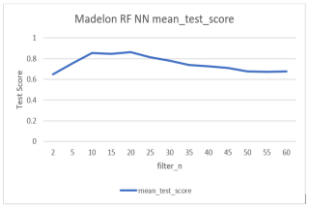
Random Forest classifiers can be used to perform dimensionality reduction as they rank the dataset features, based on its average weighted impurity reduction.

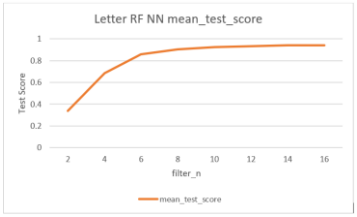
Based on the below plots, around 20 features in Madelon has higher ranking compared to the rest. As per the cross validated Neural Network performance, Madelon achieves high accuracy at 10 and 20 features and hence, we choose 10 features as the optimal value.

The below plots reconfirm our earlier statement that Madelon has lot of unpredictive features or noise and this algorithm has successfully portrayed the same.

With regards to Letter Recognition Dataset, 10 features have high ranking compared to rest and the Neural Network shows high accuracy for 10 features as well and hence the same is reckoned to be the optimal value for this dataset.

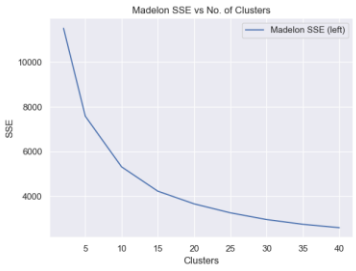


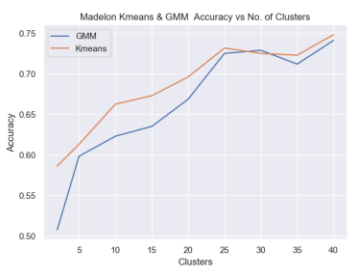


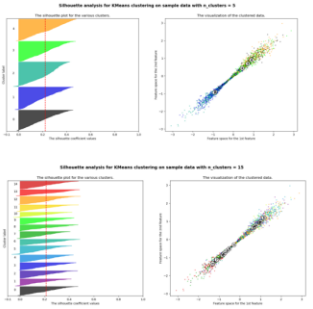


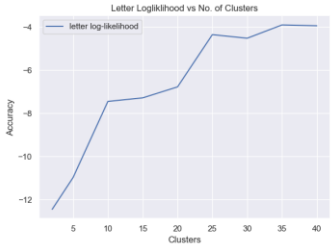
Clustering Analysis

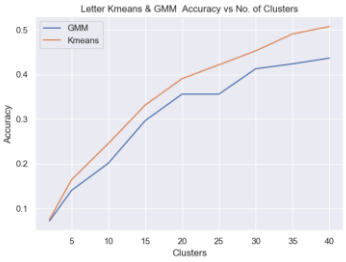
Based on the below plots , Madelon has the second best SSE improvement after PCA and Letter Rcognition has the best SSE improvement out of the rest. With 10 dimensions, the silhoutte plot cluster thickness for Madelon for cluster sizes 5 and 15 looks similar to that of PCA plots. Loglikelihood for Letter Recognition shows improvement and shows a spike around ~26, which seems to relate to it's no. of labels.

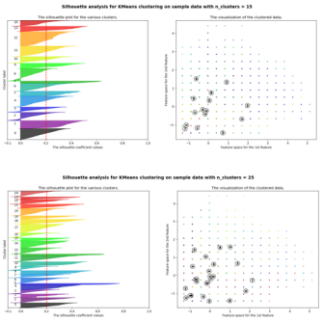










Part 3

Dimensionality Reduction Analysis on Madelon Dataset

Assignment 1 Results for Madelon - Benchmark

Dataset	Neural Network Parameters	Train Set Accuracy	Test Set Accuracy	Training Clock Time	Testing Clock Time
Madelon	Hidden Layer Sizes: (62,62), Alpha: Logistic, Activation: 1e-05	0.67	**0.56**	0.66s	0.02s

Analysis

We had achieved only ~56% accuracy for Madelon dataset in Assignment 1 , when we did not apply any dimentionality reduction techniques. After applying dimensionality reduction using the above four algorithms, we can observe that PCA was able to improve the accuracy to nearly ~83% , with just 5 principal components, while Random Forest was able to achieve ~70% accuracy with just 10 features. However, this came at the cost of increase in training time for the Neural Network for both algorithms. ICA performed decently well but couldn't achieve higher accuracy than PCA as it couldn't remove the noise completely using 45 components.Random Projection performed the worst with reduced accuracy compared to Benchmark, mostly because randomly

projecting the data points to lower dimensions broke the structure of this dataset, which was created using symmetrical clustered data points in 5 dimensions. However, Random projection took the least time compared to other algorithms.

Dimensionality Reduction Results for Madelon

Algorithm	Accuracy	Precision	Recall	F1-Score	Train Time (s)	Test Time (s)	Reduced Dimension
PCA	0.826	0.83	0.83	0.83	2.10	0.001	5
ICA	0.598	0.60	0.60	0.60	0.88	0.002	45
RP	0.492	0.49	0.49	0.49	0.40	0.001	60
RF	0.692	0.69	0.69	0.69	1.56s	0.004	10

Part 4

Clustering Analysis on Madelon Dataset

As part of this analysis, clustering algorithms – Kmeans and EM were applied to Madelon dataset and the resultant cluster attributes (excluding the dataset features) were used to run the same Neural Network used as part of Assignment 1. The K-means output, which was fed to the Neural Network, is the distance of the sample from the cluster centers (Euclidean), while for EM, it is the posterior probability of each component given the data (predict_proba). Hence, we reduce the dimension of the dataset by discarding its features and using cluster outputs as the reduced feature dimensions.

As we can observe from the below tables, both clustering techniques were able to achieve better accuracy than the benchmark, with K means providing the best with just 5 clusters and EM achieving the same with 15 clusters. EM also seems to be on the higher side in terms of the computational time required for its implementation.

K Means Clusters	Accuracy	Testing Time (s)	Training Time (s)
2	0.496336996	0.002101967	0.702190911
5	0.65018315	0.001903951	1.070042116
10	0.598901099	0.002653351	1.203867124
15	0.598901099	0.002364896	1.389344222
20	0.538461538	0.002856838	1.533476674
25	0.595238095	0.002440383	1.582200815
30	0.54029304	0.002094674	1.813100318
35	0.635531136	0.002679242	2.190123015
40	0.543956044	0.002805784	2.172790901

EM Clusters	Accuracy	Testing Time(s)	Training Time(s)
2	0.494505495	0.013445078	0.464931535
5	0.521978022	0.032007229	0.864271771
10	0.534798535	0.071874956	1.459928605
15	0.626373626	0.096274405	1.974219174
20	0.611721612	0.12615414	2.718547527
25	0.564102564	0.16772233	3.333641724
30	0.523809524	0.187220228	3.776103651
35	0.558608059	0.21898641	4.412788597
40	0.56959707	0.255054622	5.222561286

Conclusion

Clustering techniques is useful to find patterns or structure in your data and can be used as a pre-processing step to derive insights. If the data is highly dimensional, visualizing the clusters may be difficult and hence Dimensionality Reduction can help to overcome that.

Out of all Dimensionality Reduction Algorithm, Principal Component Analysis has achieved the best performance for Madelon dataset, as it was able to eliminate the unpredictable and redundant features by choosing new dimensions which preserves the maximum variance. Random projection, though having the fastest computation time, may not be suitable for well structured data as it trades accuracy over runtime.

When a dataset is noisy, clustering techniques, even though being an unsupervised learning technique, can be effectively used to identify structures in a dataset and the features of the optimal clusters could then be effectively used to reduce the dimension of the noisy dataset and further refine its supervised learning accuracy. This was observed with Madelon where it could achieve ~10% more accuracy just by using cluster features. It can also be noted that higher cluster numbers increase the time complexity of the clustering algorithms.

References

http://fourier.eng.hmc.edu/e161/lectures/ica/node4.html