# HW3 : Unsupervised Learning and Dimensionality Reduction

Sowmya Yellapragada (syellapragada3)

April 1, 2018

**Abstract**

In this paper, we implement and explore the performance of unsupervised learning algorithms. Particularly, two clustering algorithms - k-means and expectation maximization (EM), 4 dimensionality reduction algorithms - principal component analysis (PCA), independent component analysis (ICA), randomized projections (RP) and feature selection based on information gain have been implemented and discussed. Further, the neural network (established on original data) was rerun on the newly projected data, and the performance was compared.

## 1 Datasets

1. **Statlog (Landsat Satellite)**:
Soil classification is of particular importance in large and developing countries, like India, where huge percentage of the population still relies on agriculture. Broadly classifying the different soil types in the country, will help the government or organizations identify which crops would be more suitable to grow in which areas. Here we analyze the classification of soil into – red soil (1), cotton crop(2), grey soil(3), damp grey soil(4), soil with vegetation stubble(5), mixture class(6) and very damp grey soil(7). The database is a (tiny) sub-area of a scene, consisting of 82 x 100 pixels. Each line of data corresponds to a 3x3 square neighborhood of pixels completely contained within the 82x100 sub-area. Each line contains the pixel values in the four spectral bands (converted to ASCII) of each of the 9 pixels in the 3x3 neighborhood and a number indicating the classification label of the central pixel. Thus, there are 36 attributes, each of which has integer values in the range of 0 and 255. The dataset was classified into testing and training data sets, containing 4435 and 2000 examples respectively.

2. **Spam E-mail Database:**
Spam email is a major concern for most people using an email service. Important emails are often lost among the huge dump of spam emails. Simple classification based on what the user had marked spam is no longer sufficient. Machine learning algorithms have become common place for identifying trends among data that at the first sight looks random or unsuspicious. In this report we explore classification of email into spam (1) or not spam (0). Most of the 58 attributes in the data set indicate whether a particular word or character was frequently occurring in the e-mail. The dataset containing 4601 instances was classified into testing and training split containing 25%, 75% of the instances respectively.

## 2 Clustering

Clustering is a method of grouping a set of objects in such a way that objects in the same group (cluster) are more similar (in some sense or another) to each other than to those in other groups (clusters). Here, two clustering techniques - K-Means and Expectation maximization (EM) are implemented and discussed, for both the data sets

**Best K**: The performance of the clustering algorithm is analyzed for different cluster sizes by using the following metrics -

- Silhouette score : Measures the quality of a clustering. The best value is 1 and the worst value is -1. Values near 0 indicate overlapping clusters. Negative values generally indicate that a sample has been assigned to the wrong cluster, as a different cluster is more similar.
- Completeness Score : Measures the completeness metric of a cluster labeling given a ground truth. A clustering result satisfies completeness if all the data points that are members of a given class are elements of the same cluster. Score between 0.0 and 1.0. 1.0 stands for perfectly complete labeling
- Homogeneity Score : Measures the homogeneity metric of a cluster labeling given a ground truth. A clustering result satisfies homogeneity if all of its clusters contain only data points which are members of a single class. Score is between 0.0 and 1.0. 1.0 stands for perfectly homogeneous labeling.

  Notice that perfect homogeneity and clustering results in the perfectly classified clustering. However, in most cases, it is difficult to achieve to high performances on both those scores. We give greater importance to completeness score, as although a clustering may contain more than the desired clusters, but it would still predict correct outputs for any new inputs supplied to it.

## 2.1 K-Means

Simple clustering algorithm, that aims to form k clusters in the data set, by initially starting by starting at k different cluster centers, and iteratively add data points to each of the clusters based on similarity and also in the process re-evaluate centers until convergence have been achieved.

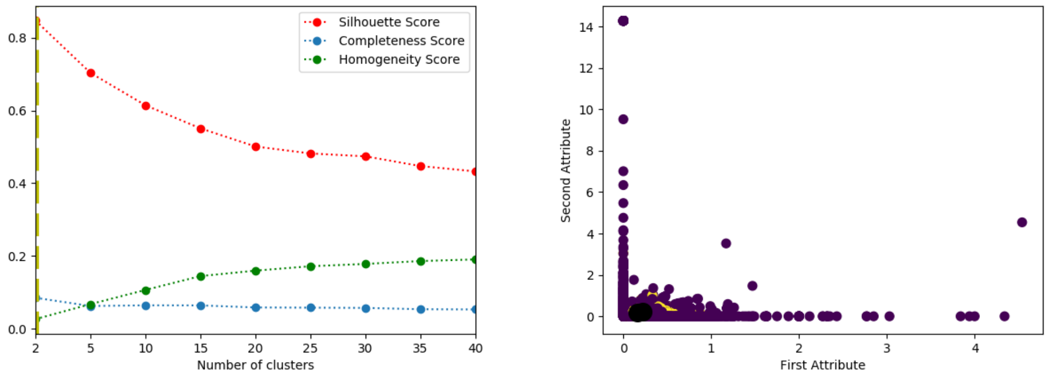
For the purpose of this analysis, this algorithm was implemented using the sklearn library's kmeans clustering algorithm. The analysis can be sensitive to the initial selection of centroids. Hence, we use the option of random restarts, that rerun this algorithm with randomly selected initially cluster centers and reports the best solution.

### 2.1.1 Spam Database

As can be observed from the graph(Figure 1a), the silhouette score, completeness score decreases as the cluster size increases. The decrease in the silhouette score is much more rapid, indicating the overlapping increases as cluster size increases, implying that the dataset is overly being classified or clustered. We'd expect the optimum cluster count to be around the number of class, which is 2 in the case of the Spam database. This can also be observed from the graph. Although the homogeneity score increases, with the cluster size, this only indicates, that the dataset is being over-fit by believing the data too much, ensuring no false positives on the train dataset. Also notice that, change in homogeneity and completeness score is not as much as the variation in silhouette score, hence it is fair in this case, to choose the ideal cluster size, based off the optimal value of the silhouette score, which in this case is 2

### 2.1.2 Statlog Database

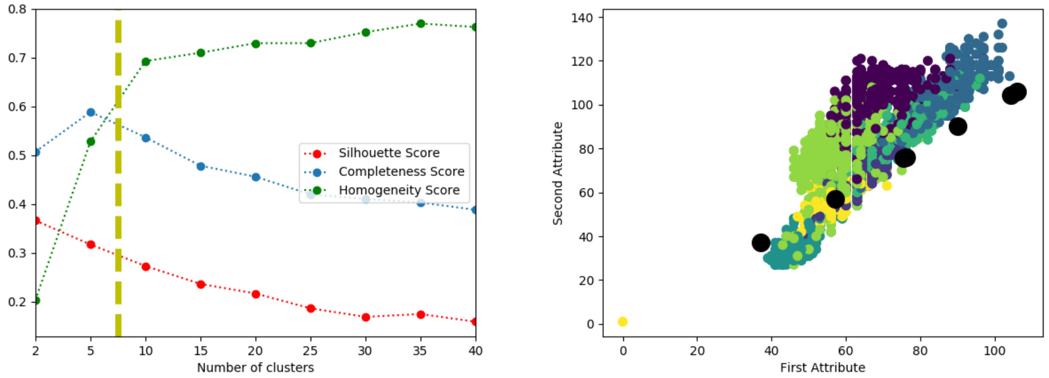
In this case (Figure 2a), the variations between the three scores are much more comparable. We would like to choose an optimal cluster size, with respect to all the three scores. This optimal value can be chosen between the cluster sizes 5 -10 , the values can be estimated to be either 6 or 7. This can be expected, as the number of classes in the Statlog is 7. Low Silhouette scores for this dataset indicate that the attributes are



(a) Comparision to choose best cluster size (Spam)

(b) Clustering visualization (k = 2)(Spam)

Figure 1: Clustering - Spam dataset



(a) Comparision to choose best cluster size (Statlog)

(b) Clustering visualization (k = 7)(Statlog)

Figure 2: Clustering - Statlog dataset

## 2.2 Expectation Maximization

Expectation maximization (EM) is a powerful algorithm that comes up in a variety of contexts within data science. This approach, similar to k-means has two steps :

- Guess some cluster centers
- Repeat until converged
  1. E-Step: assign points to the nearest cluster center
  2. M-Step: set the cluster centers to the mean

Here the "E-step" or "Expectation step" is so-named because it involves updating our expectation of which cluster each point belongs to. The "M-step" or "Maximization step" is so-named because it involves maximizing some fitness function that defines the location of the cluster centers — in this case, that maximization is accomplished by taking a simple mean of the data in each cluster.

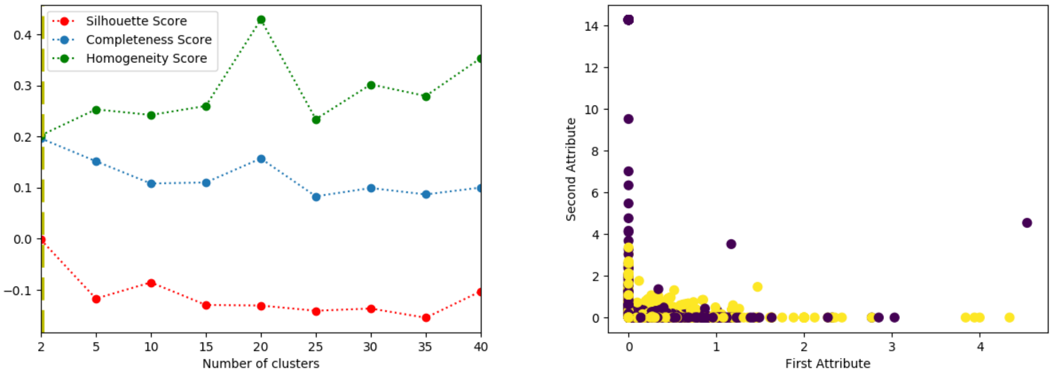
For the purpose of this analysis, this algorithm was implemented using the sklearn library's Gaussian mixture model was used. The Gaussian Mixture object implements the expectation-maximization (EM) algorithm for fitting mixture-of-Gaussian models. The analysis can be sensitive to the initial selection of centroids. Hence, we use the option of random restarts, that rerun this algorithm with randomly selected initially cluster centers and reports the best solution.

### 2.2.1 Spam Database

From Figure 3a, we can observe that the silhouette score for the spam dataset is tending to negative values as the cluster size increases, this indicates wrong classification, as discussed before. Hence, the only choice of k value would be 2. Notice that around k=10, homogeneity and completeness scores seems to attain a maximum value, but the negative silhouette score indicates wrongful classification. Hence, this cluster size can't be chosen. Figure 3b shows the visualization of the clustering with the help of two attributes of the dataset

### 2.2.2 Statlog Database

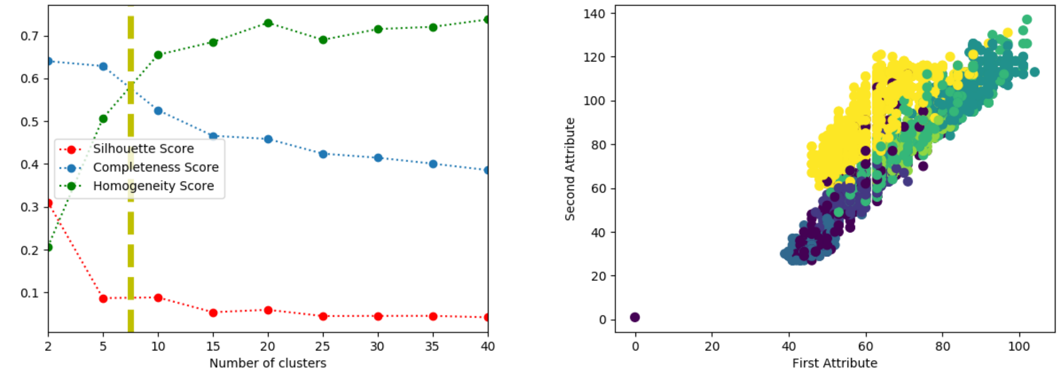
From Figure 4a, we can observe that, the algorithm produces very low silhouette scores with the increase in the cluster size. However, the completeness and homogeneous scores are high. The ideal classification would be the value where both these values are large. The intersection point of these two lines can be chosen as the best cluster size, which is observed to be 7. This is as expected equal to the number of classes.



(a) Comparision to choose best cluster size (Spam)

(b) Clustering visualization (k = 2)(Spam)

Figure 3: EM Clustering - Spam dataset



(a) Comparision to choose best cluster size (Statlog)

(b) Clustering visualization (k = 7)(Statlog)

Figure 4: EM Clustering - Statlog dataset

## 2.3 Comparison between the two methods

K-Means Clustering gives good clusters with high silhouette scores for Spam dataset, but the completeness and the homogeneity scores produced at the optimal cluster size = 2 are slightly lower than that produced by the EM algorithm. EM algorithm for Spam dataset produces only one valid clustering. This may be due to distribution and correlations between the features. The performance of both EM and K-Means algorithms are comparable for the statlog dataset.

# 3 Dimensionality Reduction and Analysis

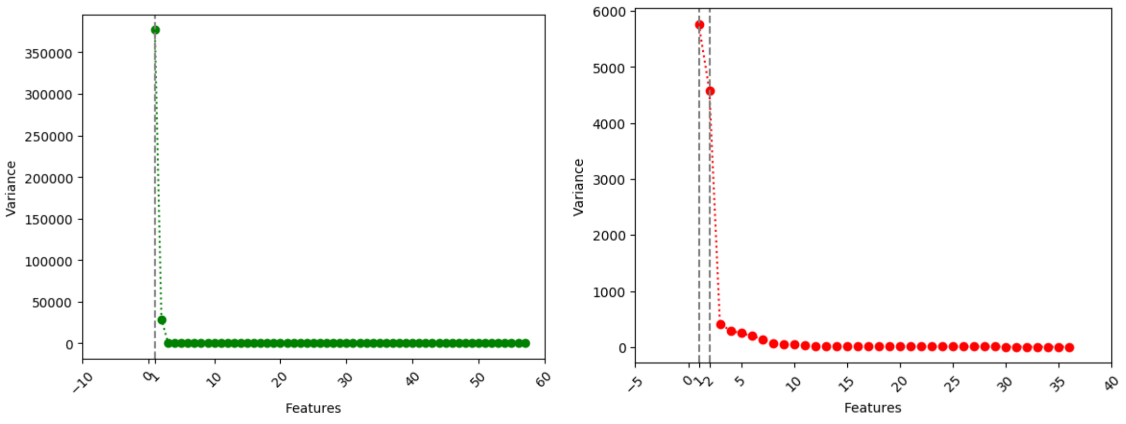
Dimensionality reduction is the process of reducing the number of random variables under consideration by obtaining a set of principal variables. Here we implement and discuss 4 methods to obtain this set of principal variables.
Further to analyze how the dimensionally reduced dataset performs, we apply clustering algorithms and compare performances with the analysis done in the above section.
**Methodology**:

- Applied dimensionality reduction algorithm to each of the dataset, and measured a metric which was used to determine the number of features to be retained.
- Feature with the max metric value was determined and then all features with metric values with at least 60% of this maximum value were considered important to be retained. We thereby determine the number of features to be kept in the reduced dataset
- We perform clustering algorithms on each of the dimensionally reduced datasets and observe the different scores measure on this new dataset, compared to the original dataset

## 3.1 Principal Component Analysis (PCA)

It is a statistical procedure that uses an orthogonal transformation to convert a set of observations of possibly correlated variables into a set of values of linearly uncorrelated. For this analysis we implemented PCA using sklearn's PCA class. Here we measure the variances which is the amount of data retained by a principal component(PC). For a good dimensionality reduction the reduced dataset should have variance that is at least 60%-80% of original variance.

Using the variance plots, we identify the feature that retains maximum variance of the original dataset, we will consider all the features that retain at least 60% of this maximum value. As can be clearly observed from the graph for Spam dataset, the only significant feature is the first attribute and no other features comes close in comparison. Hence, it is safe to choose the number of dimension to reduced to by PCA to be 1 for this dataset. Similarly for Statlog dataset, the dimensions to be retained can be set to 2. We then generate a dimensionally reduced



(a) Dimensionality Reduction on Spam Dataset

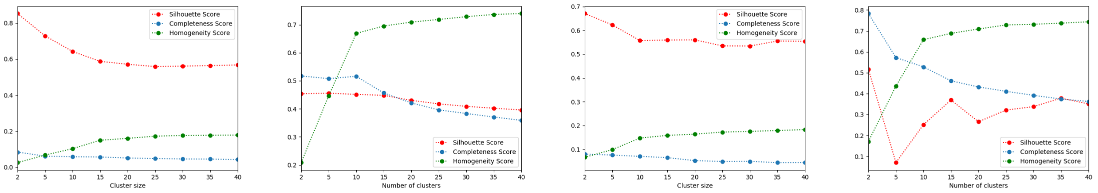
(b) Dimensionality Reduction on Statlog Dataset

Figure 5: PCA dimensionality reduction

### 3.1.1 Clustering analysis

The K means clustering graph for PCA reduced Spam dataset seems to be nearly identical to the original clustering graph. This can be expected as the explained variance was extremely high for this selected dimension and the contribution for every other feature was largely minimal. Hence here too, the optimal clusters can be chosen to be 2. The EM clustering graph in contrast to K Means for Spam dataset is quite from the original analysis. Particularly note worthy is the high Silhouette scores, indicating disjoint clusters, this can be expected because now the clustering is on just one most relevant attribute. The best cluster count would then again be 2 for this data set.

The K Means clustering graph for the Statlog dataset in contrast is quite different from the original k means graph for this dataset. This can be explained as, although we have retained the features with high explained variance, the contribution of the other features is not entirely negligible .Particularly noticeable is the silhouette score, which has improved for higher values of the cluster count, indicating. From the graph we can clearly conclude that the optimal number of clusters would be 10, in contrast to the previously chosen value, 7. The EM clustering graph for the reduced Statlog dataset shows improved Silhouette scores as well. The optimal cluster count seems to be 2, this quite contrary to our intuition as the number of classes in this dataset is 7. This might indicate that the dataset has a high bias towards a few classes.

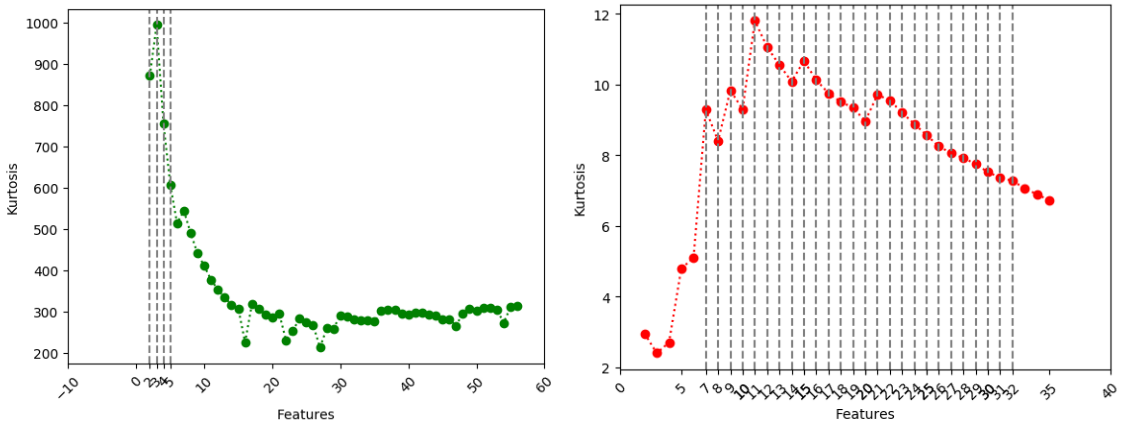


(a) Reduced Spam (KM) (b) Reduced Statlog (KM) (c) Reduced Spam (EM) (d) Reduced Statlog (EM)

Figure 6: (a, b) - KMeans Clustering Analysis; (c, d) - EM Clustering Analysis on PCA reduced dataset

## 3.2 Independent component analysis (ICA)

It is a computational method for separating a multivariate signal into additive subcomponents. This is done by assuming that the subcomponents are non-Gaussian signals and that they are statistically independent from each other. For this analysis used a sklearn's 'fastICA' library is used. In order to measure non-gaussianity we use a parameter called kurtosis. ICA does not work if the independent components have a guassian distribution. Hence we choose the features that have high kurtosis values. As can be seen, the kurtosis values of the features in Statlog dataset are close to one other than than the features in Spam dataset. The choice of features is as indicated by the vertical lines in the graphs below



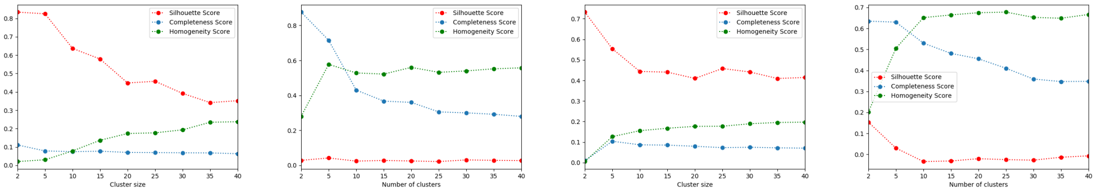
(a) Dimensionality Reduction on Spam Dataset (b) Dimensionality Reduction on Statlog Dataset

Figure 7: ICA dimensionality reduction

### 3.2.1 Clustering analysis

The trend of the scores in K Means clustering on the Spam dataset is similar to the original dataset. Observe that the features chosen in ICA includes the features selected in PCA. The k means clustering on the reduced Statlog dataset produces higher completeness scores at lower number of clusters, but at the cost of low silhouette values.

EM clustering on Spam dataset interestingly produces comparable completeness and homogeneity , but high silhouette scores, indicating correct classification of the dataset into clusters. The best cluster size would be 5 for this dataset. EM clustering on Statlog dataset produces higher completeness score but relatively lower silhouette scores in comparison to the original analysis. The optimal choice of number of clusters for this dataset would thus be 5.

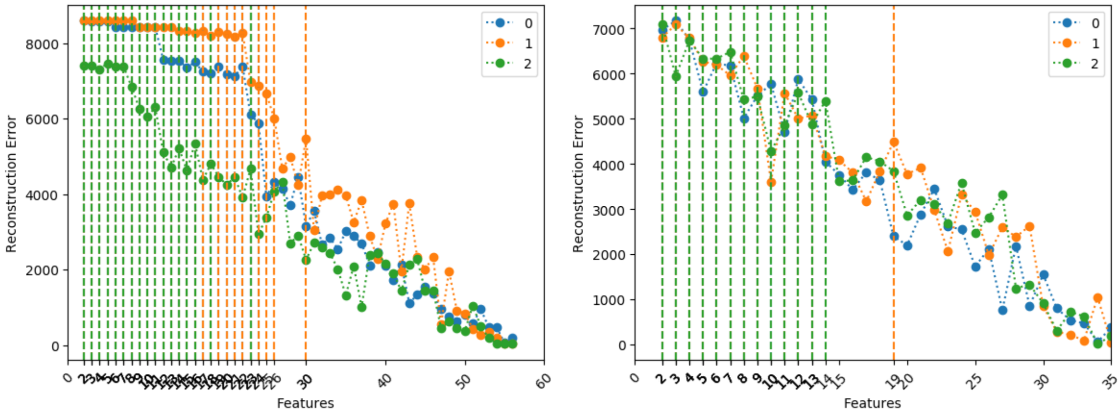


(a) Reduced Spam (KM) (b) Reduced Statlog (KM) (c) Reduced Spam (EM) (d) Reduced Statlog (EM)

Figure 8: (a, b) - KMeans Clustering Analysis; (c, d) - EM Clustering Analysis on ICA reduced dataset

## 3.3 Randomized projections (RP)

It is another dimensionality reduction method used to project 'n' total attributes in to k-dimensional space where k $\ll$ n. It is similar to PCA but the directions of projection are independent of the data. It serves as an efficient way to reduce high dimensional data while preserving distances between instances. For this analysis the randomized projections was implemented with the sklearn library's SparseRandomProjection. We use reconstruction error to choose the features that need to be retained in the process of dimensional reduction. We perform the random projection on the datasets three times over all possible choices of features that can be chosen for the sparse random projection, to avoid chance based anomalies. The best choices of the features are indicated by vertical lines in both the following graphs.



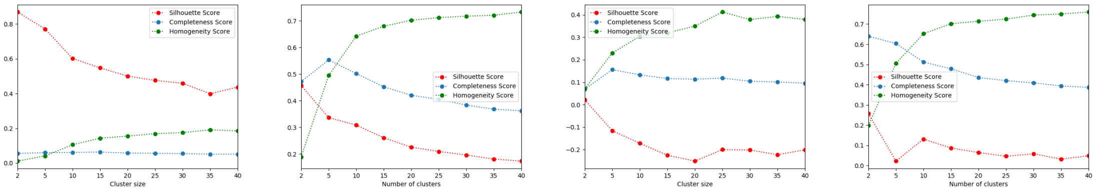
(a) Dimensionality Reduction on Spam Dataset (b) Dimensionality Reduction on Statlog Dataset

Figure 9: RP dimensionality reduction

### 3.3.1 Clustering analysis

It is interesting to note that the K-Means clustering graph produced by PCA reduced and RP reduced methods are strikingly similar. K-Means on the RP reduced Statlog dataset produces much better similar completeness and silhouette scores but at lower homogeneity scores. The optimal choice for K value for this data set would be 5.

EM analysis on Spam dataset interesting produced negative values on Silhouette score for the cluster counts $> 2$. This, as talked about previously, seems logical as the class size of spam dataset is 2. But this clustering however produces low homogeneity and completeness scores again. The EM analysis on this reduced Statlog dataset is strikingly similar to the original analysis. The best choice of the number of cluster count for this case would be 5

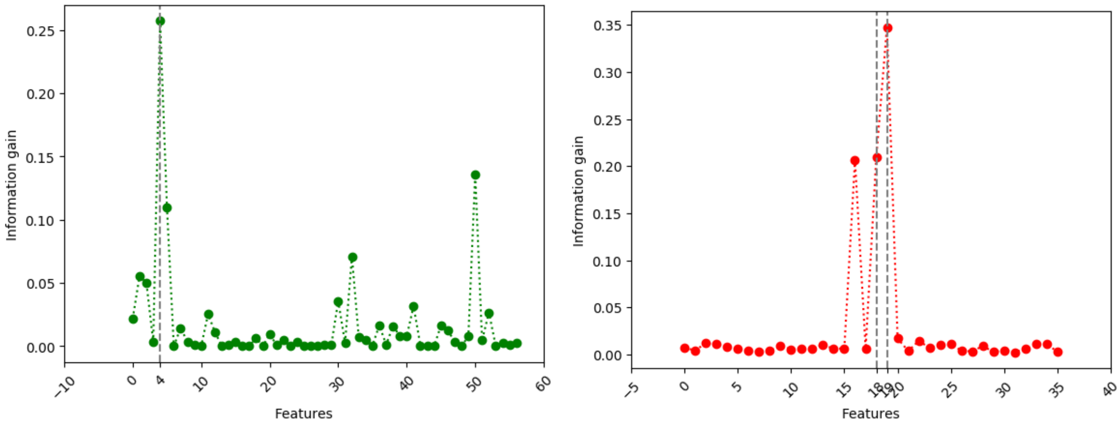


(a) Reduced Spam (KM) (b) Reduced Statlog (KM) (c) Reduced Spam (EM) (d) Reduced Statlog (EM)

Figure 10: (a, b) - KMeans Clustering Analysis; (c, d) - EM Clustering Analysis on reduced dataset using Randomized projections

## 3.4 Feature selection based on information gain

In this analysis method, we evaluate the worthiness of an attribute by measuring the information gain with respect to the class. It works on the same principal using which decision trees evaluate information gain to determine which attribute goes on top node for best splitting at every level in the tree. For both the datasets, we choose the features that have high information gain w.r.t the class. We first choose the feature with maximum values, and also choose the number of features that have an info gain atleast 60% of the maximum value.



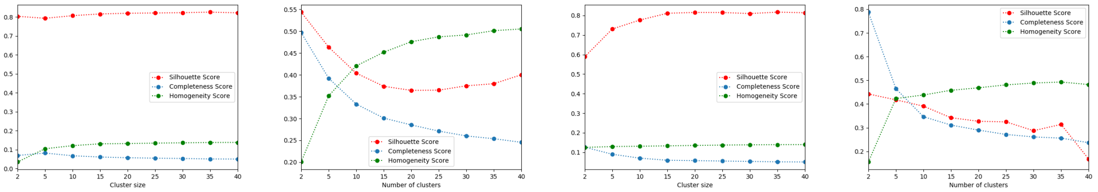
(a) Dimensionality Reduction on Spam Dataset (b) Dimensionality Reduction on Statlog Dataset

Figure 11: Information gain based dimensionality reduction

### 3.4.1 Clustering analysis

K-clustering on the reduced Spam dataset produces relatively constant silhouette, completeness and homogeneity scores over the different cluster sizes, silhouette scores is very high. The k clustering on the reduced Statlog dataset produces high values of silhouette and completeness scores for lower number of clusters. The optimal cluster choice would be 5.

EM on the reduced spam dataset produces low completeness and homogeneity scores but high silhouette scores. The EM analysis on Statlog dataset has an interesting variations of the the scores over the number of clusters. High completeness scores are observed for low cluster numbers.



(a) Reduced Spam (KM) (b) Reduced Statlog (KM) (c) Reduced Spam (EM) (d) Reduced Statlog (EM)

Figure 12: (a, b) - KMeans Clustering Analysis; (c, d) - EM Clustering Analysis on reduced dataset using information gain

### 3.5 Conclusion

Different dimensionality reductions on Spam dataset, produce similar completeness and homogeneity scores with EM clustering and K-Means Clustering, indicating an inherent bias in the dataset. EM clustering on the different dimensionality reductions on Statlog dataset produces relatives low silhouette scores. K-means on information gain based feature selection , PCA and randomized projections of Statlog dataset produces the best results. It is interesting to note that these reduction algorithms serve as an efficient way to reduce high dimensional data while preserving distances between instances, and hence it would seem consequential that K-Means works well. This implies that the Statlog dataset can be visualized a space of instances where distances are preserved. The attributes in the Statlog dataset are nothing but spectral bands of pixels in neighborhood. Hence, clearly the distance between instances is quite critical, hence we could have expected that K Means would work best for this dataset, and now this hunch has been confirmed by our analysis.

## 4 Neural Networks on projected data

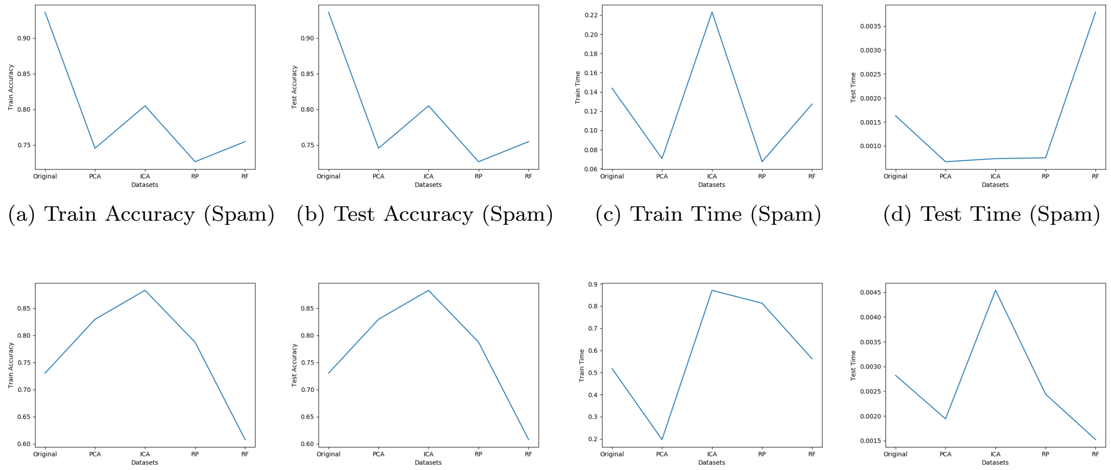
In this section, we want to compare the performances of a neural networks on the original datasets to that of the performance of the same learner on the projected data. Our analysis will be divided into two sections : first, will be on dimensionally reduced data, second on clustered data.
The neural network used for Spam data set, had 1 hidden layer of count 15. For Statlog data set, the neural network used contained 3 hidden layers of 25 in each.

### 4.1 Dimensionality Reduction

Here we compare the performance of the neural networks on the original dataset to the dimensionally reduced data set. We used the same neural network across the dimensionally reduced dataset as well.

We observe that dataset produced from PCA and randomized projects dimensionality reduction, significantly reduce the training time for the Spam dataset. However, they also produce lower the test and train accuracy. ICA reduced spam dataset has a significantly twice as large training time, but produces accuracy levels better than PCA and RP reduced datasets. This might indicate stochastic independence of the class variables, but that can't confirmed, as the accuracy levels on the original set were much higher.

The neural network performances on dimensionally reduced Statlog dataset gives us interesting insights. The train accuracy on ICA reduced dataset is far superior than the original dataset, although at the cost of increased training time. The improved accuracy levels on the ICA reduced Statlog dataset indicate that the class variable are statistically independent from one other. Hence, identifying / predicting them can be done best with the help of ICA reduction. This is similar to source identification in the cocktail party problem.



(a) Train Accuracy (Spam) (b) Test Accuracy (Spam) (c) Train Time (Spam) (d) Test Time (Spam)

(e) Train Accuracy (Statlog) (f) Test Accuracy (Statlog) (g) Train Time (Statlog) (h) Test Time (Statlog)

Figure 13: Performance of the metrics of neural network learner

## 4.2 Clustering

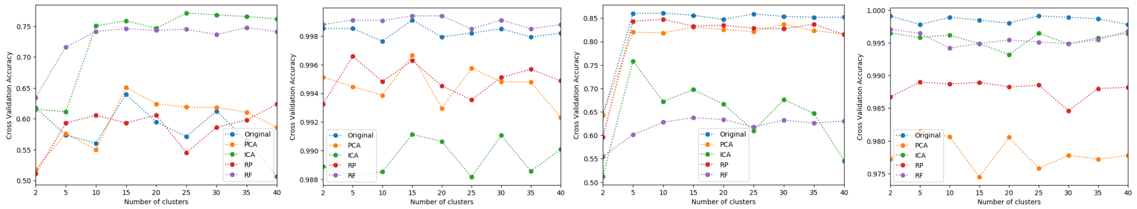
In this section, we measure the cross validation performance of the transformed datasets obtained after performing clustering on the datasets. We perform this experiment even on the dimensionally reduced datasets.

For KM clustering we use the fit transform function to compute clustering and transform X to cluster-distance space. To this transformed X, we append the clustering label additionally and then use the dataset, along with the true Y labels to measure the cross-validation errors for different K values.

For EM clustering, we use predict probna function after fitting the X samples. This function returns the probability each Gaussian (state) in the model given each sample. We use this transformed X, along with the true Y labels to measure the cross-validation errors for different number of components.

We can observe from the graphs the EM clustering performs exceptionally well for on the RF reduced Spam database, indicating that information gain is an important criteria in the prediction process on this dataset. We can thus assume that decision trees would work well on this dataset. High accuracy values at 5 and 20 cluster size, as opposed to 2 on the non transformed dataset.

For the Statlog dataset, both PCA and randomized projects seem to produced good accuracies on the KM transformed space, while on EM transformed dataset, RF and ICA give good accuracies. This seems intuitive, as both PCA and randomized projections serve as an efficient way to reduce high dimensional data while preserving distances between instances, and hence it would seem consequential that K-Means works well, as discussed before.



(a) KM Clustering on Spam (b) EM Clustering on Spam (c) KM Clustering on Statlog (d) EM Clustering on Statlog

Figure 14: Neural network performance analysis on clustered dataset

# 5 References

- Statlog data set : Ashwin Srinivasan, Department of Statistics and Data Modeling, University of Strathclyde, Glasgow, Scotland - UCI Machine Learning repository
- Spam Database : Mark Hopkins, Erik Reeber, George Forman, Jaap Suermondt Hewlett-Packard Labs, Palo Alto - UCI Machine Learning repository