Project 3: Unsupervised Learning

Andrew Scott
ascott97@gatech.edu

Abstract—This report examines several different methods of clustering and dimensionality reduction on classification data sets.

1 INTRODUCTION

This report details the analysis done on two different clustering methods and four different dimensionality reduction methods. The clustering methods include K-Means and Expectation Maximization and the dimensionality reduction methods included Principal Component Analysis, Independent Component Analysis, Random Projection and Variance Threshold feature selection. Each of these methods are analyzed in detail in the sections below for two separate classification datasets. Additionally, all the methods are used to transform the datasets and then analyze their effectiveness training a neural network to classify the data.

2 DATA SETS

Two data sets were used in the following analysis, the Wine dataset, and the Breast Cancer dataset. The Wine data set has 1599 data points which corresponds to different wines and the features for each data points includes 11 different measurements that describe the physical and chemical properties of the wine. The labels for this data set is a quality score given by a group of wine tasters on a scale of 1-7, with 7 being the highest quality and 1 being the lowest. The breast cancer data set has 682 data points which corresponds to different patient data. There are 9 different features for this dataset that are attributes for lumps/tumors observed for each patient. The labels for this dataset are either a positive or negative for a cancer diagnosis.

3 CLUSTERING

3.1 K-Means

For K-Means the KMeans API from Sklearn was used to cluster the data from both the Wine and Cancer Data sets. The plots below show the silhouette score and the within cluster sum of squared distance (WCSS) for both the wine and cancer datasets. These are plotted against the number of clusters (k values).

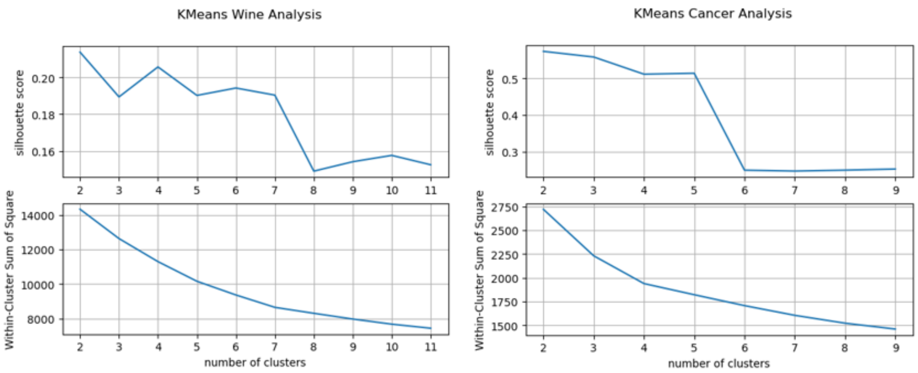


Figure 1: K-Means: K selection for Wine and Cancer Datasets

The process used to select the optimal k value was to look at both the silhouette score of the data in conjunction with using the elbow technique. The elbow technique involves identifying the point in the WCSS plot where the plot has the sharpest change in slope. This appears to occur at k = 7 for the wine data and k = 4 for the cancer data. The process for the silhouette score is simply to look at the values for k that have the highest score. While the silhouette score is not the highest for either k values for each of the data sets, the score is still relatively close to the highest score, and it is prior to the score "dropping off" in both of the plots. The plots above seem to suggest that it might be useful to use both metrics to select the correct k value since just looking at either silhouette score or the elbow method in isolation might not yield the best information. In the case of the above datasets the elbow method seemed to be sufficient for selecting the k-value; however, the silhouette score still provides the opportunity for added confidence in the selection as a crosscheck.

3.2 Expectation Maximization

For expectation maximization the GaussianMixtures API from Sklearn was used to cluster the data from both the Wine and Cancer Data sets. The metrics used to analyze the clustering performance were Akaike Information Criteria (AIC) and Bayesian Information Criteria (BIC). The plots below show the results for the both the wine and cancer datasets.

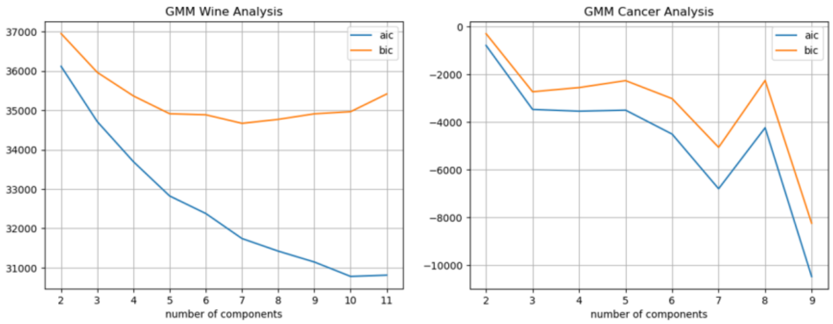


Figure 2: GMM K Selection for Wine and Cancer Data

An interesting thing to note when comparing the plots between the wine and the cancer data sets is that the AIC/BIC curves align very closely for the cancer data set while the curves do match up as closely for the wine dataset. Since the BIC curve also considers the number of elements in the dataset this indicates that the maximum likelihood value calculated for the wine dataset is not enough to compensate for the increased model complexity for higher number of clusters. The method used to select the number of clusters involves looking at the lowest scores for both AIC and BIC. In the wine dataset the value of 7 was chosen since the BIC seems to be at a minimum at this point, while AIC is still relatively low. The BIC and AIC scores were at a minimum at 9 clusters for the cancer dataset but there also appears to be another local minimum value at 7 so this point was chosen to avoid cluster overfitting.

The table below summarizes the homogeneity, completeness and v-measure scores for the wine and cancer datasets for both K-means and GMM for the selected number of clusters.

Cluster Type	Dataset	Clusters	Homogeneity	Completeness	V-Measure
K-Means	Wine	7	0.1238	0.0869	0.1021
K-Means	Cancer	4	0.8181	0.3420	0.4824
GMM	Wine	7	0.113	0.0750	0.0902
GMM	Cancer	7	0.7161	0.3367	0.4581

All three of the metrics seems to be close when using K-means and GMM for each dataset; however, all the scores are higher for the K-Means clustering than they are for the GMM clustering. It's possible the reason for this is that GMM is converging to some local optima in both datasets that it not optimal for the clustering techniques.

4 DIMENSIONALITY REDUCTION

4.1 Principal Components Analysis (PCA)

To determine how many principal components to use to reduce the dimensionality of the data the explained variance ratio was analyzed while running the PCA algorithms. The explained variance value corresponds to the eigenvalue for each component and essentially ranks how important that specific component is as well as how much variance it contributes. The plot below shows the explained variance for the wine data starting with only one component all the way up until the number of components equals the number of features from the original data set. Here, the explained variance is a cumulative sum that is added each time a new component is included. The component values along the a-axis are ordered such that the first component is the most significant while the final component is least significant, which is why the curve in both plots starts out steeper and then shallows out.

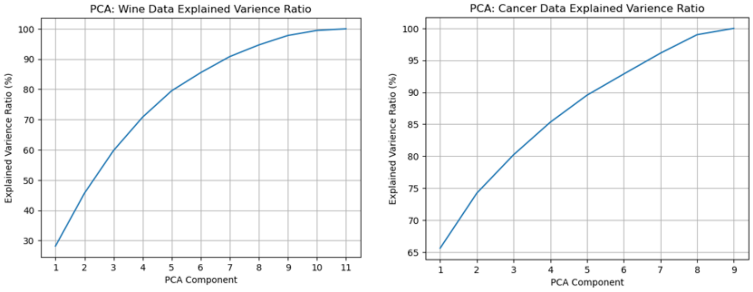


Figure 3: *PCA: Wine and Cancer Data - Explained Variance Ratio*

There are several different heuristics that can be used to select the target number of components when using the PCA algorithm. In this case, the number of components was selected by looking at the point the explained variance is greater than 90%. This threshold value was chosen to preserve as much information in the data as possible while also attempting to avoid overfitting. In this case, that occurs once there are 7 components for the wine data and 6 components for the cancer data. One interesting to note between the two datasets is that the wine data seems to have a more consistent decrease in contribution for each component which gives it a more parabolic shape. The first component of the cancer data set account for 65% of the variance, while most of other components equally contribute around 5-10% of the variance which is why the curve appears to be more linear.

After running the PCA algorithm on both datasets with the reduced number of components, both reduced datasets were both run through the K-Means and GMM clustering algorithms to assess and inspect the effect that the dimensionality reduction has on the clustering performance. The plot below shows the K-Means performance for both the wine and cancer datasets as a function of the number of clusters.

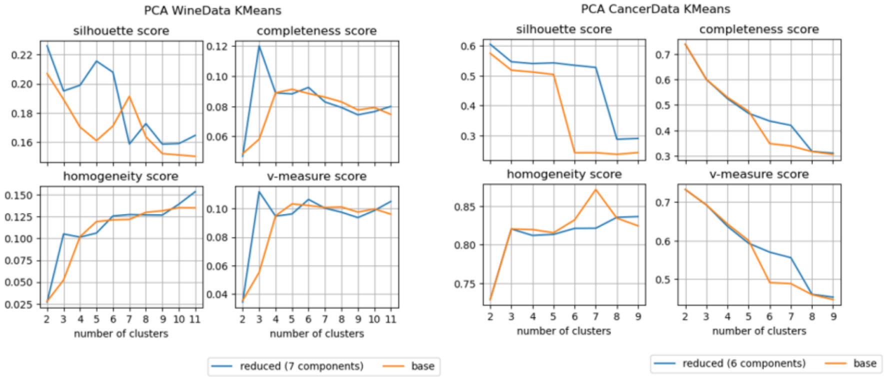


Figure 4: *PCA - Wine and Cancer K-Means Metrics*

For the wine dataset there seems to be an improvement in the clustering metrics, particularly for the lower clusters. This improvement is likely since PCA is removing features that do not contribute much to the variance in the dataset. This is sort of like removing the added noise in the data, which probably makes it easier for the clustering algorithms to have more completeness since there will be less samples with varying labels present in other clusters. The results from running the GMM data aligned closely with the k-means results when comparing the same score metrics in the plots above.

4.2 Independent Components Analysis (ICA)

The method used to choose the optimal number of components for ICA involved looking at the kurtosis of the dataset after applying ICA dimensionality reduction algorithm for various values of the number of components. Kurtosis is a measure of how closely a set of variables does not match a Gaussian distribution. In the plots below the kurtosis value shown is the average kurtosis of all the datapoints for each component. Additionally, the kurtosis value is normalized to zero so that higher numbers correspond to higher average kurtosis among all the components.

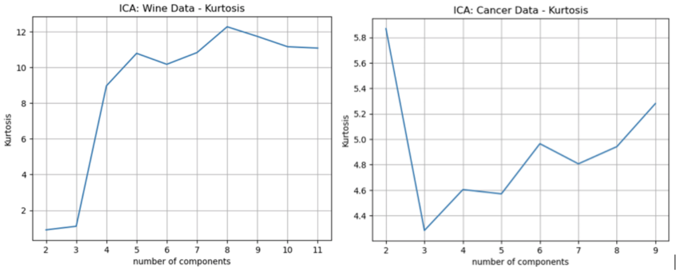


Figure 5: *ICA Wine and Cancer - Kurtosis*

The plot for the wine data indicates a clear point, when number of components = 8, where the average kurtosis is highest. This implies that reducing the number of components down from 11 (original number of features) to 8 helps increase the non-gaussian-nity of the data and thus this seems to be a useful heuristic for determining the appropriate dimensionality reduction of the wine data. The Cancer dataset is a bit tricker since the kurtosis seems to be highest when the number of components is 2. Although, the kurtosis does not seem to vary nearly as much with the cancer data as the wine dataset. Thus, another useful metric to look at with ICA is the reconstruction error as a function of the number of components.

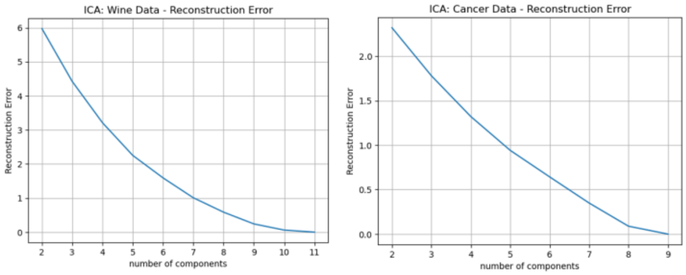


Figure 6: *ICA - Reconstruction Error with Wine and Cancer Data*

The reconstruction error essentially shows how well the data can be put back together using the transformed dataset and the ICA components. The reconstruction error starts out high for the wine data and decreases significantly for 8 components, which supports that 8 components is likely a good choice. The reconstruction error is not very large for the cancer data even for 2 components. Accordingly, given the higher kurtosis and relatively low error, 2 components might be the best choice for the cancer dataset.

The plots below show the results of running K-Means and GMM clustering for different cluster sizes on the data post ICA using the optimally chosen component values. One interesting aspect, particularly with the cancer data is that there seems to be a unanimous improvement with all the scores for GMM. This might indicate that some features in the original dataset are redundant, given that the v-measure score is consistently higher after being transformed via ICA to have only 2 components. The silhouette score was also higher for the cancer data set after running k-means, which is consistent with GMM; however, the homogeneity and completeness scores did not show the same improvement as they did for GMM. Its

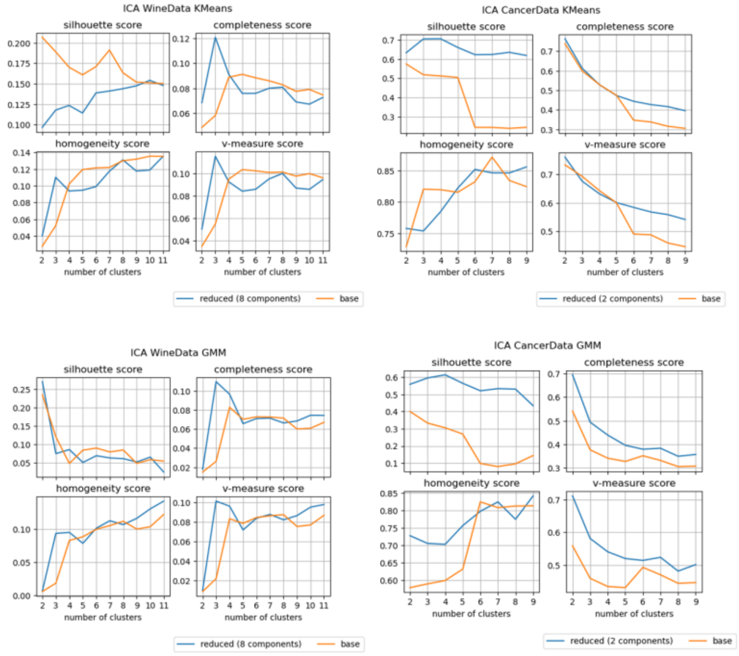


Figure 7: *ICA - Wine and Cancer GMM Metrics*

4.3 Random Projections (RP)

For Random Projections (RP) the reconstruction error was initially used as a method for choosing the optimal number of components. However, this turned out to not be the best heuristic for choosing the number of components since the reconstruction error essentially constantly decreased as the number of components were increased. As an alternative, the pairwise Euclidean distance was calculated for both the original and reconstructed data and for each component value the squared differences between the

Euclidean distance matrices were averaged. This produced a more useful curve that gave more insight as to when the reconstruction of the original data seems to look more promising. The plots below show these curves for both the wine and cancer data. Also, given the random nature of the RP method, the plots below were obtained by looking at the average Euclidean distances for 10 runs of the RP method.

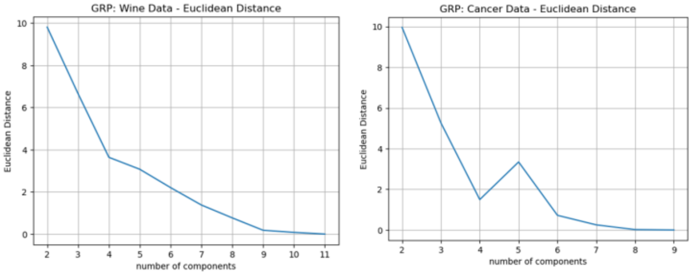


Figure 8: *RP Wine and Cancer - Euclidean Distance Comparison*

Using a heuristic like the elbow technique used for k-means clustering the number of selected components was selected as 9 and 6 for the wine and cancer data respectively. The idea behind this approach was to select a value for components that persevered the ability to reconstruct the data while still endeavoring to avoid overfitting.

The plot below shows the K-means and GMM clustering applied to the wine and cancer data after being reduced to the selected number of components through RP.

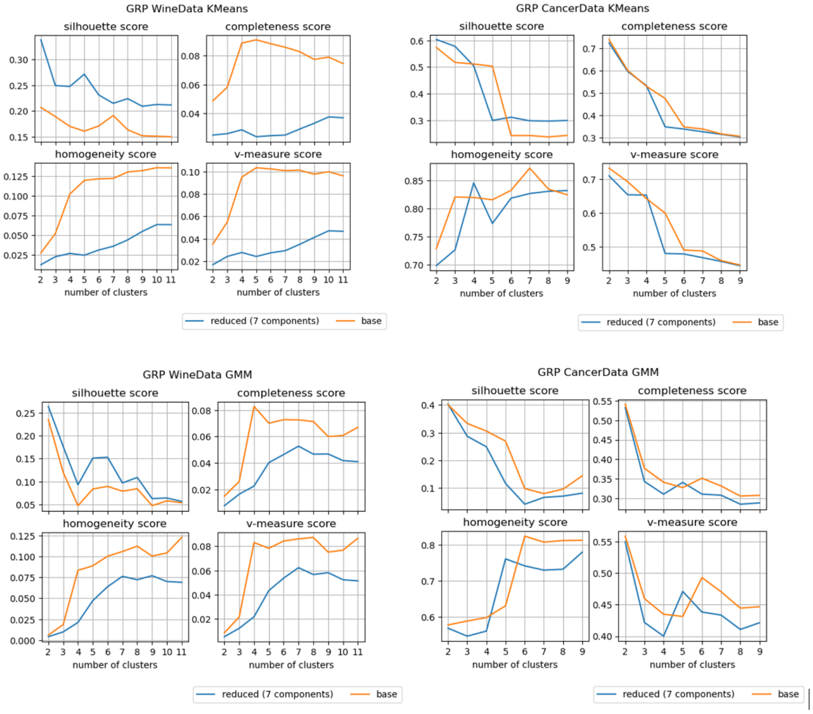


Figure 9: *RP- Wine and Cancer K-Means Metrics*

One interesting thing to note with the plots above is that the silhouette score appears much improved over the baseline K-means results while the completeness and homogeneity score seem to be worse. This implies that clustering is tighter with the RP processed data since the silhouette score is higher; however, the overall effectiveness of labeling the data with respect to the actual true labels might be diminished since the homogeneity and completeness score are reduced, meaning that the actual data points are more spread out among the clusters and each cluster likely has a several different datapoints corresponding to different labels. The same observation was noted after running the RP data through the GMM clustering algorithm too which indicates that RP seems to help with generating tighter clusters but may result in a degradation to the ability to label the data for the wine data set.

4.4 Variance Threshold (VT)

Variance Threshold is a relatively simple feature selection method that eliminates features that have a variance less than some threshold variance. For both the wine and cancer datasets a variance threshold value or "1" was used, meaning that any feature that has a variance less than 1 across all datapoints is eliminated. The value of 1 was selected since this seemed to be the only variance threshold that was successful in reducing the number of features without removing all the features in both datasets. The VT method reduced the number of features in the wine dataset from 11 to 4 while the cancer dataset was only reduced from 9 to 6. It is interesting in this case that the wine dataset had more features eliminated than the cancer data set and is probably due to the number of similar values in each feature set in the wine dataset vs the cancer dataset.

The plot below shows the GMM clustering metrics after running the reduced wine and cancer datasets with 4 and 6 features, respectively.

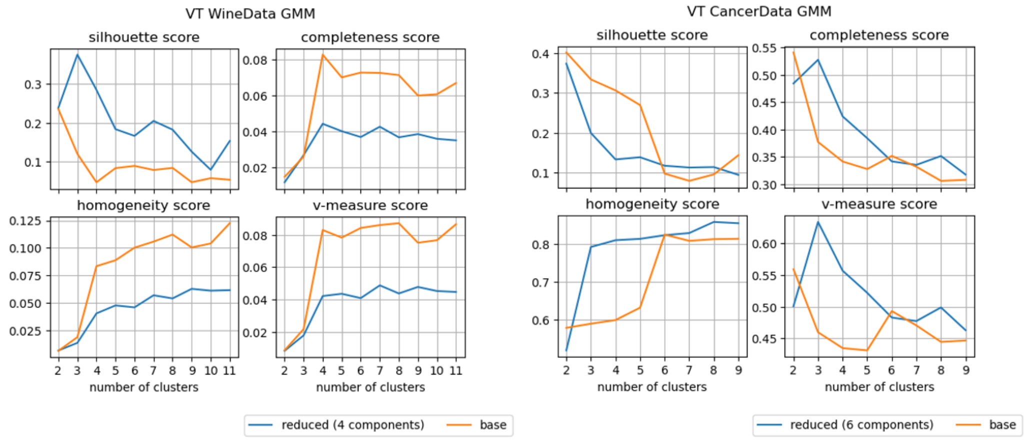
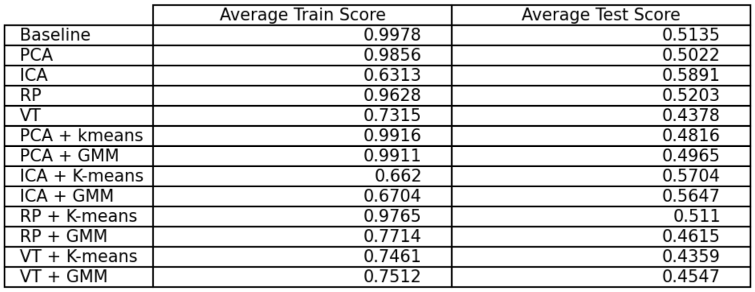


Figure 10: *VT - GMM Clustering Metrics*

Interestingly, the metrics are almost opposite between the wine and cancer datasets. The silhouette score seems to have gotten much better for the wine data, indicating tighter clusters, which is likely due to the large reduction in the number of features. The reduced number of features seems to have had an adverse effect on the homogeneity and complexness of the data, which could be because some of features removed helped keep the labeling more consistent. For the cancer dataset it appears that the VT algorithm potentially removed features that were providing some redundant information.

5 NUERAL NETWORK ANALYSIS

To analyze classification performance with clustering, dimension reduction, and feature selection the wine dataset was transformed through several different combinations of dimension reduction and clustering and used to train a neural network to classify the wine dataset. The table below captures the different transformations applied to the data and tracks the average training and test scores from the neural network evaluation. The configurations that involve both dimension reduction and clustering were obtained by first transforming the dataset using the dimension reduction technique and then applying either K-means of GMM to the reduced dataset. The cluster each data point was assigned to was appended as a new feature onto the reduced data set, thereby effectively adding new features for each cluster. Each train and test score was obtained using cross-validation with 5 folds, where the training data and test data was split 80% and 20% respectively. Additionally, to account for some of the randomness in scores the scores were averaged across 3 runs of the cross validation.



	Average Train Score	Average Test Score
Baseline	0.9978	0.5135
PCA	0.9856	0.5022
ICA	0.6313	0.5891
RP	0.9628	0.5203
VT	0.7315	0.4378
PCA + kmeans	0.9916	0.4816
PCA + GMM	0.9911	0.4965
ICA + K-means	0.662	0.5704
ICA + GMM	0.6704	0.5647
RP + K-means	0.9765	0.511
RP + GMM	0.7714	0.4615
VT + K-means	0.7461	0.4359
VT + GMM	0.7512	0.4547

Figure 11: *Neural Network Training Performance*

In general, the PCA and RP data seemed to perform similarly to the running the neural net with the original data. This indicates that the dimensions eliminated for both PCA and RP likely contained a lot of extraneous information since they did not seem to affect the training of the neural network. Interestingly, ICA seemed to perform quite a bit better for the test data but also quite a bit worse for the training data. The two plots below compare the learning curves for the original dataset (baseline) and the dataset that has been transformed with ICA.

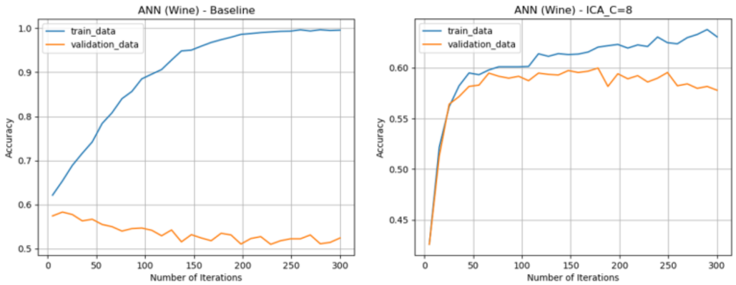


Figure 12: *Learning Curves for Baseline and ICA Data*

Judging by the learning curves, training the network on the ICA data seems to do a pretty good job at generalizing to new data vs the baseline data. This could be due to the fact that the since the ICA data has tried to partition itself into independent components this helps reduce some of the original noise in the data which helps prevent overfitting. This does appear to come at the cost of not achieving the same training accuracy though.

Another interesting observation is that for essentially all the dimension reduction techniques, adding the cluster results as new features to the dataset seemed to increase the training data score but decrease the testing data score. This is likely due to the additionally features causing some additional overfitting in the neural network, due to the added model complexity. It is possible that with more tuning the testing score could less impacted or improved with the additional features.

6 CONCLUSION

This project provided an interesting opportunity to explore many different methods for transforming datasets and provided a lot of insight into what aspects and attributes of datasets are important for machine learning. A few key takeaways are that the effectiveness of the dimensionality reduction and clustering techniques depends largely on some understanding of the original dataset so that the optimal algorithms can be used. Additionally, when trying to evaluate the performance of the clustering and dimensionality reduction algorithms it is useful to try and look at multiple different metrics to get a complete view of improvements or degradations in the dataset and how this might be good or bad for future applications.

7 CITATIONS

https://archive.ics.uci.edu/ml/datasets/Breast+Cancer+Wisconsin+%28Original%29

https://archive.ics.uci.edu/ml/datasets/Wine+Quality