Assignment 3: Clustering

Naziul Talukder
ntalukder6@gatech.edu

***Abstract*—In this paper, I utilized clustering and dimensionality reduction to project useful information from the original heart disease and cancer dataset into a lower dimension. I used clustering to identify patterns in the dataset and train a NN model using the clusters. I also used various dimensionality reduction algorithms to reduce noise without losing the underlying pattern in the dataset. For clustering techniques, I will use K Means and Gaussian Mixture Model (EM) clustering. I obtained the optimal cluster size using silhouette score and validated it using intra-cluster distances, AIC, BIC score. I ran PCA, ICA, Randomized Projections and Variance Threshold as dimensionality reduction algorithms on the datasets to project the data into a lower dimension without losing critical information. Variance Threshold was used to disregard features with lower variances. Observing the optimal components in these algorithms, I also ran clustering algorithms on the dimensionally reduced datasets. I compared the clustering observations on the reduced dataset and original datasets. I trained a Neural Network model on the cancer dataset similarly as assignment 1. Using this as baseline, I trained the model adding findings from clustering as new features. I also trained the model on dimensionally reduced cancer dataset and compare the findings. Using clustering and dimensionality reduction, I observed better performance compared to the baseline neural network that was trained on the original dataset. On the cancer dataset, NN trained on clustering data observed through GMM provided better performance and lower training and prediction time. Dataset transformed by PCA provided the lowest training time and prediction time.**

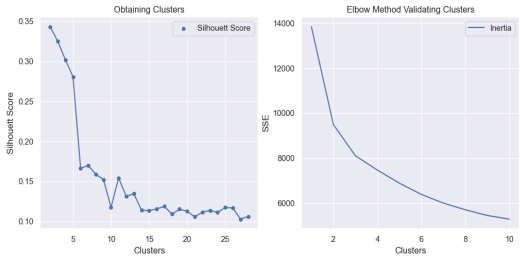

Fig. 1. Breast Cancer Clustering K means

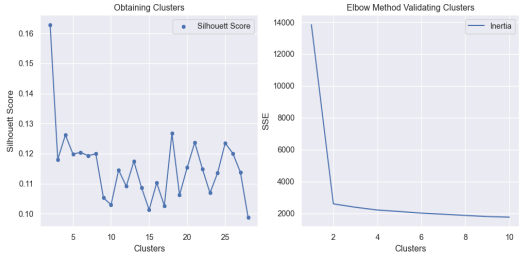

Fig. 2. Heart Disease Clustering K means

I. Clustering

To determine the best number of clusters on the breast cancer dataset: I used silhouette score. It measures similarity of elements to its own cluster and other clusters with values ranging form -1 to 1. Higher values indicate better matching and appropriate configuration. Lower values indicate to sub-optimal configuration and matching. Figure 1 shows that with increasing number of clusters the score decreases. 2 Clusters provide the maximum score for both k means for both cancer and heart disease dataset as seen in figure 1 and 2.

For the k means, I used elbow method to validate these results as seen in the right side of the figures 1 and 2. I obtained the intra cluster distances for both datasets with varying the number of clusters. The elbow method suggests: the optimal number of clusters would be at the spot where the slope of the graphs drastically change creating an elbow shape. In both figure 1 and 2 the elbow is seen when the cluster size is 2. This matches with results obtained in silhouette score.

I ran GMM on the same datasets to obtain the silhouette score. Figure 3 shows the silhouette score and cluster size relationship in the breast cancer dataset. The highest score is obtained when cluster size is 2. To validate this result I used AIC, BIC and lower bound of log likelihood results. Cluster size 2 illustrates the lowest BIC score and lower bound of the log likelihood. The AIC score is comparable to the BIC score but the AIC score decreases further with higher cluster size. But the lower bound of the log likelihood increases with higher number of clusters. BIC penalizes complex models more compared to AIC. Hence, opting to prioritize BIC score I expect that 2 clusters provide the most optimum result. This matches the result obtained by K means. It implies that the breast cancer dataset contains a strong relationship of linearity, clusters are well separated enough that both hard and soft clustering produce similar result, distinct cluster centers, Gaussian distributions to describe the probability of the clusters.

A similar analysis was carried out on the heart disease dataset. The silhouette obtained a peak at 4 suggesting 4 as the best cluster size. The lowest BIC score suggests 6 as the best cluster size. AIC score decreases with higher and higher cluster size but lower limit of the log likelihood increases with higher cluster size. AIC score also provides low score for cluster size 6. Cluster size 6 provides a comparable high silhouette score close to 0.09 even though it is not the peak. Silhouette score being close to 0 implies that object in a cluster is similar to other clusters and the clusters are poorly separated. For the heart disease dataset, K means provide more reliable clustering that is easily verifiable with the elbow method and higher silhouette score. As K means and GMM make different assumptions on the distribution of the dataset which impacts the outcome, we observed two different results

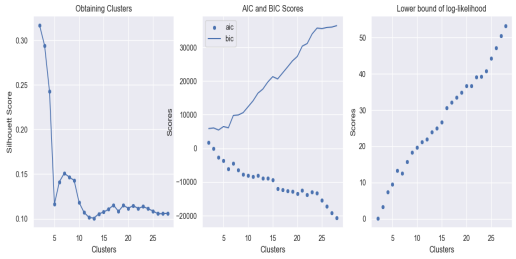

Fig. 3. Breast Cancer Clustering GMM

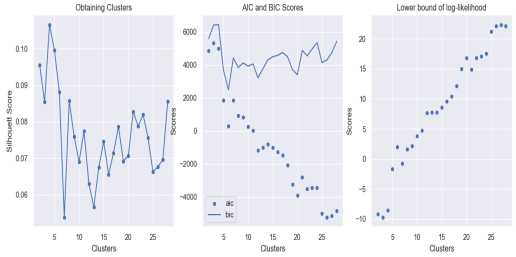

Fig. 4. Heart Disease Clustering GMM

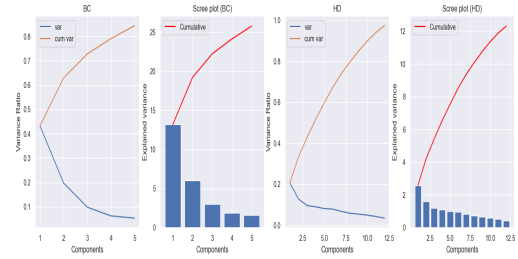

Fig. 5. PCA Variance and Eigenvalue analysis on both datasets

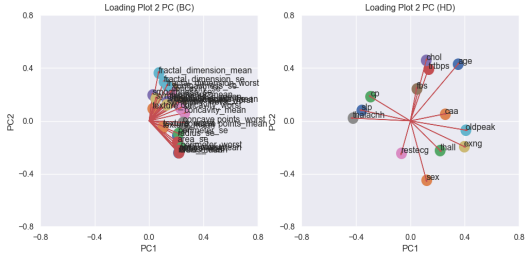

Fig. 6. PCA Loading Plot both datasets

on cluster size from those methods. The results indicate that the dataset is well separated with spherical clusters and hard clustering is more appropriate for this dataset.

II. PCA

I carried out PCA analysis on both cancer and heart disease datasets. PCA is a useful technique to reduce dimensionality of a dataset without losing accuracy. Principal components are useful to capture the variance in the dataset as it lies on the direction in which the data varies the most. Thus it can identify and capture the most important features and the underlying patterns in the dataset even if it is not in a human readable form. It also can remove noise from the dataset. During our PCA analysis we will remove components with low variance to ignore the noise and focus on the main pattern of the dataset.

Figure 5 shows how variance is captured as component increases in both datasets. The barplots show how the eigenvalues changes with increasing number of components in both datasets. For the breast cancer dataset: using 3 component accounts for higher than 70% of the variance in the dataset. The scree plot shows that the eigenvalues of components 4 and 5 are pretty negligible compared to the other ones and much of the cumulative variance can be explained by 3 components.

Similarly, for the heart disease dataset: the eigenvalue is much lower for components higher than 7 as seen in the Scree plot. In the scree plot cumulative variance curve does not change slope as it did in the breast cancer dataset. But 7 Components account for almost 80% of the variances seen in the dataset. Figure 6 shows a 2D loading plot for both datasets. The plot shows the eigenvectors with an arrow mark illustrating relationship between the features and the principal components. This plot only demonstrates the relationship of 2 PCA components as we can not make a 7 dimensional plot for the heart disease dataset and the first two components accounts for a good chunk of the variance for both datasets. The loading plot shows the correlation of the features and their directions with respect to the components.

Figure 7 shows the transformed cancer dataset and it shows how PCA is successful at determining two clusters with 3 components. Two different colors in the plot indicate the two labels in the dataset. Figure 8 shows a 3D projection of the PCA transformed dataset of heart disease dataset. The figure shows a mixture of labels in the dataset, because true separation of clustering will only be seen on 7 dimensional plot.

III. ICA

ICA attempts to obtain independent components from the distribution. Kurtosis compares a distribution with normal distribution. Because ICA attempts to obtain non Gaussian independent components they tend to have higher peaks or heavier tails than Gaussian distribution. Hence, I compared the kurtosis values of the distribution to obtain the optimal number of components. Figure 9 shows how the kurtosis score varies as the number of components change for ICA. In the breast cancer dataset, the maximum kurtosis is observed when

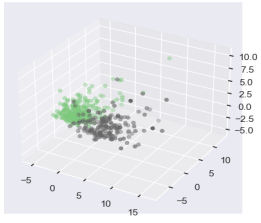

Fig. 7. PCA transformed cancer dataset

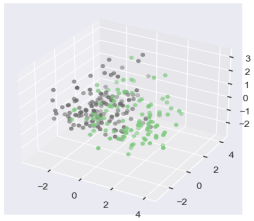

Fig. 8. 3D plot of PCA transformed heart disease dataset

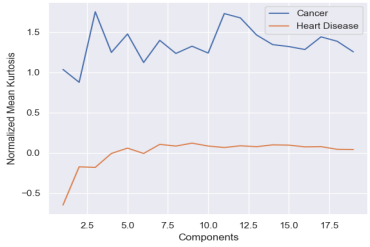

Fig. 9. Comparison of Kurtosis for various components of ICA

component is 3. For the heart disease dataset, the maximum kurtosis is observed when component is 10.

Figure 10 shows the ICA transformed cancer dataset where two colors indicate two different labels in the dataset. The plot illustrates two clusters in the dataset and obvious separation. Figure 11 shows a 3D projection of the ICA transformed heart disease dataset. The clusters are not as separated as the cancer plot shows because the plot can only capture 3 components. A 10 dimensional plot would capture the separation of clusters properly.

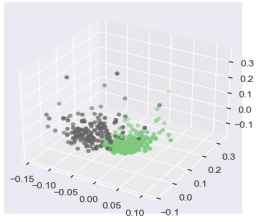

Fig. 10. ICA transformed cancer dataset

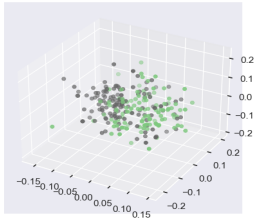

Fig. 11. 3D plot of ICA transformed heart disease dataset

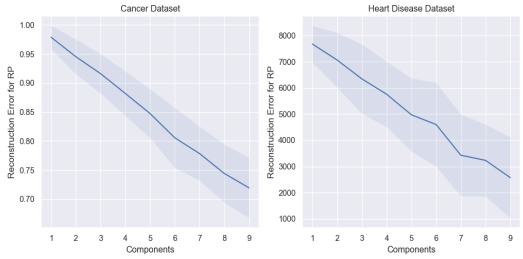

Fig. 12. RP with Reconstruction Error

IV. RP

I adopted a randomized projection to efficiently reduce the dimensionality of the data through trading accuracy for faster processing in a controlled method. I used Gaussian random matrix to describe the underlying data in the lower dimension. The pairwise distance among data points are preserved in the distribution of random projection matrices. Using the Johnson-Lindenstrauss lemma in breast cancer dataset I observed: with 455 training samples the target dimension (lowest dimension randomized projection can obtain retaining the insight from the data in higher dimensions) is 146. In our dataset we only have 30 features. Similarly on the heart disease dataset: with 242 training samples the target dimension becomes 131. Due to the smaller feature space and number of samples in both datasets, RP is not a reliable method. Figure 12 shows a linear relationship in the reconstruction error as dimension increases in RP. We can implement an elbow method to obtain a sub-optimal dimension size from the projections. For the breast cancer dataset: the slope gets flatter for a while after 6 components. Similarly, on the heart disease dataset after 7 components the slope gets flatter. If we had higher number of samples in the datasets as well as higher feature space, the elbow would be more obvious and the graph would not be as linear as it looks in figure 12

V. Threshold Variance

One popular method for feature selection is observing the features and removing the ones with low variance. Low variance data usually do not contribute to the underlying complexity as much as the high variance data. Removing low variance data can help the model learn faster, reduce the complexity while preventing overfitting. But one must be careful in not removing useful information even if that has low variance. I discarded the features which lied on the bottom 20 percentile of the variance score in the dataset. Figure 13 shows variance in breast cancer and heart disease datasets. The threshold obtained for breast cancer dataset was 0.979051067311012 and for heart disease dataset was 0.9270439466688257. Through this process in the final dataset, I will only consider features with high variance (at least higher than 20 percentile) in each dataset. I chose 20th percentile to be the threshold to be mindful about losing useful information as well as removing unnecessary noise in the dataset. The breast cancer dataset was reduced from 30 features to 19 features and the heart disease dataset was reduced from 13 features to 10 features.

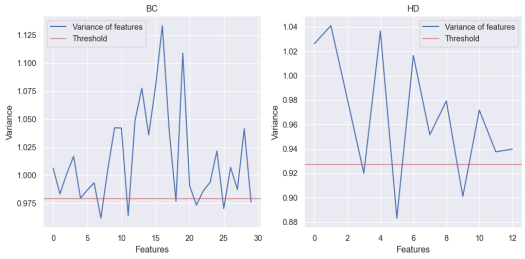

Fig. 13. Removing features with low variance

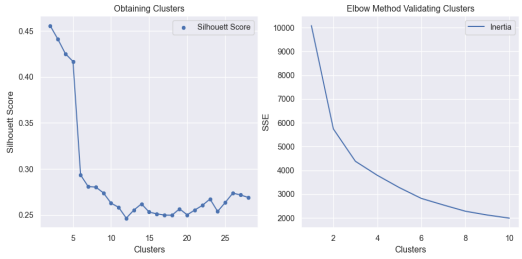

Fig. 14. K Means Clustering on BC dataset after PCA

The breast cancer dataset contained a lot more features in the low variance space than the heart disease dataset.

VI. Dimensionality Reduction + K Means

I will employ these feature selection techniques we saw in the previous section to analyze the clustering algorithms like K means and GMM. Firstly I will implement appropriate feature space using the feature selection methods and then train on the same clustering techniques as seen in the clustering section. Figure 14 shows how the clustering algorithm performs on the breast cancer dataset after implementing PCA with 3 components (obtained from figure 5) on the dataset. Comparing with figure 1, I observe a higher Silhouette score and lower inertia values in the PCA transformed dataset for smaller cluster size. The maximum score increased from 0.35 to 0.45 after the PCA transformation. The elbow method verifies that 2 is the best possible cluster size and the figure shows the highest score when cluster size is 2.

Similarly, figure 17 shows a similar analysis when dataset has been transformed by ICA. ICA used 2 components as obtained by 9. The silhouette score does not increase much in this case but the inertia decreases drastically. Lower inertia implies that the points are closer to the centroid in the cluster creating a compact tightly knit cluster. Maximum silhouette score is obtained when cluster size is 4. Looking at inertia, we observe an elbow (drastic change in the slop) when cluster size is 4. It verifies the observation from silhouette score.

Similarly, figure 19 shows when the dataset has been transformed by RP (6 components were chosen). We did not see any drastic change in the results compared to 1. With higher sample size and feature space RP would be more effective on this dataset. It is an effective technique in reducing feature space in the dataset and the figure shows that inertia decreases and the maximum silhouette score increases compared to the original dataset. Silhouette score and elbow method looking at the inertia indicates the cluster size is 2.

Figure 21 shows a similar analysis when dataset has been transformed using variance. In this method I removed the feature with low variance. I used 20th percentile to be the threshold for the variance. The clustering model only trained on features with high variances. The silhouette score is maximum when cluster size is 2. Elbow method on the inertia values confirm this. Silhouette score is similar than what we observed with RP but the inertia is much smaller using this method (indicates a more compact cluster).

I can do a similar comparison on the heart disease dataset. Figure 15 shows an increase in the silhouette score and decrease in inertia. PCA with 7 components was implemented to transform the heart disease dataset. The figure shows maximum silhouette score for 2 clusters. Looking at inertia values using the elbow method, we can confirm that 2 is the optimal cluster size.

Figure 16 shows a similar analysis using ICA with 10 components as obtained from figure 9. The silhouette score did not improve in this transformation, but the inertia decreased drastically. The maximum score was obtained for 25 clusters and the inertia values confirm this. There is also a possible elbow shape in the figure when cluster size is 13. But silhouette score for 13 is close to 0.13 whereas for 25 is close to 0.15. We will consider 25 to be the cluster in this case.

Figure 18 shows when the dataset was transformed using RP with 7 components. The best silhouette score is found for 2 clusters. Looking at the inertia values we can also find an elbow shape for 2 clusters. Another possible elbow shape is seen for 3 clusters. The silhouette score between 2 and 3 clusters varies from 2.34 to 2.28. Due to maximum silhouette score, I will use 2 clusters as the optimal condition. Using RP we observed higher silhouette score than original dataset, PCA transformed dataset, ICA transformed dataset.

Figure 20 shows when only the high variance features were used as part of the transformation. The silhouette score obtained maximum for 2 clusters and we observe the elbow in the inertia plot for 2 clusters. The silhouette score is comparable to the original dataset case. Due to training on a lower dimension feature space it easily reduces inertia. It also provides higher silhouette score when cluster size is higher compared to the original dataset.

In both datasets we observed low inertia when datasets were tranformed using ICA. PCA provided the best silhouette score in cancer dataset and RP provided the best score in heart disease dataset. ICA tended to suggest higher cluster size for both datasets (4 for cancer and 25 for heart disease). Whereas PCA, variance threshold, RP and original data tended to agree on same size (lower cluster size).

VII. Dimensionality Reduction + EM

I ran a similar analysis using GMM clustering on both datasets. Figure 22 shows the impact of PCA transformation on the breast cancer dataset. We obtain a higher maximum silhouette score for 3 clusters compared to the maximum

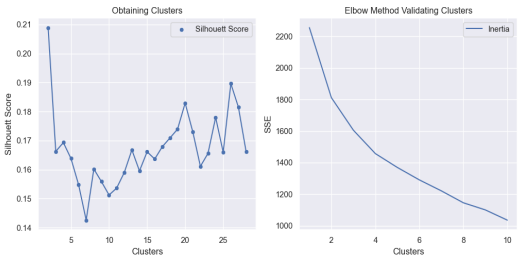

Fig. 15. K Means Clustering on HD dataset after PCA

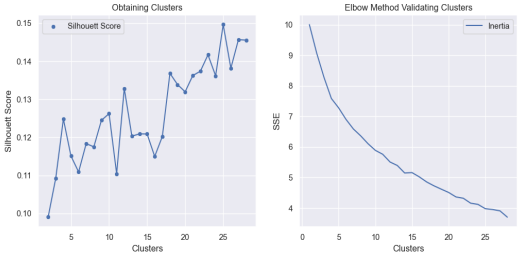

Fig. 16. K Means Clustering on HD dataset after ICA

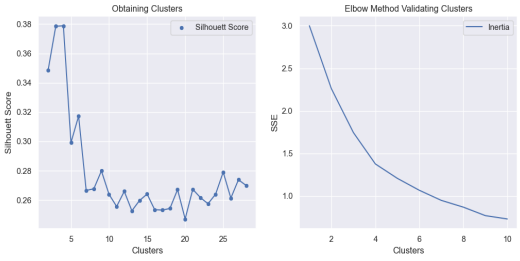

Fig. 17. K Means Clustering on BC dataset after ICA

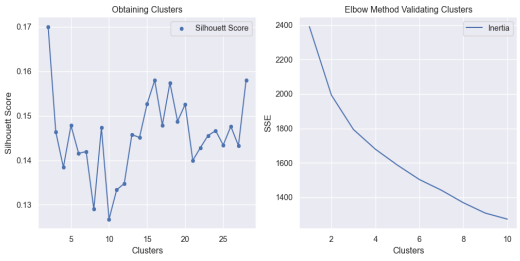

Fig. 18. K Means Clustering on HD dataset after RP

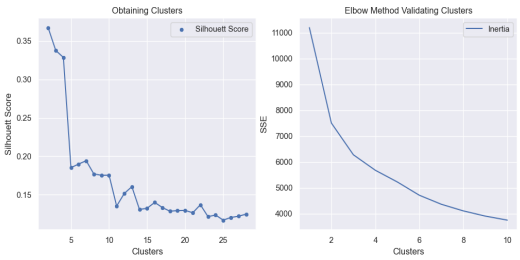

Fig. 19. K Means Clustering on BC dataset after RP

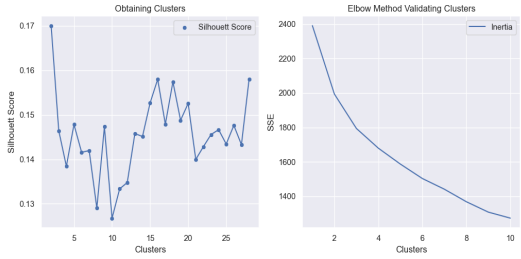

Fig. 20. K Means Clustering on HD dataset after High Variance Feature Selection

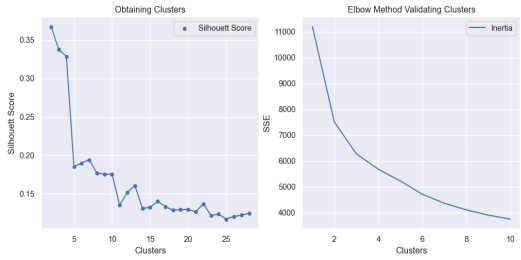

Fig. 21. K Means Clustering on BC dataset after High Variance Feature Selection

score seen in figure 3 for 2 clusters. We validate this result by looking at AIC and BIC score. A lower BIC score was also obtained after the PCA transformation. It implies that the model fits the transformed data better. The figure shows lowest BIC score for 3 clusters. AIC score keeps decreasing with higher cluster size but due to simplicity in such cases we will prioritize results from BIC score. The lower bound of log likelihood obtained after the transformation was also significantly lower than results from original data.

Figure 23 shows a similar analysis on the heart disease dataset. The highest score is obtained for 20 clusters. In the original dataset we observed this behavior for 4 clusters in figure 4. The maximum value is higher for PCA transformed dataset. The BIC score is low for 4 clusters but higher for 20 clusters. Comparatively, the AIC score is much lower for 20 clusters. As discussed in previous section, AIC score does not get penalized for the complexity of the cluster size as much as BIC. The low AIC score validates that 20 clusters would be an optimal choice for this transformed dataset.

Figure 24 shows the performance on ICA transformed cancer dataset. The figure suggests higher maximum silhouette score compared to the original dataset. The maximum score was obtained for 4 clusters. AIC and BIC score validates this result. BIC score is close to the lowest value in the figure for 4 clusters. AIC score is also pretty low for 4 clusters. AIC score keeps decreasing with higher cluster size but it shows that 4 clusters would be the optimal choice. ICA transformed dataset also shows lower value for the lower bound of log likelihood compared to original dataset. It implies the model fits the dataset better after the ICA transformation.

Figure 25 shows similar analysis on the heart disease dataset. It also observes higher silhouette score for the transformed dataseet as well as lower AIC, BIC scores and lower

bound of log likelihood. The maximum score was obtained for 22 clusters which is much higher compared to 4 that was observed in figure 4. With lower AIC and BIC scores, the model fits better on the ICA transformed data compared to original dataset. I validated the number of clusters looking at BIC and AIC scores. As BIC score gets penalized for higher cluster size the lowest BIC score was obtained for 3 clusters. For 4 clusters we observe a local peak on the silhouette score. For 22 clusters AIC score demonstrates the lowest possible score. Hence, 22 clusters is the optimal size for ICA transformed dataset.

Figure 26 shows performance when the cancer dataset has been transformed by RP. The silhouette score did not increase under this transformation. The score decreased and AIC and BIC score increased compared to the performance on the original dataset. It implies that the transformation lost crucial information to train the model using GMM. This is due to lack of samples and lower feature space on our dataset.

Figure 27 shows performance when the heart disease dataset has been transformed by RP. The silhouette score increased in this transformation. But the AIC and BIC score increased too. The lower bound of the log likelihood has also increased. It implies that the model fit the original dataset better. The maximum score is obtained when cluster size is 4. The lower bound of log likelihood confirms this as we can see a dip in the plot for 4 clusters. The BIC score has lowest value for 2 clusters. But the score for 4 clusters is comparable to the lowest value.

Figure 28 shows performance when the high variance features are selected in the cancer dataset. The silhouette score did not change much compared to the original dataset. The AIC, BIC and lower bound of log likelihood also provides a similar result as original dataset. The maximum score is obtained when cluster size is 2. BIC score and lower limit of log likelihood are also lowest for 2 clusters.

Figure 29 shows performance when the high variance features are selected in the heart disease dataset. The silhouette score increased a little compared to original dataset. The AIC and BIC scores have also decreased a little. It implies the model is trained better on the transformed dataset compared to the original dataset. The maximum score is obtained for 3 clusters. BIC score has a local minima for 3 clusters and AIC score is much lower too. BIC score obtains another lower value for 13 clusters but the silhouette score is not promising. But even for 13 clusters the score is much higher than scores from original dataset.

The PCA transformation provided the best performance across datasets under GMM clustering. ICA transformation was most effective in lowering the AIC and BIC scores. It provided datasets that the model can easily train and avoid overfitting.

VIII. Dimensionality Reduction + NN

In this section, I will analyze the NN obtained from assignment 1 on the cancer dataset and observe the performance on a dimensionally reduced dataset compared to original dataset. Figure 30, 34 and 35 shows different stages of hyper parameter

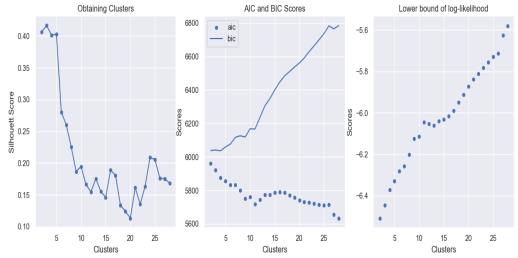

Fig. 22. GMM Clustering on BC dataset after PCA

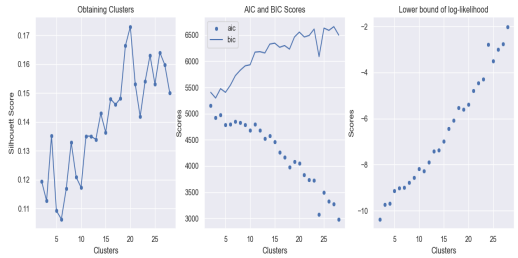

Fig. 23. GMM Clustering on HD dataset after PCA

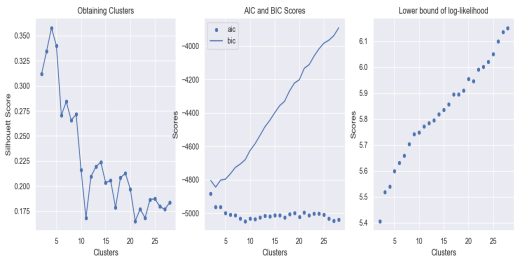

Fig. 24. GMM Clustering on BC dataset after ICA

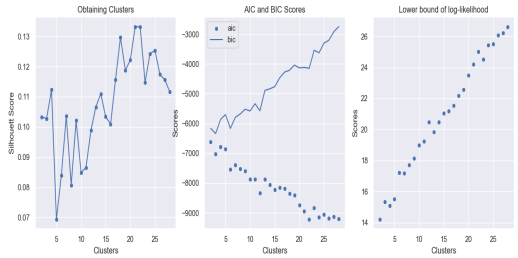

Fig. 25. GMM Clustering on HD dataset after ICA

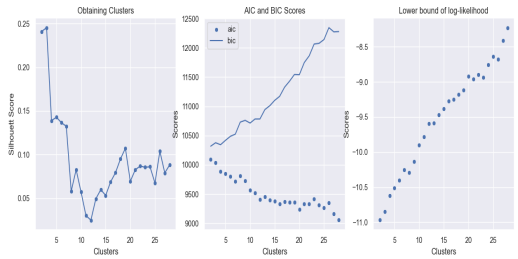

Fig. 26. GMM Clustering on BC dataset after RP

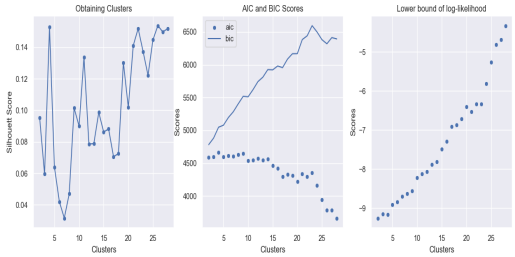

Fig. 27. GMM Clustering on HD dataset after RP

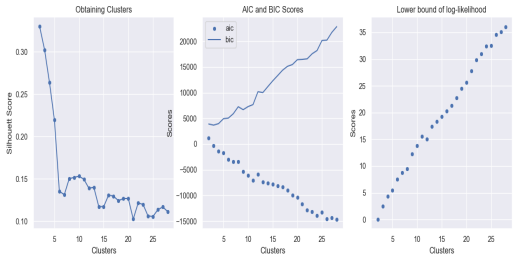

Fig. 28. GMM Clustering on BC dataset after High Variance Feature Selection

tuning for the NN on the original dataset. The simplest NN model with best performance was observed when 1 hidden layer with 1 neuron was used with identity as the activation function, adam as the solver and 100 iterations. Figures 31, 32 and 33 shows a similar hyperparameter tuning on the datasets after dimenstionality reduction has been applied. Hyper-parameter tuning on these dataset did not change the model from its original structure as seen in these figures. We continue the experiment with same network structure and activation functions as the original dataset.

Figure 36 shows the performance of the NN on each of the dimensionality reduction algorithms and the original dataset. Other than the high variance threshold method: the training time and validation time sees a significant improvement on the dataset. The threshold variance method results a much higher feature space compared to the other methods. Hence, its natural that we observe a little reduction in this method compared to the other ones. PCA provides the best reduction in training time in this dataset. Although PCA and ICA have same number of components PCA provides dataset that is quicker to train and validate. In ICA transformed data training

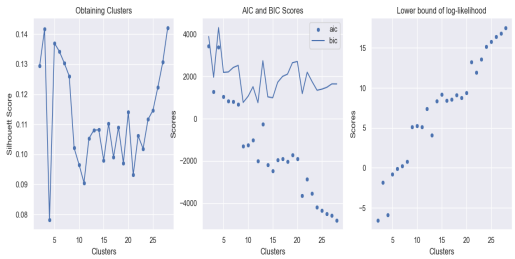

Fig. 29. GMM Clustering on HD dataset after High Variance Feature Selection

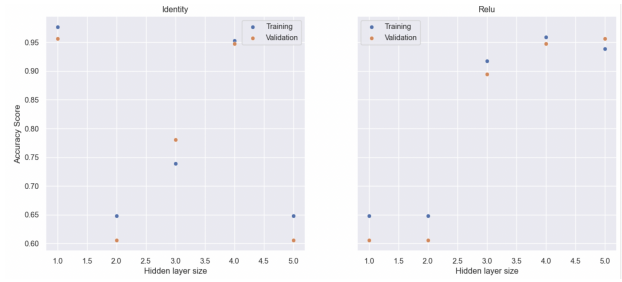

Fig. 30. Activation Function and Hidden layer size for NN

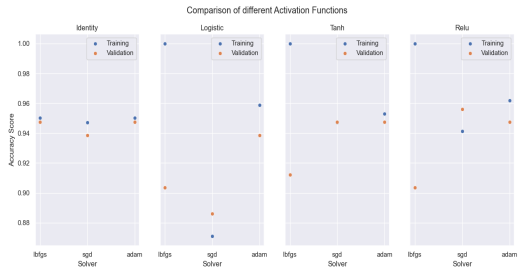

Fig. 31. Activation Function and solver for NN on PCA transformed dataset

time is similar to PCA but prediction takes a much longer time. Figure 31 shows the highest validation score compared to figures 32 and 33. It shows that PCA is the most reliable technique for dimensionality reduction in this dataset. The validation score obtained through PCA is similar to the best validation score observed in the original dataset in figure 34. It demonstrates that with PCA we can obtain better training time without losing efficacy of the learning model. ICA, RP and High Variance Features are helpful in reducing the training time, but their validation score is much lower compared to the original dataset.

IX. Clustering + NN

I will train the NN on the cancer dataset with added new features obtained through clustering. I will run K means and GMM clustering on the cancer dataset. As validated in the previous sections: 2 clusters provided the best results for those clustering methods on the dataset. I will add those 2 new features on the existing dataset and train the existing NN on this new dataset.

For the K means clustering the best performance is obtained with Relu as the activation function and adam as the solver

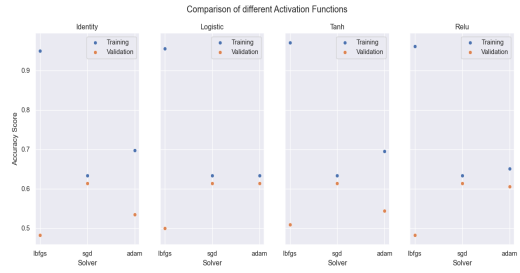

Fig. 32. Activation Function and solver for NN on ICA transformed dataset

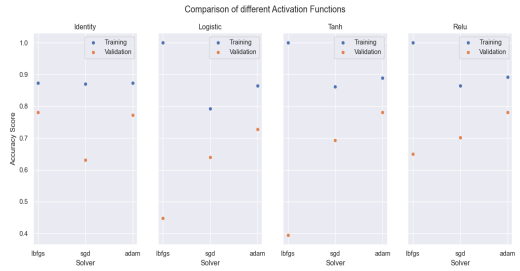

Fig. 33. Activation Function and solver for NN on RP transformed dataset

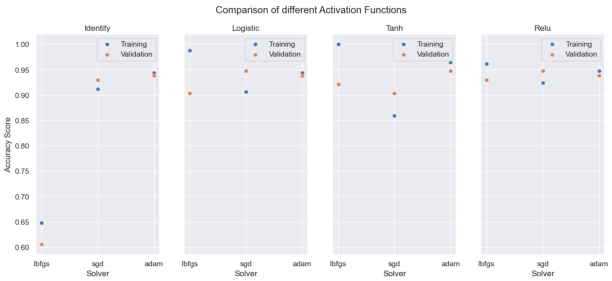

Fig. 34. Different activation functions and solver methods for hyper parameter tuning of NN

with same network architecture. The most optimal model trained on original dataset along with the features obtained using K Means clustering the score is similar to the original dataset as seen in figure 38. Figure 39 shows that with K means the training time and training and validation time is higher than original dataset scenario. It indicates that the dataset is not suitable for hard clustering.

Figure 37 shows the result on when the new features are obtained through GMM clustering. It ensures one more extra feature compared to two from K means. With identity activation function and adam solver it obtains 0.975 as validation score which is comparably higher than original dataset scenario or K means clustering scenario. Figure 39 shows

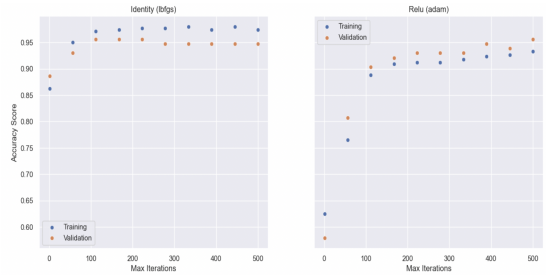

Fig. 35. NN Accuracy score and number of iterations relationship

	Features	Train Time	Train Error	Total Time	Total Error
PCA	3	1.07	0.00599	1.32	0.00672
ICA	3	1.66	0.00726	6.15	0.00545
RP	6	1.36	0.00594	1.96	0.00295
HV	13	10.20	5.30000	12.54	6.10000
Original	30	24.70	2.20000	27.30	2.81000

Fig. 36. Wall clock time in milliseconds for various dimensionality reduction on the dataset. Total time includes both time to train and predict values.

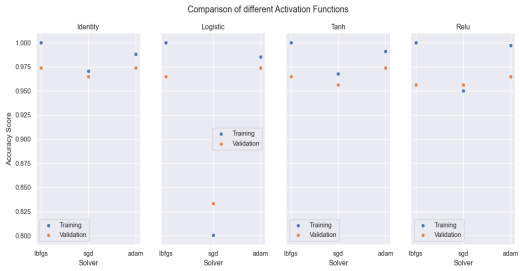

Fig. 37. NN performance of various activation functions and solvers on GMM clustered feature and original dataset

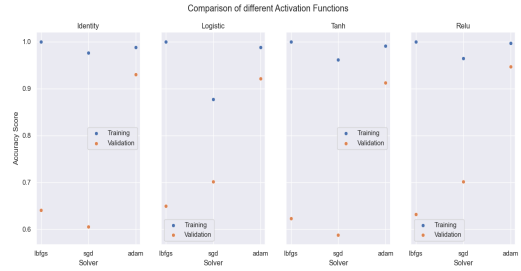

Fig. 38. NN performance of various activation functions and solvers on K Means clustered feature and original dataset

an improvement using feature obtained through GMM. GMM Clustering introduces an extra feature but still the training time and training and validation time is less than original dataset. It indicates that soft clustering on the cancer dataset is more suitable than hard clustering.

Due to better performance of GMM, one can infer that the features in the dataset can have mixed distribution and is not best represented by spherical clusters. K Means can only adopt spherical shapes whereas GMM can adapt to different shapes depending on the probability distribution of the data. It is possible that the dataset is not well separated and the soft clustering ability of GMM and uncertainty calculation helps it to outperform K means. PCA is great on the dimensionality reduction sphere which obtains great performance for very little training and validation time. But GMM outperforms PCA on validation score at the cost of high training and validation time.

X. Conclusion

I analyzed multiple clustering and dimensionality reduction techniques. I used Silhouette score to obtain the optimal number of clusters using K means and EM techniques. I used the intra-cluster distance (inertia) along with the elbow method to validate the optimal number of clusters obtained

	Features	Train Time	Train Error	Total Time	Total Error
K Means	32	139.0	21.20	162.0	4.73
GMM	31	16.2	1.93	18.3	2.19
Original	30	24.7	2.20	27.3	2.81

Fig. 39. Wall clock time in milliseconds before and after clustering. Total time includes both time to train and predict values.

from silhouette score. Mostly the elbow method was successful in verifying the cluster size. I also used the AIC and BIC score to validate cluster size. I analyzed various dimensionality reduction techniques like PCA, ICA, RP and High Variance Features. RP did not yield a successful analysis due to a small sample size and feature space on both datasets. With more data RP would be a lot more effective in reducing dimensionality. PCA outperformed ICA, RP and High Variance Features techniques with higher accuracy score and reducing training time. The model trained on PCA transformed data performed better than the model trained on the pristine data. I also used outputs from the clustering techniques as features to train the NN model. The NN model outperformed when I used GMM as clustering technique. It indicated that the cancer dataset was more suitable for soft clustering methods like GMM than hard clustering methods like K means. GMM clustering obtained higher validation score than original dataset and even on PCA tranformed dataset. GMM obtains higher validation score even though it has costly training time compared to PCA.