**Unsupervised Learning and Dimensionality Reduction Assignment**

Anurag Sahoo  
[asahoo41@gatech.edu](mailto:asahoo41@gatech.edu)

GT ID: 90368005

**Abstract**— This paper deals with different clustering and dimensionality reduction techniques and using the resultant data to train a neural network classifier on the white wine quality and Wisconsin breast cancer diagnostic data. This paper is divided into 5 parts. In part 1, we run K Means and an Expectation-Maximization (Gaussian Mixture Model) on both the data. In part 2, we use 4 different dimensionality reduction technique to reduce the high dimension data. In part 3, we run the clustering algorithms on the dimensionally reduced data. In part 4, we train the neural network classifier on the dimensionally reduced data of breast cancer data. In part 5, we train the neural network classifier on the original data with the K Means and GMM cluster label added as features. We will dig deep in the following sections.

**Introduction to Datasets**:

White Wine Quality Dataset: This data set is sourced from UCI Machine Learning Repository. Wine experts are asked to grade the quality of wine on a scale of zero (very bad) to ten (very good). Various attributes of wine including alcohol, pH, density and few others is recorded for the wine that is tested. The data comprises 4898 examples and 11 attributes. The dependent variable is the quality of the wine. All the attributes of the data are numeric. I converted the regression problem into classification by labeling the quality >= 6 as 1 (indicating good taste) and < 6 as 0 (indicating bad quality). The class distribution after re-labeling is as follows: 66.52% good quality wine (1) and 33.48% bad quality wine (0). Since the attributes/independent variables are numeric and all of them would possess different scale (or range) at which they can lie, we need to scale the value of these attributes. I used the Standard Scaler which standardizes by removing the mean and scaling to unit variance.

Breast Cancer Wisconsin (Diagnostic) Dataset: The second dataset is also sourced from UCI Machine Learning Repository. The examples are collected from the University of Wisconsin Hospitals, dealing in with the results from the diagnosis which are to be used to indicate if the person is diagnosed with Breast Cancer or not. The original dataset has 699 examples and 10 attributes excluding the predicted column (class). Class 2 indicates benign and 4 indicates malignant. I re-labelled the class 2 as 0 indicating absence of cancer cells and class 4 as 1 indicating presence of cancer cells. The class distribution is as follows: 65.52% absence of cancer cells (0) and 34.48% presence of cancer cells (1). The column “Bare Nuclei” had 16 missing instances. Since, this is breast cancer problem and we do not want our model to wrongly classify, I didn’t impute the missing values of “Bare Nuclei”. Instead, I dropped those 16 examples. Few of the attributes of the diagnosis include Clump Thickness, Uniformity of cell size and shape, Mitoses, Bland chromatin and few more. All these features are numerical and are on different scale. Hence, normalized the features to bring it to a similar scale. I used the Standard Scaler in this data as well. The data was split into training and test set in the 80:20 ratio.

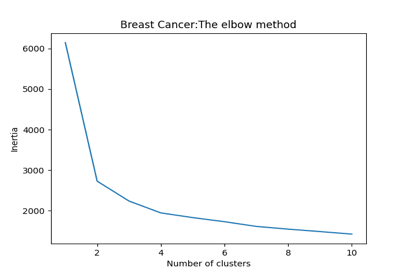
**Why these datasets are interesting?**

These datasets are interesting because they solve different real-world problems. One is related to assessing the quality of wine based on different features. The other is pertaining to the prediction of breast cancer based on the diagnosis data.

These two problems are especially interesting because of the type of data. Wine quality dataset has substantial number of examples and 11 quantitative features. Whereas the Breast Cancer data has limited number of examples. The limited examples would invite overfitting to this problem. Both the datasets are not perfectly balanced, but they are not highly imbalanced. Also, because of the different type of problem statement, we would look at different performance metrics and try to optimize it. Wine quality problem involves classifying the quality of wine as good or bad. So, using balanced accuracy as a metric would work fine. When working on an imbalanced dataset where positives are as important as negatives, balanced accuracy is a better metric over F1 score. In case of Breast Cancer problem, we would not want the breast cancer actual positive cases to go unnoticed. Hence, we would be using recall as the metric to optimize. Because of the difference in size of data, some algorithms would perform comparatively, and some would take more time to train. We will analyze the performance of different algorithms on these datasets.

**Part 1: Clustering**

K Means is a distance-based algorithm where we find the distances (popularly Euclidean distance) between a data point and cluster centroids to assign the data point to a cluster. K Means algorithm tries to minimize the sum of distance between the points and their respective cluster centroids. The clusters formed in the K Means clustering have a circular shape. When the distribution of points is not in a circular form, the K Means might fail to identify the right clusters. Instead of using the distance-based metric, we will have to use a distributed based model. That’s when Gaussian Mixture Model (Expectation-Maximization) algorithm comes into the picture. Gaussian Mixture Model attempts to find a mixture of multi-dimensional Gaussian probability that best fit any input dataset. Gaussian Mixture Model follows the Expectation-Maximization approach where it finds the weight implying the probability of membership of each cluster (E-step) and update the cluster based on the weights (M-step).

 Chart, line chart

Description automatically generated Chart

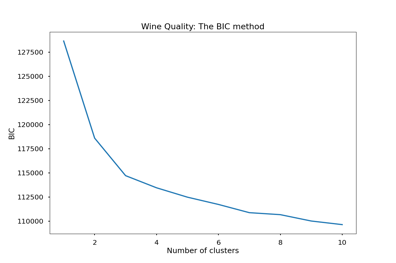
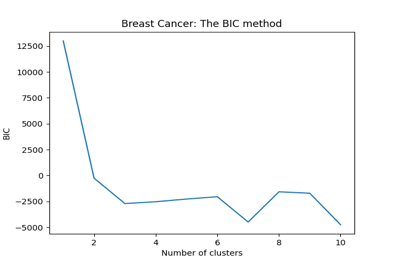
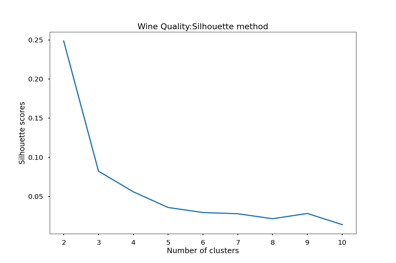
Description automatically generatedChart, line chart

Description automatically generated

**Fig 1:** Elbow & Silhouette Analysis for determining optimal number of clusters for K Means algorithm

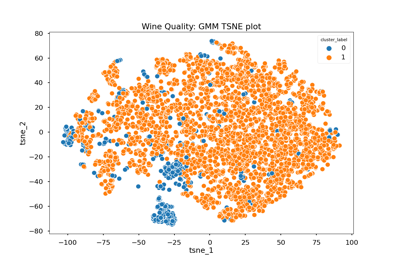
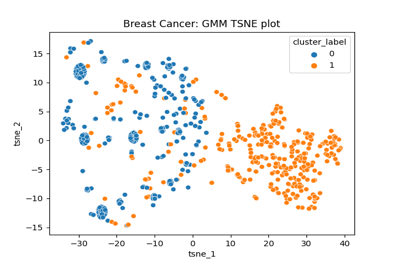
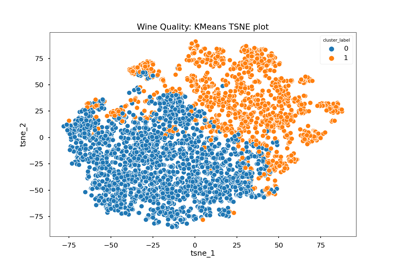
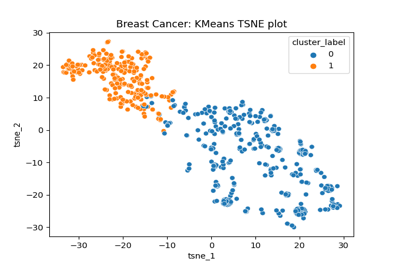
Elbow and Silhouette method was used for determining the optimal number of clusters for K Means algorithm. In Elbow method, we need to find the minimum number of clusters after which the within cluster sum of squares (inertia) flats out. Silhouette scores measure how similar a point is to its own cluster compared to other clusters. Silhouette scores lies in the range of -1 to 1. Better the silhouette scores, better the segmentation. The major difference between elbow and silhouette method is that the elbow only calculates the Euclidean distance whereas the Silhouette value considers the variance, skewness and high-low differences. It is evident from the above figures that 2 is the optimal number of clusters for both the datasets.

Chart, line chart

Description automatically generated

**Fig 2**: Silhouette & BIC method for Gaussian Mixture Model

For GMM, we used both Silhouette values and Bayesian Information Criteria (BIC) for determining the optimal number of clusters for GMM**.** BIC is an estimate of a function of the posterior probability of a model being true under a certain Bayesian setup. Lower BIC indicates that the model is likely to be a true model. From the charts, it is evident that 2 should be the optimal number of clusters for both the datasets.



**Fig 3**: TSNE plot for K Means and GMM clustering on Breast Cancer and Wine Quality datasets

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Breast Cancer | | Wine Quality | |
|  | K Means | GMM | K Means | GMM |
| Homogeneity score | 0.73 | 0.54 | 0.04 | 0.06 |
| AMI score | 0.72 | 0.55 | 0.04 | 0.04 |
| Fowlkes Mallows score | 0.92 | 0.79 | 0.57 | 0.7 |

Table 1: Clustering Validation on original data

TSNE plots which helps us to visualize the higher dimensional data, was used to visually validate the clustering along with other metrics like Homogeneity score, Adjusted Mutual Information (AMI) score and Fowlkes Mallows score. Homogeneity measures if all of its clusters contain only data points that are members of a single class. Adjusted Mutual Information score measures the agreement between the actual label and cluster label ignoring permutations. Fowlkes-Mallows score is the geometric mean of pairwise precision and recall. From the TSNE plots, it looks like K Means on the Breast Cancer data segments really well followed by GMM on the Breast Cancer data although there seems to be some overlap between clusters. The K Means and GMM doesn’t seem to cluster well on the Wine Quality data possibly because the problem was originally a regression problem and was forcibly converted to a classification one. Originally, the output label had values from 3 to 10 and it was re-labelled to 1 (good) if quality was greater than equals to 6. The wine with quality 5 would have hardly any difference in attributes than the ones with quality 6. But after re-labelling, there is a stark difference in the output label but the attributes may not have that significant difference. There seems to be a lot of overlap between the clusters on the Wine Quality data for both approaches indicating that the clustering wasn’t good and the Homogeneity, AMI and Fowlkes-Mallows score tells the same story.

K Means forms a circular/spherical cluster based on the Euclidean distance whereas the Gaussian Mixture Model creates clusters based on probabilistic model and the shape of the clusters can be of any form. K Means creates a hard separation between clusters and GMM creates a soft partition between clusters. K Means is based on the Euclidean distance between two points and GMM is based on the Mahalanobis distance to centers. K Means is a special case of Gaussian Mixture Model with equal covariance per component.

Table

Description automatically generatedTable

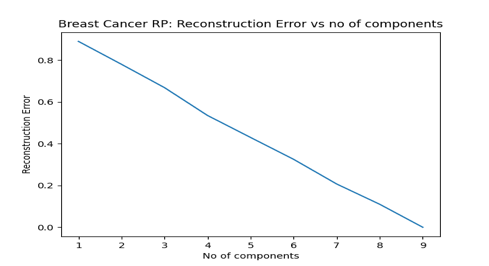
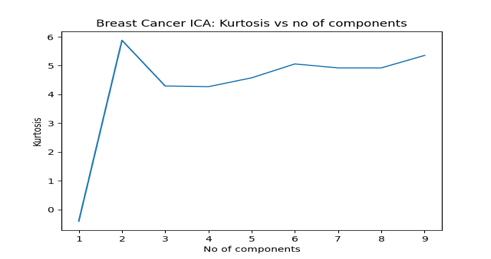
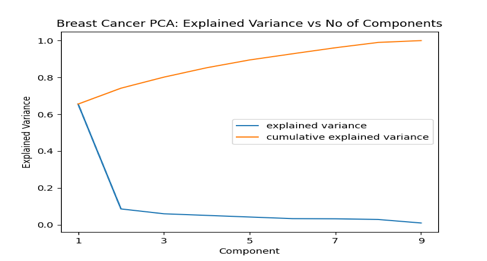
Description automatically generated

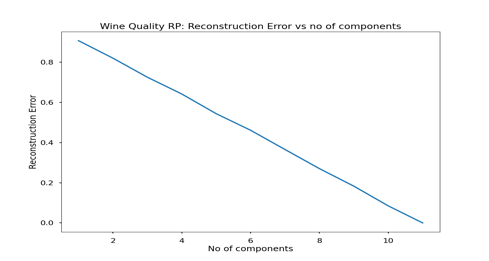
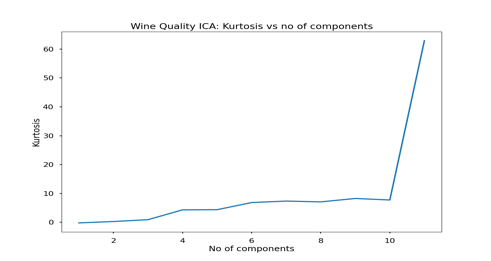
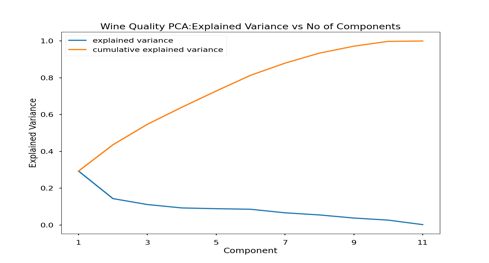
Table 2: Clustering as a means of isolation of positive and negative cases

The clustering labels generated using K Means and Gaussian Mixture Model were cross-checked with actual labels. For Breast Cancer data, K Means did a fantastic job in segregating the data points with cluster label 0 having 95% of negative cases and cluster label 1 having 96% of positive cases. Whereas GMM did a decent job in Breast Cancer data with cluster label 0 having 74% of positive cases and cluster label 1 having all negative cases. Clustering doesn’t significantly do well in Wine Quality data because it is a regression problem and was forcibly converted to a classification problem. All the positive and negative cases seem to be almost equally distributed among each cluster for both K Means and GMM.

**Part 2: Dimensionality Reduction**

In this part, we will look at multiple feature transformation algorithms to convert the higher dimensional spaces to lower. Principal Component Analysis converts a higher set of correlated features to a lower set of uncorrelated features based on the singular value decomposition. PCA determines the direction of maximum variation in a dataset and then project onto them. To apply PCA we need to process the data prior to computing the projection. This may be computationally expensive if the dataset is large and impossible if we are streaming data in real time. Random Projection chooses the projection before seeing the data. ICA is a technique to separate independent sources from a mixed signal. We use feature importance of an Extra Trees (Extremely Randomized) classifier to find the most important features out of higher dimensional spaces as another form of dimensionality reduction technique.



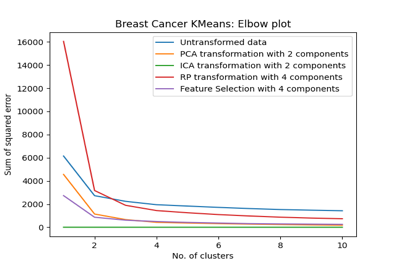


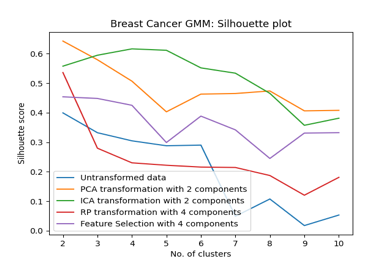
**Fig 4**: Deciding number of components for Dimensionality Reduction techniques

For Breast Cancer data, the subsequent components after first 2 components give us minimal explained variance. Hence, 2 components which explained a major part of variance were selected for PCA on Breast Cancer data. Similarly, 6 components were selected for PCA on Wine Quality data. For ICA, 2 components for Breast Cancer and 11 components for Wine Quality data gives us the highest Kurtosis value. Arguably, ICA didn’t help much for Wine Quality dataset as there were 11 features originally and there was no reduction in dimensionality. With Random Projection, the drop in reconstruction error was so gradual that it gets difficult to decide optimal number of components with a naked eye. Taking a closer look, we decided upon selecting 4 and 6 components for Breast Cancer and Wine Quality data respectively. The Extra Tree classifier selects the best 4 features (comprising 69% of feature importance) and 2 features (comprising 28% of feature importance) from the Breast Cancer and Wine Quality dataset respectively. Final data were generated after decomposing with each technique.

**Part 3: Clustering with Dimensionality Reduction**

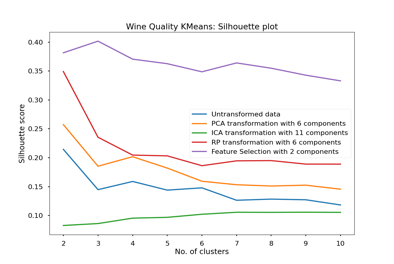
K Means and Gaussian Mixture Model were applied on the reduced data generated in the part 2. A combination of elbow plot and silhouette plot were used to determine optimal number of clusters for K Means and a combination of Silhouette plot and BIC plot were used to determine optimal number of clusters for Gaussian Mixture Model. A k value of 2 was chosen for PCA, RP and Feature Selection reduced data and 3 for ICA reduced data for running K Means on Breast Cancer data. 4 optimal clusters for ICA and 2 optimal clusters for PCA, RP and Feature Selection reduced data was chosen for GMM on Breast Cancer data. A k value of 2 was chosen for PCA and RP reduced data and 3 for Feature Selection reduced data for running K Means and GMM on Wine Quality data. A k value of 7 was chosen for ICA reduced data for K Means and GMM on Wine Quality data.

Chart, line chart

Description automatically generatedChart, line chart

Description automatically generated

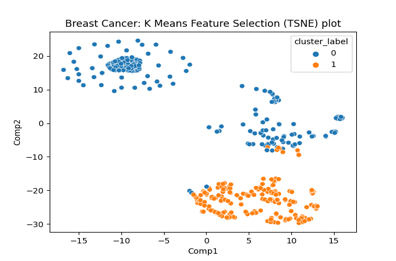
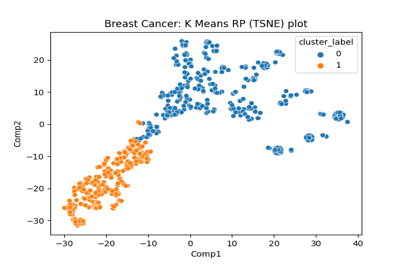
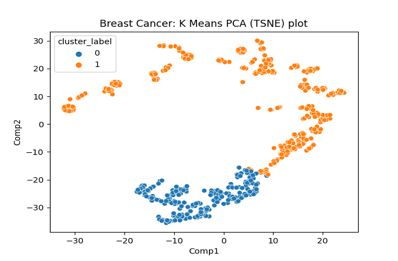
Chart, line chart

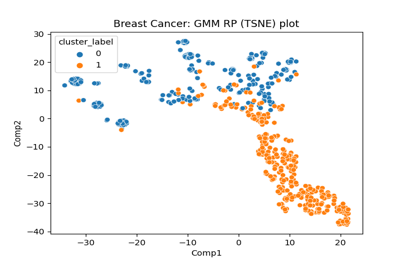
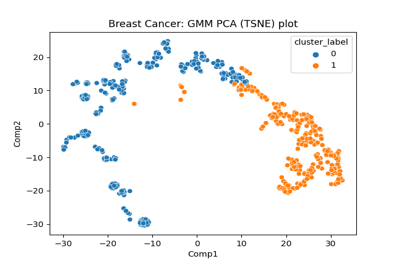
Description automatically generatedChart, line chart

Description automatically generatedA picture containing graphical user interface

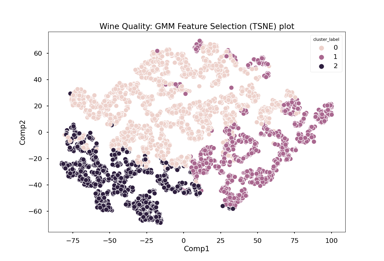
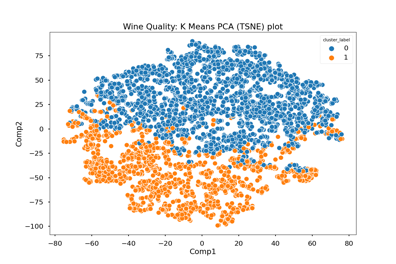
Description automatically generated

**Fig 5**: Determining optimal number of clusters for K Means and GMM on dimensionally reduced data





**Fig 6**: Notable clustering plots on the dimensionally reduced data of Breast Cancer



**Fig 7**: Notable clustering plots on the dimensionally reduced data of Wine Quality

Figure 6 and 7 presents us with few clustering plots where we see some separation between the clusters. In most cases, any clustering algorithm (K Means or GMM) works great (as good as the original clusters) on PCA reduced data and we could see a clear separation between the clusters. Although in Wine Quality, there is some overlap of clusters for K Means on the PCA reduced data because of the regression problem as mentioned earlier. Clustering on ICA reduced data works the worst out of all 4 dimensionality reduction techniques in most cases. Independent Component Analysis works poorly in reality as such as it only works with variables that have uniform distribution. For the Breast Cancer data, the Random Projections work great as well considering the fact that it is quite inexpensive compared to PCA and ICA. Clustering on Random Projections reduced data also does a good job in creating a good separation between the clusters. Random Projections implements a simple and computationally efficient way to reduce dimensionality by compromising a controlled amount of accuracy. Interestingly, the Feature Selection of Extra Trees classifier does a great job in reducing the dimensions by a great magnitude and creating a good separation between the clusters. This has partly to do with the fact that it selects features based on the output label than the other dimensionality reduction techniques. The Homogeneity, Adjusted Mutual Information and Fowlkes-Mallows scores for all the possible combinations in the below table tells us similar story. Metrics on PCA reduced data for both K Means and GMM are as good as the original data (in the previous table). Metrics on ICA reduced data looks worst in most cases. Random Projections does a decent job in getting reduced dimensional data with decent metrics but in real quick time. Metrics for Feature Selection data looks good as well because it peeks into the output label to dimensionally reduce the data.

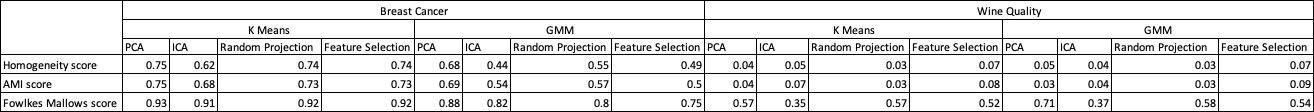


Table 3: Clustering Validation on projected data with respect to actual class label

The above table 3 summarizes the metrics to validate the clustering algorithm with respect to the actual class label on both the datasets.

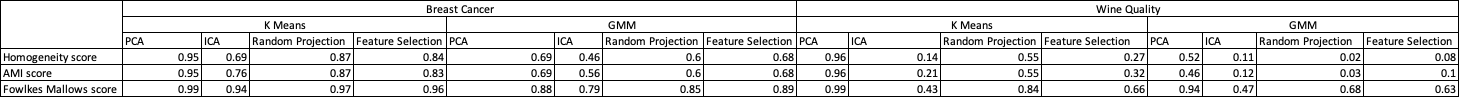
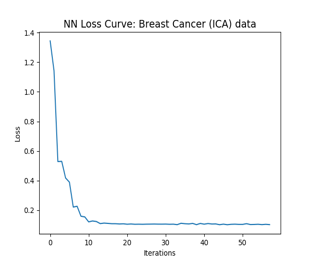
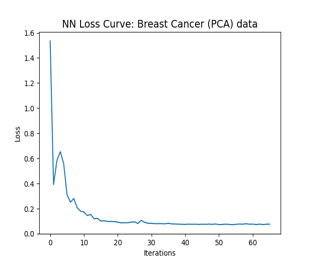
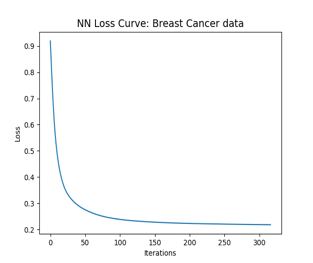


Table 4: Validation of clusters on projected data with respect to clusters on original data

After having compared the clusters on the projected data with the actual class label, we then assess whether performing clustering on the projected data made any changes to the cluster label as compared to that performed on the original data. From the table it is evident that if we perform K Means clustering on PCA transformed data, most of the data points gets similar cluster label as when K Means is performed on the original data. This implies that PCA even after decomposing the data conserves all the variance associated with the data such that there is little change in cluster labels. Random Projections with 4 components when clustered using K Means also preserves the cluster label for most of the data points. K Means on Feature Selection transformed data of Breast Cancer carrying about 69% feature importance also produces clusters which is similar to the clustering on original data. In general, Gaussian Mixture Model clustering produces output which are quite different from that of original data. GMM fits based on maximum likelihood and can fit overlapping clusters with a very soft partition between clusters. Thus, interchanging most of the cluster labels at or near the overlapping cluster region. Clustering on ICA transformed data gives us very different results as it is not a reliable technique to reduce dimensionality. ICA works only with variables that have uniform distribution. The low value of metrics in the Wine Quality dataset is attributed to the type of problem it is. Wine Quality problem is a regression problem converted to classification one.

**Part 4**: **Neural Network with Dimensionality Reduction**

Neural Network classifier was trained on each of the dimensionally reduced data of the Breast Cancer dataset and compared to the results on the original dataset. We are trying to optimize the Recall score for this problem. We used Grid Search CV to find the best hyperparameters. The loss curves and learning curves were generated for the tuned model.

Chart

Description automatically generatedChart, line chart, histogram

Description automatically generated

**Fig 8**: Loss Curves of Neural Network on original and dimensionally reduced data

Chart, line chart

Description automatically generatedChart, line chart

Description automatically generatedChart, line chart

Description automatically generatedChart, line chart

Description automatically generatedChart, line chart

Description automatically generated

**Fig 9**: Learning Curves of Neural Network on original and dimensionally reduced data

The loss curves are as expected with the loss values reducing with iterations. One interesting thing to notice in Fig 8 is that loss curve flattens out after around 100 iterations with the original data. However, with all the four dimensionally reduced data, the loss curve flattened out after about 20 iterations indicating that the model is achieving faster convergence. This implies that the dimensionality reduction techniques, although reducing the dimensions to a much lower dimension spaces, capture most of the variance of the data.

All the learning curves except one that of PCA reduced data show a general trend. For the initial part (up to 20% of data), there is a big gap between the training and cross-validation scores indicating high variance and the gap closes in with more data. With more training data, the training and cross-validation score curves are moving in tandem and converging finally implying that the model is well built. No additional data is required for improving the model. In case of PCA reduced data, although the training and cross-validation curves are moving in tandem, they still have some gap indicating presence of some variance and haven’t converged fully. Additional data would have been helpful for PCA reduced data to build a better model.

Table

Description automatically generated

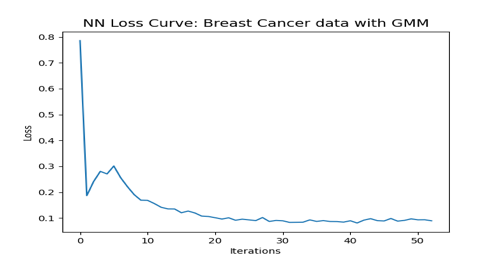
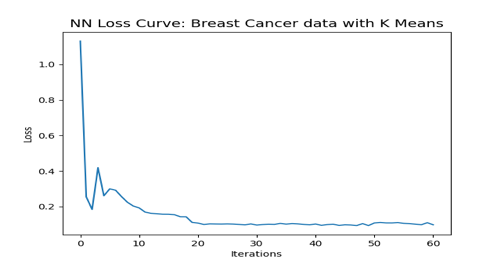
Table 5: Performance Metrics on dimensionally reduced data

The above metrics were generated after running tuned classifier on different dimensionally reduced data. The neural network classifier returns a test recall of 0.96 on PCA reduced data, tad bit lower than the benchmark set by original data. Firstly, from the learning curve it was evident that there was some variance as the training and cross-validation curves were not converging. Secondly, by reducing dimensions with PCA (or any other dimensionality reduction technique), we are getting rid of some information from the data. If the removed data is redundant, we might get better score and if the removed data is slightly important, we would get worse results. PCA doesn’t capture all the explained variance but captures most of it. PCA guarantees that each consecutive dimension will explain lesser and lesser variance. Hence, we are getting slightly lower recall value for PCA reduced data. We might have seen similar results with other dimensionality reduction techniques, but the problem statement is so simple that the classifier is able to build a good decision boundary.

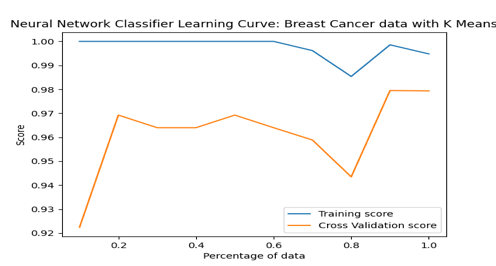
From the training time charts, it is evident that the dimensionality techniques like PCA, ICA and Random Projections produce data with good explained variance in lower dimensional spaces that the classifier takes much lesser time to train compared to the original data. This might be partly due to the fact that the hidden layer sizes got changed after tuning. However, the classifier on Feature Selection transformed data takes more time to train. Other dimensionality reduction techniques squeeze in most of the information in lower dimension spaces, but the Feature Selection takes the best features (based on feature importance) out of all. Although it does a decent job, but it does not capture the variance from the variables which we have lost and may perform poorly with complex problems. That’s why the classifier takes more time to train with limited features.

**Part 5: Neural Network with Clustering**

The cluster labels generated from the K Means and Gaussian Mixture Model with a k value of 2 were used as the feature along with the original features to train a neural network on the Breast Cancer data. The hyperparameters were tuned using Grid Search CV and loss curves and learning curves were generated.



**Fig 10**: Loss Curves with cluster labels as features

Chart, line chart

Description automatically generated

**Fig 11**: Learning curves with cluster labels as features

The loss curves shows that the loss values keep on decreasing with increasing iterations as expected. From the learning curves, it looks like the cross-validation curve has lot of variability indicating error due to variance. Although the gap between training and cross-validation curve is diminishing with training data, it hasn’t converged fully. More data might be helpful for a better fit of the model.

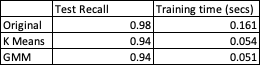


Table 6: Performance Metrics on original data with clustering label as input feature

The test recall metric and training time were recorded for both the data with K Means clustering label and with Gaussian Mixture Model clustering label. The fact that the clustering (both K Means and GMM) didn’t do an excellent job in creating clusters that could isolate the data points with their actual labels completely. The Homogeneity, Adjusted Mutual Information and Fowlkes-Mallows scores in the Part 1 prove that. There would be some contradictions because of which the classifier is unable to create a better decision boundary. Hence, the lower recall value.

References:

1. UCI Machine Learning Repository – Wine Quality Data Set: https://archive.ics.uci.edu/ml/datasets/wine+quality

2. UCI Machine Learning Repository – Breast Cancer Wisconsin (Diagnostic) Data Set: https://archive.ics.uci.edu/ml/datasets/breast+cancer+wisconsin+(diagnostic)

3. Learning Curve: <https://www.scikit-yb.org/en/latest/api/model_selection/learning_curve.html>

4. sklearn: <https://scikit-learn.org/stable/>

5. ICA: <http://www.sci.utah.edu/~shireen/pdfs/tutorials/Elhabian_ICA09.pdf>

6. Random Projections: <https://cims.nyu.edu/~cfgranda/pages/OBDA_spring16/material/random_projections.pdf>

7. TSNE: <https://danielmuellerkomorowska.com/2021/01/05/introduction-to-t-sne-in-python-with-scikit-learn/>

8. Silhouette Analysis: <https://scikit-learn.org/stable/auto_examples/cluster/plot_kmeans_silhouette_analysis.html>

9. Feature Selection: <https://scikit-learn.org/stable/modules/feature_selection.html>

10. Clustering Performance Evaluation: <https://scikit-learn.org/stable/modules/clustering.html#clustering-performance-evaluation>