Unsupervised Learning and Dimensionality Reduction

Darsh Thakkar (dthakkar33)

Code and Readme.txt: https://github.com/darshthakkar/Unsupervised-Learning-and-Dimensionality-Reduction

Introduction:

This report details the implementation and analysis of the project based upon two clustering algorithms namely K-means and Gaussian Mixture Maximization (GMM) via expectation maximization which will be applied to the two datasets from the first project. The other algorithms applied are dimension reduction algorithms - PCA, ICA, Random Projection and Feature Selection using Chi Square, these will be combined with the clustering algorithms and neural networks. The datasets used here are:

Wine Quality Dataset: This data consists 1599 samples with 11 features that give data and description of the wine and the classifier needs to categorize it based on quality (1- very poor, 10 - excellent). This dataset is interesting as the features are continuous, so we need to discretize the values for our classifier, some of the labels have no samples so the algorithm needs to adapt to that and the number of features is also large with all the features being almost equally important for classification.

Breast Cancer Dataset: This dataset consists 684 samples with 9 features that describe the mass to classify whether it's malignant or not. The dataset in on Kaggle and is provided by the Wisconsin breast cancer cytology report. This dataset unlike the wine dataset is cleaned and hence we might expect to get much better results. Moreover, the quality of the wine can be a subjective measure whether a mass is malignant is defined very clearly.

Both datasets don't have missing values, are normalized with min max scaling and are publicly available.

K Means

K-means clustering is a method of vector quantization, originally from signal processing, that is popular for cluster analysis in data mining. k-means clustering aims to partition n observations into k clusters in which each observation belongs to the cluster with the nearest mean, serving as a prototype of the cluster.

GMM via Expectation Maximization

In statistics, an expectation–maximization (EM) algorithm is an iterative method to find maximum likelihood or maximum a posteriori (MAP) estimates of parameters in statistical models, where the model depends on unobserved latent variables. GMM's are in a class of soft clustering algorithms where each instance is assigned a probability distribution across a set of discovered categories (probabilities of all categories must sum to 1).

Principal Component Analysis (PCA)

Principal component analysis (PCA) is a statistical procedure that uses an orthogonal transformation to convert a set of observations of possibly correlated variables (entities each of which takes on various numerical values) into a set of values of linearly uncorrelated variables called principal components.

Independent Component Analysis (ICA)

Independent component analysis (ICA) is a statistical technique for revealing hidden factors that underlie sets of random variables, measurements, or signals. ICA attempts to decompose a multivariate signal into independent non-Gaussian signals. The data variables are assumed to be linear mixtures of some unknown latent variables, and the mixing system is also unknown. The latent variables are assumed non-Gaussian and mutually independent, and they are called the independent components of the observed data. These independent components, also called sources or factors.

Random Projection (RP)

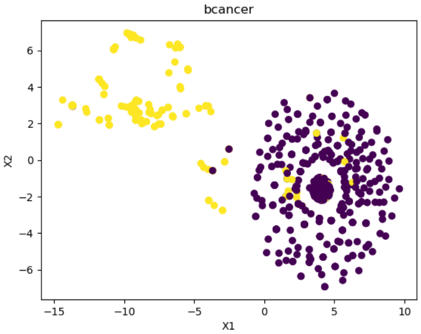
Random Projection is a technique used to reduce the dimensionality of a set of points which lie in Euclidean space. Random projection methods are powerful methods known for their simplicity and less erroneous output compared with other methods. The RP is implemented using the Python's Scikit Library.

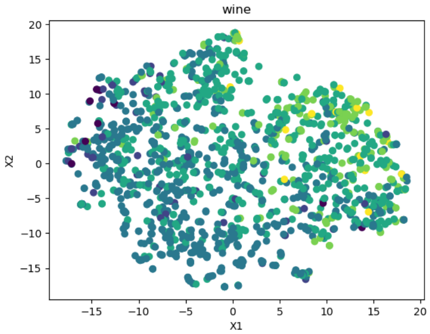
Feature Selection with Chi Square metric

Chi-square test is used for categorical features in a dataset. We calculate Chi-square between each feature and the target and select the desired number of features with best Chi-square scores. It determines if the association between two categorical variables of the sample would reflect their real association in the population.

Part 1: Clustering Algorithms on Datasets

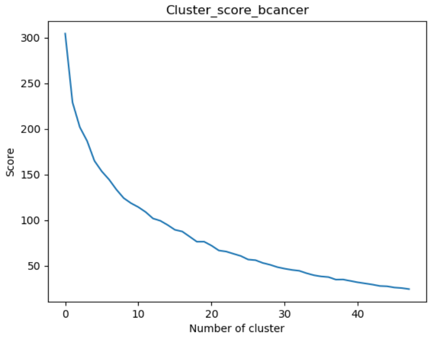
First the clustering algorithms will be applied on the classification datasets and the results will be compared with the actual labels. For the visualization of the clustering TSNE has been used to reduce the number of dimensions to 2 so that it can be plotted on a graph. The clusters with the colors according to the actual labels for both the datasets after applying TSNE are given below. The breast cancer dataset shows clear separation between the classes even in two dimensions, but the wine dataset labels are very mixed up in this 2-dimensional representation. The labels might be better separable in a 11-dimensional space but it's not possible to visualize it.

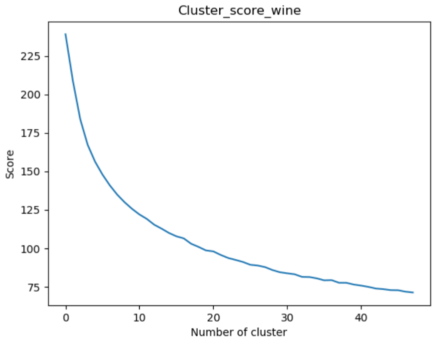




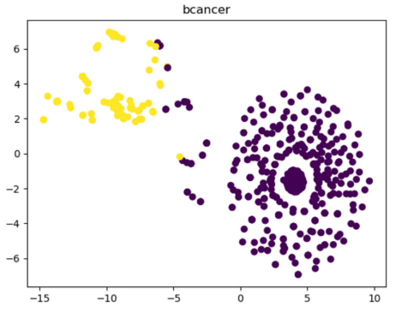
K Means

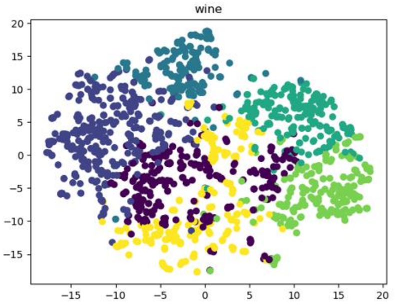
For k means, the k value for the wine dataset is 6 and for the breast cancer dataset is 2 as that is the number of labels in the classification. The algorithm was implemented with the scikit learn library in Python. Having fewer clusters than the number of labels doesn't make sense as there is bound to be overlap and on the other hand having more clusters might be helpful but here as we know what the labels are having the same number of clusters as labels, helps us compare the results better. Also from the graphs given below for both the datasets we can see that there is no clear elbow curve to select a value of k from. The graph shows the inertia score vs the number of clusters where, the inertia refers to within cluster sums or can be recognized as a measure of how internally coherent clusters are.





For these experiments, K means is run on the wine dataset which has 11 features and the breast cancer dataset which has 9 features. Below are the results for clustering using k-means and the visualized results with TSNE on both datasets.

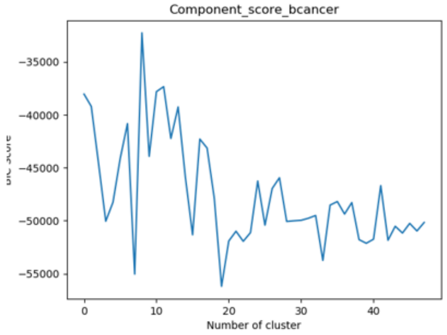


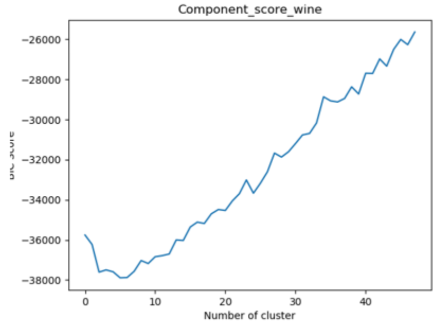


The results on the breast cancer dataset show the separation between the 2 classes after applying k-means apart from a few examples. On the wine dataset, k means also shows nice separation apart from some examples in the center of the cluster plot but it is when compared to the classification labels in each clusters the algorithm doesn't perform as well. There is a definite scope of improvement in the results, which I will try to achieve by introducing dimensionality reductions in later sections of the report.

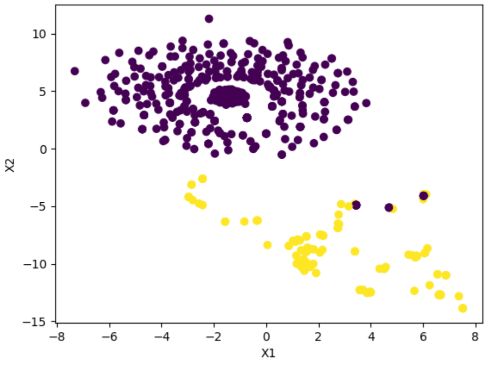
GMM with Expectation Maximization

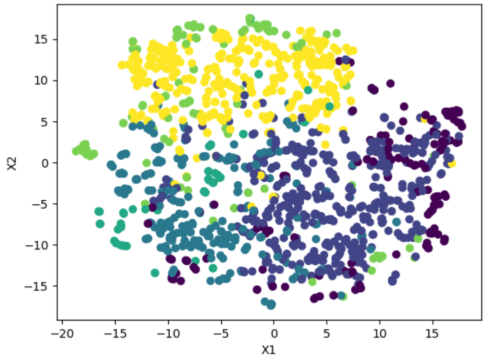
This algorithm unlike k-means assigns probabilities to each data point for which cluster they should belong to and the sum of the probabilities for each point is one. While plotting these clusters though the point is assigned to the cluster with the highest probability. To select the number of components for the GMM, BIC score is used and the graphs for the score vs the number of clusters are given below.





The graphs show that there is no clear winner for number of clusters in the breast cancer dataset so the number of label 2 is selected as the number of components. In the wine dataset the lowest BIC score is around 6 which is equal to the number of labels so that is selected from here on further. The implementation is done using the scikit learn library and the visualization of the cluster after applying the GMM algorithm to both datasets is given below.



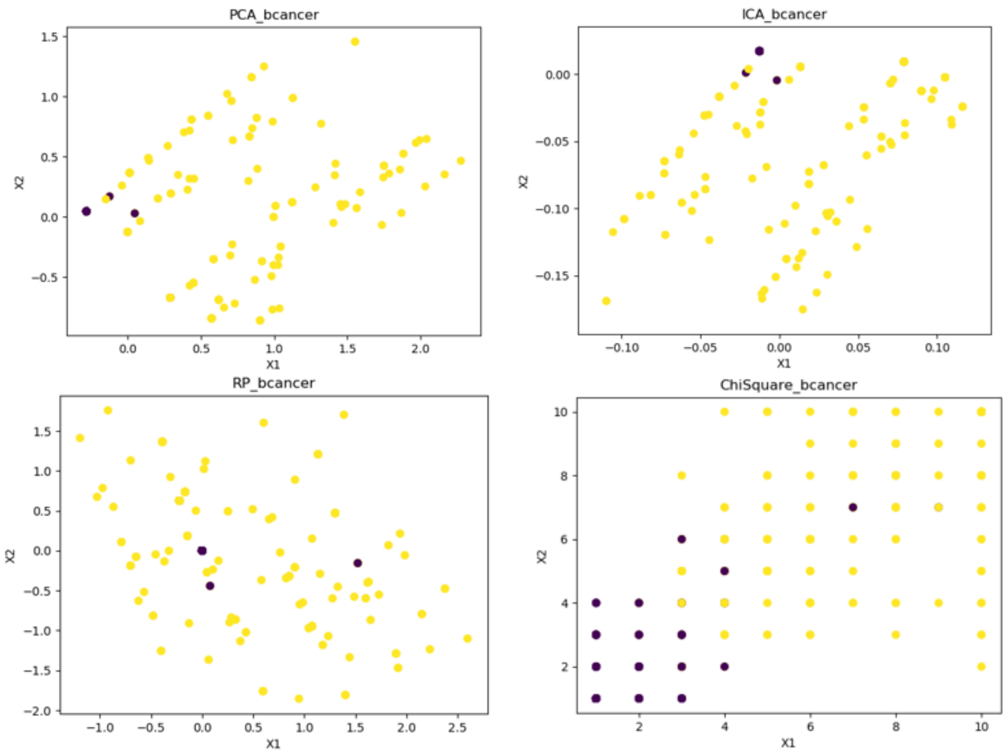


The visualization shows us that GMM works slightly better than k-means on the breast cancer dataset with a few more correctly labelled points and quite a lot better on the wine dataset as it has the same kind of mix of labels as the original visualization. Again as this representation is in two dimensions the wine dataset cannot be visualized well.

Part 2: Applying Dimensionality Reduction

This part deals with dimensionality reduction algorithms which are used widely in machine learning to reduce the dimensions of the feature space for any machine learning problem. Here we reduce the feature space from 11 dimensions and 9 dimensions to 2 dimensions for the wine dataset and the breast cancer dataset respectively. The algorithms are implemented using the scikit learn library implementation in Python.

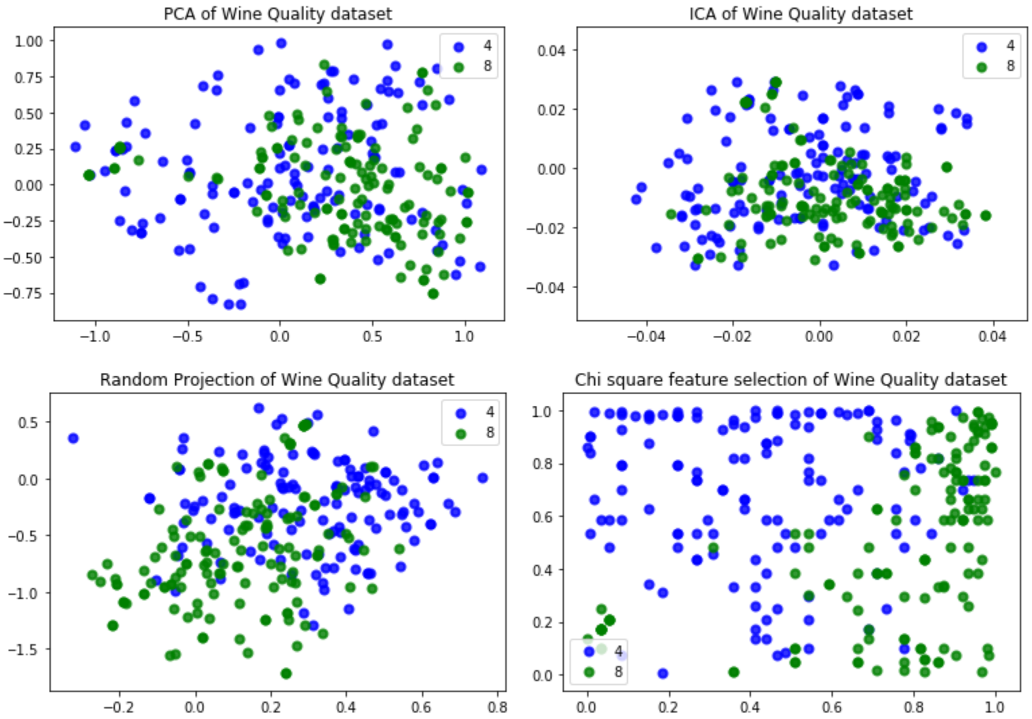
Breast Cancer Dataset



The results show that in case of PCA, ICA and Random Projection, the labels of one class are overlapped and can only be 3 points in the cluster plot. This uneven distribution of clusters for the breast cancer data set leads to problems in clustering with respect to the labels as can be seen in later sections of the report. This poor performance in dimensionality reduction may be due to the fact that we are converting a large number of dimensions 9 to 2 dimensions and thus the performance might be better for more dimensions but they cannot be visually represented in 2-D. Also, we can observe a pattern in PCA and ICA since their internal functioning is a bit similar but the patter observed in Random Projection is quite different. In case of Chi square feature selection, we get a very clear spread and distribution of points probably because this is a feature selection and not feature transformation method, plus it is also a supervised method and thus gets an edge over the other 3 techniques.

Wine Quality Dataset

For Wine Quality dataset, only labels of class 4 and 8 were plotted in order to receive a clearer visualization of how different algorithms are working, plotting everything together didn’t give much idea. The classes 4,5,6 cannot be separated by using both PCA and ICA, which is the reason why graph becomes very messy if those classes are also visualized. This seems to show that doing feature transformation, algorithms may not just separate the classes directly, but ideally the transformed values can provide better classification results when applying a feasible classifier. The min max normalized data was also used, but it doesn't change the separation of points for both algorithms. It's understandable since the variability of samples were kept the same while only the magnitudes of values were changed, so the percentage of variance explained by each component is the same.



From the graphs above, we can see that PCA definitely does a better job compared to ICA in separating the classes. On running Random Projection, not much difference was seen even after multiple trials (All points converging to x=y or x=1/y line). This is probably because random state=10 for all runs. After that, a univariate feature selection technique was used for selecting 5 base features based on chi square score. This method was selected as there seemed to be clear 4-5 features which impacted the result and thus provide better separation of classes. Mutual information score was also tried but chi square provided better results.

Overall

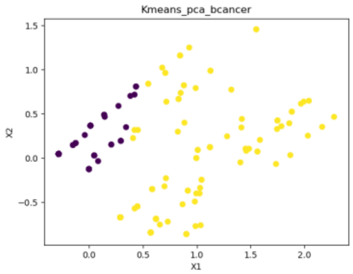
By looking at the graphs, we can observe that feature selection gives the best output, after which RP seems to be better than both ICA and PCA. However, PCA and ICA takes less time to compute. To compare the feature selection and transformation methods is not fair because the known classes information was used in the feature selection. The additional information will help with the feature selection, and it might lead to the selected and transformed features to do better in classification. But this supervised method also requires to check for overfitting problem, that's a big drawback of the supervised feature selection methods.

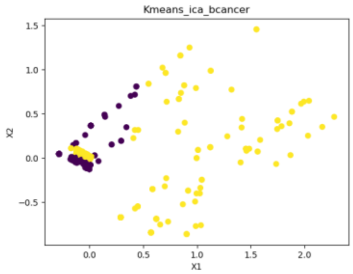
Part 3: Clustering Algorithms after Dimensionality Reduction

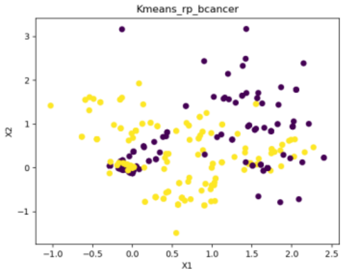
In part 3, we will first apply dimensional reduction on data and then using the transformed data we will do clustering and analyze the results that are obtained. In this section for Wine Quality dataset I used NMI score to determine which clustering works better, this is because from the graphs, it might not be clear which algorithm is working better as they look similar. For breast cancer dataset however, the data points are much more distinct and thus we can analyze the results from the graph.

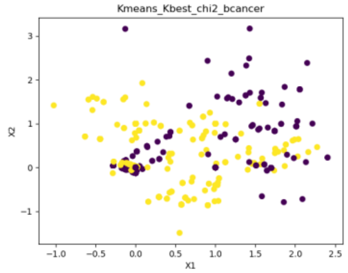
K Means

For breast cancer dataset, we can compare these graphs with our original clustering graph. We can observe that the results obtained from K means after PCA and ICA are not credible, this is probably because of what we saw in the previous part, that the data was not transformed properly in this case by PCA and ICA. These might be improved by increasing the number of reduced features. For RP and feature selection however, the algorithm seems to perform slightly better compared to the other two techniques and seems closer to the actual result.

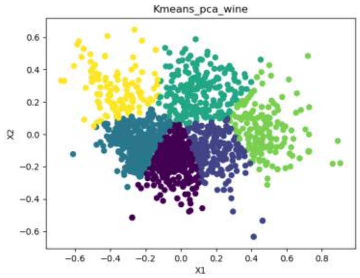


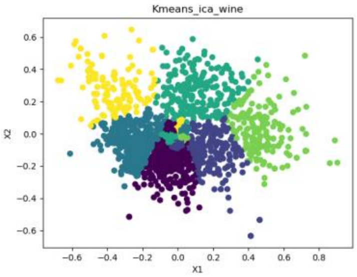


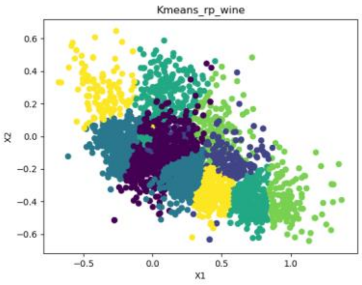


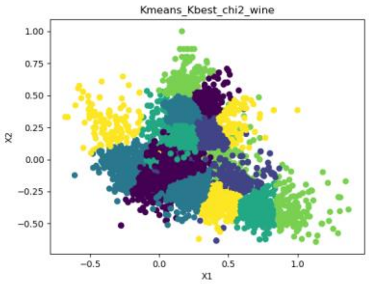


For Wine Quality dataset, NMI scores for the following 4 methods are 0.073, 0.088, 0.058, 0.071. The NMI scores did improve comparing to the raw feature values, but not much improvement when comparing to simple min max normalized features. The original NMI score was 0.056, thus all 4 algorithms do better than the original case probably because the outliers and sparsity might have reduced when the data was transformed. Overall, in this case data transformation using ICA and then clustering does a good job. In all these cases, however, for both datasets, runtime is much lower than the original run time but information is also lost.



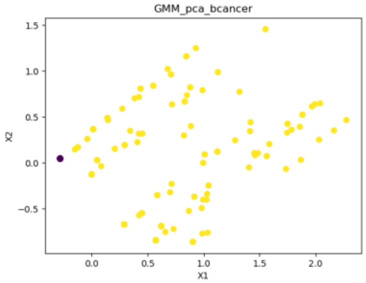


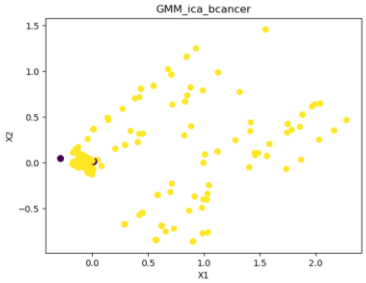


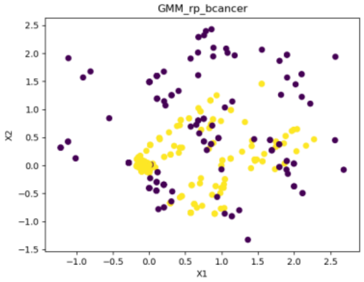


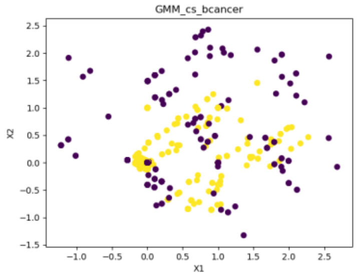
GMM via Expectation Maximization

In Breast Cancer dataset, the scenario is similar to K Means, however, we see some difference in labels in all figures, this is probably because GMM utilizes probabilities for classifying, and the probability difference between labels which are changed would be very less compared to the labels that haven't changed between the two algorithms.

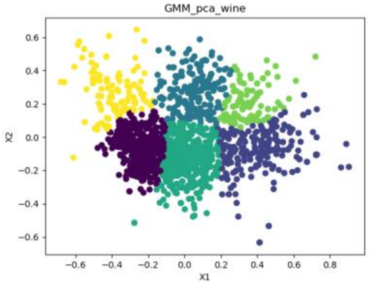


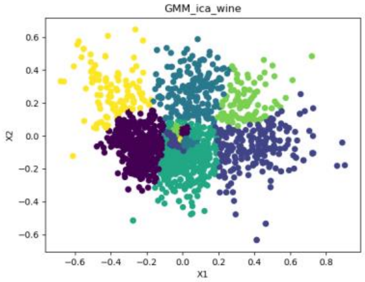


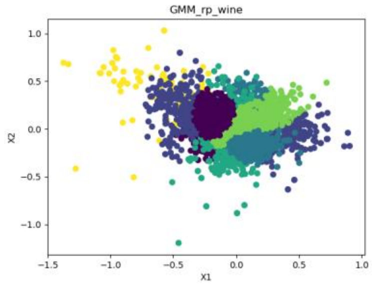


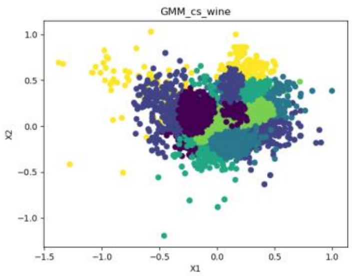


For the Wine Quality dataset, the NMI score of 4 transformed data set are 0.074, 0.092, 0.052, 0.074. The ICA transformed data get the best mutual information score again, the reason why ICA performs better might be because of some independent relations between the features, ICA can extract that independent information better than other ranking based methods.



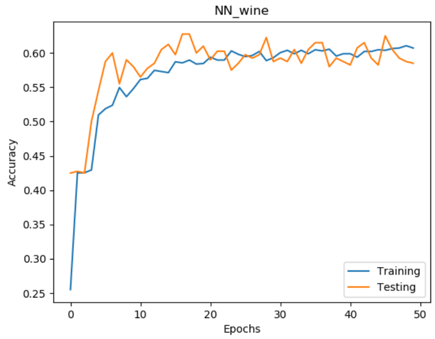


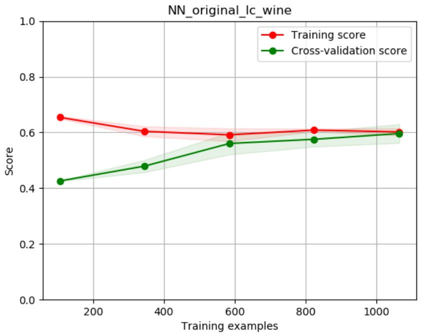


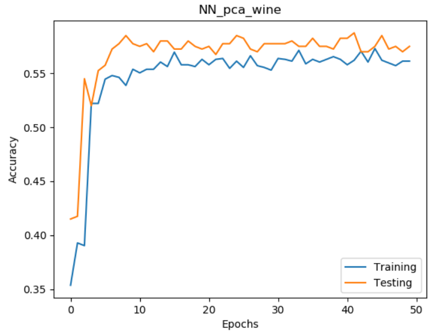


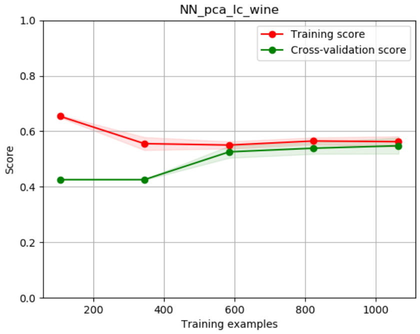
Part 4: Neural Network after Dimensionality Reduction on Wine Dataset

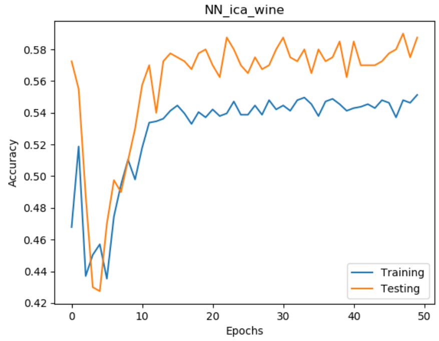
For this part, we use neural networks to train on the datasets after applying PCA to the dataset and reducing the number of dimensions in the feature space to 2. The neural network used here is the same as in the first project. It has two hidden layers with 64 and 32 units respectively and the outputs for training are one hot vectors. The neural network uses relu activation and is trained for 50 epochs as that is sufficient to get all the models to converge. The experiments below are done on the wine quality dataset. First graph row is of the original neural network graph without doing any dimensionality reduction on the dataset.

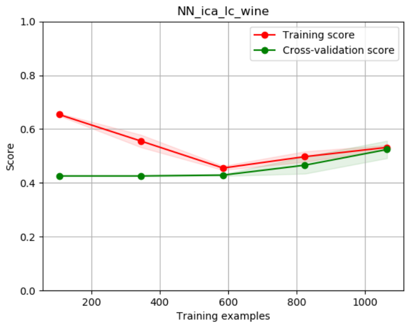


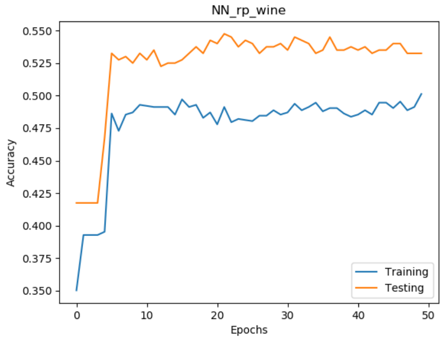


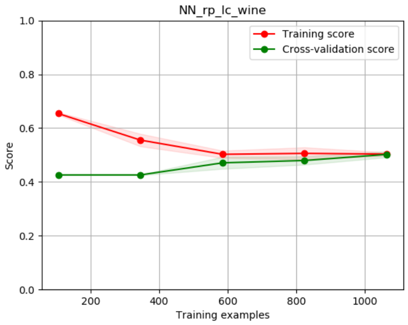


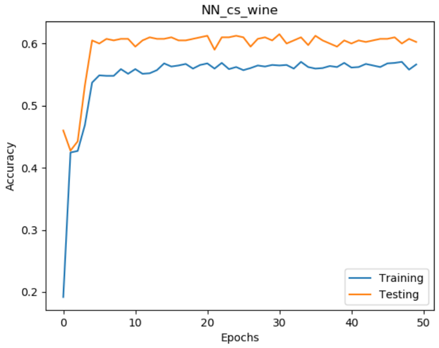


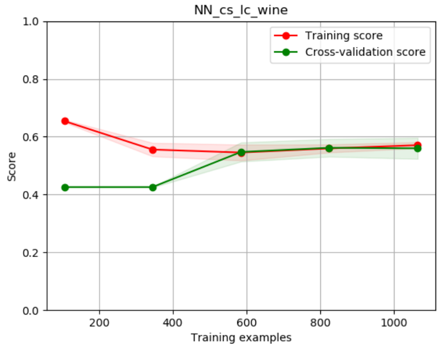












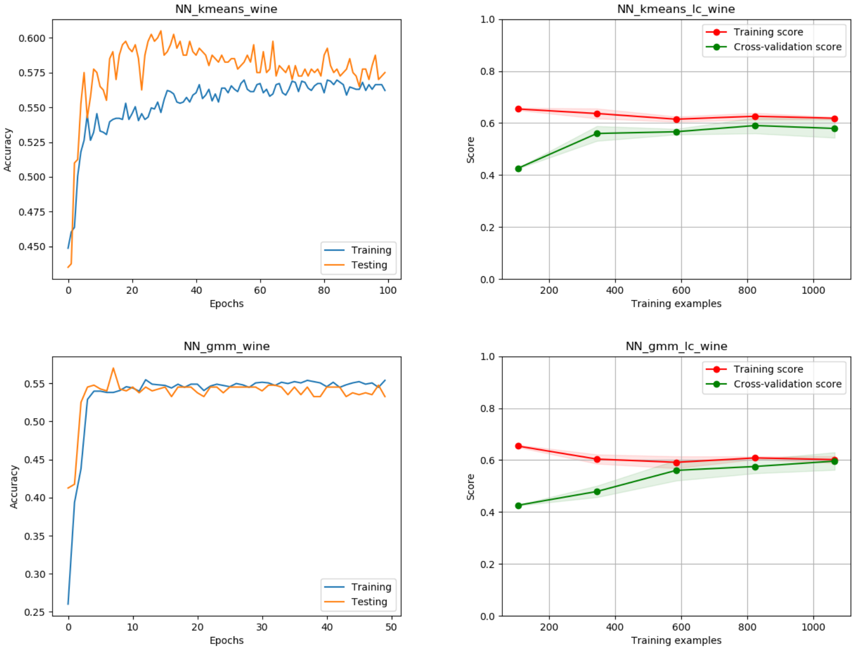
Using the 4 transformed data set on the previous optimal ANN model. The PCA and RP transformed data suffered a little bit on overfitting as the final training score is about 0.03 higher than CV score. Though the training score of different data set are different, the CV score is very similar at around 0.55-0.58. Comparing with previous ANN result of 0.6 optimal accuracy, these results do not show significant improvements. This is due to the fact that when the wine quality dataset is reduced to 2 features the labels are harder to separate leading to lower performance on the classification task. This method might improve if instead of reducing features to 2 dimensions, we reduce it to 4-6 dimensions, this is because we saw in Part 2 in Wine Quality dataset, we got a good separable distribution for feature selection when we selected 4-5 features, maybe that same analogy would be applied here.

Overall in this case, original ANN has the best accuracy and from the transformed data ANN, PCA and Feature Selection have better accuracy than the other two and RP has the worst accuracy. RP having the worst accuracy might be because RP makes it difficult to learn the sparse data points after transformation, thus if some classes had very few data points, they would not be learned properly after transformation.

Part 5: Neural Network after Dimensionality Reduction using clustering algorithms on Wine Dataset

In this final part of the report the aim to train the same type neural network used in part 4 but instead of using PCA, ICA, RP or Feature Selection as a means of dimensionality reduction, the clustering algorithms are used to achieve a similar result. Here a new set of features will be extracted from the clusters. The graphs for the performance of the neural network without any other algorithm is already given in the previous section which will be used for comparison in this section.

Going according the specification of the question first we run the K means algorithm on the raw data and set the location of the newly calculated clusters centers. Next a distance matrix is calculated, using scikit-learn, from all the clusters centers to each point in the feature space. For the wine dataset this reduces the dimensions from 9 to 6, as 6 is the respective number of clusters for the datasets. As explained previously, GMM are a soft clustering method and o each point is assigned a probability to being in a certain cluster. So first GMM is built from the data points and the soft clustering for each point is used as a feature to train the neural network. Results for both the tests are provided below.



Performance for the wine dataset dropped as expected because the reduction in dimensions leads to a new feature space where the labels are not neatly separable. This is an example of how dimensionality can hurt performance for classification tasks. In case of GMM, performance for the wine dataset dropped lower than the base result and even more than for the k-means experiment and the soft clustering or probability for points belonging to the cluster can be very close leading to some bad assignments. The general reason behind this technique not working affectively for this dataset is probably because clusters are used as features but the clusters themselves are not accurately formed, thus when neural network runs, it utilizes the wrongly clustered features.

Conclusion

1. Dimensionality reduction should be implemented on if there is no loss of spatial information, that is even in a lower dimensional feature space the data the algorithm should be able to find clusters that are differentiable.
2. Visualizing high dimensional problems can be difficult as they have to be reduced to 2 or at most 3 dimensions.
3. The breast cancer dataset has higher accuracy for all the experiments when compared to the wine dataset as even in lower dimensions there is a separating hyperplane for classification.
4. The advantage of dimensionality reduction is that it speeds up training time but as the datasets here are smaller the speed up isn’t significant enough.
5. K-means clustering works better with Binary classification where as GMM works better with Multi-Class classification.
6. Applying neural nets with Dimensionality reduction is advisable when number of dimensions are too high, otherwise normal neural nets works well.
7. Number of reduced features impacts the results achieved form transformed data.