CS7641 Assignment 3

Sahil Soni -ssoni41@gatech.edu

Unsupervised Learning and Dimensionality Reduction

CS7641 Machine learning, Assignment 3 , Dimensionality Reduction,Unsupervised Learning

Abstract: Kmeans,Expectation Maximization

Index Terms: Clustering,PCA,ICA,Randomized Projections,LDA,Neural Networks,Clustering Features

1. Introduction

This is third assignment is part of OMSCS CS7641(Machine Learning) for Spring 2020. In this task we will explore unsupervised learning algorithms.We will used clustering and dimensionality reduction algorithms .

2. Data Sets

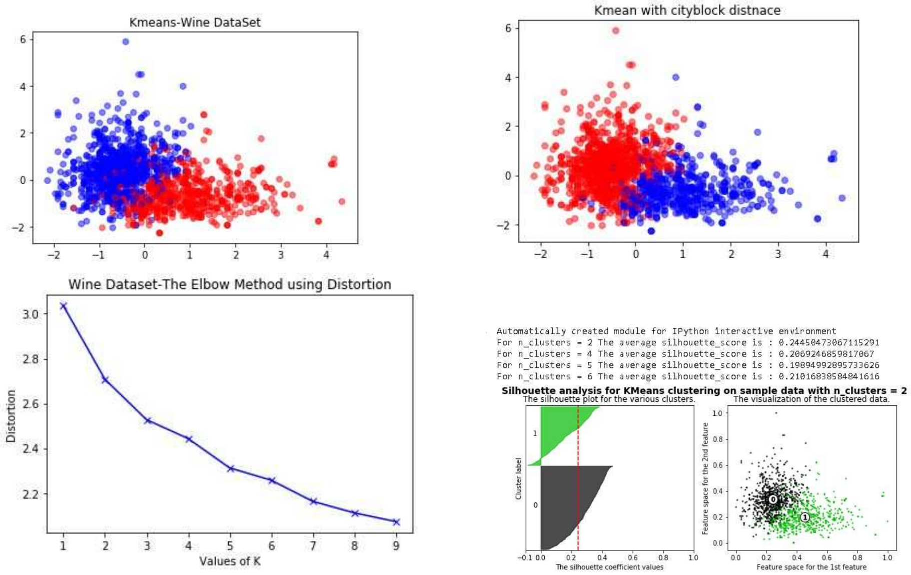
I have used both data sets as i used in first and second assignment. I have used Wine and Heart Data sets.We already have applied Supervised learning algorithms to these datasets . Now we will apply unsupervised learning algorithms to them .Wine dataset has 12 properties, and instances of 4898. For applying unsupervised learning we have removed Output label ”quality” from the dataset which we predicted on Supervised learning. Wine data set is unbalanced ,its target values are distributed unequally .Second Dataset , heart dataset consists of 14 attributes for which we have removed ”target” Output variable so we can apply unsupervised learning algorithms on this.Target values are 1:95 distributed equally.

3. Data Acquisition and Tools used

I have used Sklearn library along with Phton 3.7 . I have used Jupiter notebook and take help of Google cloud to run my experiments faster via Google colab . Datasets taken from Kaggle and UCI.

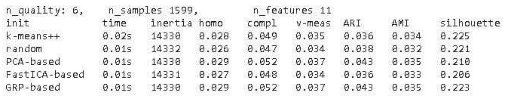
4. Metrics and Methodology

I have used Distortion and Inertia metric to find K Values of cluster for Kmeans . For Expectation Maximization i have used AIC and BIC metrics to find K Clusters. For dimensional reduction i have used Snsplot,Histograms and Variance to find no. of ideal dimensions for their algorithms .Thereafter we have used kurtotic and eigenvalues to find PCA and ICA components . Later we applied PCA/ICA/RP/LDA to neural network for Wine DataSet then used Clustering algorithms as features .



```
Automatically created module for IPython interactive environment
For n_clusters = 2 The average silhouette_score is : 0.24450473067115291
For n_clusters = 4 The average silhouette_score is : 0.2069246859817067
For n_clusters = 5 The average silhouette_score is : 0.19894992895733626
For n_clusters = 6 The average silhouette_score is : 0.21016838584841616
```

Fig. 1. Wine Problem-Kmean



```
n_quality: 6,    n_samples 1599,         n_features 11
init            time     inertia homo     compl    v-meas   ARI      AMI      silhouette
k-means++       0.02s    14330   0.028    0.049    0.035    0.036    0.034    0.225
random          0.01s    14332   0.026    0.047    0.034    0.038    0.032    0.221
PCA-based       0.01s    14330   0.029    0.052    0.037    0.043    0.035    0.210
FastICA-based   0.01s    14331   0.027    0.048    0.034    0.036    0.033    0.206
GRP-based       0.01s    14330   0.029    0.052    0.037    0.043    0.035    0.223
```

Fig. 2. Wine Problem-Metrics

5. Unsupervised Learning-Clustering

In comparison to supervised learning, Unsupervised Learning finds undetected correlations in a data collection without labeling. Clustering is one of the approaches used in Unsupervised Learning to combine events so that identical artifacts shape the same cluster.

5.1. Kmeans

Kmean is one of the easiest algorithms to split the dataset into K non overlapping clusters or groups. This would ensure that inter-cluster data points are as close as possible but still maintaining clusters as distinct (far) as possible.

5.2. Kmeans-Wine Analysis

Choosing distance and no. K is one of the most significant parameters. So first, we're going to test what distance parameters to use in Kmeans. I used Cosine, City Line, Euclidean distance for

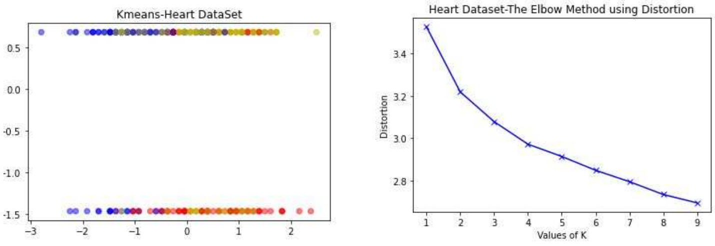


Fig. 3. heart Problem-Kmean

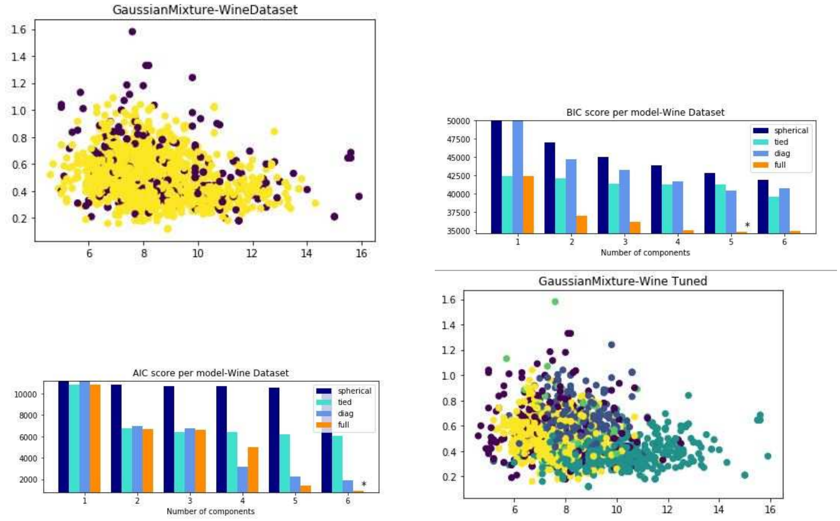


Fig. 4. Wine Problem-GaussianMixture

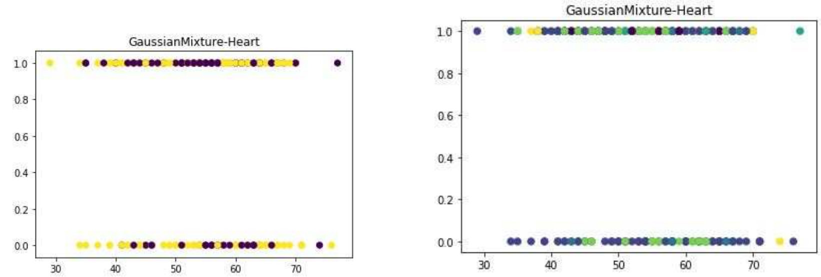


Fig. 5. heart Problem-GaussianMixture

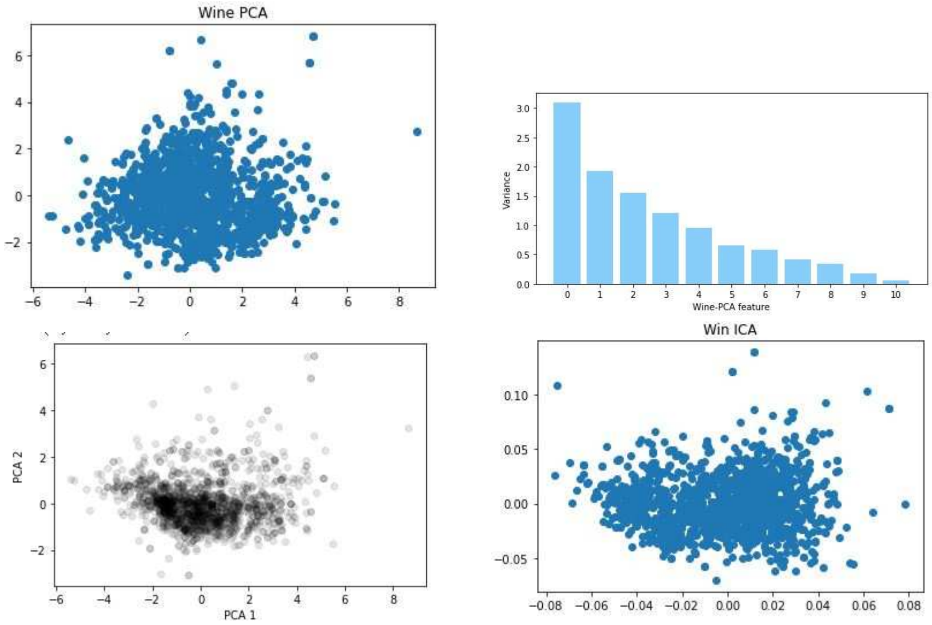
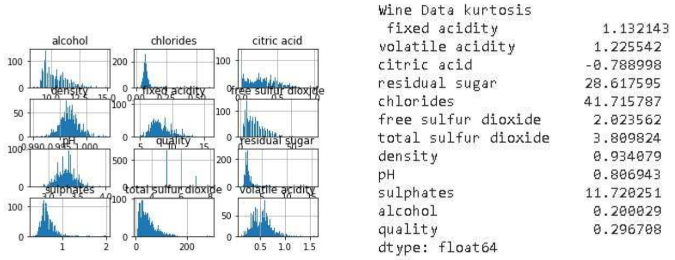


Fig. 6. Wine Problem-Dimensionality Reduction



```
Wine Data kurtosis
 fixed acidity            1.132143
volatile acidity         1.225542
citric acid             -0.788998
residual sugar          28.617595
chlorides               41.715787
free sulfur dioxide      2.023562
total sulfur dioxide     3.809824
density                  0.934079
pH                       0.806943
sulphates               11.720251
alcohol                  0.200029
quality                  0.296708
dtype: float64
```

Fig. 7. Wine Problem-ICA kurtotic

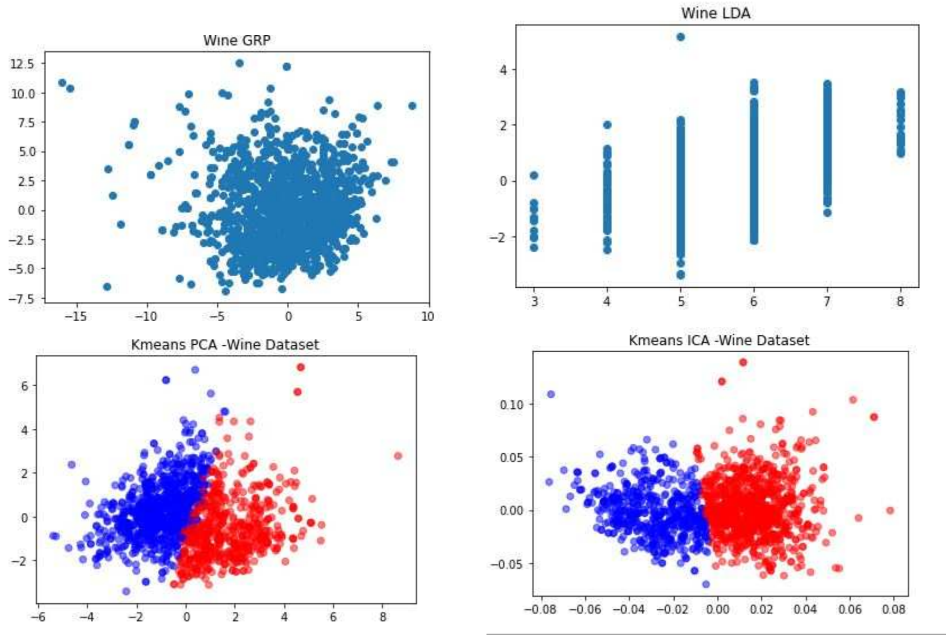


Fig. 8. Wine -Dimensionality Reduction Clustering

distance. I've played with all these three distances on both Wine and Heart Data set as seen in Figure 1.As we can see Euclidean distance cluster better than City Block radius. Sklearn often used the Euclidean interval by convention.

The reason for preferring Euclidean distance is that we have to be sure that the centroid of the cluster is equivalent to the sum of the pairwise squared Euclidean distances, which is Euclidean geometry itself. The specification of the euclidean interval between data points is thus consistent with K-means.

Then, for selecting no clusters, we used the Elbow approach utilizing Distortion and Inertia metrics. As seen in the Wine Dataset graphs, we have a strong Elbow at no. 2 clusters and we have selected no. 2 clusters for Wine Results.

To select the ground reality, we used the homogeneity metric, completeness ranking. Also for the option of a similarity indicator between two clusterings, we used Rand Metrics.

5.3. Kmeans-Heart Dataset Analysis

Similar to the wine study, we used the Elbow Test to determine a cluster number for core research. As we can see with the elbow form, there are also 2 perfect clusters with heart examination. Silhouette research was also used to test the inter-cluster gap between clusters.It is necessary to

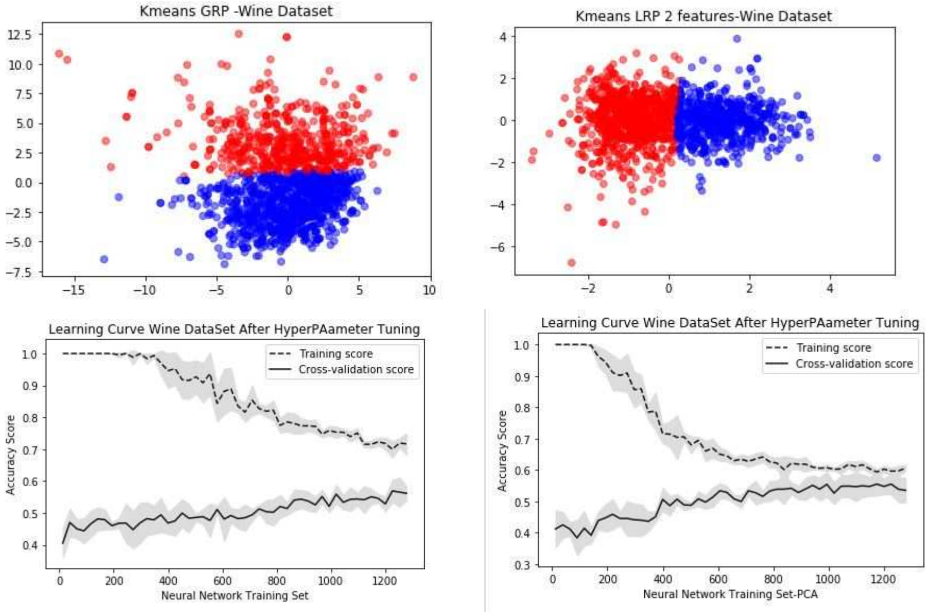


Fig. 9. Wine Problem-Dimensionality Reduction Clustering

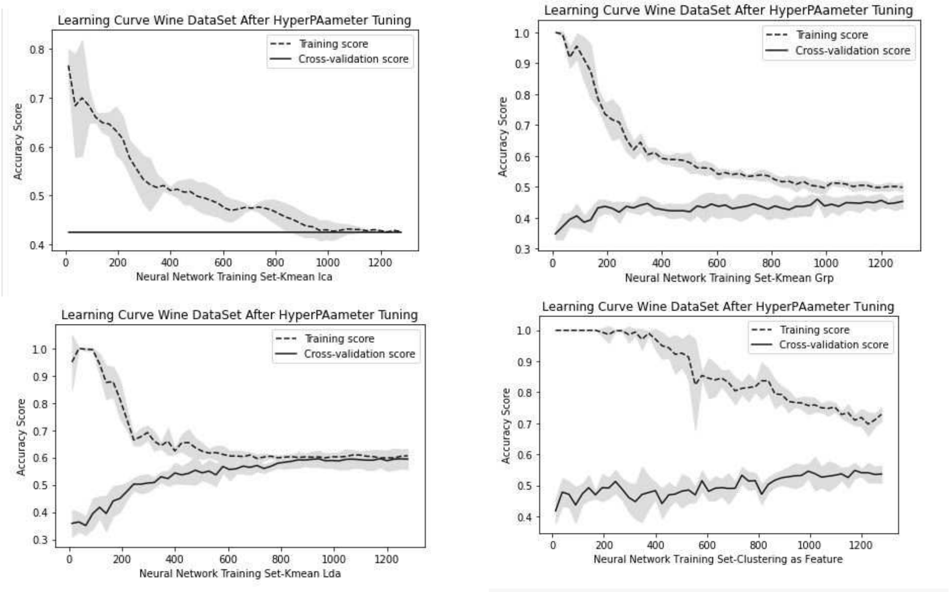
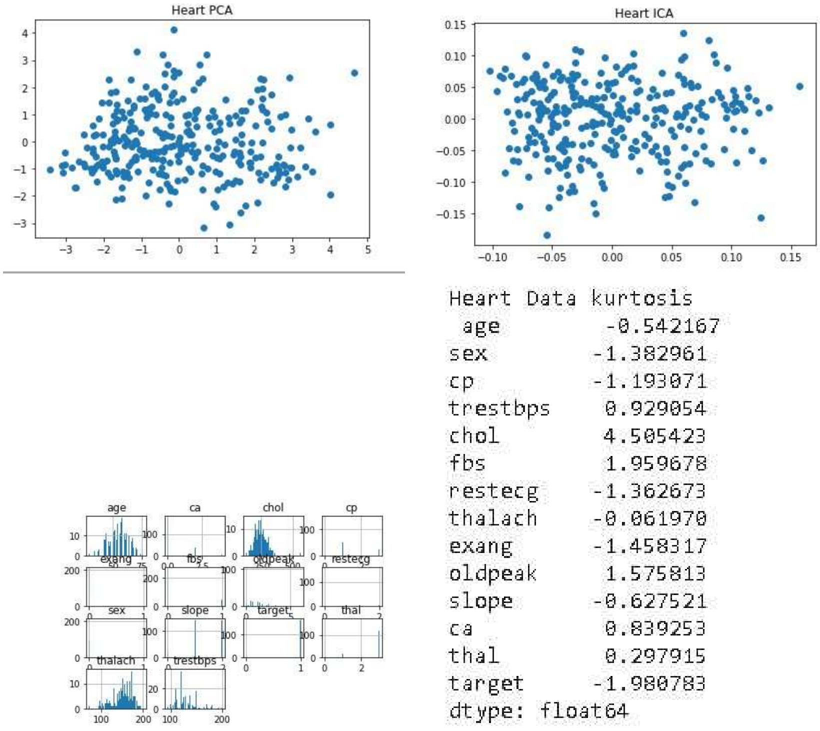


Fig. 10. Wine Problem-Neural Network



```
Heart Data kurtosis
 age         -0.542167
sex         -1.382961
cp          -1.193071
trestbps     0.929054
chol         4.505423
fbs          1.959678
restecg     -1.362673
thalach     -0.061970
exang       -1.458317
oldpeak      1.575813
slope       -0.627521
ca           0.839253
thal         0.297915
target      -1.980783
dtype: float64
```

Fig. 11. heart Problem-Dimensionality Reduction

see how near each point in a cluster is to the points in the neighboring clusters where no clusters can be located.

5.4. Expectation Maximization

Some of the issues with sklearn models is that we don't realize which cluster points come from other modules. That's when Expectation Maximization modules come to save this issue through iterative method.This maximized the likelihood of data points by estimating the likelihood of each variable. Continue to iterate to reach the expected local equilibrium.

5.5. Analysis-EM Wine Dataset

I used the Python Sklearn Gaussian mixture to apply EM Unsupervised learning on both datasets. As seen in the figure for choosing cluster number and whether to use(' spherical," tied," diag," full) templates, we used AIC and BIC metrics. As for the wine dataset, we have "Star" which has a low BIC score of 5 and a complete model which we have applied on the final Cluster.

5.6. Analysis-EM Heart Dataset

Similar to Wine analysis we have used BIC and AIC Score to find the no. of clusters and homegentiry metrics to find the ground truth among clusters. For heart dataset we have low score of metric at no. of cluster 6 and Spherical model metric so we have chosen same for heart dataset.Please see heart data set has almost flat and linear cluster and not grouping things properly .Same is true for kmean for heart data set. Almost flat curves there . Also Inters-tingly we have low score on Wine dataset for AIC/BIC metric compared to heart dataset so looks like Wine dataset more suitable for applying clustering . May be Heart dataset not suitable for Clustering. Let's apply Dimensionality reduction and find more on this.

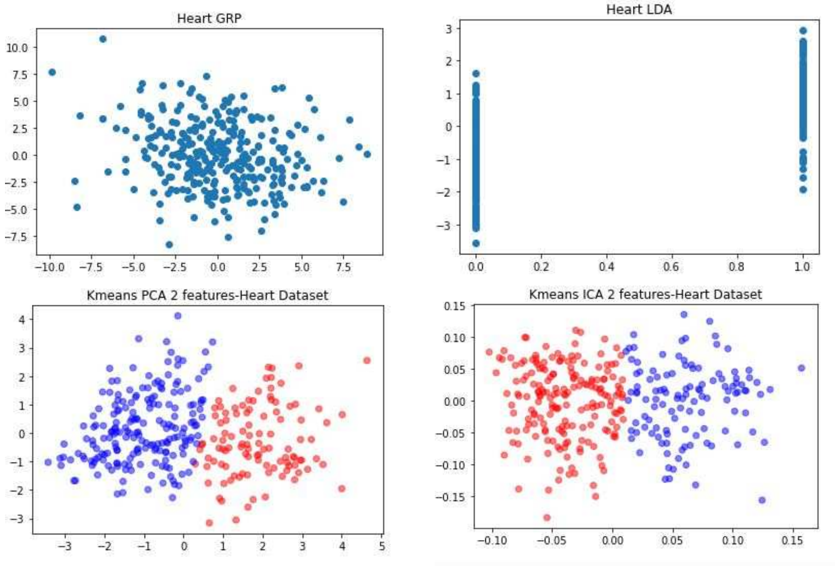


Fig. 12. heart Problem-Dimensionality Reduction

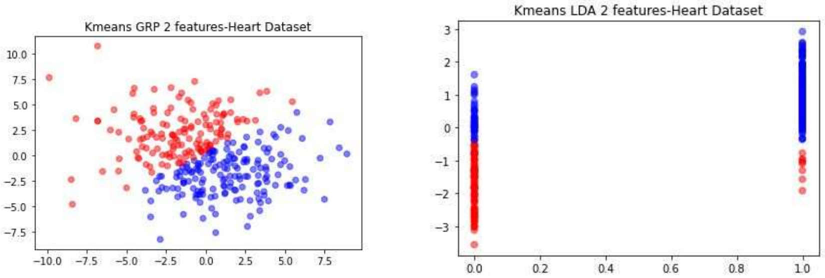


Fig. 13. heart Problem-Dimensionality Reduction Clustering

6. Dimensionality Reduction

As we know from previous assignments, ”Dimensionality Curse” is one of the main challenges on the datasets. It implies an exponential rise in the size of data induced by a huge number of measurements, features and attributes corresponding to the time and space complexity of the algorithms. Solution for this is Dimensionality Reduction.It is process of reducing the number dimensions using principal variables.

6.1. PCA

Principal Component Analysis(PCA) is most used and popular method of Dimensionality Reduction.PCA seeks the strongest linear variations of the initial variables in order to reduce the variation or the distribution over the current component. In PCA, the proprietary vector is the location or axis, and the resulting proprietary meaning is the difference along the proprietary vector. The higher the value in the patented track, the greater the variance in that section.

6.2. PCA-Analysis

As we have used Sklearn library which uses itself EigenValues built in . As we can in figure , to determine no. of dimensions for PCA-Wine Analysis we have used Varaince. As For Wine dataset more than 75 percentage of data has been on 4 dimensions so for wine dataset we have used n_4 for PCA Analysis. We also shown how much variance has 1st component of PCA to Second component. Please see dark portion for PCA1 and PCA2. More dark on middle means more strong co-orelation .Similar to Wine Data set , we used same approach to find PCA ideal dimensions . For heart looks like ideal components are 2 .Since PCA produces a subspace function that maximizes variation along the axes, it makes sense to standardize the data, particularly if it is calculated on different scales. That's what we did for both our datasets .

6.3. ICA-Analysis

ICA stands for Independent Components Analysis.Here we tried to find independent components The process for reduction of linear measurements. ICA kurtosis recovered individual components with a 1.5 kurtosis suggesting that they are less Gaussian than their origin. This may be the case, because the ICA is seeking to automate non-Gaussianism.So if we have more kurtosis score greater than (3) it means those components should be part of ICA.We have plot histograms of datasets and where tails are higher and higher kurtosis score to decide the ICA. so As we increased more dimensions our clusters data points become more scattered.

6.4. GMM-Randomized Projections-Analysis

Randomized Projections offers an simple and computationally effective way to reduce the dimensionality of the data .We used Gaussian random projection, which reduces dimensionality by putting the initial input space on a matrix generated randomly. There is also Sparse random projection on sklearn, which reduces dimensionality by utilizing a sparse random matrix to extend the original input space.We experimented with different dimensions and find for both datasets no. of dimensions should be 2. As we increased more dimensions our clusters data looks more ugly . It is also not sensitive to Noise.Without standard scalar it shows sphere shape for heart dataset .

6.5. LDA-Analysis

LDA stands for Linear discriminant analysis. This is meant to describe a linear combination of features that separates or characterizes two or more forms of objects or occurrences.As this is based upon linearity, so as the no. of its dimensions increased it should find linear combinations. We have for our both datasets as no. of components increased beyond 2 it performed worst. Both Wine and heart dataset its tried to separate data by linearity function .I think it should be more suitable for supervised learning or linear models.

7. Dimensionality reduction on Clustering

Here we will apply all above Dimensionality reduction on our both Kmean and EM clustering Technique and see how it performed.

7.1. Kmean/EM -PCA/ICA/RM/LDA Analysis

As we can see in figure , when we apply PCA to Wine datasets , now both clusters has better centriods and homogeneity score improved. Although still there are no impact on outliners. As we increased no. of dimensions it has little impact on PCA.Basically it improved shapes of clusters on all experiments. When we apply ICA to wine dataset variance among data points reduced. But here as we increased the dimensions data points scatters and model is not good. Same behaviour has been noticed on other algorithms although shaped of data points are according to model applied.When EM clustering applied to wine Data set only PCA performed better . ICA has negative AIC and BIC Score . All other Dimensionality reduction techniques worst on EM clustering on wine dataset.

Although initially Kmeans on heart data sets were linear structures and clustering were vertical line. When we apply PCA/ICA on heart dataset we can see better cluster shaped formed with good metrics score. Interesting was GaussianRandomProjection , when we apply on heart data set without Scaling it formed sphere shape cluster on Kmean . While after applying Standard

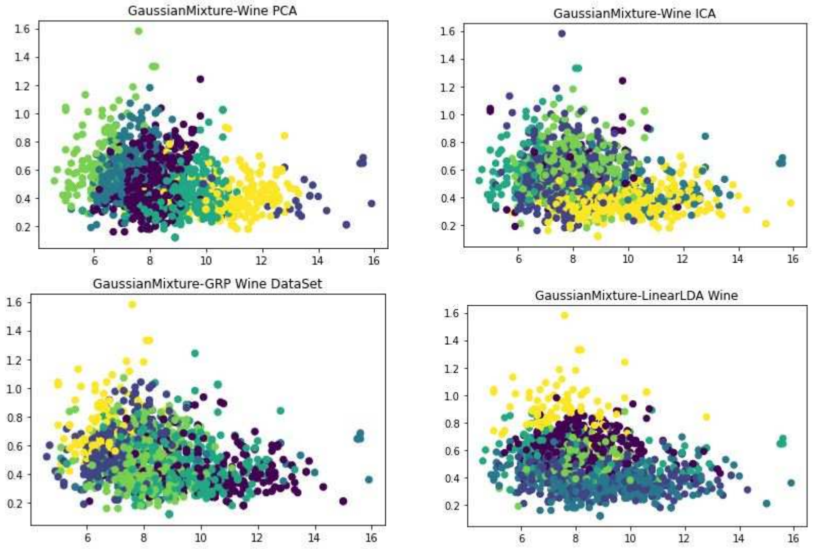


Fig. 14. Wine Problem-GaussianMixture

Scalar sphere shaped was no more. As expected , LDA resulted same linear cluster once applied to heart dataset. In case of heart EM clustering , dimension reduction has little impact and shape of cluster were same linear and almost no improvements on metrics and shape of cluster.

8. Neural Network -Wine Dataset

Here we will apply Neural network which we used on Assignment 1 . But our Data-set will be reduced using Dimensionality reduction.

8.1. PCA/ICA/RP/LDA on Neural Network

Here as we can see after applying Dimensionality reduction technique our neural network models took less time . We will used Wine data sets here .When we apply PCA to Neural network its run time has been improved but accuracy for train and test is almost same.But when we apply ICA to Neural network test accuracy almost flat . It means after Neural network do not learn anything once ICA applied on datasets.In case of Gaussian reduction , training and accuracy has been reduced compared to original and PCA Reduction method.although after applying LDA both train and test accuracy was improved but model was under fitting after 1000 samples. So finally only PCA was best dimension reduction method when we applied to Neural network it helped to reduce wall time and provide us same training and test accuracy.

9. Neural Network -Kmeans/EM Clustering Feature

Here we have apply first Kmeans/EM Clustering on our dataset and then chose truth values of our clustering algorithm as feature . We insert this feature to our data sets and then run neural networks models on this.

When we apply kmean as feature to neural network its training accuracy improved but test accuracy was same. Gap between training and validation curve suggests model was over fitting.Compared when we apply Expectation Maximization on neural network gap was low between training and validation curved compared to kmean. But still not as good as we see with original and PCA Datasets. May be if we have more data those features might have helped or may be i need increased weights to improved test accuracy. But again that's more testing we needed to confirmed which i will do in future if time permits .

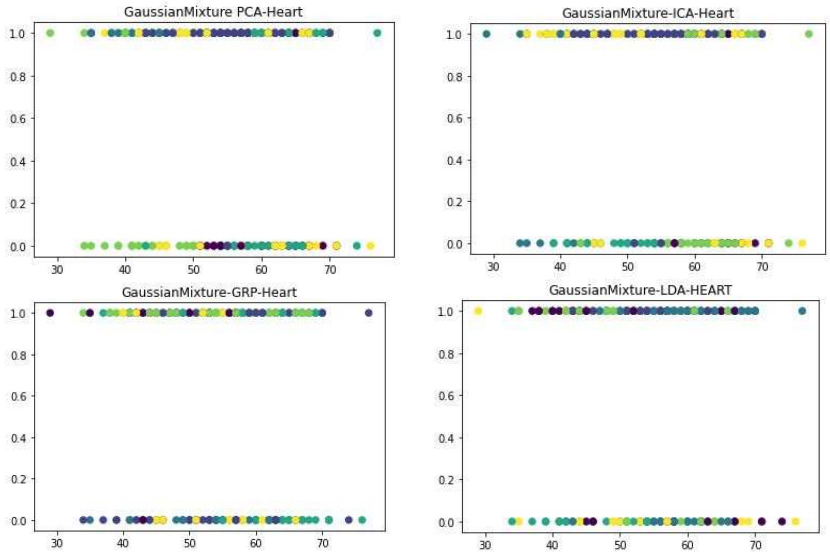


Fig. 15. heart Problem-GaussianMixture

10. Conclusion

Dimensional reduction improve the shape of cluster but based upon domain knowledges and more analysis we need to chose right method. For Wine dataset PCA works great but for Heart data set ICA gives better shape. No clustering algorithms is perfect it depends upon data as we see EM algorithm does not works well with heart dataset as dataset has linear features but for wine dataset after tuning PCA , EM algorithms provides better clustering shapes and scores.Finally for neural network PCA and GRP dimensions reduction method provides almost same accuracy with lower wall time . ICA and LDA introduces more bias and variance issue with less accuracy .While GRP almost similar to PCA but its shape are spherical .So its suited where PCA failed because of shape of cluster is sphere ot tiled.

References

https://www.kaggle.com/bburns/iris-exploration-pca-k-means-and-gmm-clustering

https://towardsdatascience.com/linear-discriminant-analysis-in-python-76b8b17817c2

https://stats.stackexchange.com/questions/81481/why-does-k-means-clustering-algorithm-use-only-euclidean-distance-metric

https://scikit-learn.org/stable/modules/generated/sklearn.cluster.KMeans.html