Predicting Diabetes and Heart Disease Using Features Resulting from KMeans and GMM Clustering

Kunal Sharma

CS 4641 Machine Learning

Abstract

Clustering is a technique that is commonly used in unsupervised learning in order to find structure in data without labels. A supporting technique that is used to reduce the dimensionality of higher dimensional data is dimensionality reduction, often with a method called Principle Component Analysis, or PCA. This paper will explore the use of two clustering algorithms K-Means and GMM as well as PCA on data related to the diagnosis of diabetes and heart disease.

Both datasets will be normalized so that features with higher magnitude will not weigh considerably more and so that training time will be reduced. Cross-validation will be used to determine the accuracy of the model. The dataset will be split randomly to create a training set with 80% of the data and a test set with 20%. Particular attention will be given to hyper parameter choices, various performance indicators, and clustering visualizations.

The datasets have been obtained from kaggle.com and their web addresses will be included in the references section. Python and Jupyter Notebook was used for development. Packages used for data manipulation and visualization include Numpy, Matplotlib, Pandas. Sklearn was used to implement machine learning methods.

This paper will have the following sections:

Section 1	Data Exploration
Section 2	Dimensionality Reduction and Clustering
Section 3	Clustering After PCA
Section 4	Neural Network Classification After PCA
Section 5	Neural Network Classification Using Extracted Features from Clustering

1. Data Exploration

1.1 Diabetes Diagnosis Introduction

Originally, this dataset is from the National Institute of Diabetes and Digestive and Kidney Diseases. The aim of this dataset is to use a pre-selected set of

diagnostic measurements to predict whether a patient has diabetes. The patients in this dataset are females of Pima Indian heritages that are at least 21 years of age.

This dataset has a total of 768 rows, a single target column (outcome) and seven medical predictor column detailed in the chart below.

Feature Description

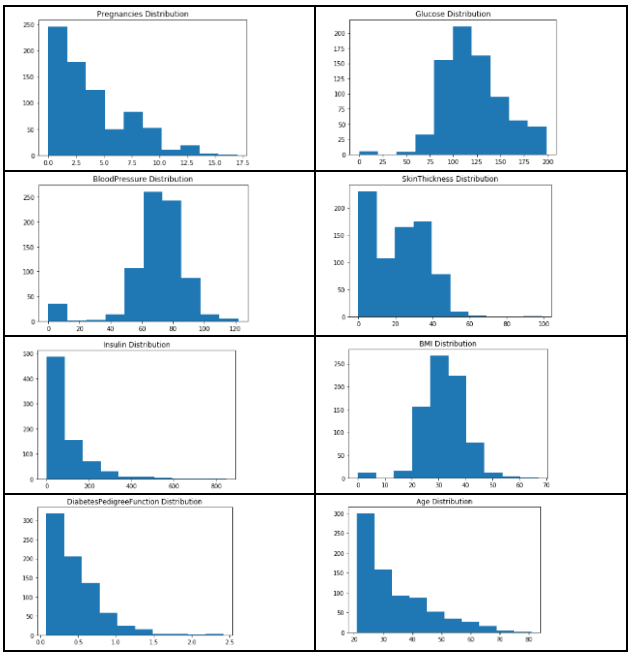
Pregnancies	Number of times pregnant
Glucose	Plasma glucose concentration a 2 hr in an oral glucose tolerance test
Blood Pressure	Diastolic blood pressure (mm Hg)
Skin Thickness	Triceps skin fold thickness (mm)
Insulin	2-Hour serum insulin (mu U/ml)
BMI	Body mass index (weight in kg/(height in m)^2)
Diabetes Pedigree Function	Diabetes pedigree function
Outcome	Class variable (0 or 1)

Outcome Distribution

65% of the data represents patients diagnosed with diabetes and 35% represent patients without diabetes. This dataset is therefore unbalanced and we should expect our accuracies should be above 65% if we expect our models to be predictive.

Feature Data Distribution

Glucose, Blood Pressure, and BMI follow a roughly normal distribution while the remaining features are typically skewed to the right.



Left to Right: Pregnancies, Glucose, Blood Pressure, Skin Thickness, Insulin, BMI, Diabetes Pedigree Function, Age

1.2 Heart Disease Diagnosis Introduction

Originally, this dataset is from the Cleveland database. The aim of this dataset is to also use a pre-selected set of diagnostic measurements to predict whether a patient has diabetes. The original dataset contained 76 attributes, however, this dataset contains the 13 most statistically significant attributes.

This dataset has a total of 303 rows, a single target column (outcome) and thirteen medical predictor columns detailed in the following table.

Feature Description

Age	Patient Age
Sex	1 = Male, 0 = Female
Chest Pain Type	4 Values associated with chest pain types
Resting Blood Pressure	Units in mm Hg
Serum Cholesterol	Units in mg/dl
Fasting Blood Sugar	>120 mg/dl

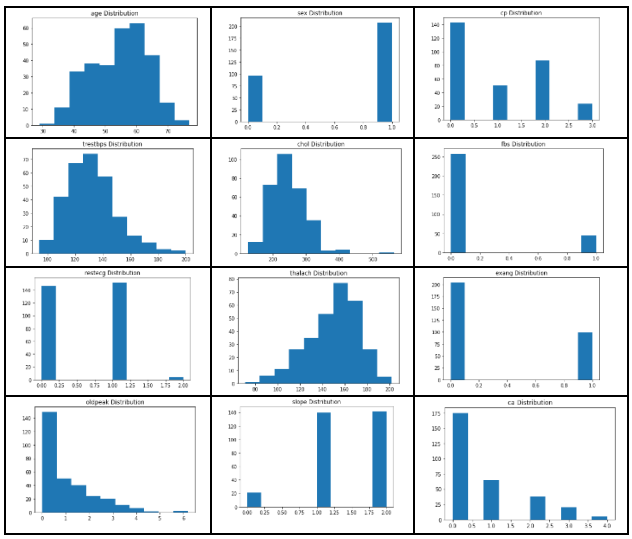
Resting Electrocardiographic results	Values 0,1,2
Maximum Heart Rate Achieved	Beats/Min
Exercise Induced Angina	numerical
Oldpeak	ST depression induced by exercise relative to rest
Peak exercise slope	Slope of ST segment
Number of Major Vessels	(0-3) Colored by flourosopy
Thal	3 = normal, 6 = fixed defect, 7 = reversible defect
Heart Disease	1 = Diagnosed, 0 = Not Diagnosed

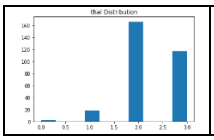
Outcome Distribution

46% of the data represents patients diagnosed with heart disease and 54% represents patients without. This dataset is nearly balanced and we should expect our accuracies should be above 54% if we expect our models to be predictive.

Feature Data Distribution

Left to Right: Age, Sex, CP, Trestbps, Chol, fbs, Restecg, Thalach, Exang, Old Peak, Slope, Ca, thal





2. Dimensionality Reduction and Clustering

In this section we will be applying Principle Component Analysis, K-Means Clustering, and Gaussian Mixture Model Clustering to our diagnostic datasets.

PCA is a dimensionality reduction technique that is used to summarize multiple features into only a few retaining as much variance as possible. More specifically, this method uses orthogonal transformations to convert the data into a set of values of linearly uncorrelated variables called principal components.

K-Means clustering is a method that attempts to separate the given data into K different groups that each have roughly the same amount of variance. Inertia or within cluster sum of square distance is essentially minimized in order to ensure that the total distance of all the data points to their respective cluster centroid is as small as possible. There are three main steps involved in the K-Means clustering algorithm. First, the algorithm initializes the cluster centers known as centroids. Each sample is then assigned to the closest centroid. Once all of the points have been assigned to a cluster, the centroid positions are recalculated based on the new groups. This process repeats until there is little to no change in the iterations.

The Gaussian mixture model is method that attempts to fit Gaussian distributions to clusters of the unlabeled data using the expected maximization algorithm. The expected maximization algorithm is an

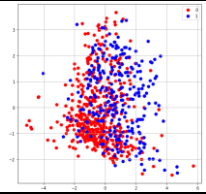
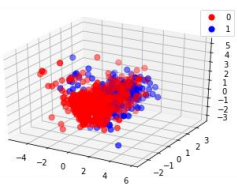
iterative process that ensures that the algorithm converges. Expected maximization first assumes random components and then calculates the probability of each point being generated by every cluster of the model and then the parameters are adjusted to maximize the likelihood of the data. This process repeats until convergence.

The elbow method is commonly used in order to find the optimal number of clusters. This method looks at the inertia value for several K values. One should select the k value after which the decrease in inertia is not as significant. It is important to note that this method is not always reliable in finding the best number of clusters.

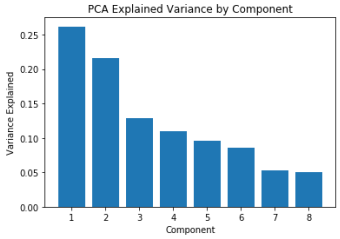
Key indicators include homogeneity, completeness, and silhouette. Homogeneity receives a high score when each cluster only contains members of a single class. Completeness receives a high score when all members of a given class are assigned to the same cluster. Silhouette explains how similar an instance is to its own cluster as compared to other nearby clusters.

2.1 Applying PCA to Diabetes Dataset

We will use principle component analysis to reduce the dimensionality of our feature set from 8 to 2 and 3. The data can now be visualized:

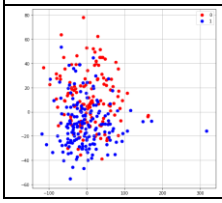
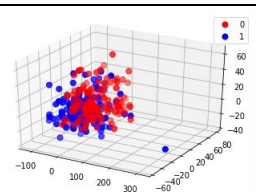
2D	3D
	

We can see that the negative instances are clustered mainly to the left and the positive to the right. There is however, significant overlap of the data which may make it harder for clustering algorithms to separate. It is also important to note that the principle components in 2 Dimensions only represent 47% of the variance in the original data, and in 3 dimensions only represent 59%. This means we must be extra careful when interpreting visualizations for this data.

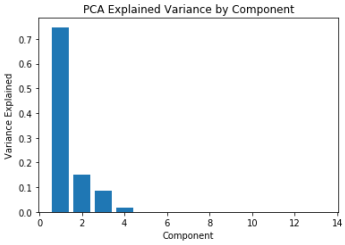


2.2 Applying PCA to Heart Disease Dataset

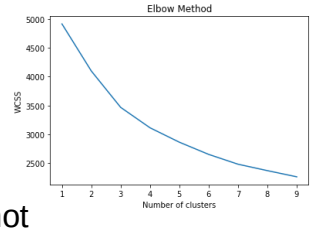
We will use principle component analysis to reduce the dimensionality of our feature set from 13 to 2 and 3.

2D	3D
	

With this dataset as well, we can see clusters have significant overlap. The 2d and 3d principle components of this dataset capture 87% and 96% of the variance in the data respectively. These results are significantly better than those of the diabetes dataset. This means that fewer features in this dataset contribute significantly more to the overall variance.

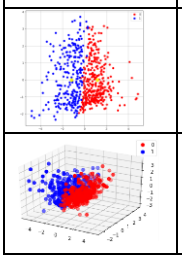
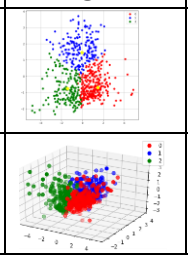
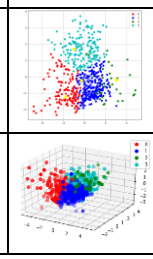
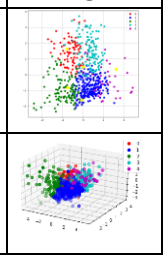


2.3 Clustering the Diabetes Dataset using K-Means



After graphing the elbow curve to determine the optimal number of clusters we find that the answer is not directly clear. This suggests that the data is not cleanly separable and this is likely given that it is in such a high dimension.

While it is not obvious, the elbow seems to be at 3 clusters. We can plot the clusters at nearby values to determine if 3 is a good choice.

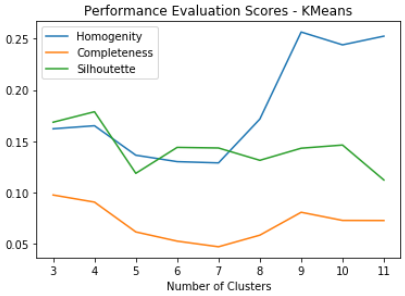
K = 2	K = 3	K = 4	K = 5
			

The plot at K = 2 seems most like the original dataset even though it seems to be divided vertically while the originally dataset seemed to be divided horizontally from the center. The accuracy of K = 2 in labelling the original points is very high.

```
KMeans Training accuracy for K = 2: 71.01
KMeans Testing accuracy for K = 2: 70.78
```

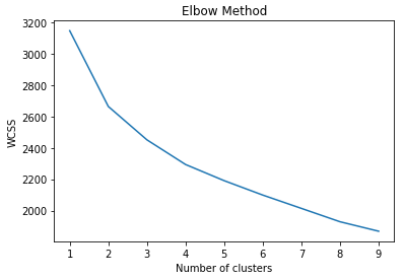
At K = 4 and K= 5 there is a significantly high amount of overlap between the data. K = 3 seems like the best choice.

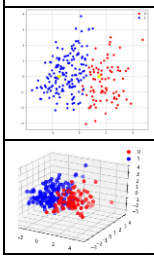
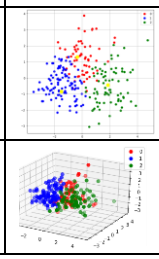
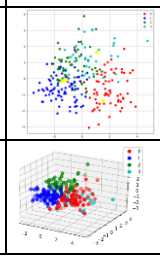
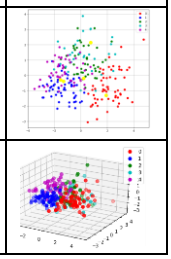
Homogeneity increases because there are more clusters, at around K = 3 or K = 4 all of the performance scores are high than the others.

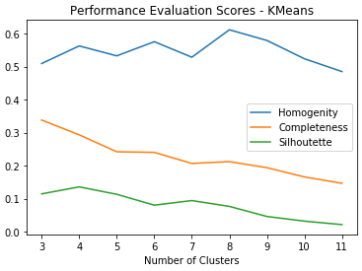


2.4 Clustering the Heart Disease Dataset using K-Means

Similar to the Diabetes data K = 3 seems to be the best choice. Performance metrics more or less stay constant.



K = 2	K = 3	K = 4	K = 5
			

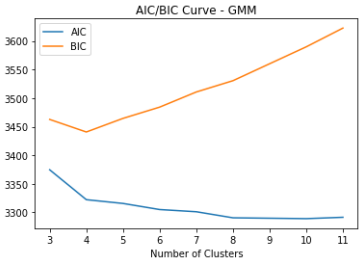


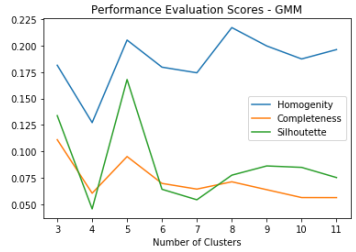
2.5 Clustering the Diabetes Dataset using GMM

```
Training accuracy for GMM for K = 2:
66.61
Testing accuracy for GMM for K = 2:
69.48
```

This accuracy is slightly lower than the accuracy for K-Means. This may be due to overlapping cluster.

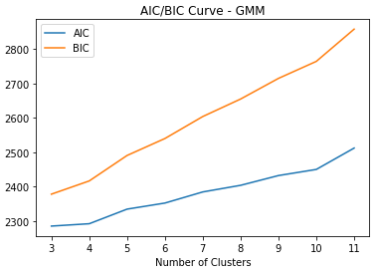
The AIC/BIC curve is the lowest at K = 4. All performance indicators jump at K = 5. This may be the best choice for K here, even though BIC is increasing at K = 5.

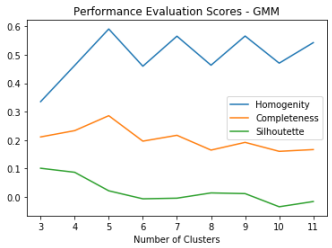


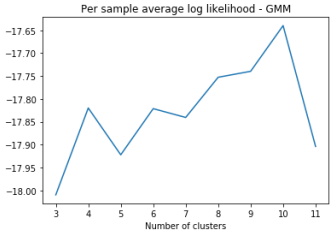


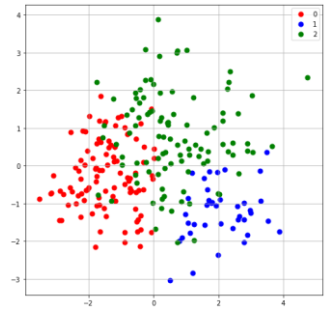
2.6 Clustering the Heart Disease Dataset using GMM

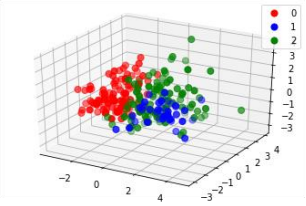
The AIC/BIC curve is the lowest for 3 and 4. Performance indicators stay relatively constant. There is a small bump in the log likelihood at K = 4.











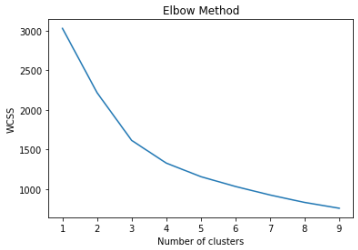
The above is a visualization of the GMM for K = 3 clusters.

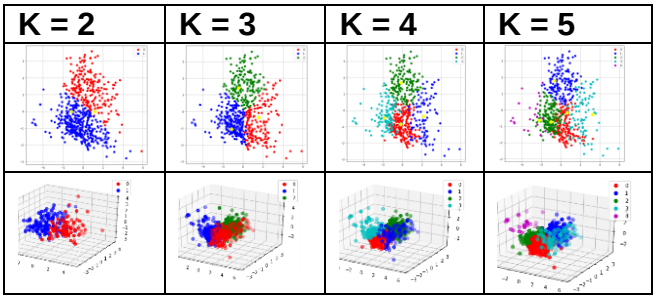
3. Clustering After PCA

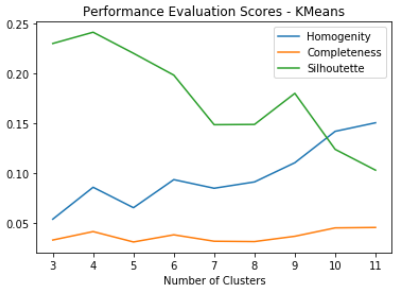
3.1 Clustering the Diabetes Dataset using K-Means after PCA

Since silhouette decreases after K = 3 and K = 4, we can say that K = 3 or K = 4 are the

best K values. The elbow method is more clearly 3 after PCA is conducted. The visualizations look the same.

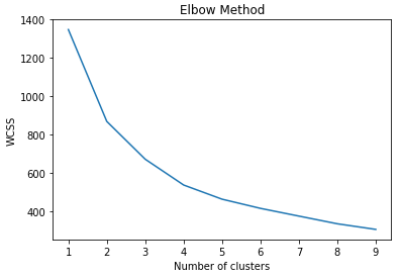


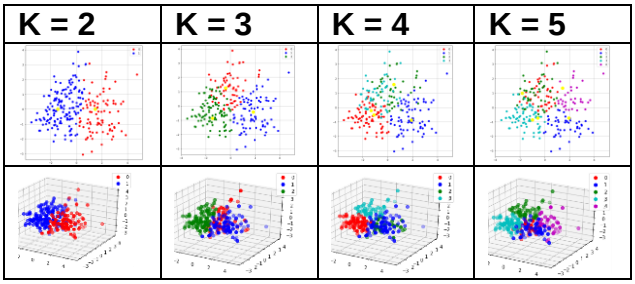




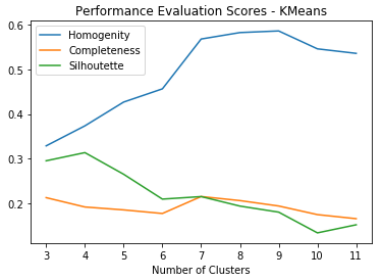
3.2 Clustering the Heart Disease Dataset using K-Means after PCA

Unlike for the diabetes dataset, after PCA the elbow method does not clearly indicate that the best value is 3. The visualizations still look very similar.



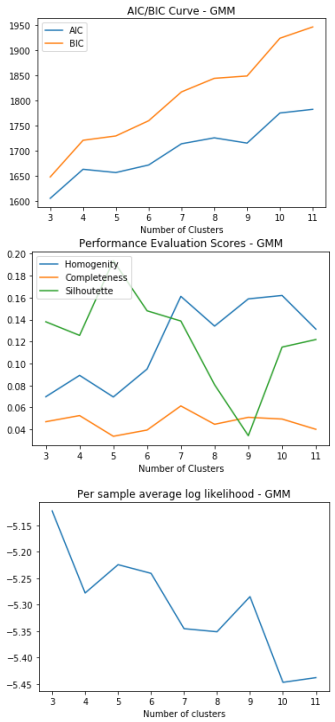


It is important to note that all of the performance indicators have higher values after PCA. This is likely because the dimensionality has been reduced so the points do not face 'the curse of dimensionality'.



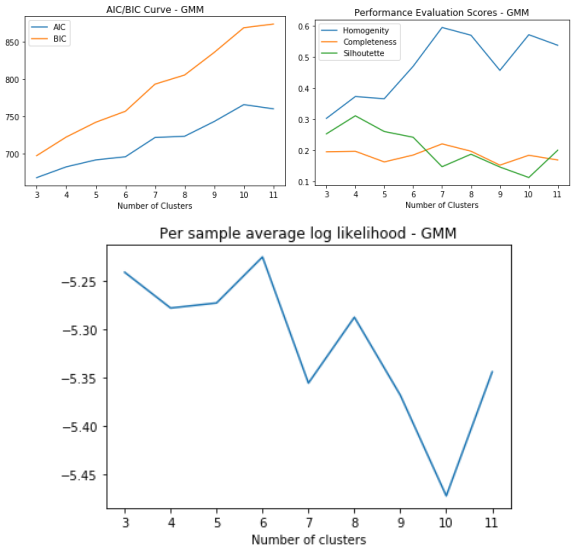
3.3 Clustering the Diabetes Dataset using GMM after PCA

The AIC/BIC curves are increasing. Performance evaluation scores are higher after PCA. The average log likelihood decreases as the number of clusters increases meaning greater accuracy (less loss). Silhouette decreases as K values increase while homogeneity goes up. This means that smaller K values are optimal here.



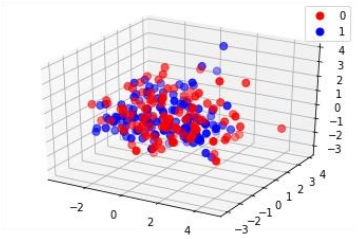
3.4 Clustering the Heart Disease Dataset using GMM after PCA

This data sets has the same results as the diabetes dataset after PCA on GMM.



4 Neural Network Classification After PCA

Instead of using the original data to train our neural network we can use the data after the original 13 dimension data is reduced to 3 dimensions using principal component analysis. The data can be visualized below:



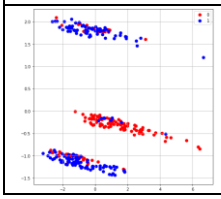
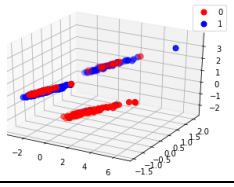
The neural network obtained an accuracy of 62.2% which is higher than the baseline prediction of 54% which means that our model is predictive. Although, its performance is still far less than when the neural network is trained on the original data directly (91% accuracy). This might be because the original data includes 100% of the variance in the data, while after post-pca data variance is lost.

5 Neural Network Classification Using Extracted Features from Clustering

5.1 Neural Network Classification Using Extracted Features from K-Means Clustering

Instead of using the original data we can use features from our K-Means clustering as a form of dimensionality reduction to train our model.

We will select the following features: Cluster classification, distance to closest cluster, distance to farthest cluster, average distance to all clusters. These features were chosen as they likely capture the relationship between the data. The feature set can be visualized below can clearly see three clusters.

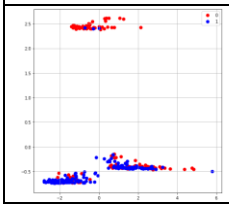
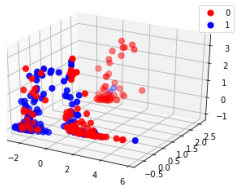
2D	3D
	

The neural network obtained an accuracy of only 55.7% after the training. This is much less than the baseline prediction accuracy of 65% and thus can be concluded as not predictive. This implies that the features extracted by the K-Means algorithm do not adequately represent the original data. It may be worth exploring different cluster values to improve the accuracy. We could also try inputting additional features from the K-Means model.

5.2 Neural Network Classification Using Extracted Features from GMM Clustering

Similar to what we did with K-Means, instead of using the original data we can use features from our GMM clustering as a form of dimensionality reduction to train our model.

We will select the following features: GMM classification, highest probability of belonging to a cluster, average probability of belonging to a cluster, and the weighted log probability. These features were chosen as they likely capture the variance and relationship between the data. The feature set can be visualized below, again we clearly see clusters. Since 87% of the variance is explained by the 2D pca we can be confident that our data is clearly clustered.

2D	3D
	

The neural network obtained an accuracy of 68.8% which is significantly higher than the baseline prediction of 54%. This implies that the features extracted by the GMM algorithm captured a considerable amount of variance of the original dataset. This performance is better than when the neural network was trained on the pca results (only 62.2% accuracy). This accuracy however is still far less than the accuracy of 91% when the neural network is trained on the orignal data. Again, to improve the accuracy we can explore different cluster values and try adding addittional features from the Gaussian Mixture Model.

6 Conclusion

Dimensionality reduction using PCA is a powerful way to represent data in lower dimensions. After PCA, performance metrics consistently increased for the clustering methods. However, training with data that has undergone dimensionality reduction results in lower accuracy likely because much of the variance from the original data is lost. While features extracted by clustering algorithms like K-Means and GMM can be used as a dimensionality reduction technique, the classification performance does not match when neural networks are trained on the original data by a large margin. Future work can attempt to additionally include the features extracted by clustering into the original feature set in an attempt to increase classification accuracy.

7 References

Diabetes Dataset:
https://www.kaggle.com/uciml/pima-indians-diabetes-database

Heart Disease Dataset:
https://www.kaggle.com/ronitf/heart-disease-uci