Machine Learning Assignment 2

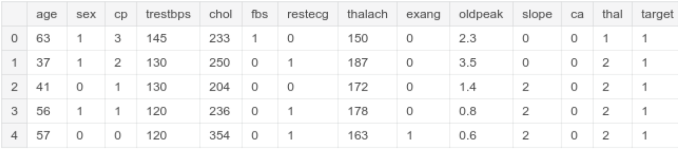
Chandragupta
M.Tech Computer Technology
2019EET2341

Abhishek Roy
M.Tech Computer Technology
2019EET2337

***Abstract*—This document has been submitted towards partial fulfillment of the course ”Introduction to Machine Learning” taught by Dr Prathosh A.P during fall 2019 under the code ELL 784.**

I. Heart Disease Dataset

A. Dataset Description



	age	sex	cp	trestbps	chol	fbs	restecg	thalach	exang	oldpeak	slope	ca	thal	target
0	63	1	3	145	233	1	0	150	0	2.3	0	0	1	1
1	37	1	2	130	250	0	1	187	0	3.5	0	0	2	1
2	41	0	1	130	204	0	0	172	0	1.4	2	0	2	1
3	56	1	1	120	236	0	1	178	0	0.8	2	0	2	1
4	57	0	0	120	354	0	1	163	1	0.6	2	0	2	1

Fig. 1. First Five Data

B. Methods used

- GMM and PCA.
  - GMM with 2 cluster
  - PCA(2 component) and GMM with 2 cluster
  - PCA(2 component) and GMM with 3 cluster
  - PCA(3 component) and GMM with 2 cluster
- KMean and PCA
  - KMean with 2 feature and 2 clusters
  - KMean followed by PCA with 2 cluster
- SVM

C. GMM and PCA

The Heart disease dataset is Labelled dataset and hence to apply to unsupervised algorithms , we removed the labels and apply various models. Principal component analysis involved reducing the dimensions to K principal components which is selected according to how much variance it account for on which later a classification algorithm has been used. It involved following steps:

1) Calculation of mean: The mean of every column is being calculated.
2) Centering the data: Each column is subtracted from its mean to position the data centered around its mean.
3) Calculation of covariance matrix: Here covariance matrix is calculated using the numpy library itself.
4) Eigen values and Eigen vectors: Using the covariance matrix, singular value decomposition is performed to calculate eigen values and eigen vectors.
5) Selecting the Principal components: After the calculation of eigen vectors, the principal components are selected. Each column of eigen vectors can be a principal component.

In GMM to initialize μ, Σ, ϕ,we picked random data point as μ ,data covariance as Σ and 1/k as ϕ where
then we iterate through E-step and M-step which is terminated by maximum no of iteration. EM typically converges to a local optimum, not necessarily the global optimum, with no bound on the convergence rate in general. It is possible that it can be arbitrarily poor in high dimensions and there can be an exponential number of local optima.

$$W_j^{(i)} = \frac{\phi_j \mathcal{N}(x^{(i)}; \mu_j, \Sigma_j)}{\sum_{q=1}^{k} \phi_q \mathcal{N}(x^{(i)}; \mu_q, \Sigma_q)}$$

$$\phi_j = \frac{1}{N} \sum_{i=1}^{N} W_j^{(i)}$$

$$\mu_j = \frac{\sum_{i=1}^{N} W_j^{(i)} x^{(i)}}{\sum_{i=1}^{N} W_j^{(i)}}$$

$$\Sigma_j = \frac{\sum_{i=1}^{N} W_j^{(i)} (x^{(i)} - \mu_j)(x^{(i)} - \mu_j)^T}{\sum_{i=1}^{N} W_j^{(i)}}$$

In Kmeans we initialized means by randomly picking data according to no of cluster.then we cluster the entire data according to norm distance. we then take the mean of cluster as new centroide.this continuous till the change in centroide is insignificant or it reaches maximum no of iteration.

1) GMM with 2 cluster:
   We applied Gaussian Mixture Model with two and three clustering, taking two features having maximum variance(age and thalach).in fig.2 top left image shows original input with labelled data,top right shows input to GMM and bottom shows GMM output with two cluster and three cluster respectively.The image shows that GMM with K=2 cluster the data quit well but with K=3 it causes overfitting.
2) PCA(2 component) and GMM with 2 cluster:
   we apply PCA to dataset and picked two components max. variance PCA1=.72 pca2=.14in fig.3 top left image shows original input with labelled data,top right shows input to GMM and bottom shows GMM output with two cluster and three cluster respectively
3) PCA(3 component) and GMM with 2 cluster:
   We choose 3 principal components with variance ra-

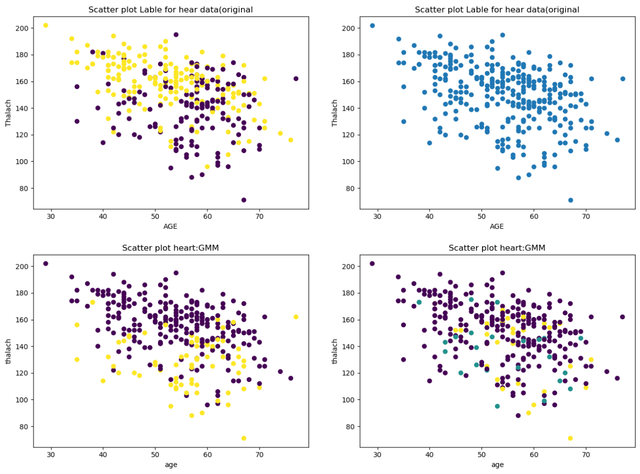


Fig. 2. GMM with 2 and 3 cluster

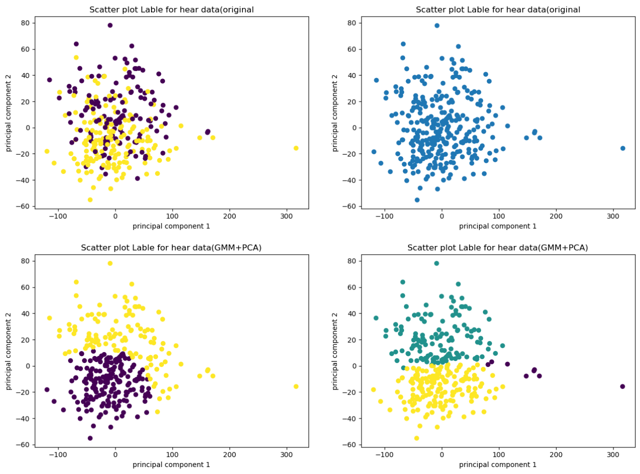


Fig. 3. PCA(2 component) and GMM with 2 cluster and 3 cluster

tio PCA1=.72 PCA2=.15 PCA3=0.084.then we applied GMM with 2 cluster.fig 4

D. KMean and PCA

1) KMean with 2 feature and 2 clusters
We applied KMean to the dataset, for visualization we pick two feature with most variance i.e Age and Thacal original class and input is shown in fig 2, output is shown in fig5 left
2) KMean followed by PCA with 2 cluster
in this , we reduce the data to 2 dimension using PCA with variance PCA1=.72 PCA2=.14.,then applied Kmean to find cluster(k=2).
CONCLUSION: GMM gives soft boundary in contrast to Kmean.GMM is more accurate than Kmean but GMM is prone to have singular co variance matrix during Maximization step. We try few methods like adding small value to diagonal element,keeping track of covariance matrix and having a copy of previous covariance and update to previous one if the new one is singular, these work but not always.

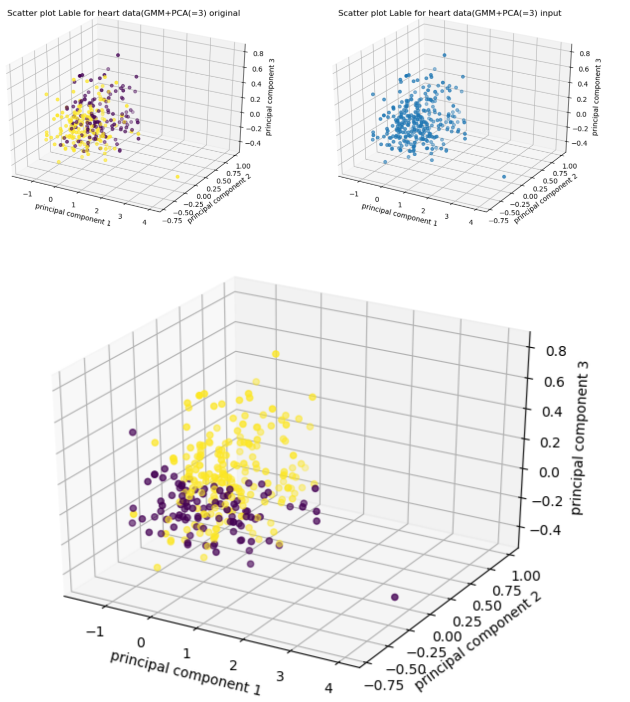


Fig. 4. PCA(2 component) and GMM with 2 cluster and 3 cluster 1.Actual 2.Input 3.Output

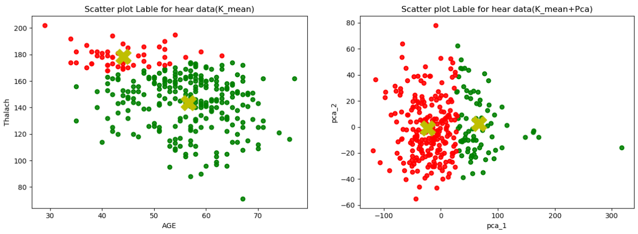


Fig. 5. KMean with 2 feature/PD and 2 clusters

E. Support Vector Classifier

We performed PCA using inbuilt library to visualize the data after dimension reduction. Unlike the Fashion dataset (in the following pages) where we used pipelined SVM and PCA, here we separately used SVM classifier.We trained the SVM model for different kernels : Linear, Gaussian and Polynomial to comapre the result. The best accuracy was 80 percent.

Kernel	Accuracy
polynomial	73
Gaussian	54
linear	80

II. MNIST Digit Dataset

A. Dataset Description

MNIST dataset is labelled data consist of 784 features which correspond 10 different classes. in this dataset we applied

Kmean and GMM to form cluster.we reduce the dimension to 2 using PCA.Each image is 28 pixels in height and 28 pixels in width, for a total of 784 pixels in total. Each pixel has a single pixel-value associated with it, indicating the lightness or darkness of that pixel, with higher numbers meaning darker. This pixel-value is an integer between 0 and 255.In fig6 top image show plotting of pca1 and pca2 with labled data,bottom left show GMM output, bottom right shows kmean output.

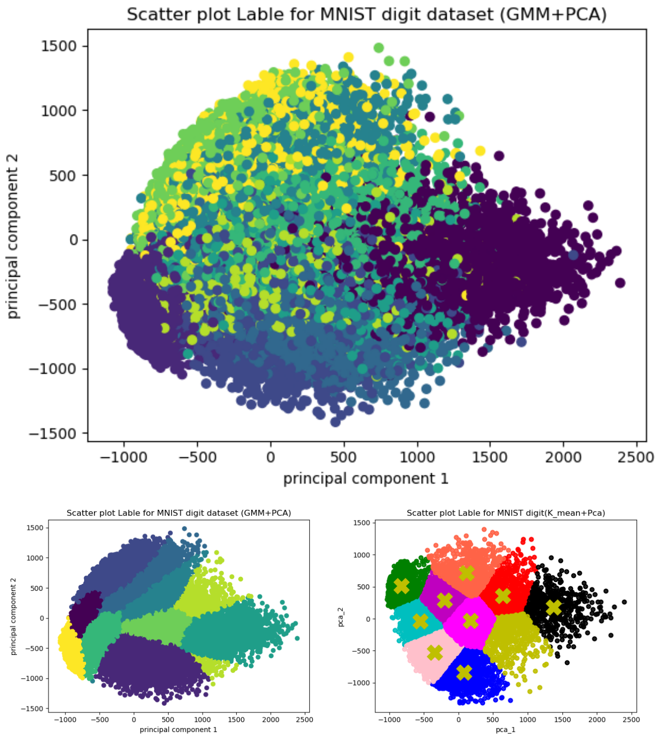


Fig. 6.

- GMM and PCA.
  - PCA(2 component) and GMM with 10 cluster: we selected 2 component which account for the variance of [0.29011353 0.17727668].Then applied GMM to create 10 clusters shown in fig6
- KMean and PCA
  - KMean followed by PCA with 10 cluster: we selected 2 component which account for the variance of [0.29011353 0.17727668].Then applied Kmeans to create 10 clusters shown in fig6

III. MNIST FASHION Dataset

A. Dataset Description

Each image is 28 pixels in height and 28 pixels in width, for a total of 784 pixels in total. Each pixel has a single pixel-value associated with it, indicating the lightness or darkness of that pixel, with higher numbers meaning darker. This pixel-value is an integer between 0 and 255. The training and test data sets have 785 columns. The first column consists of the class labels, and represents the article of clothing. The rest of the columns contain the pixel-values of the associated image.

B. Methods used

Pipelined PCA and Support Vector Classifier: PCA was used to reduce the dimensions and SVM Classifier has been used to train the model. Different kernels were used to classify and accuracy corresponding to each of them was recorded. We found that Gaussian kernel gives the best possible accuracy.

Kernel	Accuracy
Linear	51
Gaussian	56
Sigmoidal	14

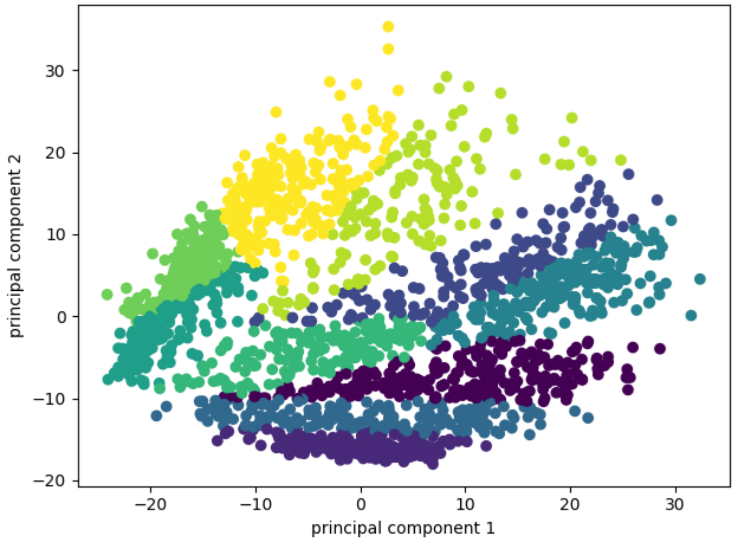


Fig. 7. Plot of PCA on the predicted values

we also used GMM and Kmeans to find cluster in this dataset

Acknowledgment

We would like to thank Dr Prathosh A.P for his continuous guidance and motivation to keep learning even after the failed attempts. We have used kaggle for data sets.

.