CS669: Pattern Recognition

Programming Assignment 2

Instructor: Dileep A.D.

By-

Group 12

Ayush Garg (B13111)

Himanshu Singal (B13312)

Hitesh Tarani (B13139)

Part-I: Introduction

In earlier assignment we assumed the data to be coming from unimodal Gaussian distribution but this method wasn't sufficient to classify certain data sets well like the spiral, real data sets. But, now in this assignment we are going to assume that the data is coming from a multimodal Gaussian distribution hence we will see how this new assumption affects the classification accuracy of our classifier.

In this programming assignment our target is to observe the accuracy of the three given data sets of image, real (speech data) and overlapping data and infer the distribution of the respective Gaussians and their density clusters configurations using Gaussian mixture model technique for different number of mixtures.

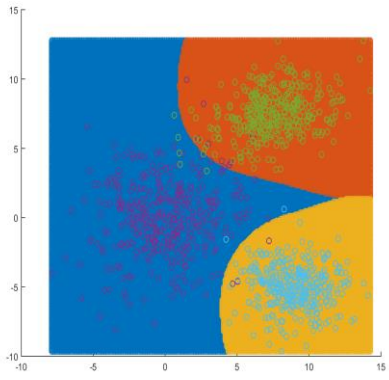
Gaussian Mixture Model

A Gaussian mixture model is a probabilistic model that assumes all the data points are generated from a mixture of a finite number of Gaussian distributions with unknown parameters. We use k means method to calculate the parameters for different clusters to be formed within various classes. Then we use E-M (Estimation-Maximization) technique to maximize the probability of each data point to belong to a particular cluster.

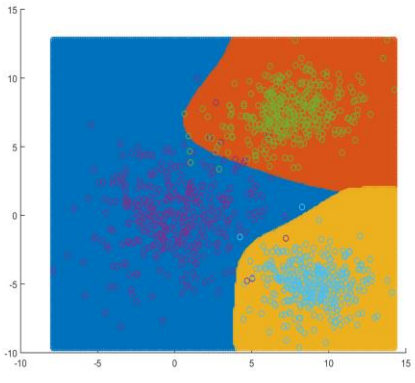
Part-II: Experiments and Observations

Overlapping dataset

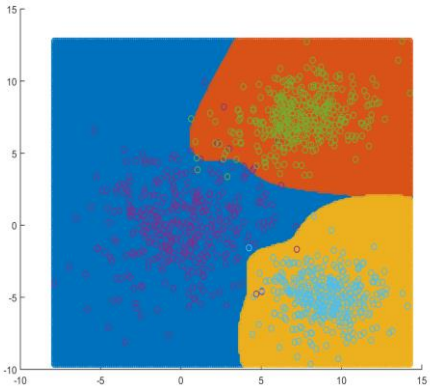
For 2 clusters



For 4 clusters



For 8 clusters



Number of clusters	Confusion matrix	Efficiency
2	122 1 2 2 123 0 1 0 124	98.4
4	121 2 2 1 124 0 1 0 124	98.4
8	120 3 2 1 124 0 1 0 124	98.133

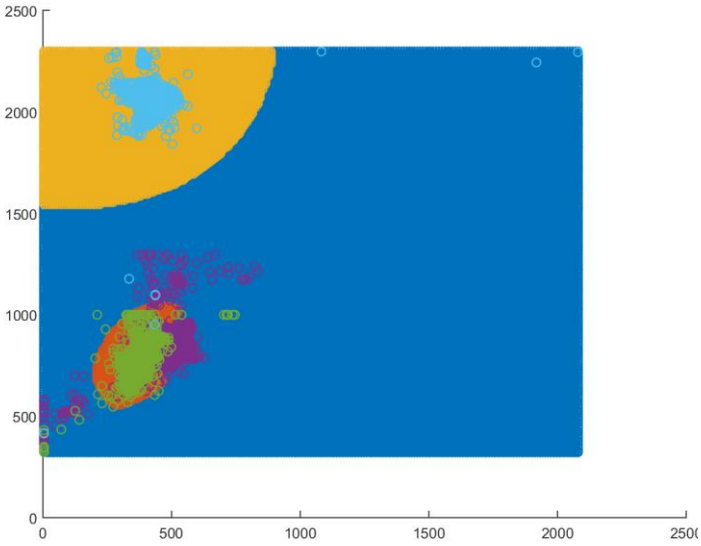
We, observe that in comparison to Bayes classifier with different comparison this method is working equally well as it had an accuracy of 98.4% too. But, the decision surface has changed a lot if we look for the case of 8 and 4 clusters.

Here, we observe that when the clusters are smaller in number the decision surface is smoother but less accurate and as we increase the number of clusters in a class the sharp edges are introduced. This, is because increase in number of clusters leads to overlapping of more and more Gaussian surfaces which enables us to classify more accurately with the help of sharp edges, this wasn't possible in case of smooth boundaries.

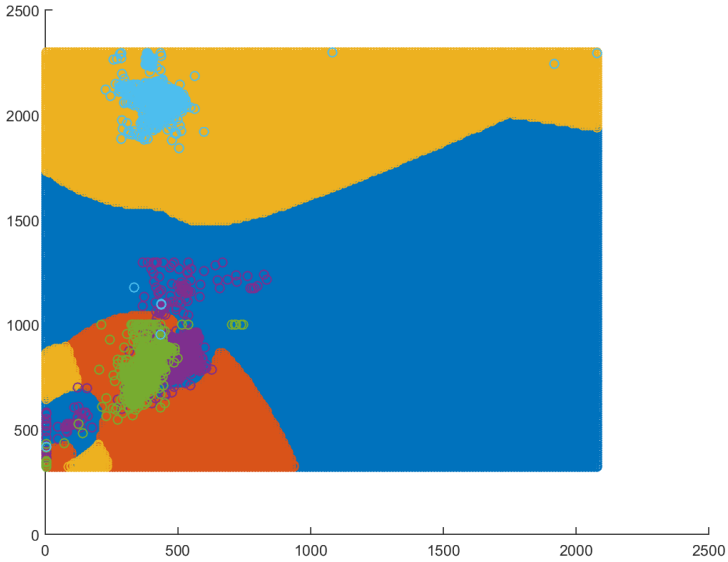
The accuracy is not getting any better even if the clusters are increased because there is overlapping data and if we want to classify those points accurately, we will have to increase the number of clusters such that point will itself be one cluster and classifying with that amount of clusters is out of our computation limit.

Real dataset

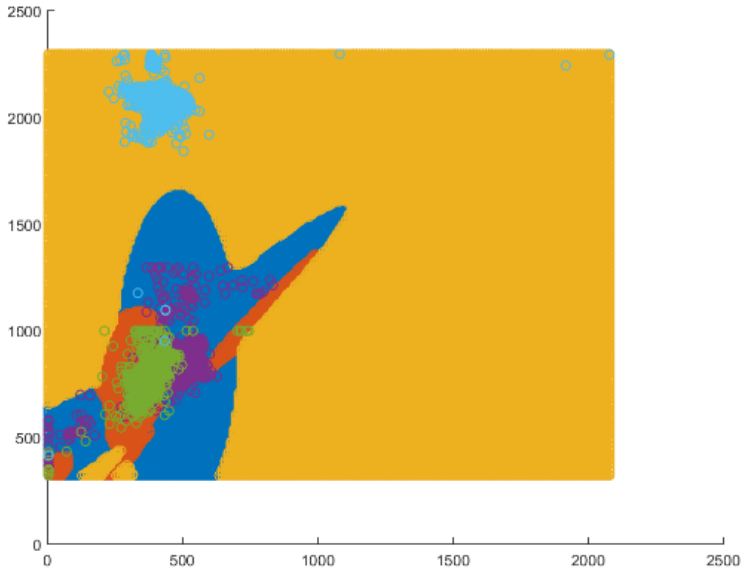
For 2 clusters



For 4 clusters



For 8 clusters



Number of clusters	Confusion matrix	Efficiency
2	429 193 0 56 566 0 19 8 545	84.802
4	384 236 2 38 584 0 9 13 550	83.59
8	404 214 4 45 573 4 7 13 552	84.196

Here, we observed that there is slight improved accuracy as compared to the Bayes classifier with the different covariance matrices as it had an accuracy of about 82.69%. Also, we see that there is drastic change in decision boundary compared to the Bayes classifier case.

Just like the previous case the decision boundaries are smooth for fewer number of clusters also it can be clearly seen that in case of two clusters the boundaries look like a single Gaussian curve but as the clusters increase in number the boundaries get more and more complex and are a result of mixture of multiple Gaussian surfaces.

Here, we notice that the efficiency is not changing much as the clusters are increased but the decision boundary is changing drastically that's because the as the number of cluster increases the possibility better mixture model arises and hence the Gaussian surfaces change.

Image dataset

In the given image data, we are given 3 sets of images from real world namely coast, forest and inside city. It is given such that it's broken down into 36 parts and then each part is extracted in form of a feature vector of 23 dimensions.

Here, we know that every part of the image is independent of the other and hence we concatenate the data from all files of a class into a single file and make our model over that file.

Number of clusters	Confusion matrix	Efficiency
2	66 7 17 24 46 12 10 10 57	67.871
4	79 3 8 3 75 4 8 4 65	87.952

Classification accuracy of image dataset is very low compared to 2-class datasets. As we increase the clusters, the accuracy increases greatly as observed from 67.8% to 87.9% but the time taken also increases by great amount. This is why we even tried one more method to classify the image data set. The confusion matrix for that method is mentioned below.

Number of clusters	Confusion matrix	Efficiency
2	80 3 7 2 80 0 5 2 70	92.369
4	83 2 5 4 78 0 3 3 71	93.173
8	84 3 3 1 80 1 4 6 67	92.771
16	80 3 7 2 80 0 3 2 72	93.173

In this method we took means of the feature vectors of all 36 blocks of an image and used it to classify the image to the right class hence the complexity reduced to a great extent and time also reduced in accordance to that. We observed the second method worked better for the given number of clusters but the first algorithm should work better as we increase the number of clusters as is evident from its confusion matrix.

We also observed that most images are such that all 36 blocks of the image or most of them are spanning the class in consideration. Hence, the second method is giving quite accurate results.