PATTERN RECOGNITION ASSIGNMENT-2
REPORT

21th October 2018

GROUP - 11

ABHIGYAN KHAUND	PRATYUSH GAURAV	SHASHWAT GARG
B16082	B16026	B16034

Faculty
Dr. A.D Dileep
Indian Institute of Technology, Mandi

Table of Contents

Introduction

K means clustering:

K-means clustering is a type of unsupervised learning, which is used when you have unlabeled data (i.e., data without defined classes). The goal of this algorithm is to find groups in the data, with the number of groups represented by the variable K. The algorithm works iteratively to assign each data point to one of K groups based on the features that are provided. Data points are clustered based on feature similarity. The results of the K-means clustering algorithm are:

1. The centroids of the K clusters, which can be used to label new data
2. Labels for the training data (each data point is assigned to a single cluster(Hard Clustering))

The algorithm works as follows:

1. First we initialize k points, called means, randomly.
2. We categorize each item to its closest mean and we update the mean's coordinates, which are the averages of the items categorized in that mean so far.
3. We repeat the process until convergence and at the end, we have our clusters.

Gaussian Mixture Model:

Gaussian mixture models (GMM) are often used for data clustering. Usually, fitted GMMs cluster by assigning query data points to the multivariate normal components that maximize the component posterior probability given the data. That is, given a fitted GMM, cluster assigns query data to the component yielding the highest posterior probability.GMM clustering is more flexible because you can view it as a *soft clustering* method. Soft clustering methods assign a score to a data point for each cluster. The value of the score indicates the association strength of the data point to the cluster. When clustering with GMMs, the score is the posterior probability.
Let class C_i have K Gaussian clusters then the probability of $\mathbf{x}$ given class C_i is,

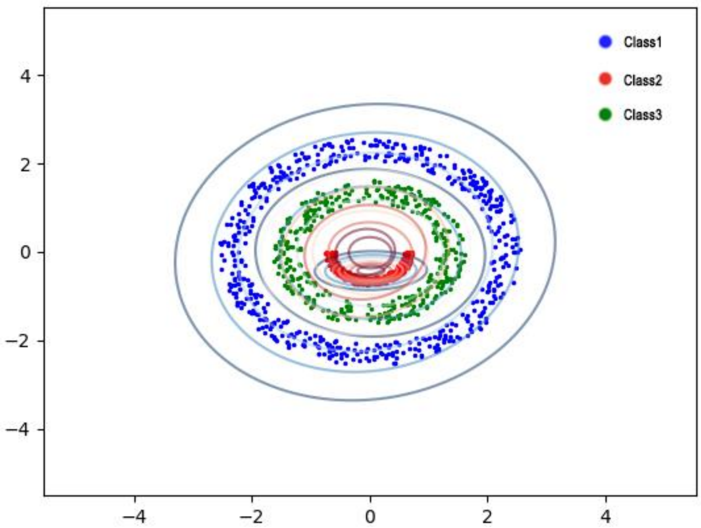
$$p(\mathbf{x} \mid C_i) = \Sigma_{k=1:K}\ \Pi_k\ p_k(\mathbf{x} \mid \boldsymbol{\mu}_{\mathbf{ik}}, \Sigma_{ik})$$

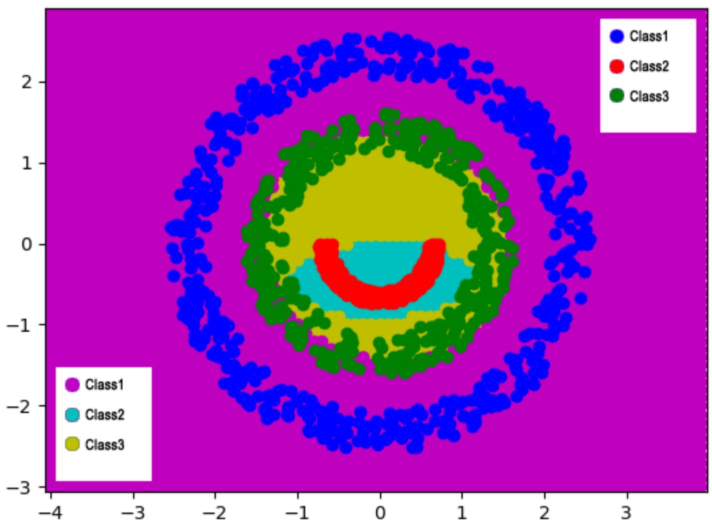
Here,
Π_k : mixture coefficient of k^{th} cluster
$p_k(\mathbf{x} \mid \boldsymbol{\mu}_{\mathbf{ik}}, \Sigma_{ik}) = N(\mathbf{x} \mid \boldsymbol{\mu}_{\mathbf{ik}}, \Sigma_{ik})$

Experimental Observations

2.1 Non-Linearly Separable dataset

2.1.1 1 cluster





K = 1

Classification Accuracy (%)	87.294%
Precision for Class1 Precision for Class2 Precision for Class3	1.0 0.908 0.791
Mean Precision	0.899
Recall for Class1 Recall for Class2 Recall for Class3	0.936 0.632 1.0
Mean Recall	0.856
F-measure for Class1 F-measure for Class2 F-measure for Class3	0.966 0.745 0.883
Mean F-measure	0.864

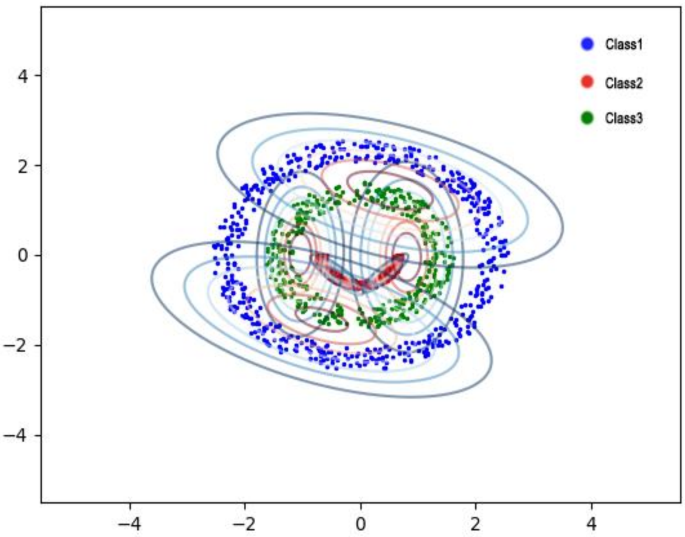
Confusion Matrix :

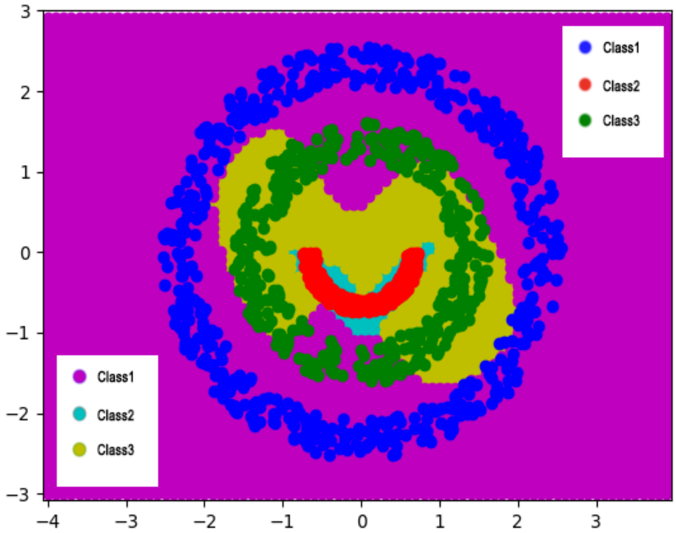
$$C = \begin{matrix} 117 & 8 & 0 \\ 0 & 79 & 46 \\ 0 & 0 & 175 \end{matrix}$$

Observations :

- In this case we have total of 3 Gaussian distributions(1 per class)
- 81.294% accuracy is obtained for K=1.
- 1 Gaussian doesn't represent the distribution of the data in the classes.

2.1.2 2 clusters



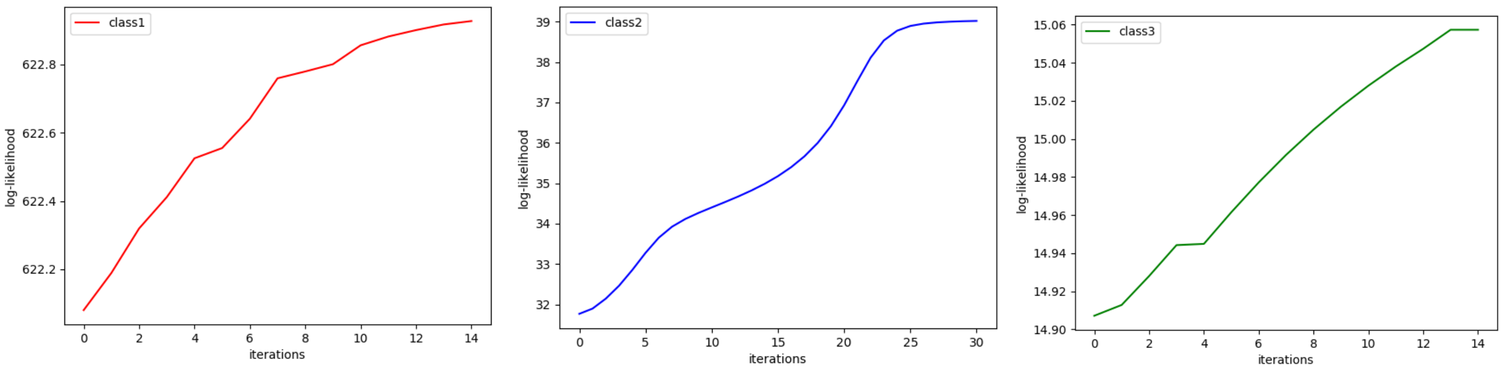


For K = 2

Classification Accuracy (%)	88.706%
Precision for Class1 Precision for Class2 Precision for Class3	1.0 0.905 0.810
Mean Precision	0.905
Recall for Class1 Recall for Class2 Recall for Class3	1.0 0.688 0.949
Mean Recall	0.879
F-measure for Class1 F-measure for Class2 F-measure for Class3	1.0 0.782 0.874
Mean F-measure	0.885

Confusion Matrix :

$$C = \begin{matrix} 125 & 0 & 0 \\ 0 & 86 & 39 \\ 0 & 9 & 166 \end{matrix}$$

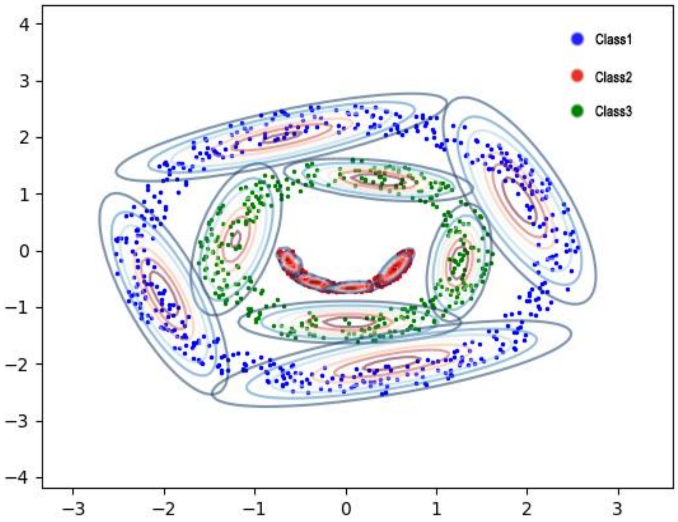


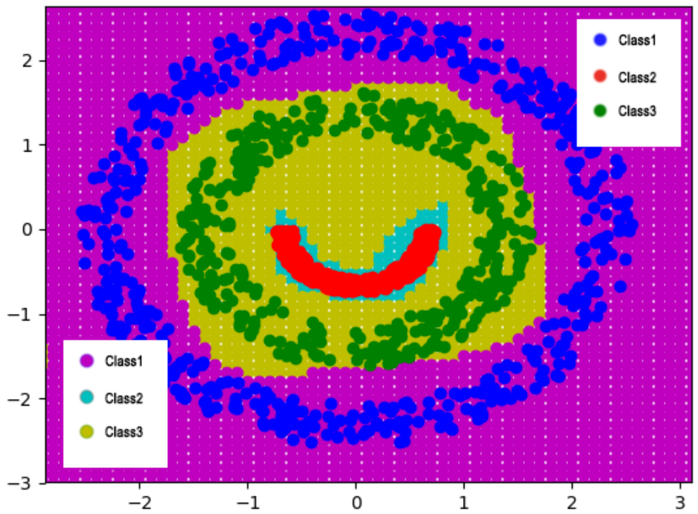
Log-likelihood vs Iterations graphs for K=2 on non-linearly separable dataset

Observations :

- In this case we have total of 6 Gaussian distributions(2 per class)
- 88.706% accuracy is obtained for 2 value of K.
- The packing of data is not sufficiently represented by 2 Gaussians.

2.1.3 4 clusters



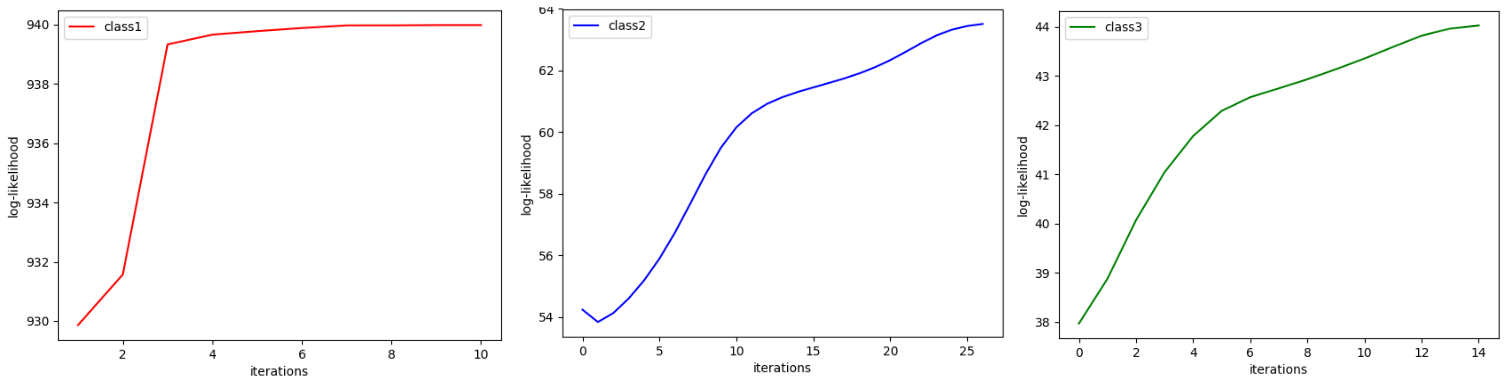


For K = 4

Classification Accuracy (%)	100%
Precision for Class1 Precision for Class2 Precision for Class3	1.0 1.0 1.0
Mean Precision	1.0
Recall for Class1 Recall for Class2 Recall for Class3	1.0 1.0 1.0
Mean Recall	1.0
F-measure for Class1 F-measure for Class2 F-measure for Class3	1.0 1.0 1.0
Mean F-measure	1.0

Confusion Matrix :

$$C = \begin{matrix} 125 & 0 & 0 \\ 0 & 125 & 0 \\ 0 & 0 & 175 \end{matrix}$$



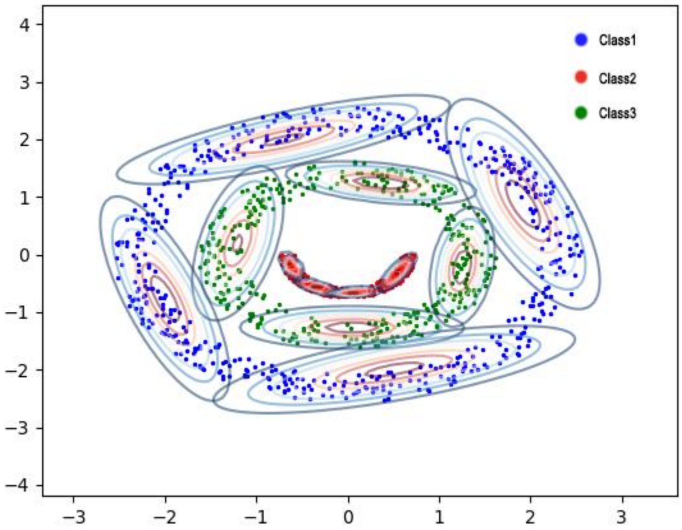
Log-likelihood vs Iterations graphs for K=4 on non-linearly separable dataset

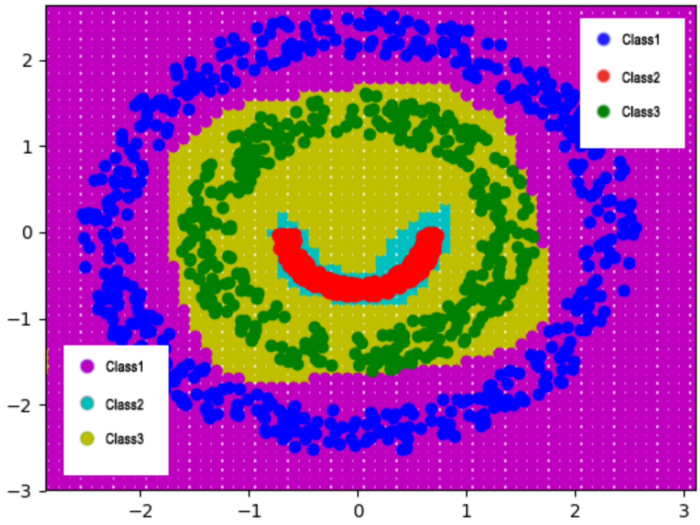
Observations :

- In this case we have total of 12 Gaussian distributions(4 per class) and class Gaussians define the decision boundary at their point of intersection with other class contours.

- 100% accuracy is obtained for 4 value of K.
- The packing of data is more precise for K = 4 as compared to K = 1, 2.

2.1.4 8 clusters



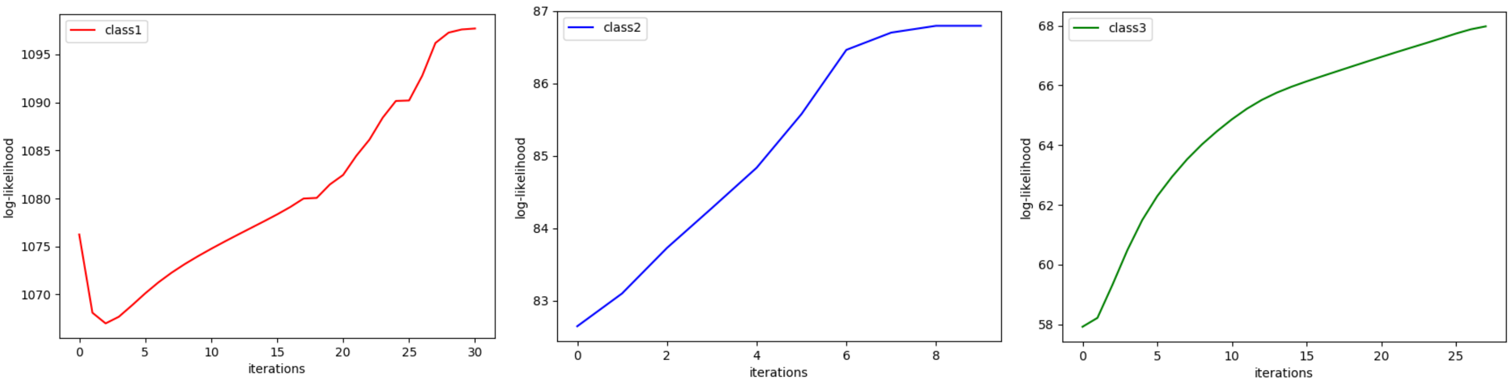


For K = 8

Classification Accuracy (%)	100%
Precision for Class1 Precision for Class2 Precision for Class3	1.0 1.0 1.0
Mean Precision	1.0
Recall for Class1 Recall for Class2 Recall for Class3	1.0 1.0 1.0
Mean Recall	1.0
F-measure for Class1 F-measure for Class2 F-measure for Class3	1.0 1.0 1.0
Mean F-measure	1.0

Confusion Matrix :

$$C = \begin{matrix} 125 & 0 & 0 \\ 0 & 125 & 0 \\ 0 & 0 & 175 \end{matrix}$$

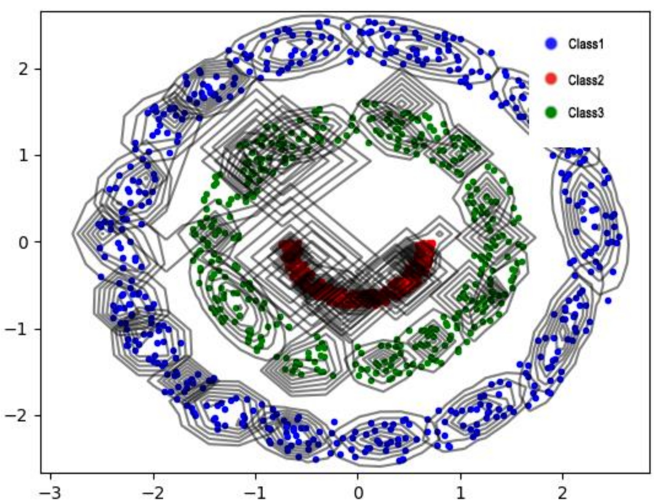


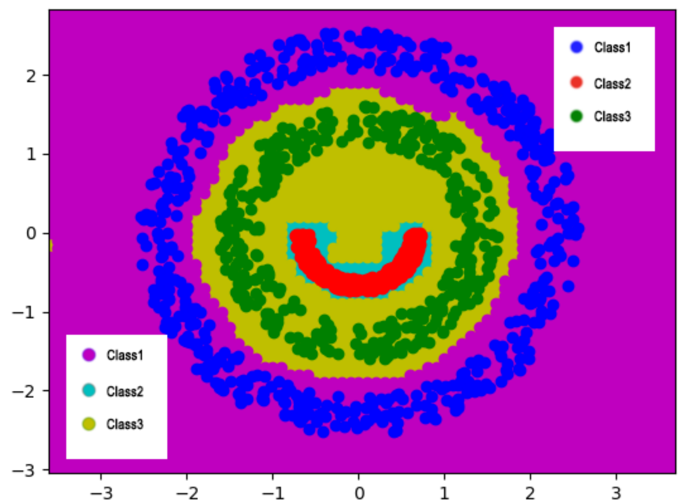
Log-likelihood vs Iterations graphs for K=8 on non-linearly separable dataset

Observations :

- In this case we have total of 24 Gaussian distributions(8 per class).
- 100% accuracy is obtained for 8 value of K.
- The packing of data is more precise for K = 8 as compared to K = 1, 2. 4.

2.1.5 16 clusters



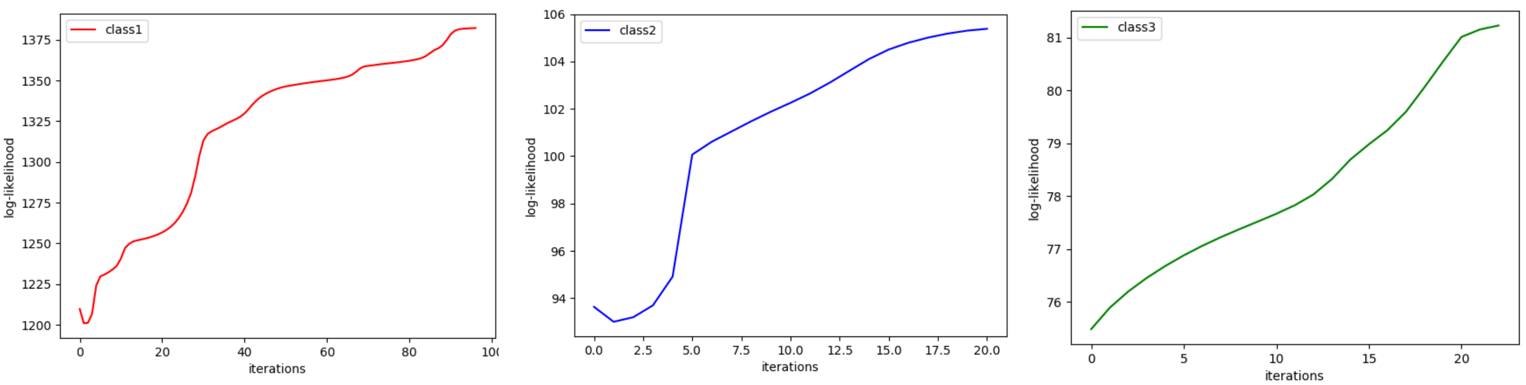


For K = 16

Classification Accuracy (%)	100%
Precision for Class1 Precision for Class2 Precision for Class3	1.0 1.0 1.0
Mean Precision	1.0
Recall for Class1 Recall for Class2 Recall for Class3	1.0 1.0 1.0
Mean Recall	1.0
F-measure for Class1 F-measure for Class2 F-measure for Class3	1.0 1.0 1.0
Mean F-measure	1.0

Confusion Matrix :

$$C = \begin{matrix} 125 & 0 & 0 \\ 0 & 125 & 0 \\ 0 & 0 & 175 \end{matrix}$$

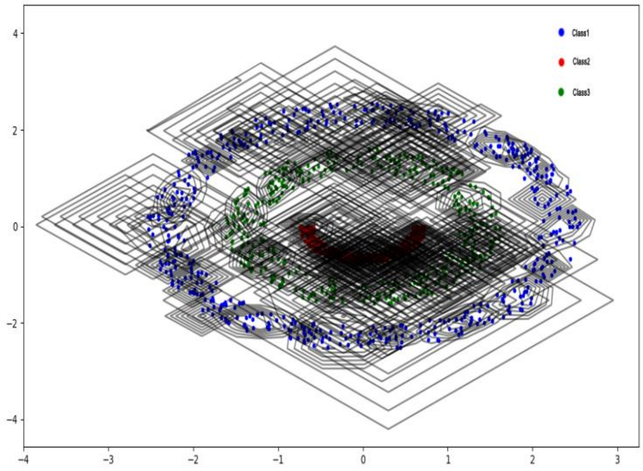


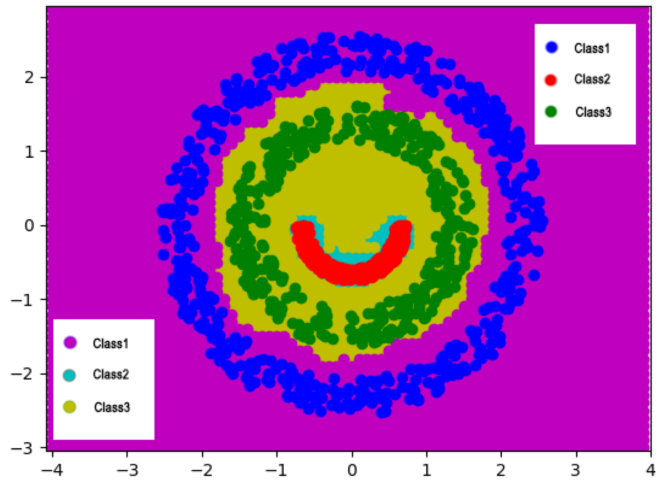
Log-likelihood vs Iterations graphs for K=16 on non-linearly separable dataset

Observations :

- In this case we have total of 48 Gaussian distributions (16 per class)
- 100% accuracy is obtained for 16 value of K.
- The packing of data is very precise for K = 16 than others.

2.1.6 32 clusters



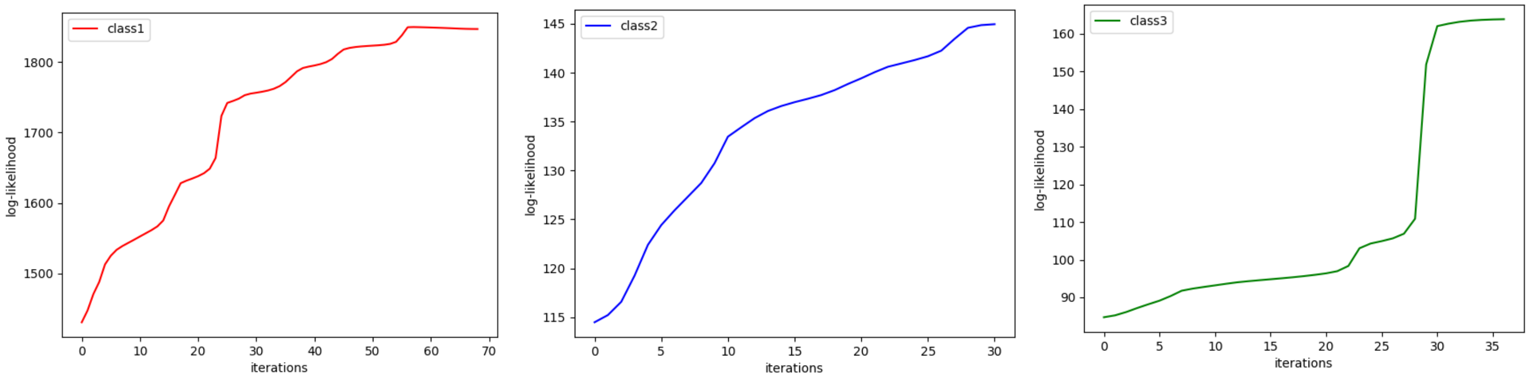


For K = 32

Classification Accuracy (%)	100%
Precision for Class1 Precision for Class2 Precision for Class3	1.0 1.0 1.0
Mean Precision	1.0
Recall for Class1 Recall for Class2 Recall for Class3	1.0 1.0 1.0
Mean Recall	1.0
F-measure for Class1 F-measure for Class2 F-measure for Class3	1.0 1.0 1.0
Mean F-measure	1.0

Confusion Matrix :

$$C = \begin{matrix} 125 & 0 & 0 \\ 0 & 125 & 0 \\ 0 & 0 & 175 \end{matrix}$$



Log-likelihood vs Iterations graphs for K=32 on non-linearly separable dataset

Observations :

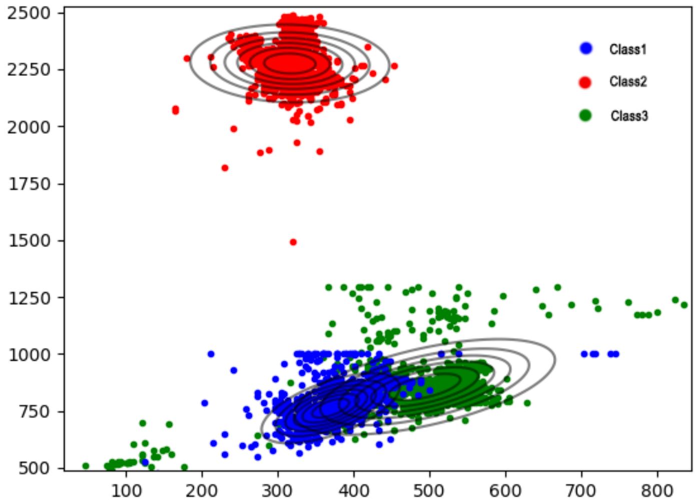
- In this case we have total of 96 Gaussian distributions (32 per class)
- 100% accuracy is obtained for 32 value of K.
- The data is over packed with more number of Gaussian then actually in the data for K = 32.

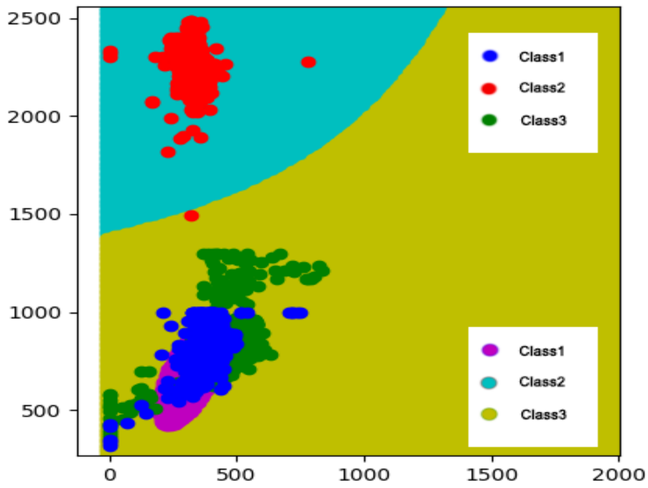
<u>Overall Observations:</u>

- The accuracy obtained in GMM is far better than in unimodal Gaussian model.
- In unimodal case, the decision surfaces obtained were linear and elliptical depending on the covariance matrices of classes.
- In GMM we assume that every class has several Gaussian clusters which leads to a non linear envelope of the constituent Gaussian components. This explains the vast difference in the accuracies obtained in unimodal case and GMM.
- The contours for outer classes are lower than those for inner class because the variance for the outer class is very high due to which the peaks are lower in height.
- The contours of inner class define the decision boundary at the points of intersection with the contours of outer class.
- The nonlinear decision boundary in the graph is an envelope of the Gaussian components.
- The accuracy value increases with K and the maximum accuracy obtained is 100%.

2.2 Real World dataset

2.2.1 1 cluster



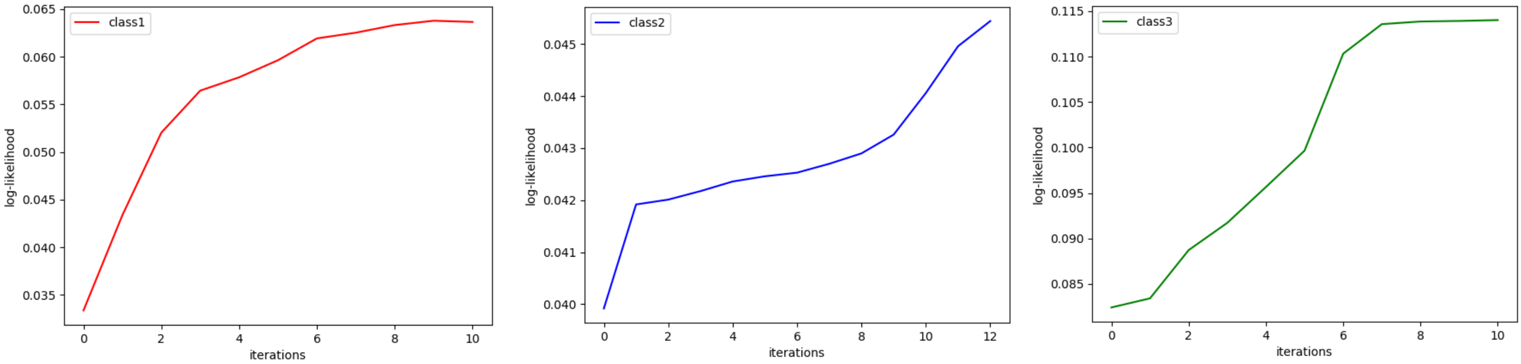


K = 1

Classification Accuracy (%)	82.815%
Precision for Class1 Precision for Class2 Precision for Class3	1.0 0.794 0.717
Mean Precision	0.837
Recall for Class1 Recall for Class2 Recall for Class3	0.973 0.661 0.854
Mean Recall	0.829
F-measure for Class1 F-measure for Class2 F-measure for Class3	0.986 0.722 0.779
Mean F-measure	0.829

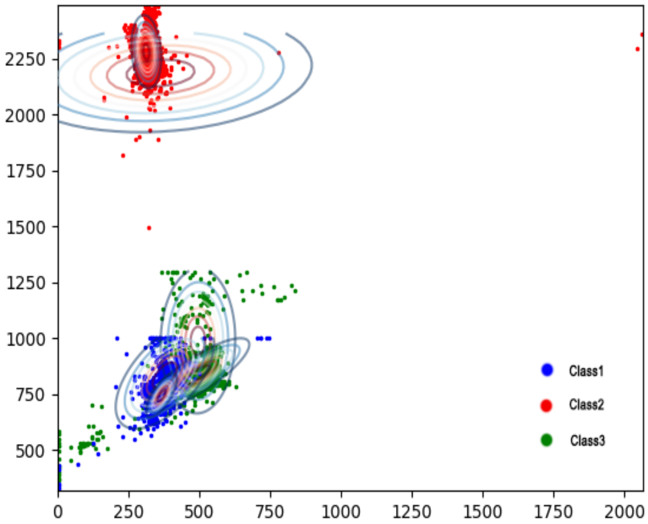
Confusion Matrix :

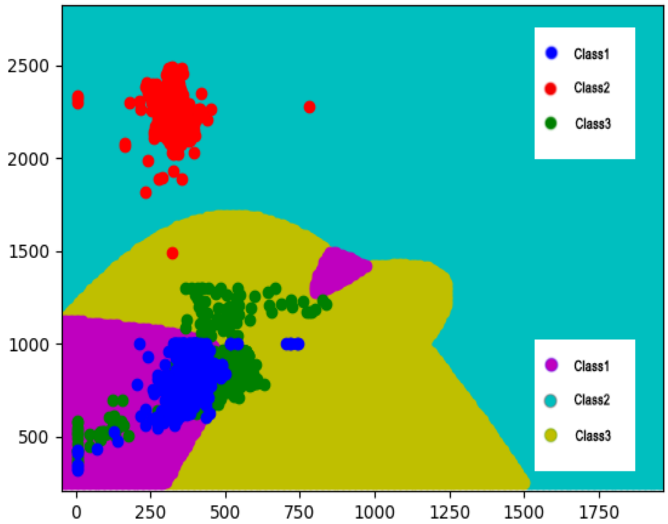
$$C = \begin{matrix} 581 & 14 & 2 \\ 0 & 406 & 208 \\ 0 & 91 & 531 \end{matrix}$$



Log-likelihood vs Iterations graphs for K=1 on real world dataset

2.2.2 2 clusters



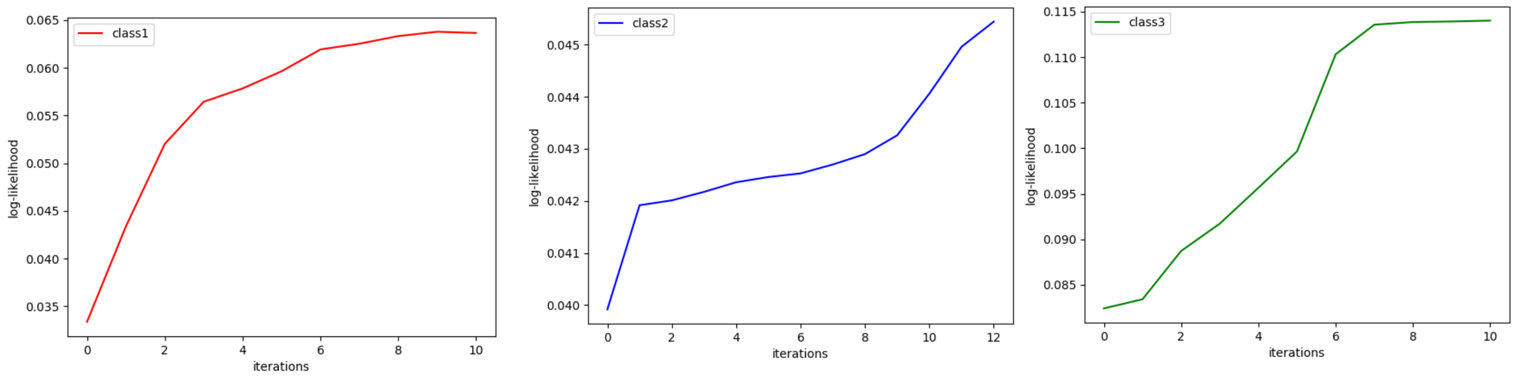


For K = 2

Classification Accuracy (%)	85.543%
Precision for Class1 Precision for Class2 Precision for Class3	0.998 0.884 0.730
Mean Precision	0.871
Recall for Class1 Recall for Class2 Recall for Class3	0.995 0.658 0.916
Mean Recall	0.856
F-measure for Class1 F-measure for Class2 F-measure for Class3	0.997 0.754 0.812
Mean F-measure	0.834

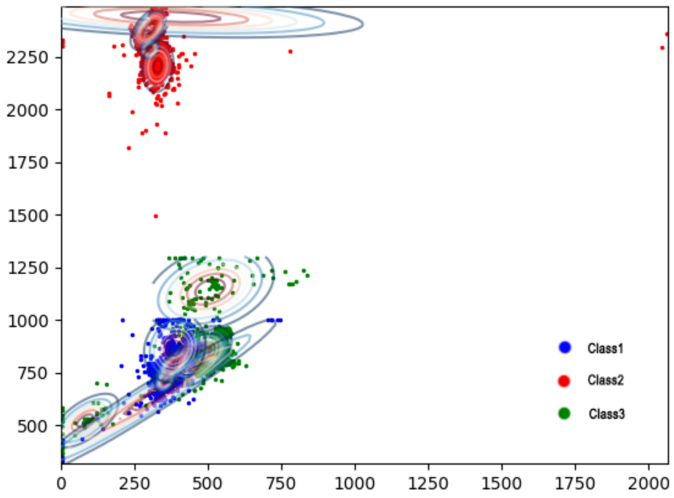
Confusion Matrix :

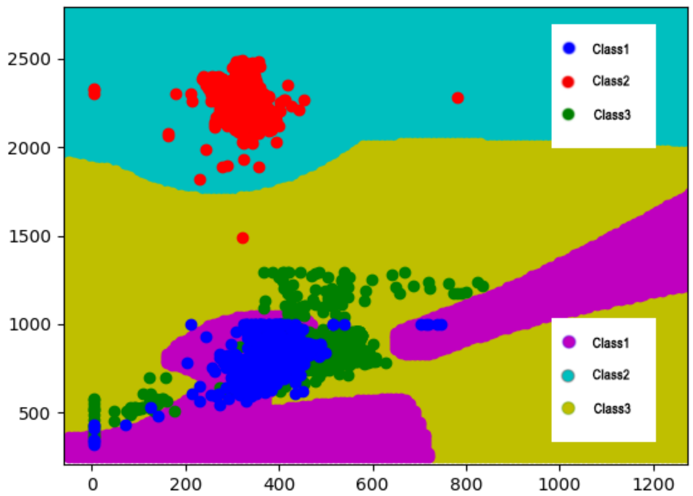
$$C = \begin{matrix} 594 & 1 & 2 \\ 1 & 404 & 209 \\ 0 & 52 & 570 \end{matrix}$$



Log-likelihood vs Iterations graphs for K=2 on real world dataset

2.2.3 4 clusters



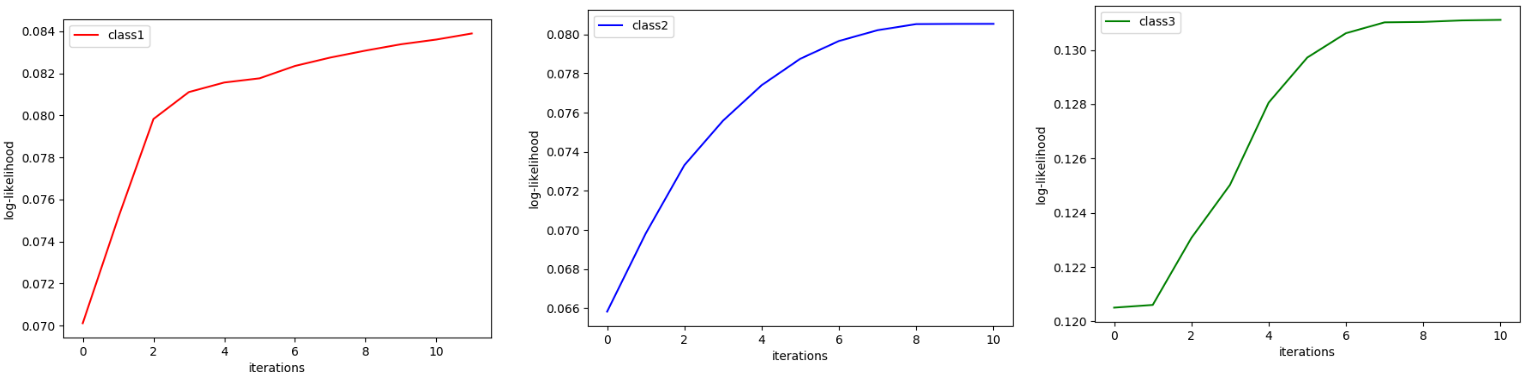


For K = 4

Classification Accuracy (%)	84.070%
Precision for Class1 Precision for Class2 Precision for Class3	1.0 0.877 0.706
Mean Precision	0.861
Recall for Class1 Recall for Class2 Recall for Class3	0.987 0.616 0.923
Mean Recall	0.842
F-measure for Class1 F-measure for Class2 F-measure for Class3	0.993 0.723 0.8
Mean F-measure	0.839

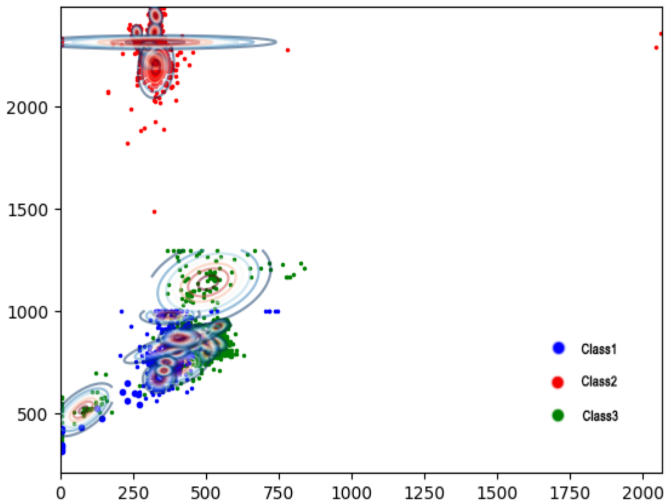
Confusion Matrix :

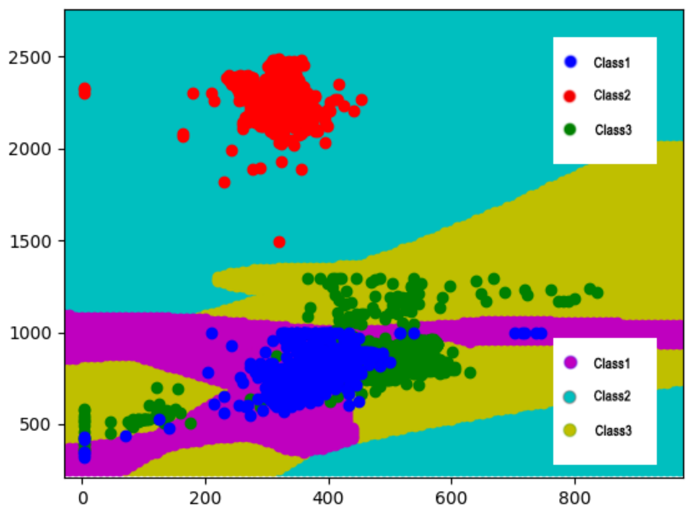
$$C = \begin{matrix} 589 & 5 & 3 \\ 0 & 378 & 236 \\ 0 & 48 & 574 \end{matrix}$$



Log-likelihood vs Iterations graphs for K=4 on real world dataset

2.2.4 8 clusters



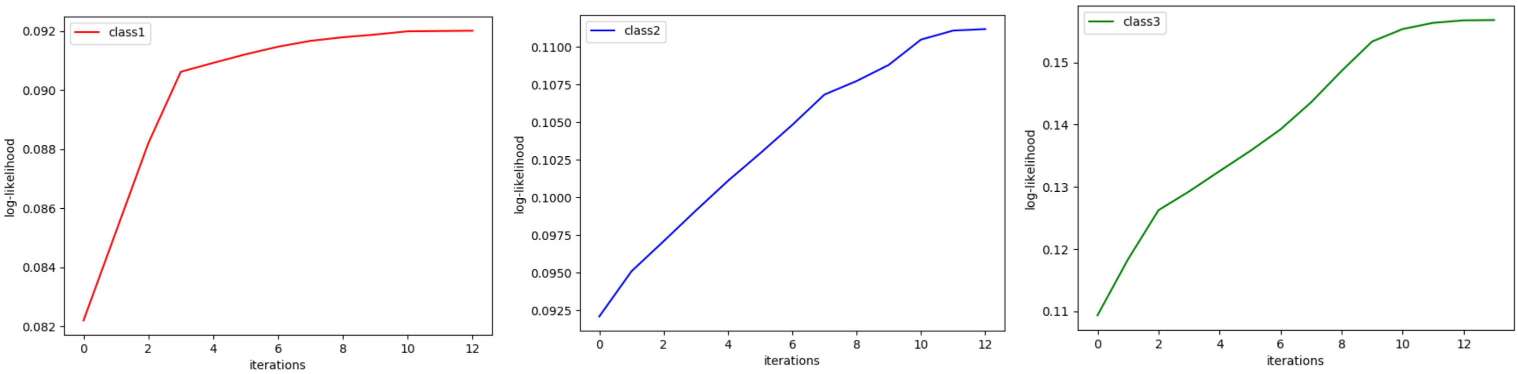


For K = 8

Classification Accuracy (%)	85.870%
Precision for Class1 Precision for Class2 Precision for Class3	0.998 0.918 0.726
Mean Precision	0.881
Recall for Class1 Recall for Class2 Recall for Class3	0.990 0.640 0.948
Mean Recall	0.860
F-measure for Class1 F-measure for Class2 F-measure for Class3	0.994 0.754 0.822
Mean F-measure	0.857

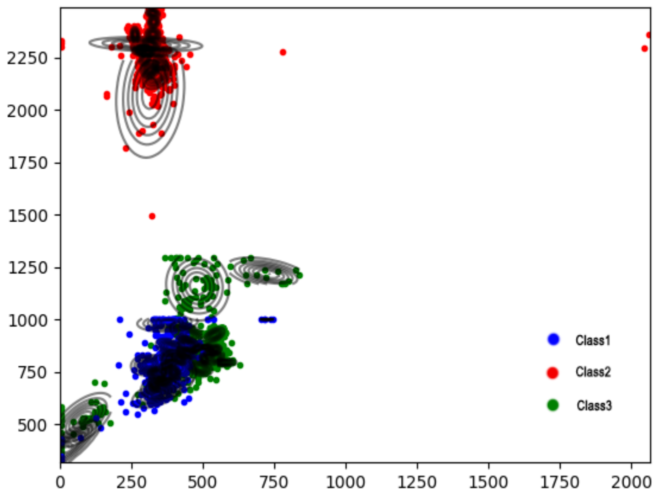
Confusion Matrix :

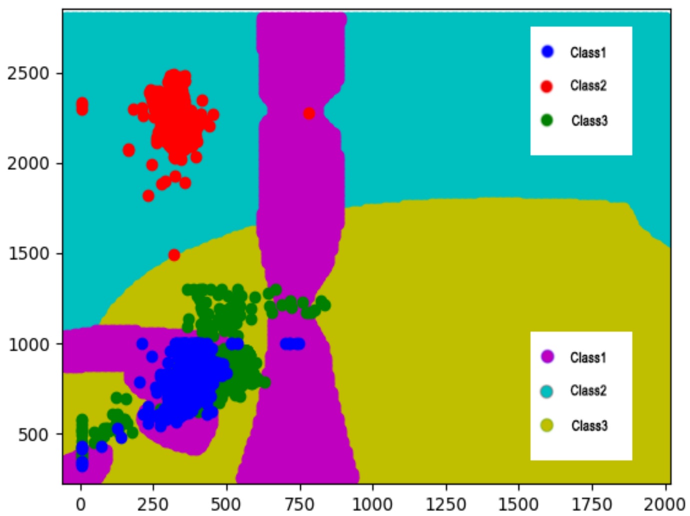
$$C = \begin{matrix} 591 & 3 & 3 \\ 1 & 393 & 220 \\ 0 & 32 & 590 \end{matrix}$$



Log-likelihood vs Iterations graphs for K=8 on real world dataset

2.2.5 16 clusters



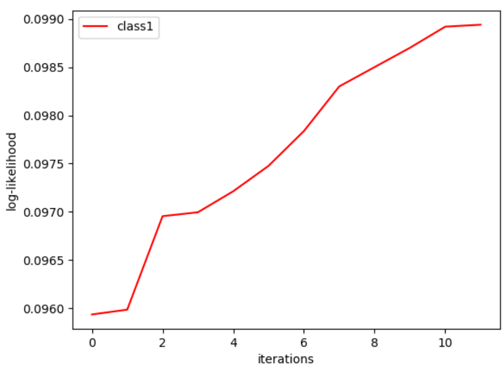


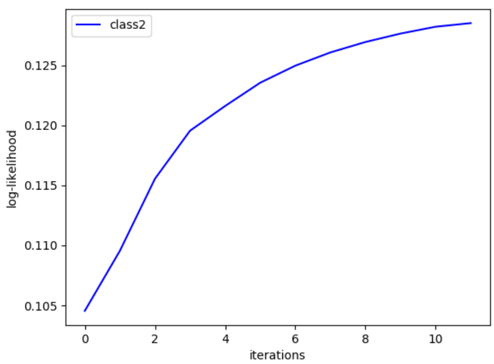
For K = 16

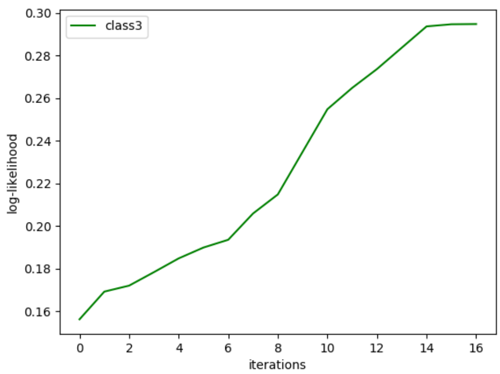
Classification Accuracy (%)	46.645%
Precision for Class1 Precision for Class2 Precision for Class3	1.0 0.968 0.388
Mean Precision	0.785
Recall for Class1 Recall for Class2 Recall for Class3	0.244 0.147 0.995
Mean Recall	0.462
F-measure for Class1 F-measure for Class2 F-measure for Class3	0.393 0.255 0.559
Mean F-measure	0.402

Confusion Matrix :

$$C = \begin{matrix} 146 & 0 & 451 \\ 0 & 90 & 524 \\ 0 & 3 & 619 \end{matrix}$$







Log-likelihood vs Iterations graphs for K=16 on real world dataset

Observations :

- From K=1 to K=8 accuracy marginally increases from 82.8 %to 85.8% as the in-creasing clusters doesn't better depicts the data distribution by much.
- From K = 8 to K = 16 over clustering leads to less accuracy as more data points get misclassified.
- Maximum accuracy in unimodal Gaussian case is a bit more than the highest accuracy obtained in GMM.

2.3 Image Dataset

2.3.1 Color Coded Histogram

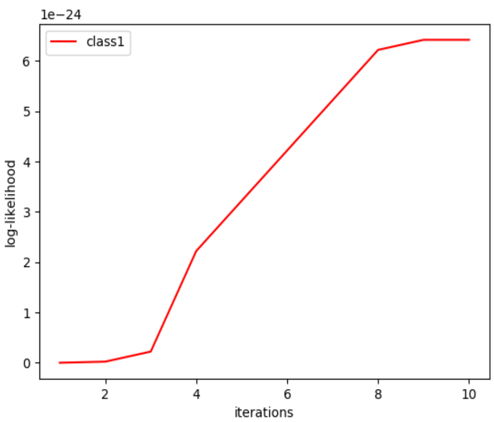
2.3.1.1 1 cluster

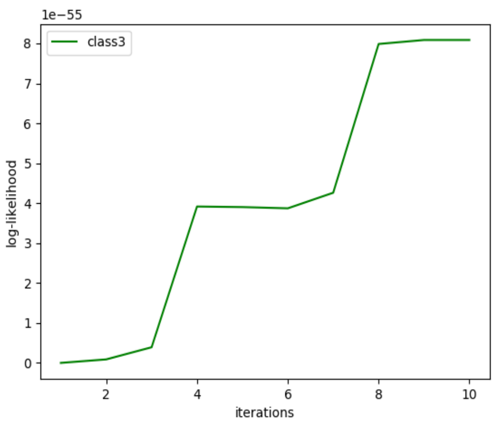
K = 1

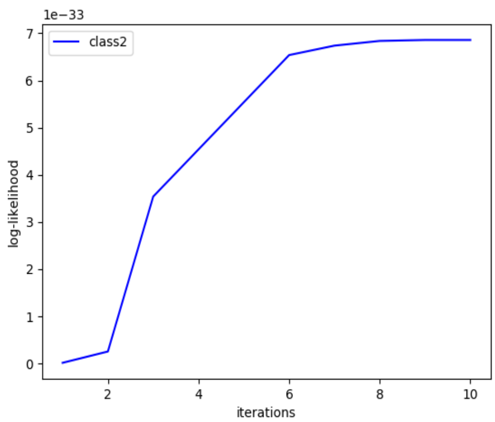
Classification Accuracy (%)	56%
Precision for Class1 Precision for Class2 Precision for Class3	0.654 0.479 0.627
Mean Precision	0.587
Recall for Class1 Recall for Class2 Recall for Class3	0.34 0.7 0.64
Mean Recall	0.56
F-measure for Class1 F-measure for Class2 F-measure for Class3	0.447 0.569 0.634
Mean F-measure	0.550

Confusion Matrix :

$$C = \begin{matrix} 17 & 22 & 17 \\ 7 & 35 & 8 \\ 2 & 16 & 32 \end{matrix}$$







Log-likelihood vs Iterations graphs for K=1

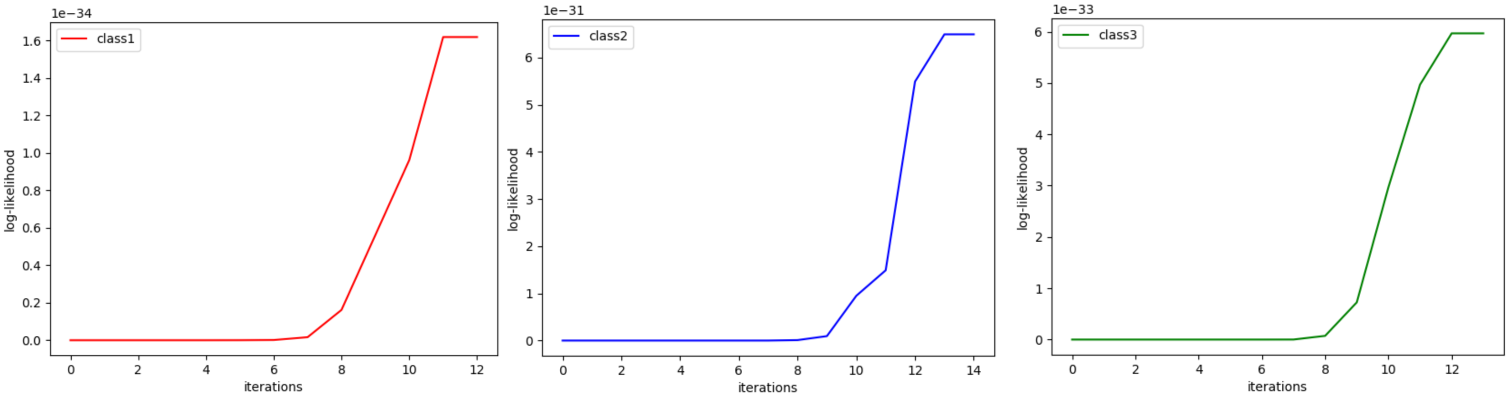
2.3.1.2 2 clusters

K = 2

Classification Accuracy (%)	62%
Precision for Class1 Precision for Class2 Precision for Class3	0.724 0.514 0.706
Mean Precision	0.648
Recall for Class1 Recall for Class2 Recall for Class3	0.42 0.72 0.72
Mean Recall	0.62
F-measure for Class1 F-measure for Class2 F-measure for Class3	0.532 0.6 0.713
Mean F-measure	0.614

Confusion Matrix :

$$C = \begin{matrix} 21 & 22 & 7 \\ 6 & 36 & 8 \\ 2 & 12 & 36 \end{matrix}$$



Log-likelihood vs Iterations graphs for K=2

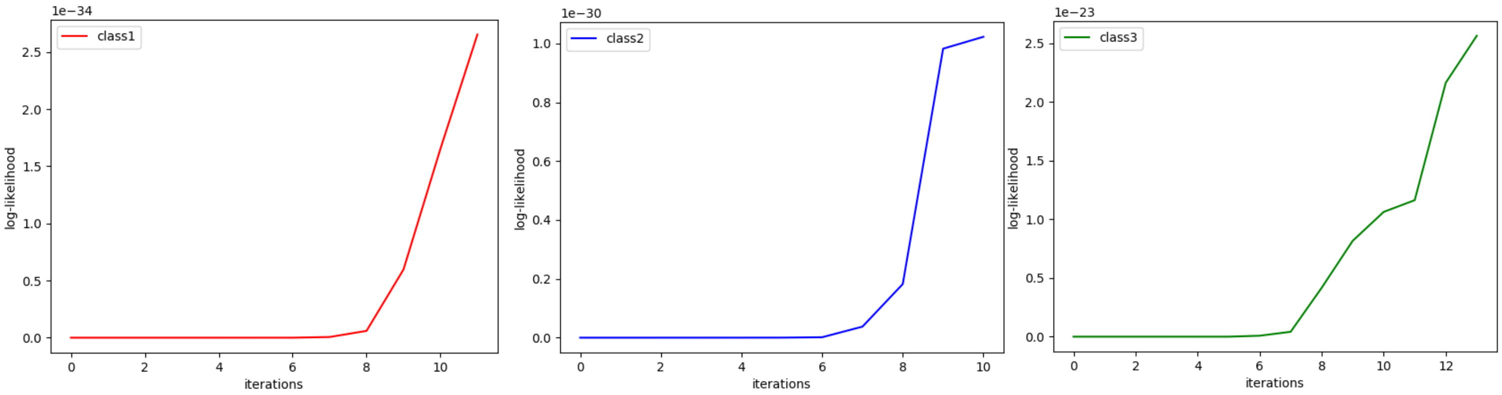
2.3.1.3 4 clusters

K = 4

Classification Accuracy (%)	62.333%
Precision for Class1 Precision for Class2 Precision for Class3	0.629 0.532 0.692
Mean Precision	0.618
Recall for Class1 Recall for Class2 Recall for Class3	0.55 0.70 0.71
Mean Recall	0.633
F-measure for Class1 F-measure for Class2 F-measure for Class3	0.550 0.591 0.717
Mean F-measure	0.620

Confusion Matrix :

$$C = \begin{matrix} 30 & 17 & 3 \\ 11 & 30 & 9 \\ 11 & 12 & 27 \end{matrix}$$



Log-likelihood vs Iterations graphs for K=4

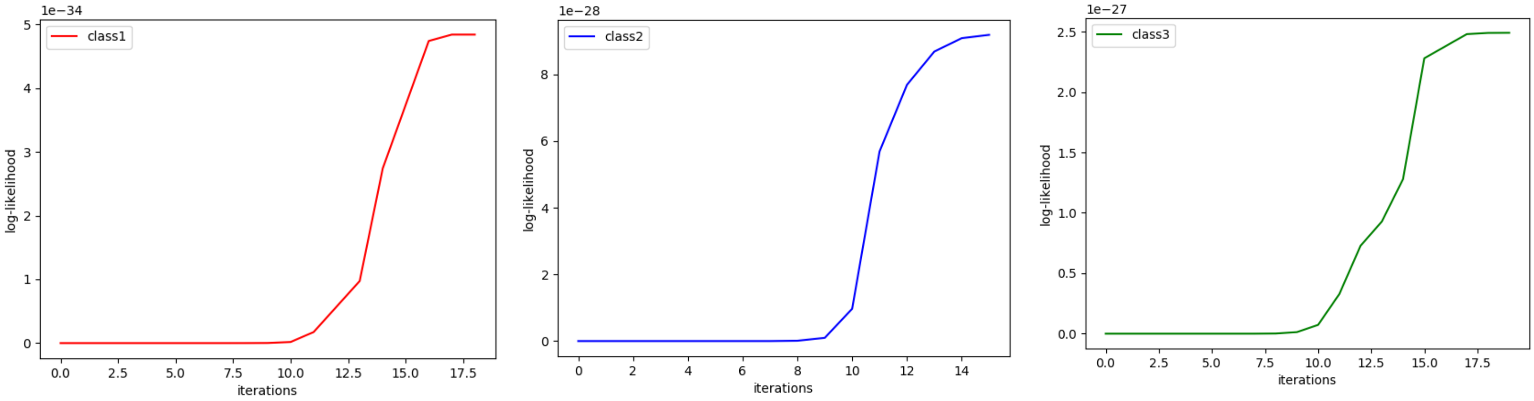
2.3.1.4 8 clusters

K = 8

Classification Accuracy (%)	62%
Precision for Class1 Precision for Class2 Precision for Class3	0.605 0.542 0.729
Mean Precision	0.625
Recall for Class1 Recall for Class2 Recall for Class3	0.52 0.64 0.7
Mean Recall	0.62
F-measure for Class1 F-measure for Class2 F-measure for Class3	0.559 0.587 0.714
Mean F-measure	0.620

Confusion Matrix :

$$C = \begin{matrix} 26 & 18 & 6 \\ 11 & 32 & 7 \\ 6 & 9 & 35 \end{matrix}$$



Log-likelihood vs Iterations graphs for K=8

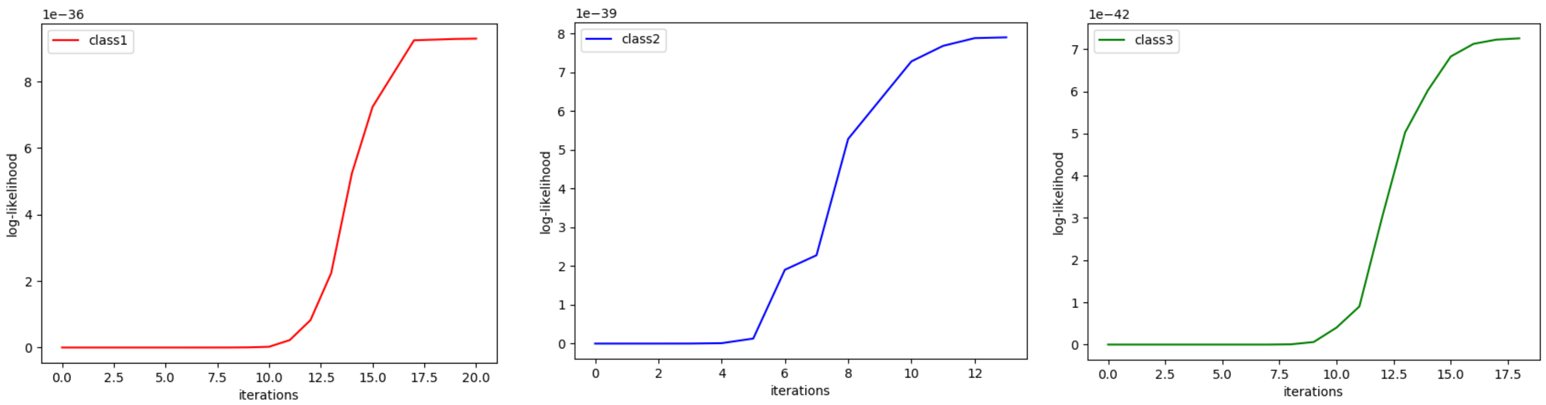
2.3.1.5 16 clusters

K = 16

Classification Accuracy (%)	64.667%
Precision for Class1 Precision for Class2 Precision for Class3	0.623 0.58 0.769
Mean Precision	0.657
Recall for Class1 Recall for Class2 Recall for Class3	0.76 0.58 0.6
Mean Recall	0.647
F-measure for Class1 F-measure for Class2 F-measure for Class3	0.657 0.647 0.323
Mean F-measure	0.646

Confusion Matrix :

$$C = \begin{matrix} 38 & 10 & 2 \\ 14 & 29 & 7 \\ 9 & 11 & 30 \end{matrix}$$



Log-likelihood vs Iterations graphs for K=16

2.3.1.6 32 clusters

K = 32

Classification Accuracy (%)	63.333%
Precision for Class1 Precision for Class2 Precision for Class3	0.610 0.547 0.778
Mean Precision	0.645
Recall for Class1 Recall for Class2 Recall for Class3	0.5 0.7 0.7
Mean Recall	0.633
F-measure for Class1 F-measure for Class2 F-measure for Class3	0.549 0.614 0.737
Mean F-measure	0.633

Confusion Matrix :

$$C = \begin{matrix} 25 & 20 & 5 \\ 10 & 35 & 5 \\ 6 & 9 & 35 \end{matrix}$$

2.3.1.7 64 clusters

K = 64

Classification Accuracy (%)	60%
Precision for Class1 Precision for Class2 Precision for Class3	0.590 0.527 0.735
Mean Precision	0.618
Recall for Class1 Recall for Class2 Recall for Class3	0.72 0.58 0.5
Mean Recall	0.6
F-measure for Class1 F-measure for Class2 F-measure for Class3	0.649 0.552 0.595
Mean F-measure	0.598

Confusion Matrix :

$$C = \begin{matrix} 36 & 12 & 2 \\ 14 & 29 & 7 \\ 11 & 14 & 25 \end{matrix}$$

Observations :

- The maximum accuracy of 65% is obtained at K = 16 clusters per class and decreases thereafter.
- The classification accuracy of the image dataset is highly dependent on the feature extraction process.
- In this case, every image is represented by 32 fixed size blocks and each block was a feature vector set of 24-dimentional color histogram feature vectors.

2.3.2 Bag of Visual Words

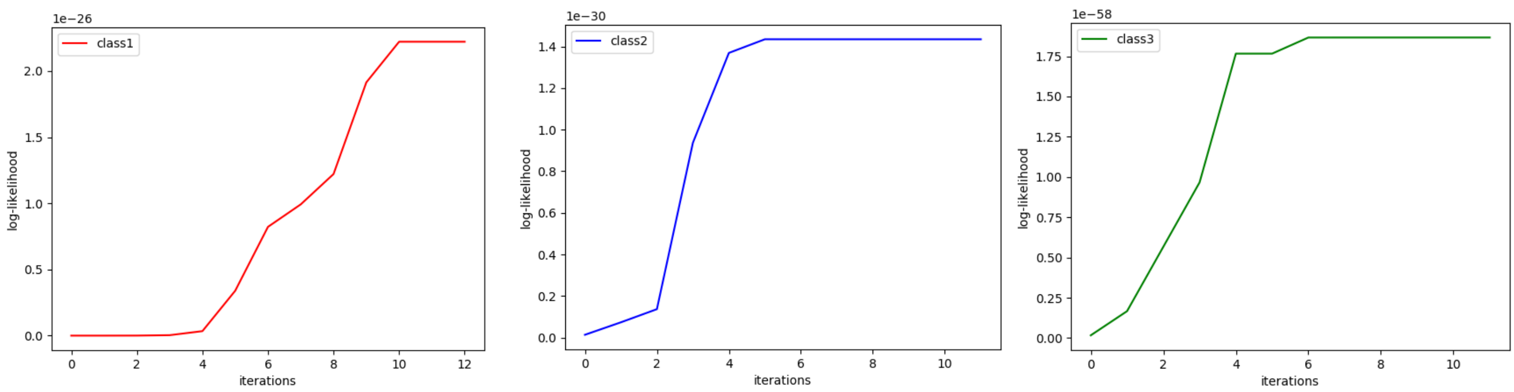
2.3.2.1 1 cluster

K = 1

Classification Accuracy (%)	70.667%
Precision for Class1 Precision for Class2 Precision for Class3	0.656 0.718 0.766
Mean Precision	0.713
Recall for Class1 Recall for Class2 Recall for Class3	0.84 0.56 0.72
Mean Recall	0.707
F-measure for Class1 F-measure for Class2 F-measure for Class3	0.737 0.629 0.742
Mean F-measure	0.702

Confusion Matrix :

$$C = \begin{matrix} 42 & 3 & 5 \\ 16 & 28 & 6 \\ 6 & 8 & 36 \end{matrix}$$



Log-likelihood vs Iterations graphs for K=1

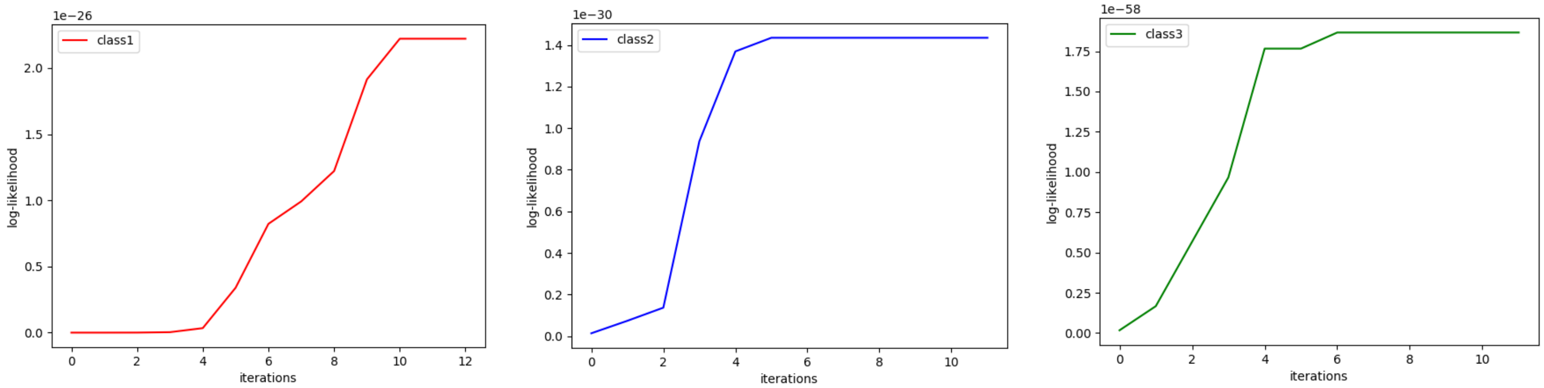
2.3.2.2 2 clusters

K = 2

Classification Accuracy (%)	64.667%
Precision for Class1 Precision for Class2 Precision for Class3	0.566 0.68 0.786
Mean Precision	0.677
Recall for Class1 Recall for Class2 Recall for Class3	0.94 0.34 0.66
Mean Recall	0.647
F-measure for Class1 F-measure for Class2 F-measure for Class3	0.707 0.453 0.717
Mean F-measure	0.625

Confusion Matrix :

$$C = \begin{matrix} 47 & 1 & 2 \\ 26 & 17 & 7 \\ 10 & 7 & 33 \end{matrix}$$



Log-likelihood vs Iterations graphs for K=2

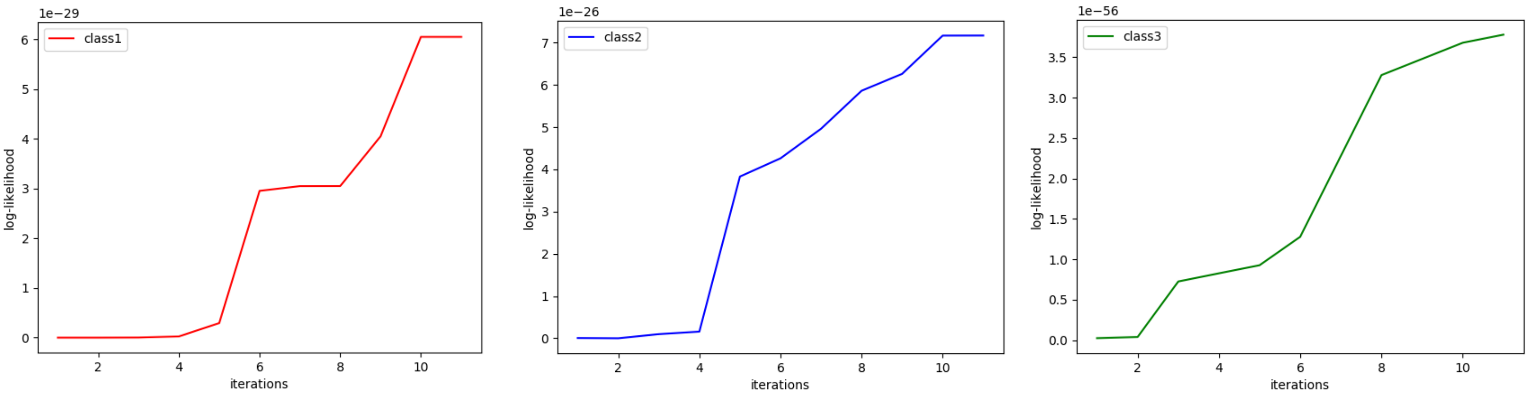
2.3.2.3 4 clusters

K = 4

Classification Accuracy (%)	61.333%
Precision for Class1 Precision for Class2 Precision for Class3	0.56 0.6 0.691
Mean Precision	0.617
Recall for Class1 Recall for Class2 Recall for Class3	0.84 0.24 0.76
Mean Recall	0.613
F-measure for Class1 F-measure for Class2 F-measure for Class3	0.672 0.343 0.724
Mean F-measure	0.579

Confusion Matrix :

$$C = \begin{matrix} 42 & 1 & 7 \\ 28 & 12 & 10 \\ 5 & 7 & 38 \end{matrix}$$



Log-likelihood vs Iterations graphs for K=4

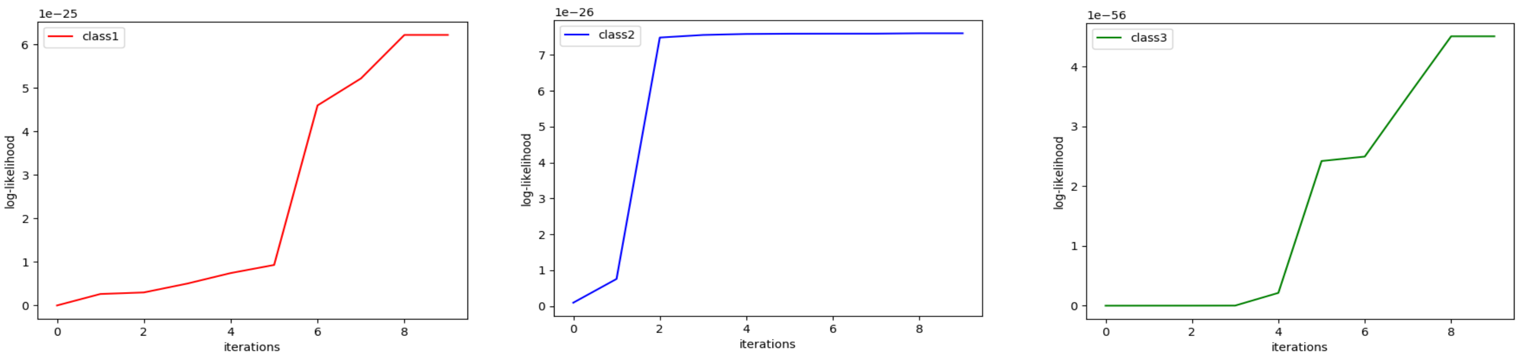
2.3.2.4 8 clusters

K = 8

Classification Accuracy (%)	66%
Precision for Class1 Precision for Class2 Precision for Class3	0.702 0.646 0.636
Mean Precision	0.661
Recall for Class1 Recall for Class2 Recall for Class3	0.66 0.62 0.7
Mean Recall	0.66
F-measure for Class1 F-measure for Class2 F-measure for Class3	0.680 0.633 0.667
Mean F-measure	0.659

Confusion Matrix :

$$C = \begin{matrix} 33 & 8 & 9 \\ 8 & 31 & 11 \\ 6 & 9 & 35 \end{matrix}$$



Log-likelihood vs Iterations graphs for K=8

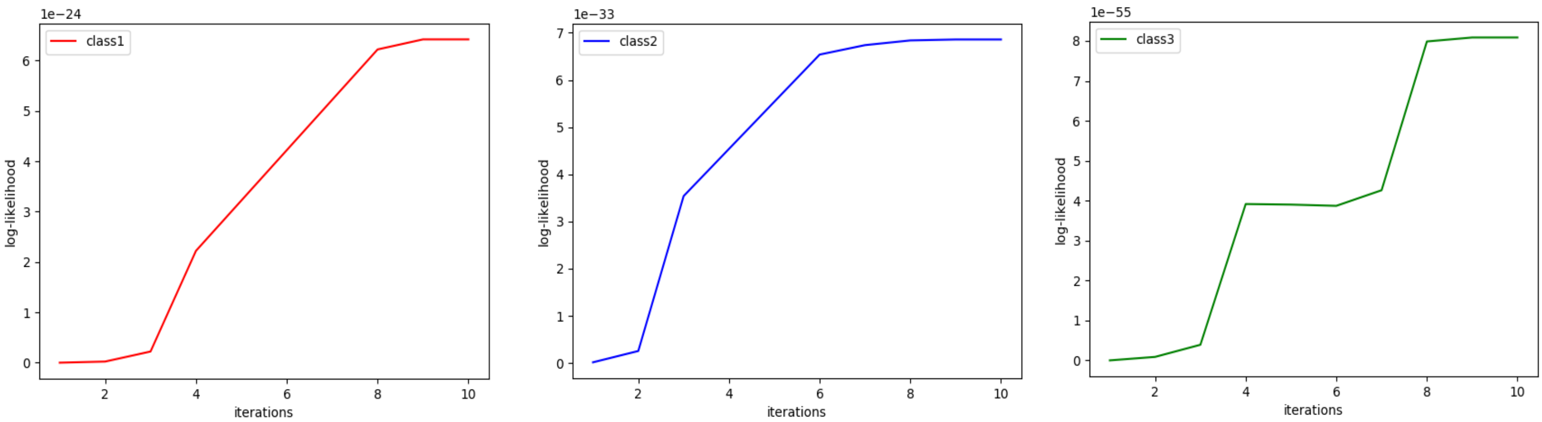
2.3.2.5 16 clusters

K = 16

Classification Accuracy (%)	60.667%
Precision for Class1 Precision for Class2 Precision for Class3	0.529 0.619 0.75
Mean Precision	0.633
Recall for Class1 Recall for Class2 Recall for Class3	0.9 0.26 0.66
Mean Recall	0.607
F-measure for Class1 F-measure for Class2 F-measure for Class3	0.667 0.366 0.702
Mean F-measure	0.578

Confusion Matrix :

$$C = \begin{matrix} 45 & 2 & 3 \\ 29 & 13 & 8 \\ 11 & 6 & 33 \end{matrix}$$



Log-likelihood vs Iterations graphs for K=16

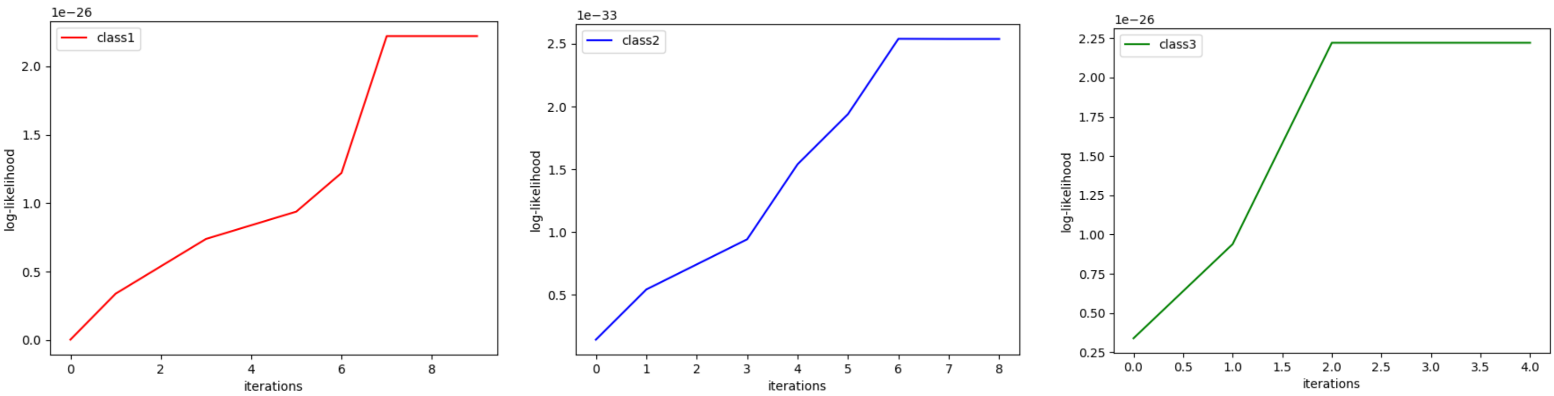
2.3.2.6 32 clusters

K = 32

Classification Accuracy (%)	70.667%
Precision for Class1 Precision for Class2 Precision for Class3	0.690 0.667 0.766
Mean Precision	0.707
Recall for Class1 Recall for Class2 Recall for Class3	0.8 0.6 0.72
Mean Recall	0.707
F-measure for Class1 F-measure for Class2 F-measure for Class3	0.741 0.632 0.742
Mean F-measure	0.704

Confusion Matrix :

$$C = \begin{matrix} 40 & 6 & 4 \\ 13 & 30 & 7 \\ 5 & 9 & 36 \end{matrix}$$



Log-likelihood vs Iterations graphs for K=32

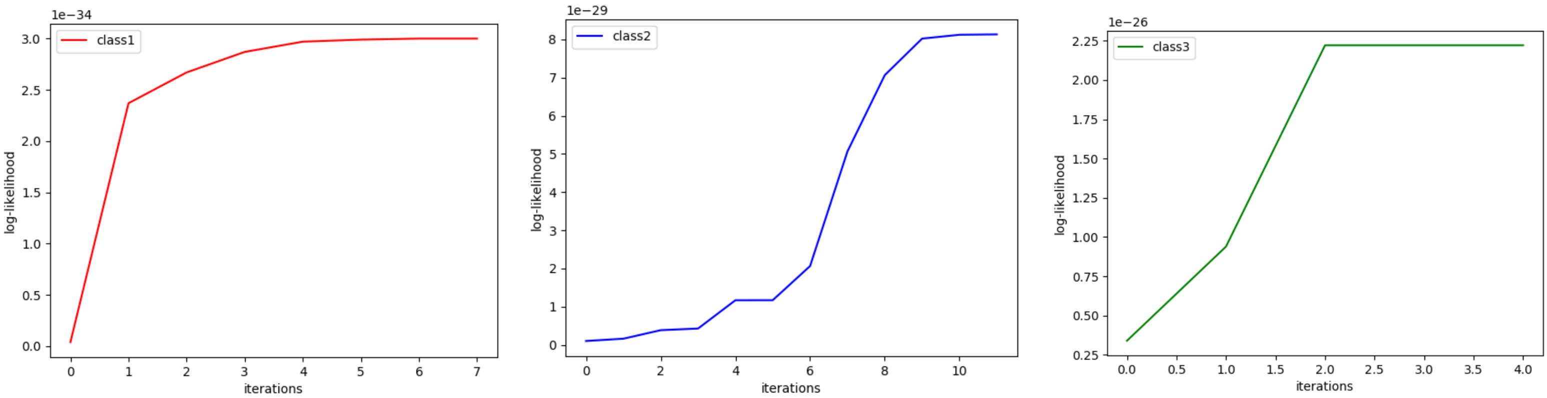
2.3.2.7 64 clusters

K = 64

Classification Accuracy (%)	66.667%
Precision for Class1 Precision for Class2 Precision for Class3	0.618 0.703 0.711
Mean Precision	0.677
Recall for Class1 Recall for Class2 Recall for Class3	0.84 0.52 0.64
Mean Recall	0.667
F-measure for Class1 F-measure for Class2 F-measure for Class3	0.712 0.598 0.674
Mean F-measure	0.660

Confusion Matrix :

$$C = \begin{matrix} 42 & 3 & 5 \\ 16 & 26 & 8 \\ 10 & 8 & 32 \end{matrix}$$

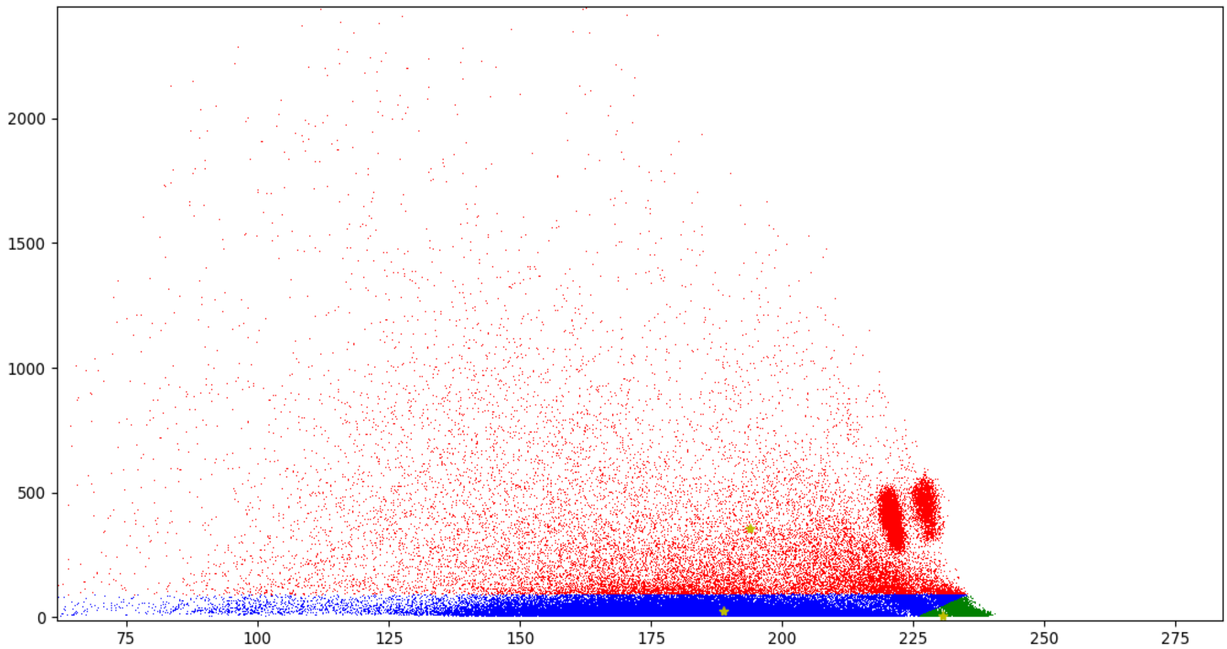


Log-likelihood vs Iterations graphs for K=64

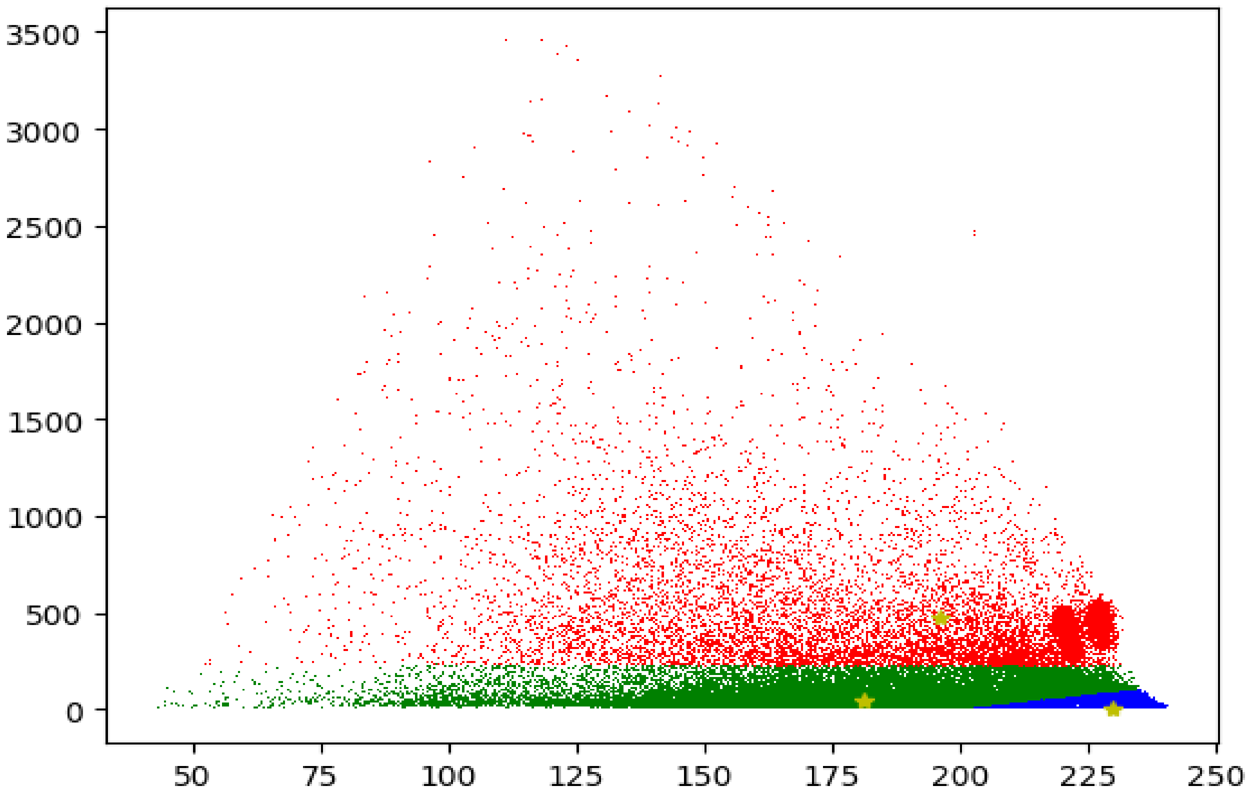
Observations :

- The maximum accuracy of 70.667% is obtained at K = 1 clusters per class.
- In this case, every image is represented by 32 dimentional Bag-of-vitual-words (BoVW) obtained by 32 clustering of the colour coded histograms.
- The accuracy in case of BoVW representation is better than that obtained by colour histogram feature vector.

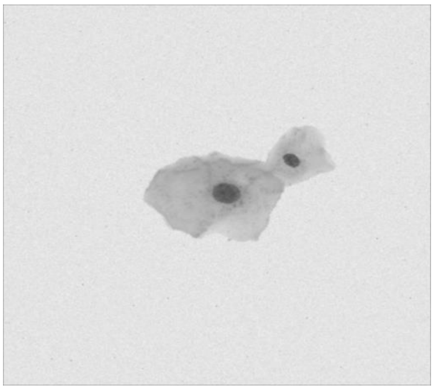
2.3 Cell Dataset



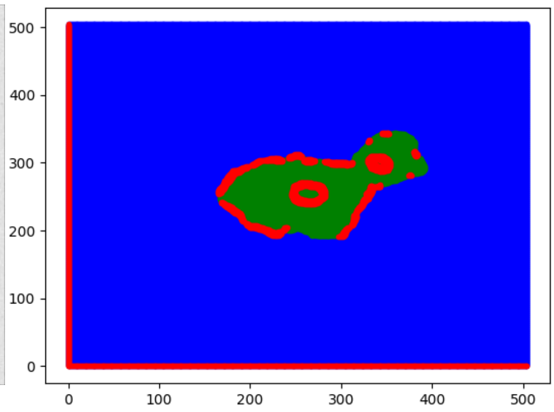
K-Means Clustering Scatter Plot



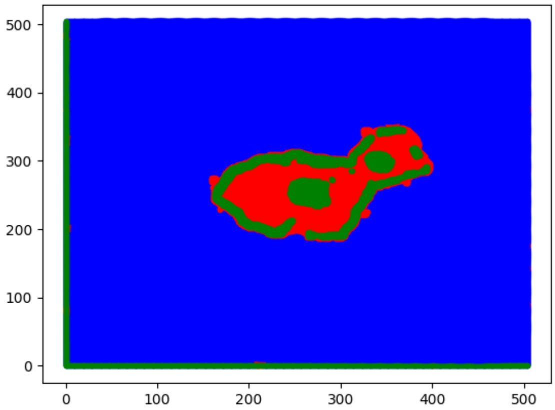
Clustering using GMM Scatter Plot



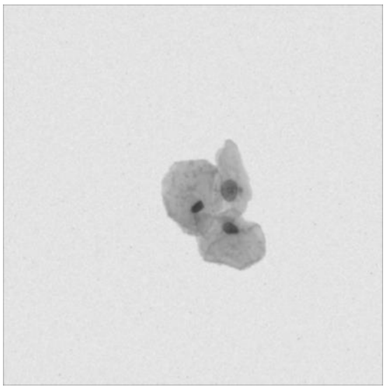
Cell Image



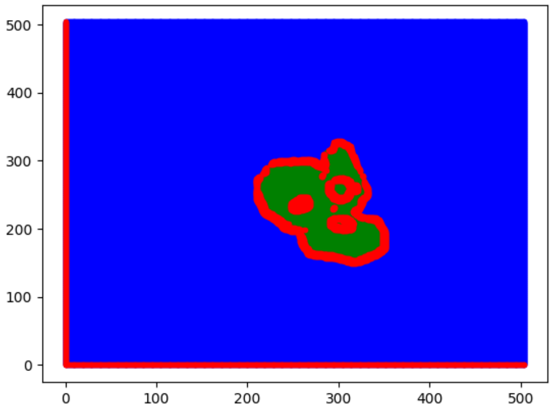
K Means



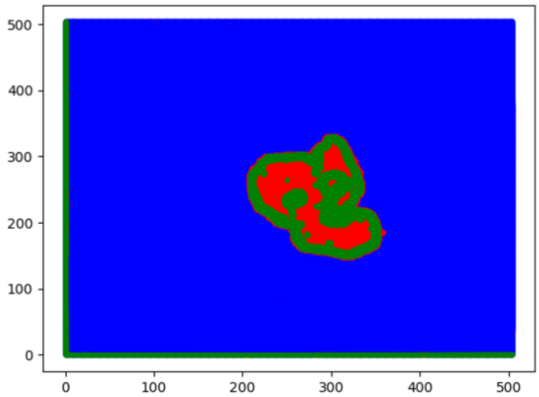
GMM



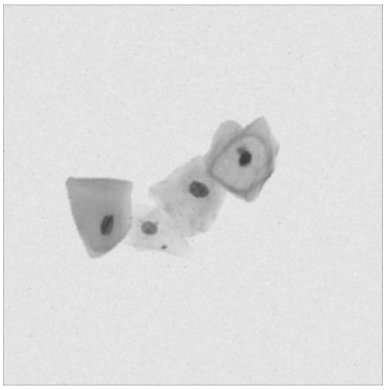
Cell Image



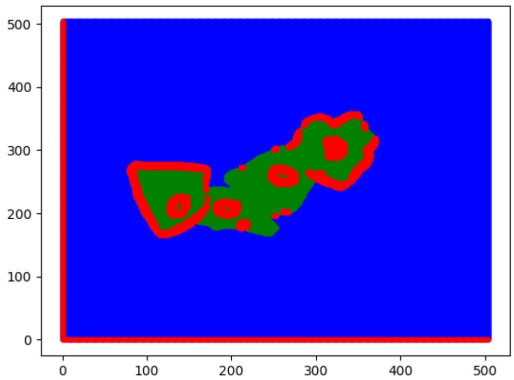
K Means



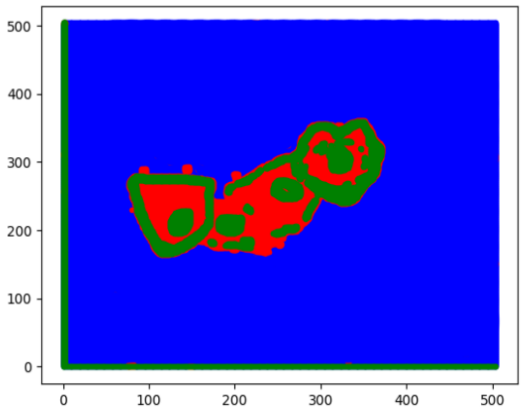
GMM



Cell Image

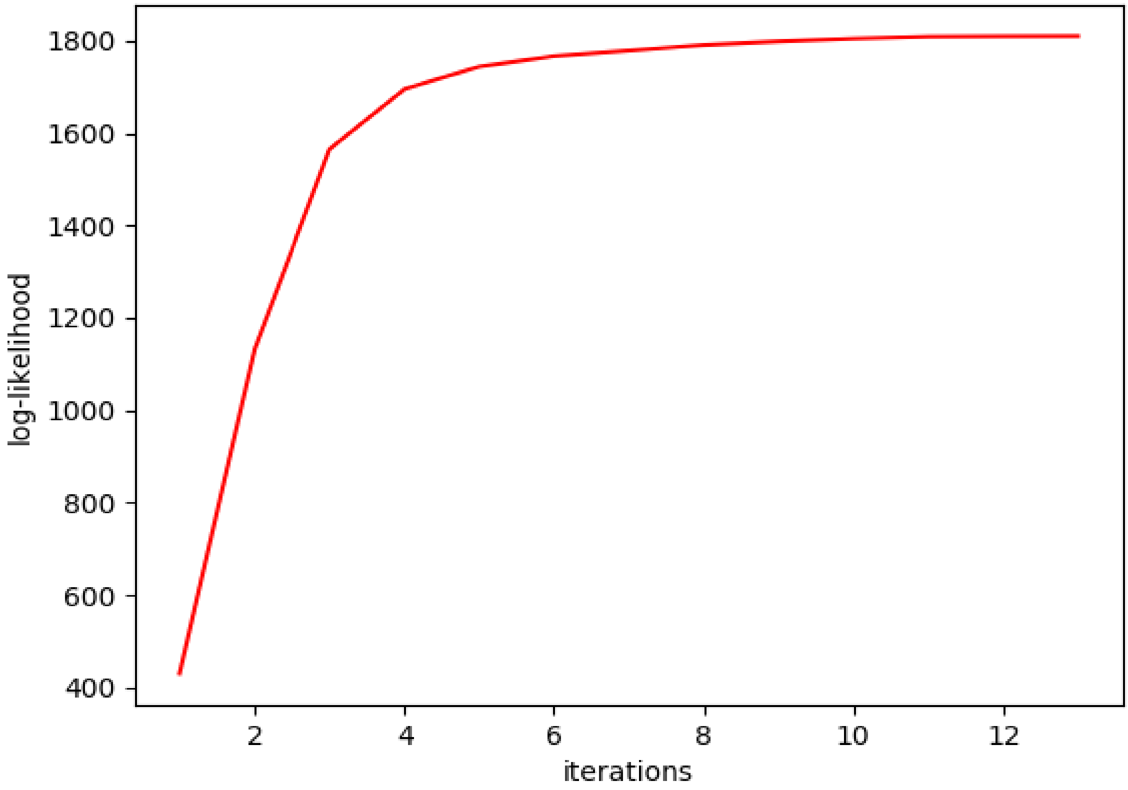


K Means



GMM

Log likelihood vs Iterations graph:



Observations :

- Better segmentation is observed in case of GMM as compared to that of Kmeans clustering.
- GMM allows non linear decision boundary therefore better segmenting between the three parts.
- Clustering accuracy depends on the way image feature vectors are extracted. In this case 7x7 non overlapping patches were used to form feature vectors of stack 2 dimentional feature vectors of mean and variance of the patch of the training set.
- The test images were represented by overlapping 7x7 such patches.

3 Conclusion and Observations:

- The decision boundaries are more precise when the data is modeled using mixture of multiple Gaussians as compared to unimodal Gaussian.
- Different classification accuracies are obtained when number of clusters K are varied.
- Although the accuracy tends to increase with the number of clusters assumed for a class, but due to overlapping data, over clustering may cause the class to cover non-belonging points as well, this is evident in the real world data.
- The boundary between 2 clusters in K-Means is linear because distance measure used in K-means classification is Euclidean distance.
- As the number of iterations increase, cluster assignment of the data points changes and approaches convergence.
- The accuracies obtained for the nonlinear datasets are far better than that obtained in unimodal case. Accuracies for real dataset are behind that obtained in unimodal case.
- The value of log likelihood increases with the number of iterations till it reaches the local / global optima.