Pattern Recognition

Assignment-2

Group: 10

October 21 , 2018

Team members:

1.Randheer kumar (B16139)

2.Dhrubodeep Basumatary (B16017)

3.Dilip Kumar Chauhan (B16018)

Contents:

Objective

1.To Build the Bayes Classifier Using GMM and Classify the following dataset:

- 2D dataset
  - 1.Non-linearly separable dataset(Artificial)
  - 2.Real World Dataset(Speech dataset)
- 24D Histogram Representation of of Scene Images
- 32D Bag of Visual Word (BoVW)Representation of Scene Images
- Parameters of the GMM to be initialised using K-means Clustering
- To calculate Classification accuracy, precision, recall for every class and mean recall, mean precision ,F-measure for classifier.
- To find the Confusion matrix based on the performance for test data
- To plot the Constant density contour for all the classes with the training data superposed Only for 2D dataset
- To plot the Decision region with the training data superposed only for 2D dataset

2. To cluster Cell images' data into 3 clusters
   - ❖ To Cluster the Data into 3 clusters using K-Means and Clustering using GMM
   - ❖ To Segment the test Images using both the clustering techniques i.e K-means Clustering and Clustering using GMM

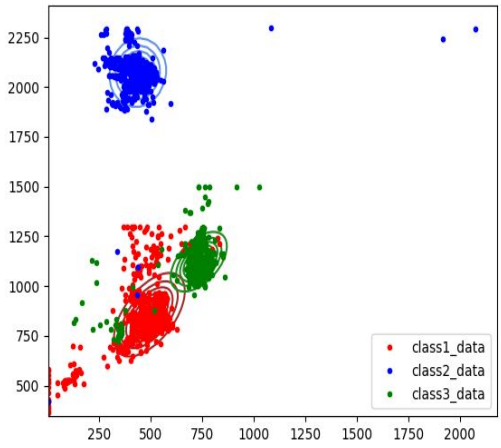
Procedure

- Data for each class is partitioned into two text files 75% for training and 25% for 2D dataset.
- For image Scene ,considering the 32*32 patch ,we extract the 24-dimensional colour histogram feature vectors from the testing and training images of all the classes.
- From the 24-dimensional colour histogram feature vectors ,we extract the Bag of visual Word (BoVW) feature vector for each image of each class from training and testing dataset.
- Considering the 7*7 non overlapping patch of cell image ,we extract the mean and Variance of the pixels which represents one feature vector,and stack all the feature vectors of all Patches of all the training cell images.
- Assuming that the class conditional density is coming from mixtures of Gaussian We estimate the parameters for each components of GMM for each class using the training dataset.
- In the case of Scene image we estimated the parameters for each Representation i.e Histogram and BoVW
- After the parameter is obtained we classify the testing dataset of each class using Bayes Classifier and find the confusion matrix
- Using the confusion matrix we calculate the accuracy,precision,recall for each class and mean precision, mean recall and F-measure for Classifier
- Cluster the cell image data into three clusters using k-means clustering and clustering using GMM
- Segment the test cell images using both Techniques K-means and GMM and compare the result.

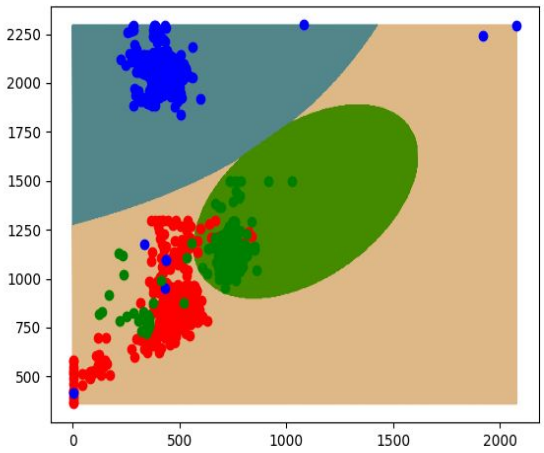
3.Observations:

3.1 Real World Dataset (speech Data) :

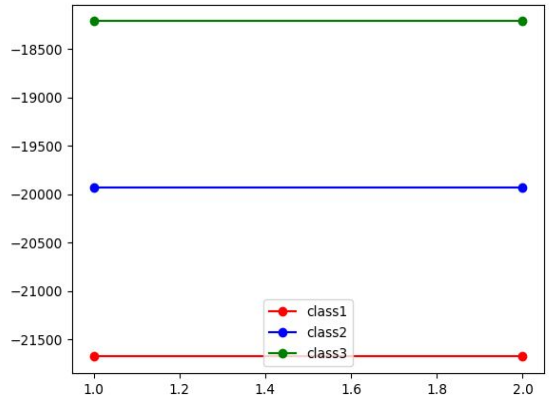
3.1.1 K=1



Contour plot

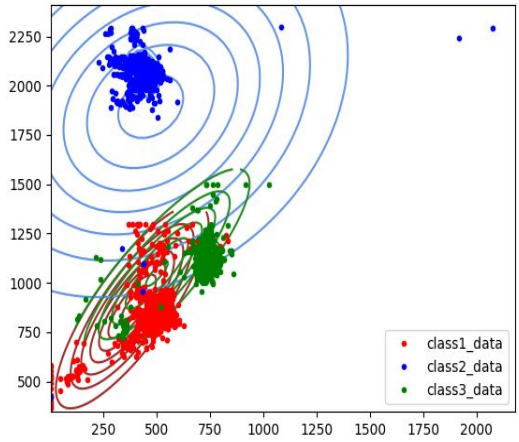


Decision Region plot

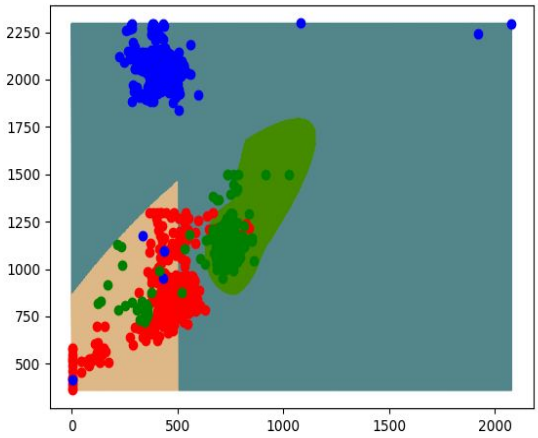


Log(Likelihood) vs Iteration

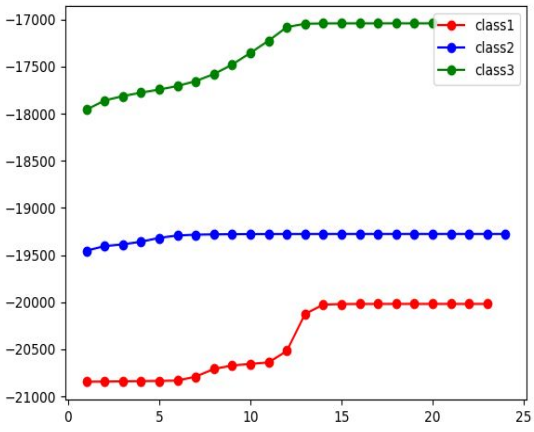
3.1.2 K=2



Contour plot

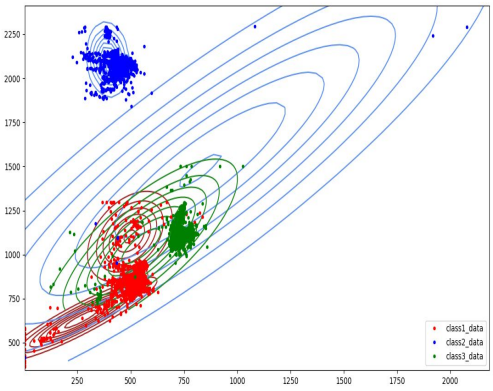


Decision Region Plot

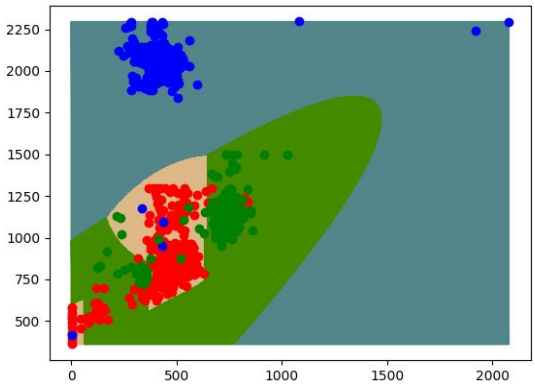


Log(likelihood) vs Iteration

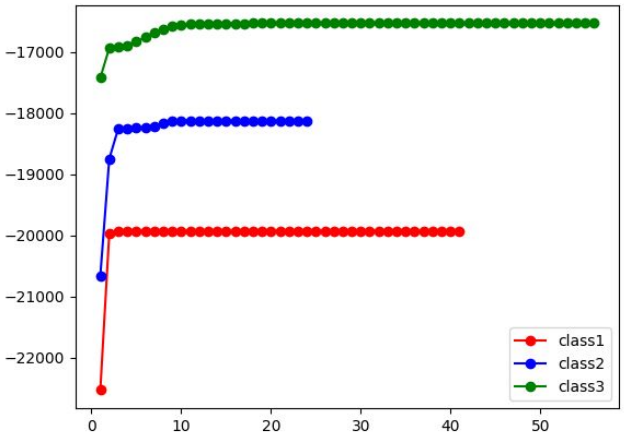
3.1.3 K=3



Contour plot

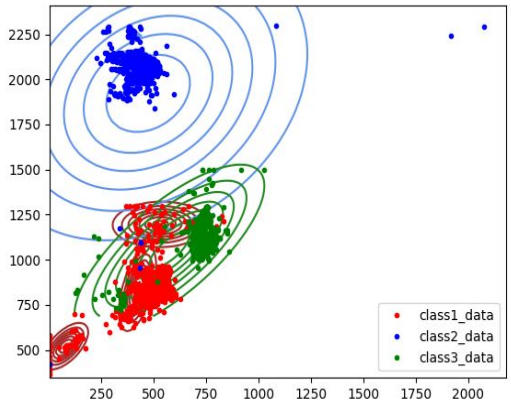


Decision Region Plot

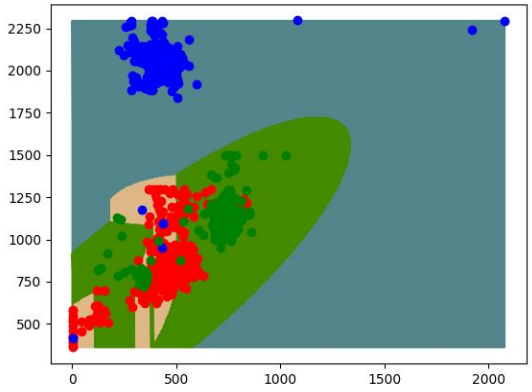


Log(Likelihood) vs iterations

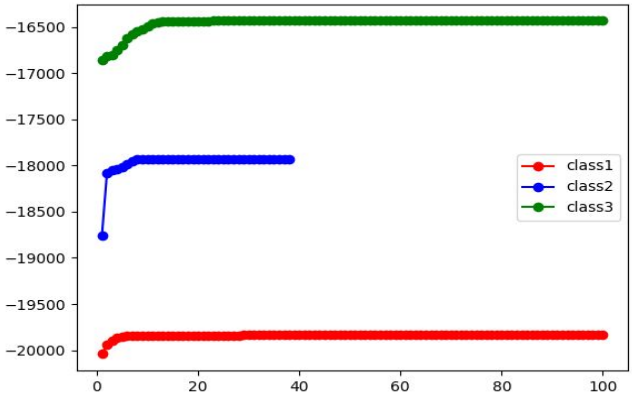
3.1.4 K=4



Contour plot

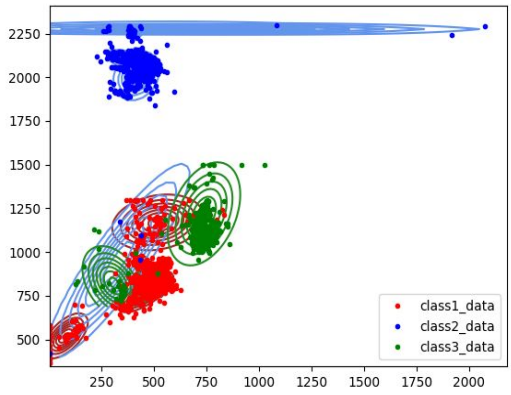


Decision Region

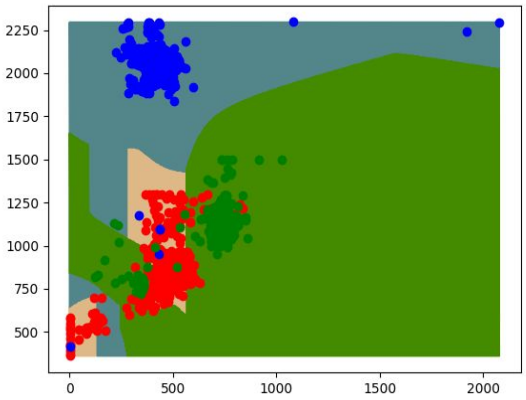


Log(Likelihood) vs iterations

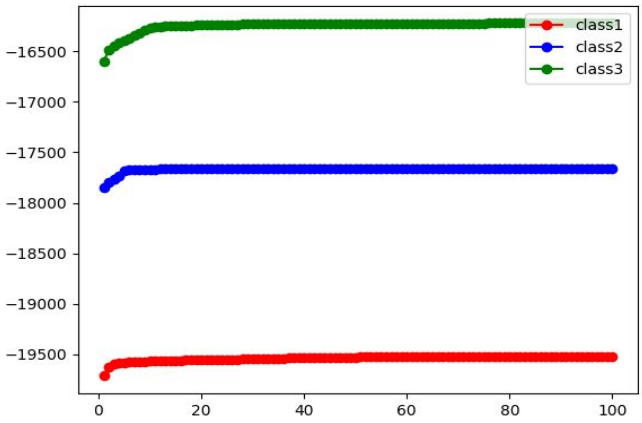
3.1.5 K=8



Contour Plot

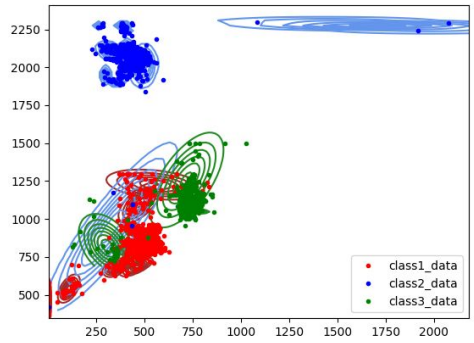


Decision Region

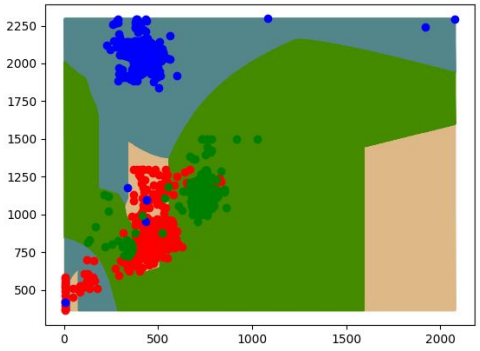


log(Likelihood) vs Iteration

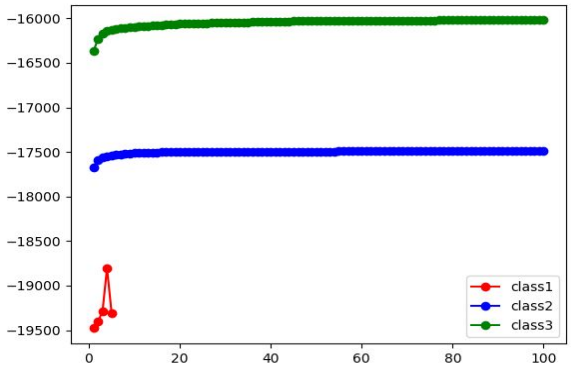
3.1.6 k=16



Contour Plot



Decision Region Plot



Log(Likelihood) vs Iteration

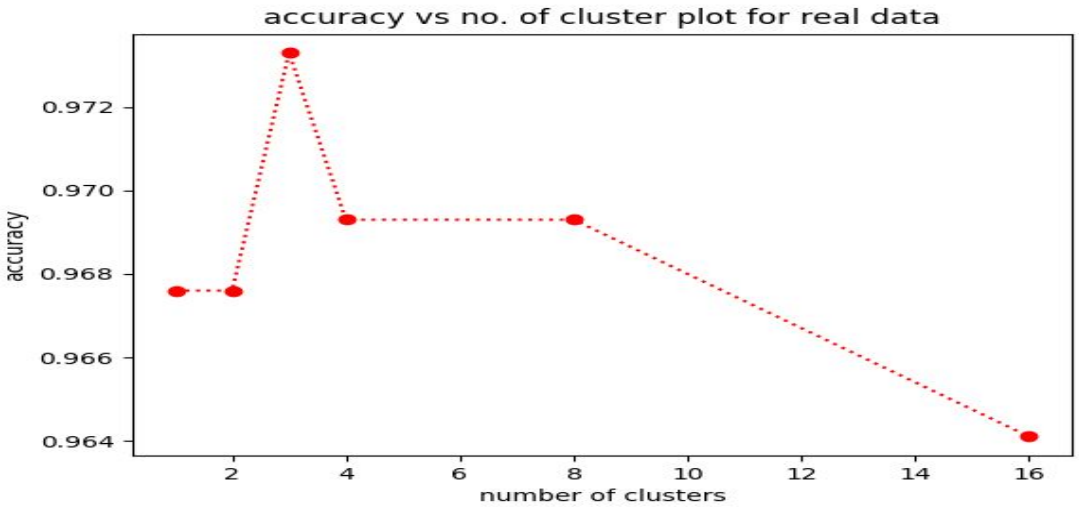
Confusion Matrix for Real data :

k=3	**Class1**	**Class2**	**Class3**
Class1	609	1	4
Class2	15	557	1
Class3	23	2	516

Analysis:

k=3	**Class1**	**Class2**	**Class3**	**Mean**
Precision	0.941267	0.994643	0.990403	0.975438
Recall	0.991857	0.972077	0.953789	0.972574
F-measure	-	-	-	0.974004

Accuracy : 97.3%

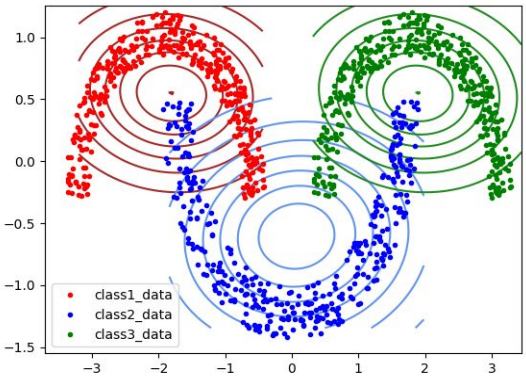


3.1.7: Inference :

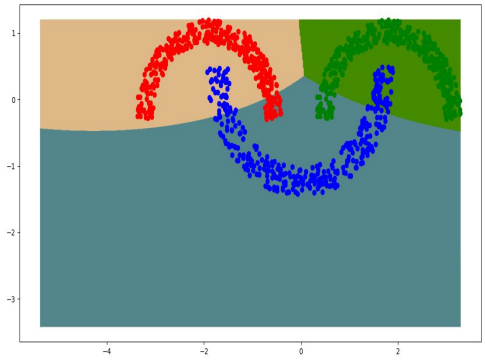
- As we increase the number of clusters , accuracy increases upto 3 clusters , then it decreases for further increase in number of cluster.The mixture component for which accuracy is coming maximum is fully dependent on the distribution of data.
- We got best result for k=3 , because data of each class best fit into mixture of 3 gaussian as it can be seen from contour plot . For k=1 (one cluster) we get exactly same result as we have got in previous assignment (considering unimodal Gaussian) .
- Total likelihood converges instantaneously because initial parameter is obtained from k-means clustering technique.
- Most of the data-points of each class is not falling in respective decision region because we are assuming that data is coming from a weighted superposition of gaussian distributions . But here gaussian distribution of a class is overlapping with gaussian distribution of other class as It can be seen from contour plot that's why the accuracy is little low(i.e not 100%)

3.2. Non-Linear DataSet :

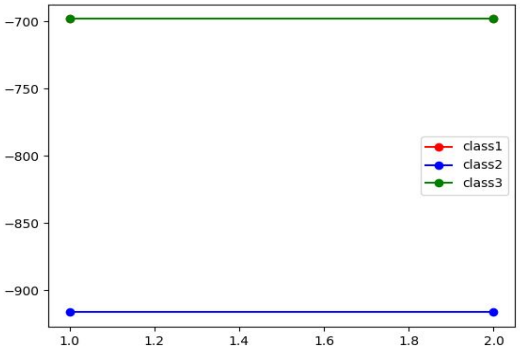
3.2.1 k=1



Contour Plot

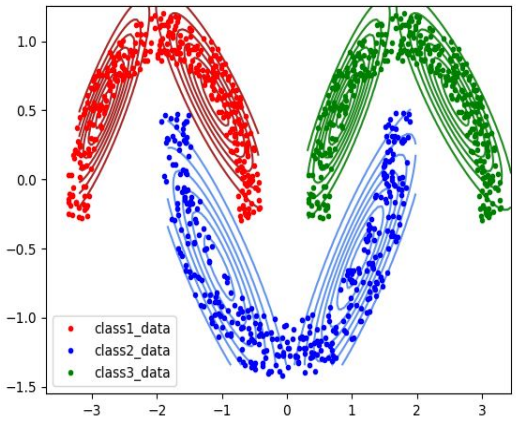


Decision Region

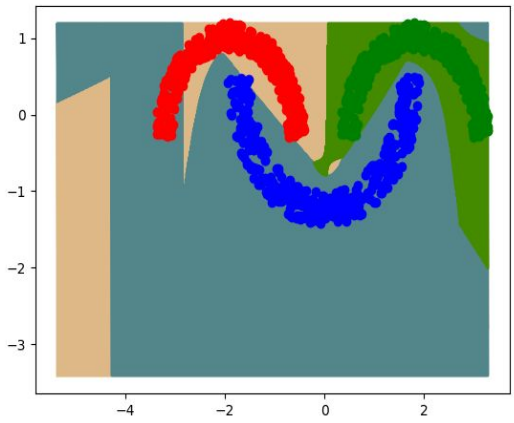


Log(Likelihood) vs iterations

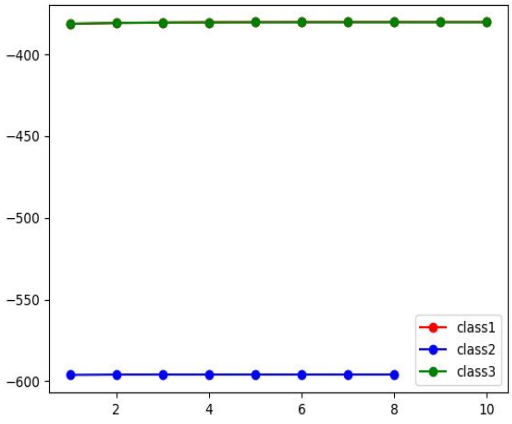
3.2.2 k=2



Contour Plot

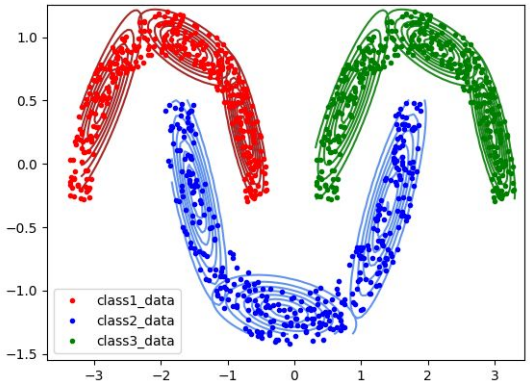


Decision Region Plot

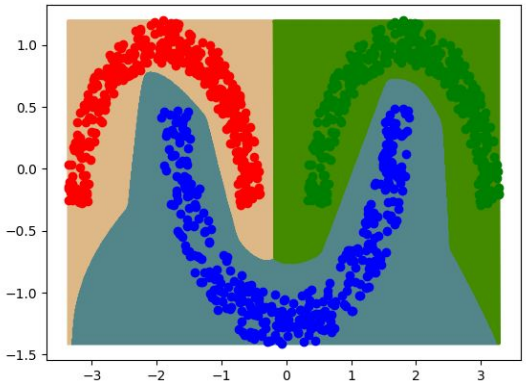


Log(Likelihood) vs Iterations

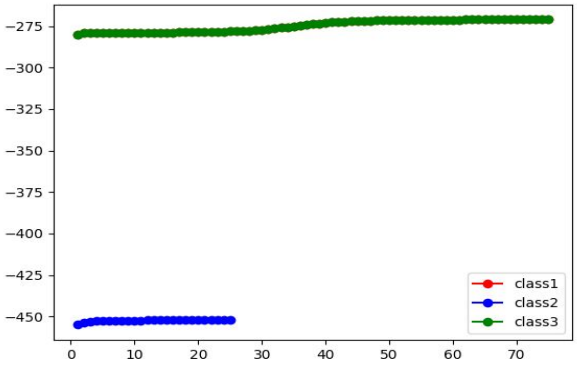
3.2.3 k=3



Contour Plot

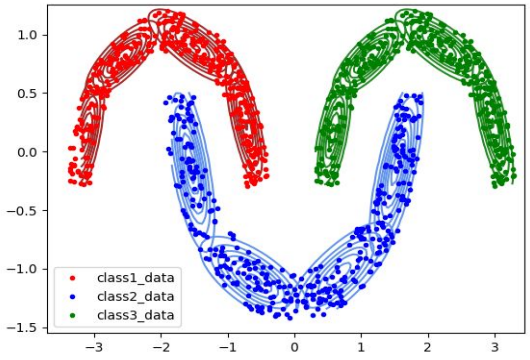


Decision Region Plot

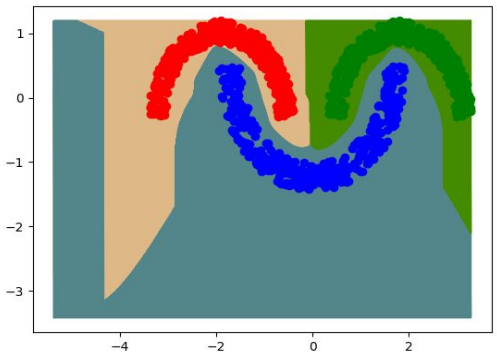


Log(Likelihood) vs Iterations

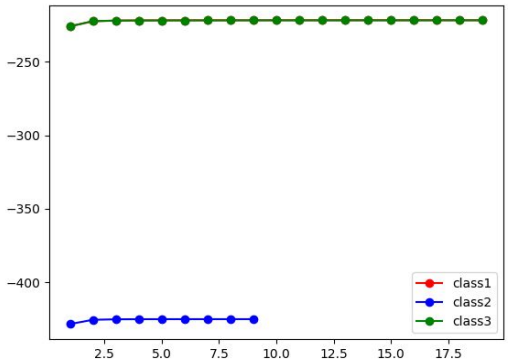
3.2.4 k=4



Contour Plot

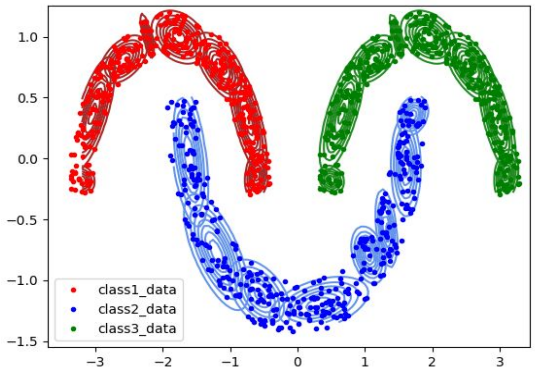


Decision Region

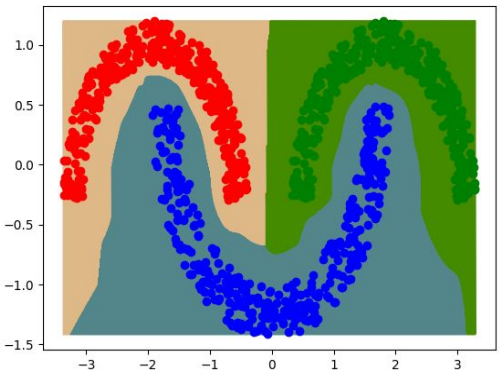


Log(Likelihood) vs iteration

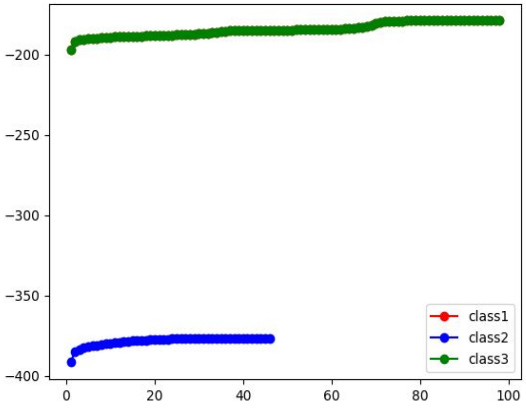
3.2.5 k=8



Contour Plot

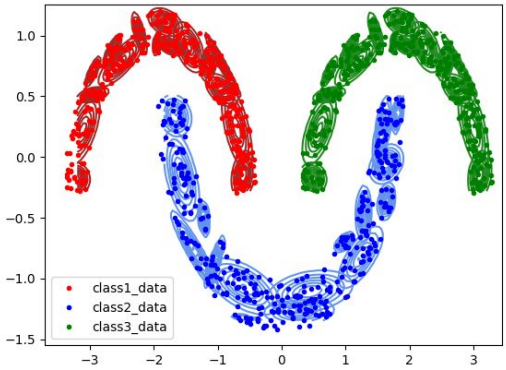


Decision Region

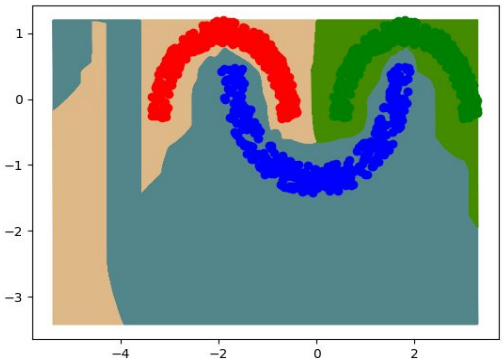


Log(Likelihood) vs Iteration

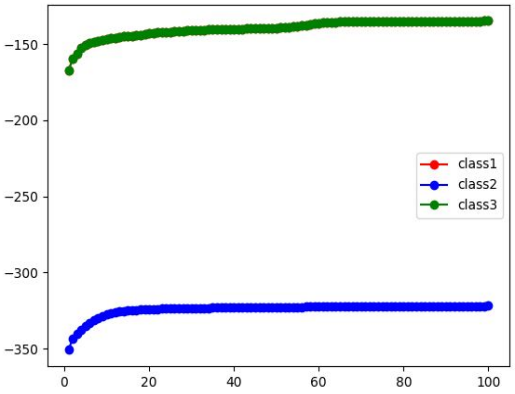
3.2.6 k=16



Contour Plot

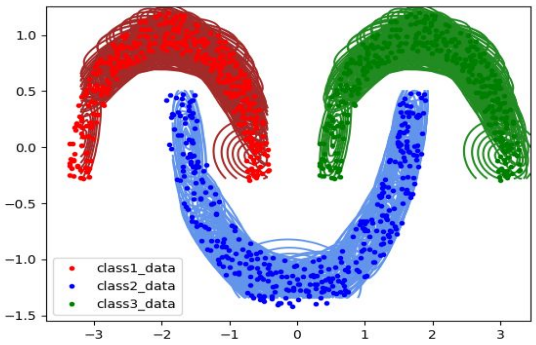


Decision Region Plot

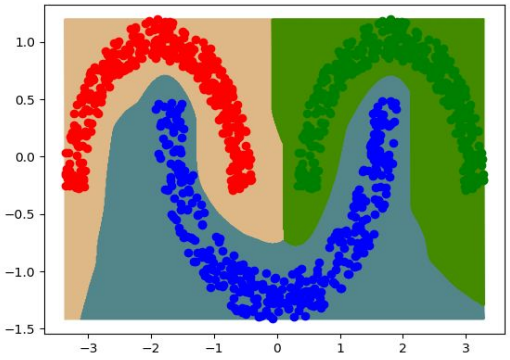


Log(Likelihood) vs Iteration

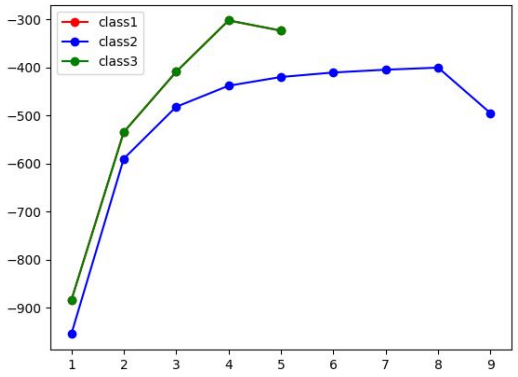
3.2.7 k=32



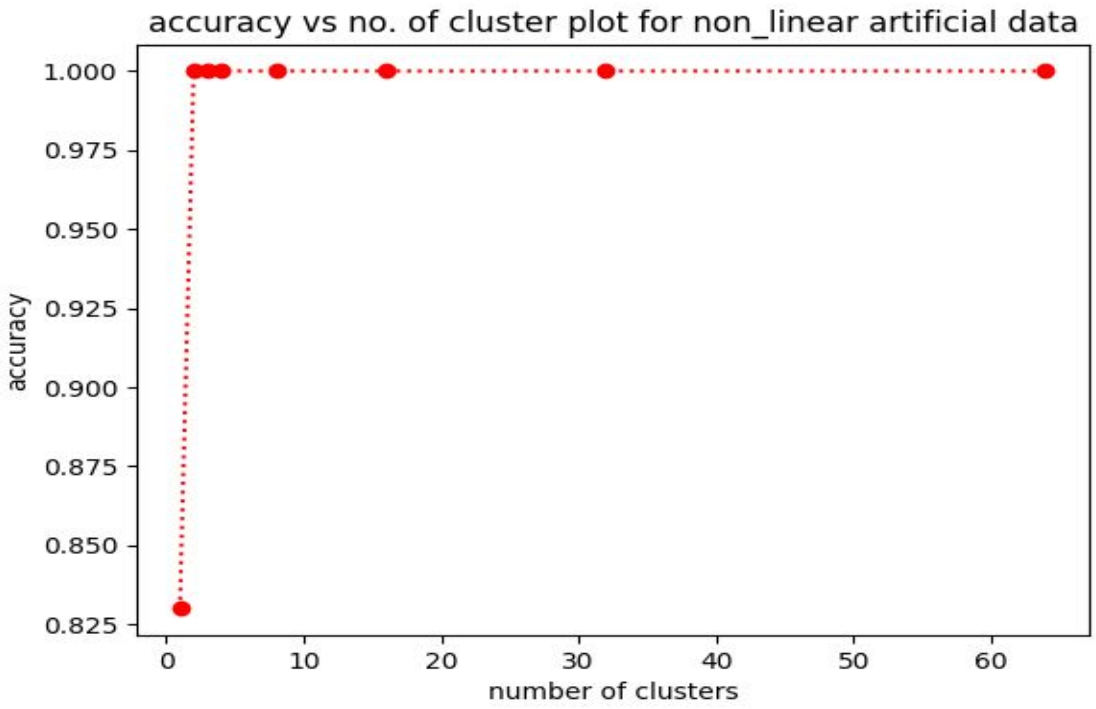
Contour Plot



Decision Region Plot



Log(Likelihood) vs Iteration



Confusion Matrix for Non-Linear data :

k=8	**Class1**	**Class2**	**Class3**
Class1	126	0	0
Class2	0	126	0
Class3	0	0	126

Analysis

k=8	**Class1**	**Class2**	**Class3**	**Mean**
Precision	1	1	1	1
Recall	1	1	1	1
F-measure	-	-	-	1

Accuracy : 100%

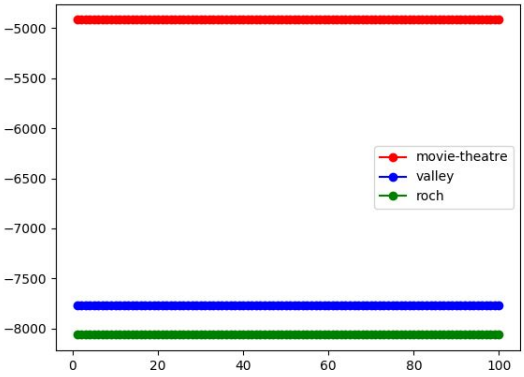
3.2.8 : Inference :

- We get similar result (in terms of accuracy and f-measure) for k=2,3,4,8,16,32. Because data-points of each class can be obtained from a mixture of 2,3,4,8, or 16 gaussian distribution (using estimated mean and covariance matrix), as it is clear from contour plot .
- Total data likelihood does not converge instantly for k=32 because initial parameter is obtained from k-means clustering and it does not work accurately for non-linearly separable clusters .
- For k=1 , most of data-points of each class is not falling in each respective decision region because we are assuming that data points of a class is coming from unimadal gaussian distribution .
- For k > 1, almost all data-points of each class is falling in its respective decision region, because data-points of each class can be obtained as weighted superposition of gaussian distributions and gaussian distributions of a class is not overlapping with other class .

3.3 Image Scene Dataset:

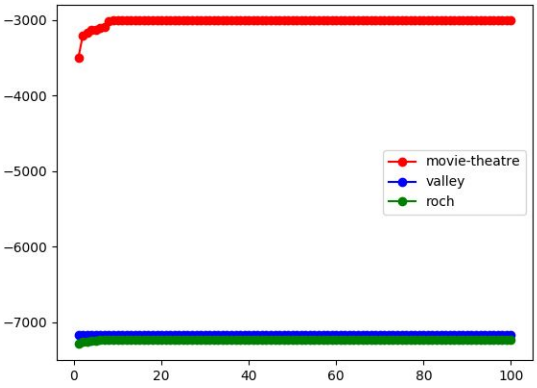
3.3.1 Bag of visual words :

3.3.1.1 K=1



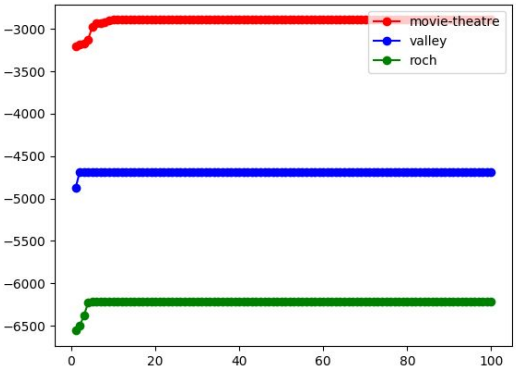
Log(likelihood) vs iterations

3.3.1.2 K=2



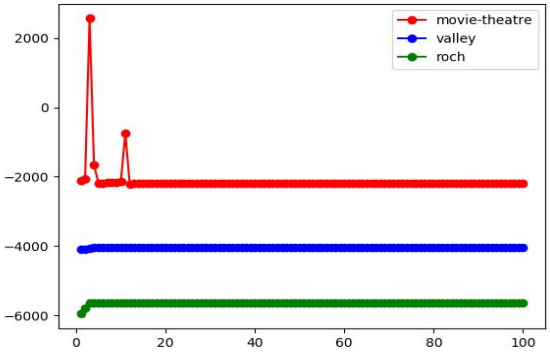
Log(Likelihood) vs iterations

3.3.1.3 K=4



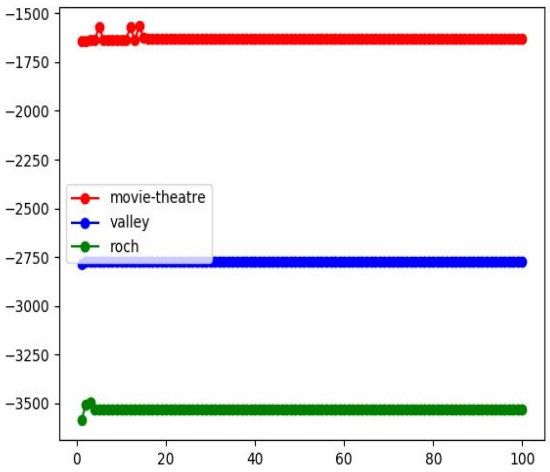
Log(likelihood) vs iterations

3.3.1.4 k=8



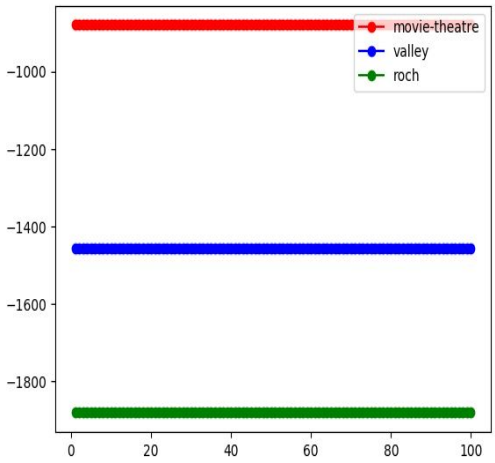
Log(Likelihood) vs iterations

3.3.1.5 k=16

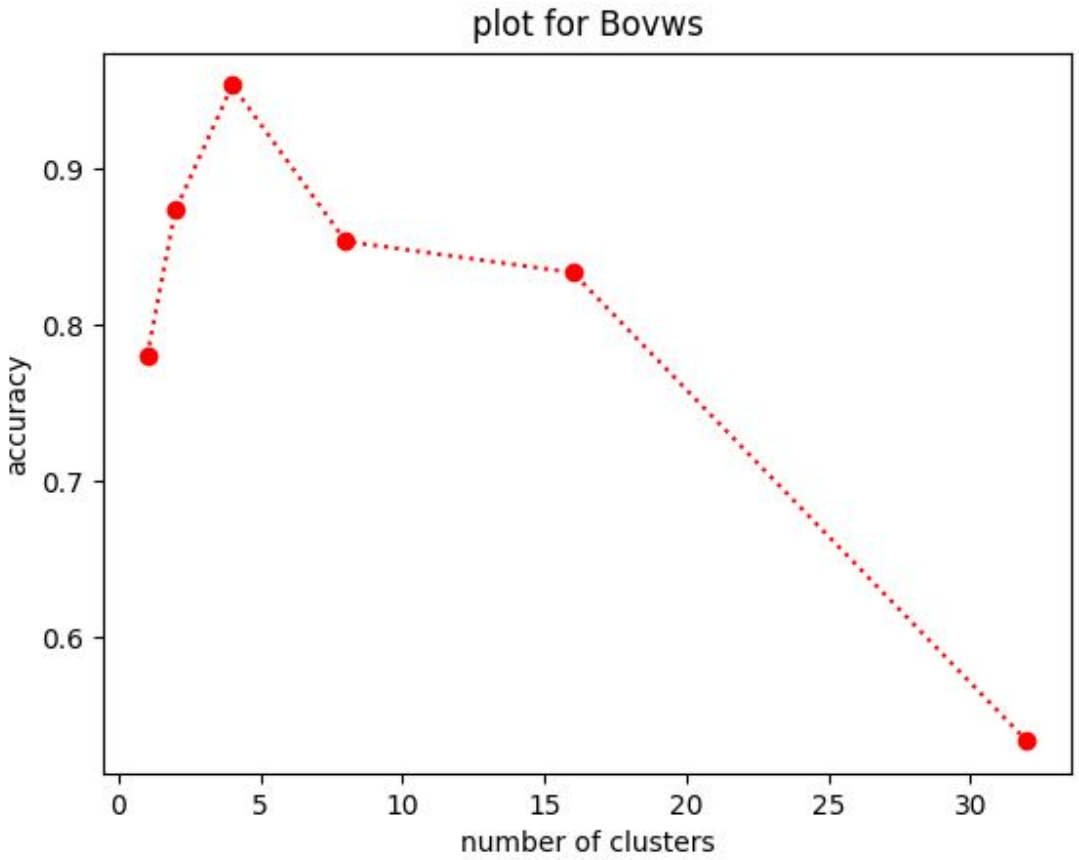


Log(Likelihood) vs Iterations

3.3.1.6 K=32



Log(Likelihood) vs Iterations



k=4	**Class1**	**Class2**	**Class3**
Class1	49	0	1
Class2	7	39	4
Class3	11	2	37

Analysis

k=4	**Class1**	**Class2**	**Class3**	**Mean**
Precision	0.979592	0.924528	0.958333	0.954151
Recall	0.96	0.98	0.92	0.953333
F-measure	-	-	-	0.953742

Accuracy : 95.3%

3.3.1.7: Inference

- Sometimes total data likelihood decreases with large amount with increase in iterations because we are forcefully making covariance matrix diagonal in each iteration . It remains almost constant for the most of the iteration due to same reason .
- Initially accuracy increases with increase in clusters , best accuracy is obtained for k=4 and it decreases with further increase in clusters . It further decreases due to overlapping gaussian distribution of a class with other class.

3.4 Cervical cytology (cell) image dataset

3.4.1 K-Means Clustering

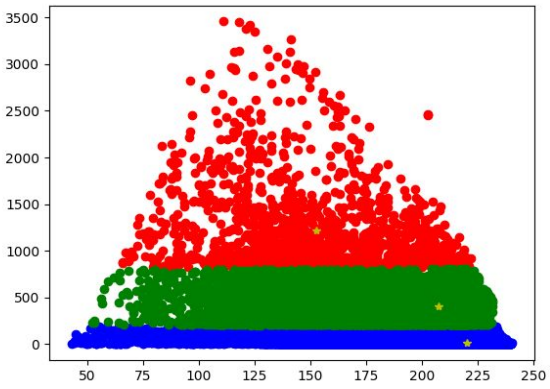


Fig.3.4.1.1 Three clusters

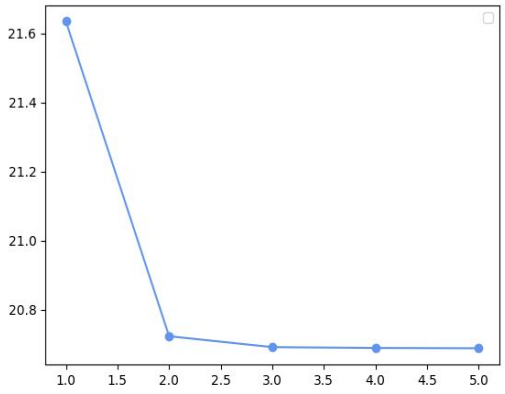


Fig 3.4.1.2 Log(distortion) vs iteration

3.4.2.Cluster Using GMM

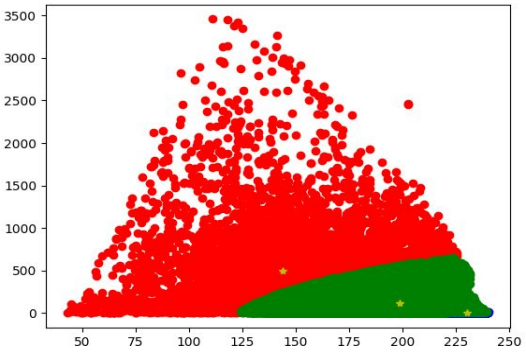


FIg3 .4.2.1 Three clusters

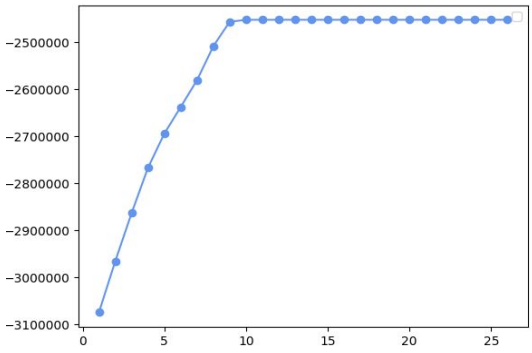


FIg 3.4.2.2 Log(Likelihood) vs iteration

3.4.3 Segmentation of test images

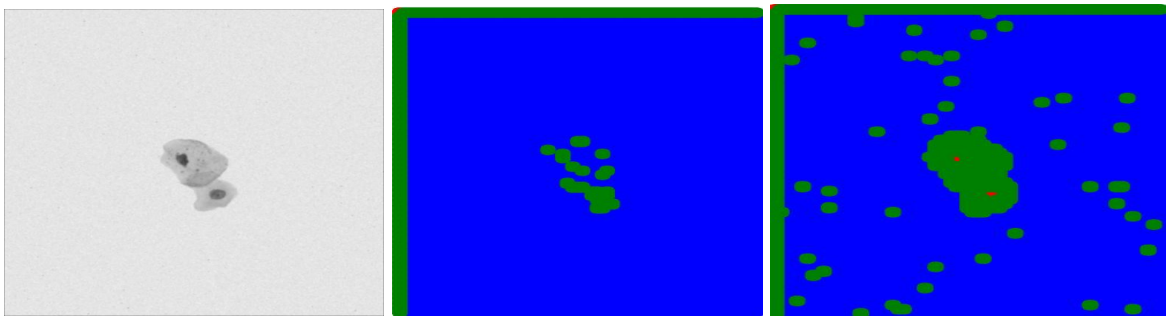


Image1 K-Means GMM

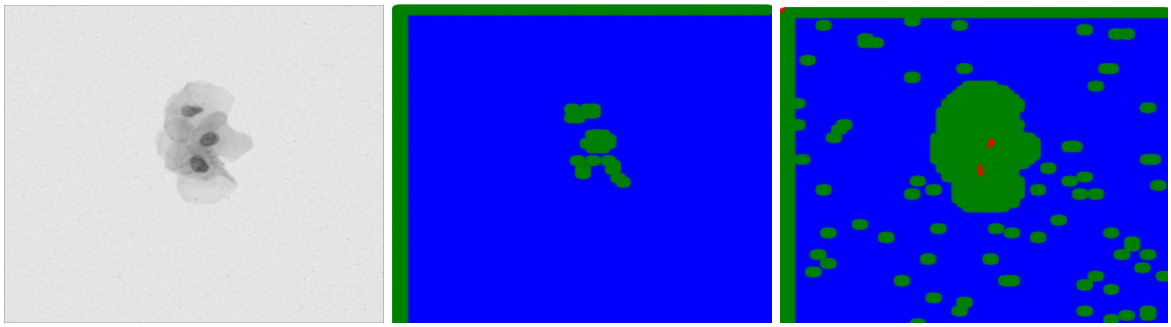


Image2 K-Means GMM

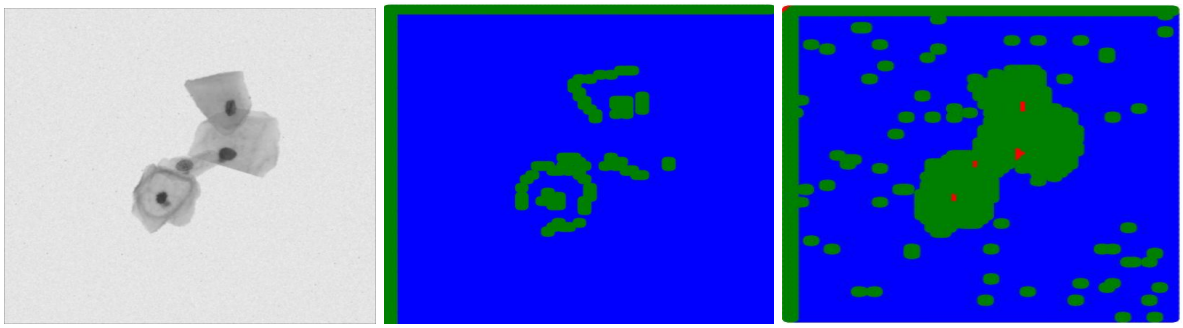


Image2 K-Means GMM

3.4.4. Inference

- When K-Means clustering is used on cell images' data to cluster into 3 cluster ,the boundary between the clusters is coming out to be linear(because of hard clustering) which we already know mathematically.
- We continue reducing the distortion function until converges(J_old-J_new <thress) and that is what we can see in Fig 3.4.1.2
- After Applying Clustering Using GMM ,the boundary between the clusters is non linear (because of soft clustering) due to the the way GMM is implemented i.e considering the likelihood of each data points to each cluster.
- In GMM,we continue maximising the Likelihood in M-step(graph of Log(Likelihood) vs iterations is shown in FIg 3.4.2.2).
- When we Segment image1 using K-Means Clustering the result is coming out to be very poor as nucleus is not detected(no.of data points falling in nucleus is very small that's why we are not be able to observe here) .this is happening because the no. feature vectors is not so large as we are considering non overlapping patch of7*7.(If we would have considered 7*7 overlapping patch the result would have been better).
- We can see that the Segmentation using GMM is giving better result as compare to segmentation using K-Means because K-Means is Hard clustering Technique and GMM is soft Clustering Technique.GMM is detecting Nucleus, Cytoplasm, and cell body but some of the cytoplasm data points are also coming in cell body this due to less number of data points using which GMM was built

- Segmentation of image2 and image3 using K-means is giving poor result and using GMM giving better result.The explanation of this is same as what we discussed for image1 .

4.conclusion

- ❖ When we consider that the data of each class is coming from the mixture of Gaussian(multimodal Gaussian),whatever be the distribution of data of each class the accuracy is better than the case when we consider unimodal Gaussian distribution of the data.
- ❖ We can see from the plot of accuracy vs no.of clusters,the accuracy is increases first and further starts decreasing and the cluster number for which the accuracy is maximum is different for different dataset.

- Here we have considered the same number of gaussian for each class and calculated the accuracy but in general different number of Gaussian for different class may give the best accuracy.
- Mathematically We know that the Soft clustering is better way of clustering which we can verify from the clustering of cell images data, as GMM being soft clustering technique gives better result than K-Means Clustering which is hard clustering
- As we know that pattern Recognition algorithm gives better result when the algorithm is applied on large dataset,here we can see that even GMM is soft clustering technique but in case of cell images giving little poor result because of less data points on which GMM clustering was built.