CS 669 Assignment 2

Rohit Patiyal
Devang Bacharwar
Mani Kumar

October 15, 2015

Contents

1 Objective

To build bayes classifier using Gaussian Mixture Model. Gaussian Mixture Model is built using the K-means clustering to initialize the parameters. The following data sets are used :

1. 2-D artificial Data of 3 or 4 classes that are nonlinearly separable.
2. Real world data of 3 classes: The real world data sets correspond to the formant frequencies F1 and F2 for vowel utterances.
3. Scene image data corresponding to 3 different classes A 23-dimensional feature vector is extracted from local blocks of an image for a particular scene. The 23-dimensional features include color histogram, edge directed histograms and entropy of wavelet coefficients. Each scene image is represented as a collection of 23-dimensional local feature vectors. Each file in a folder of a class indicates a scene image.

2 Procedure

1. Data for each class is partitioned into 75 % for training and 25 % for testing
2. The Data Set of each class is assumed to be Gaussian distribution but with multimodes. Now the distribution is assumed to be a mixture of k Gaussian distributions
3. Mean, Covariances and mixing coefficients are initialized for each of the mixtures for a class using k-means clustering, using the training data. Different values of k are taken. k ϵ { 2, 4, 8, $\cdots$ }
4. Starting from the initialized parameters from k-means, Means, Covariances and mixing coefficients for each mixture are estimated by maximizing the log likelihood which works on the optimization technique Expectation Maximization.
5. For points in a grid, likelihood is calculated for each class and is labeled as of the class with the maximum likelihood probability.
6. Labels for test data are calculated and confusion matrix with '/.
7. These labelled points are plotted with different colors to visualize the different regions separated by the decision boundaries.
8. The training data is also plotted over the regions, and observations are made.

3 Observations

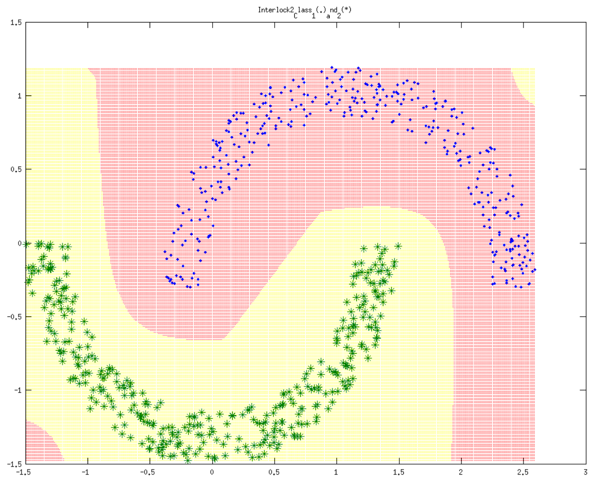
Here the cluster components are with respect to each class.The assumption taken is that each class have same number of cluster components.

3.1 Interlock Data Set

Interlock dataset has two classes with different means and similar covariances.

3.1.1 2 clusters components

2 clusters per each class gives 4 gaussian distribution which have spreads in different directions. The upper peak on the distribution determines the class for that region so giving out 4 intersection decision boundaty surfaces for gaussian distributions.



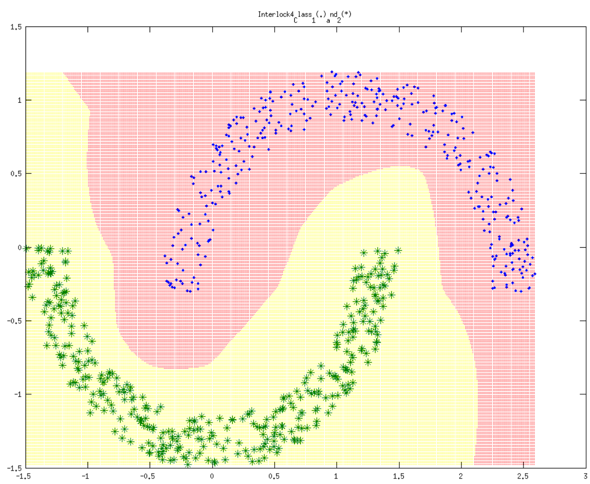
Correct : 250
Incorrect : 0
Accuracy : 100.00

		Predicted	
		Class 1	Class 2
Act.	Class 1	125	0
	Class 2	0	125

Figure 1: Decision region plot with the training data

3.1.2 4 clusters components

4 clusters per each class gives 8 gaussian distribution which have spreads in different directions. The upper peak on the distribution determines the class for that region so giving out 8 intersection decision boundary surfaces for gaussian distributions.



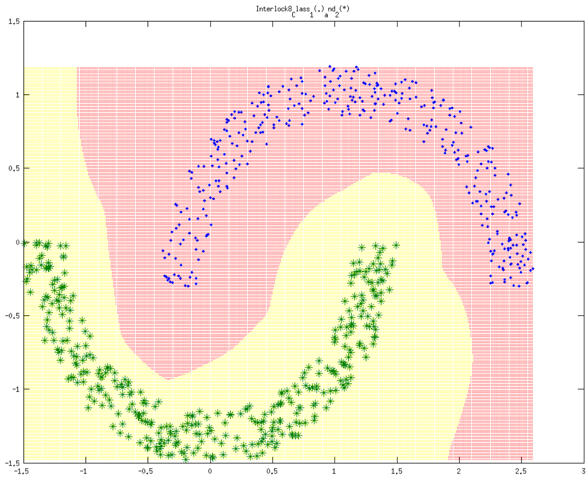
Correct : 250
Incorrect : 0
Accuracy : 100.00

		Predicted	
		Class 1	Class 2
Act.	Class 1	125	0
	Class 2	0	125

Figure 2: Decision region plot with the training data.

3.1.3 8 clusters components

8 clusters per each class gives 16 gaussian distribution which have spreads in different directions. The upper peak on the distribution determines the class for that region so giving out 16 intersection decision boundary surfaces for gaussian distributions.



Correct : 250
Incorrect : 0
Accuracy : 100.00

		Predicted	
		Class 1	Class 2
Act.	Class 1	125	0
	Class 2	0	125

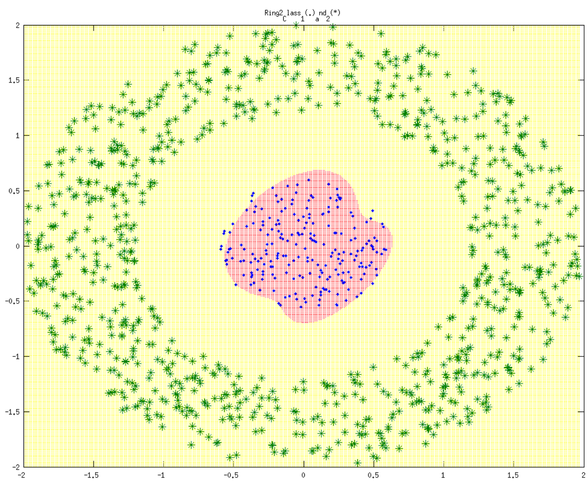
Figure 3: Decision region plot with the training data.

3.2 Ring Data Set

Ring data set includes 2 classes, both with same mean and with different covariances.

3.2.1 2 clusters components

2 clusters per each class gives 4 gaussian distribution which have spreads of different magnitude. The upper peak, class 1 peaks of the distribution determines the regions for center region giving out 3 intersection surfaces for gaussian distributions.



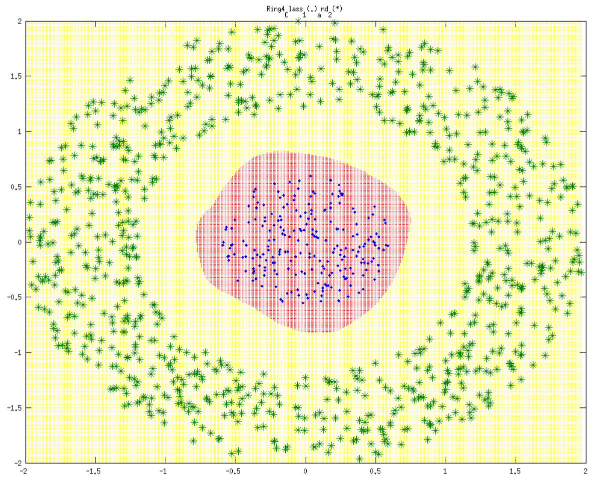
Correct : 369
Incorrect : 6
Accuracy : 98.4

		Predicted	
		Class 1	Class 2
Act.	Class 1	69	6
	Class 2	0	300

Figure 4: Decision region plot with the training data.

3.2.2 4 clusters components

4 clusters per each class gives 8 gaussian distribution which have spreads of different magnitude. The upper peak, class 1 peaks of the distribution determines the regions for centre region giving out 5 intersection surfaces for gaussian distributions.



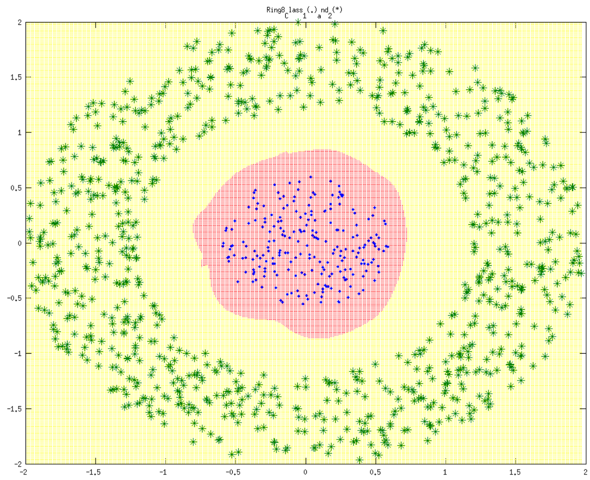
Correct : 375
Incorrect : 0
Accuracy : 100.00

		Predicted	
		Class 1	Class 2
Act.	Class 1	75	0
	Class 2	0	300

Figure 5: Decision region plot with the training data.

3.2.3 8 clusters components

8 clusters per each class gives 16 gaussian distribution which have spreads of different magnitude. The upper peak, class 1 peaks of the distribution determines the regions for center region giving out intersection surfaces for gaussian distributions.



Correct : 375
Incorrect : 0
Accuracy : 100.00

		Predicted	
		Class 1	Class 2
Act.	Class 1	75	0
	Class 2	0	300

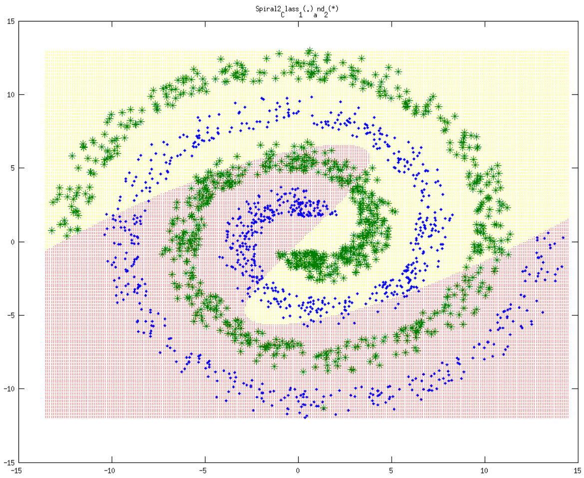
Figure 6: Decision region plot with the training data.

3.3 Spiral Data Set

Spiral data set contains 2 classes in form of spirals having slightly seperated mean and similar covariances.

3.3.1 2 clusters components

2 clusters for each class gives 4 gaussian distribution. The intersection of gaussian distribution for those clusters gives out the decision boundaries.



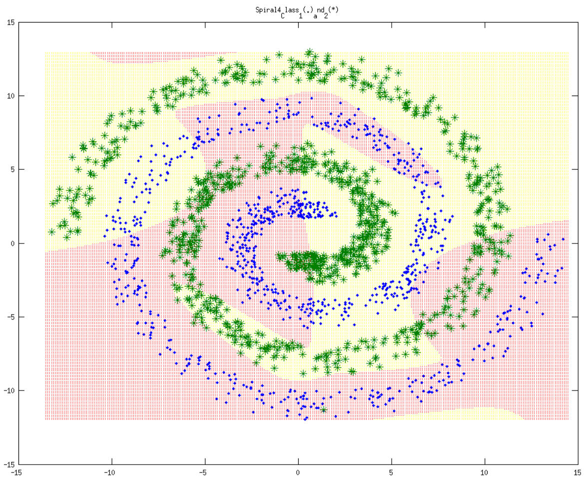
Correct : 400
Incorrect : 252
Accuracy : 61.34

		Predicted	
		Class 1	Class 2
Act.	Class 1	200	126
	Class 2	126	200

Figure 7: Decision region plot with the training data.

3.3.2 4 clusters components

4 clusters for each class gives 8 gaussian distribution. The intersection of gaussian distribution for those clusters gives out the decision boundaries.



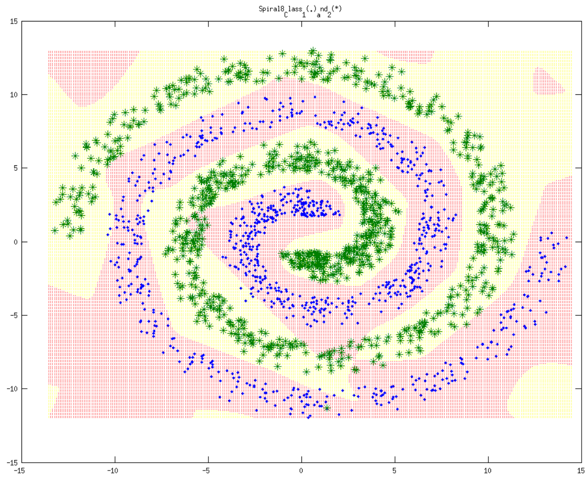
Correct : 449
Incorrect : 203
Accuracy : 68.865

		Predicted	
		Class 1	Class 2
Act.	Class 1	224	102
	Class 2	101	225

Figure 8: Decision region plot with the training data.

3.3.3 8 clusters components

8 clusters for each class gives 16 gaussian distribution. The intersection of gaussian distribution for those clusters gives out the decision boundaries.



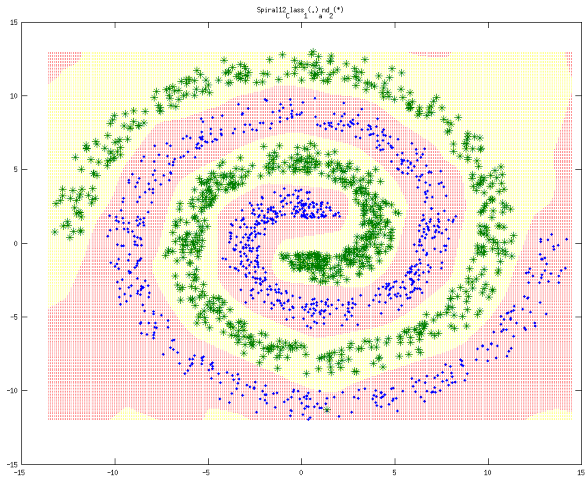
Correct : 641
Incorrect : 11
Accuracy : 98.31

		Predicted	
		Class 1	Class 2
Act.	Class 1	319	7
	Class 126	4	322

Figure 9: Decision region plot with the training data.

3.3.4 12 clusters components

12 clusters for each class gives 16 gaussian distribution. The intersection of gaussian distribution for those clusters gives out the decision boundaries. Interesting thing to note is 100 percent accuracy for this case.



Correct : 652
Incorrect : 0
Accuracy : 100.00

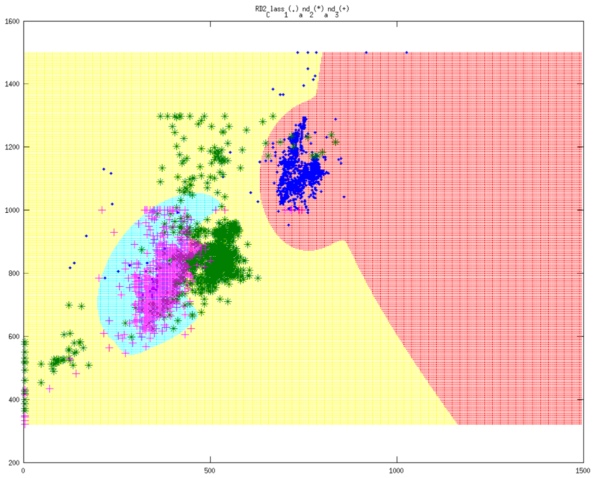
		Predicted	
		Class 1	Class 2
Act.	Class 1	326	0
	Class 126	0	326

Figure 10: Decision region plot with the training data.

3.4 Real World Data Set

3.4.1 2 clusters components

We assume two cluster components for each class i.e. two gaussian distributions for a class. The green data points on bottom left corner are not classified correctly because the two distributions center in the region between the pink and blue class, and hence could not classify the green points at corner.



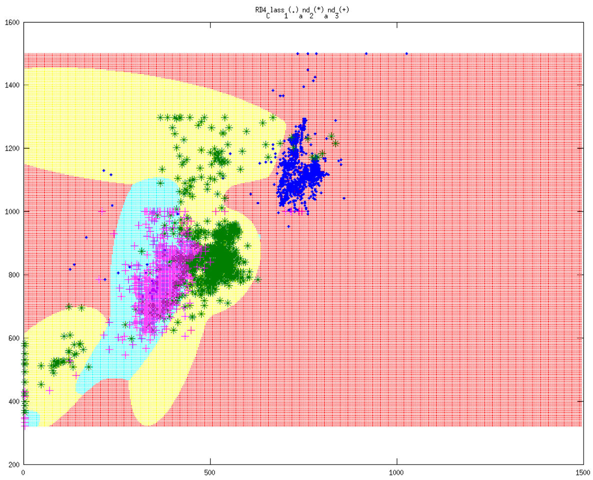
Correct : 1494
Incorrect : 284
Accuracy : 84.026

		Predicted		
		Class 1	Class 2	Class 3
Act.	Class 1	506	18	17
	Class 2	1	419	193
	Class 3	5	50	567

Figure 11: Decision region plot with the training data.

3.4.2 4 clusters components

The green points are classified correctly because of two more cluster peaks, which very well accommodate those. The region for blue points also expands as the one for yellow point class narrows. The blue class seems very well spread, probably because of large non diagonal covariance matrices.



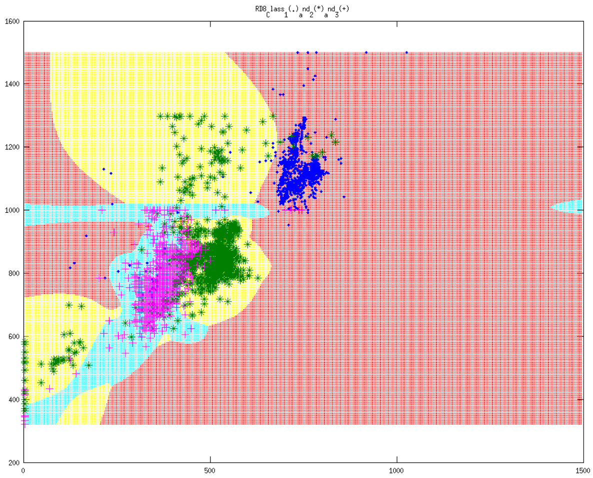
Correct : 1453
Incorrect : 325
Accuracy : 81.721

		Predicted		
		Class 1	Class 2	Class 3
Act.	Class 1	512	14	15
	Class 2	4	368	241
	Class 3	12	37	573

Figure 12: Decision region plot with the training data.

3.4.3 8 clusters components

Adding more clusters for each class makes the surfaces more precise and narrow. This is because of the increased probability for the data point to be in the respective clusters of the assigned class.



Correct : 1533
Incorrect : 245
Accuracy : 86.220

		Predicted		
		Class 1	Class 2	Class 3
Act.	Class 1	515	9	17
	Class 2	2	437	174
	Class 3	10	31	581

Figure 13: Decision region plot with the training data.

3.5 Image Scene Data Set

3.5.1 8 clusters components

Correct : 213
Incorrect : 66
Accuracy : 76.344

		Predicted		
		Class 1	Class 2	Class 3
Act.	Class 1	68	10	4
	Class 2	3	77	14
	Class 3	12	23	63

3.5.2 32 clusters components

Correct : 224
Incorrect : 55
Accuracy : 80.286

		Predicted		
		Class 1	Class 2	Class 3
Act.	Class 1	64	12	6
	Class 2	2	83	9
	Class 3	6	20	77

3.5.3 64 clusters components

Correct : 234
Incorrect : 45
Accuracy : 83.87

		Predicted		
		Class 1	Class 2	Class 3
Act.	Class 1	65	10	7
	Class 2	2	85	7
	Class 3	3	16	84

4 Conclusion

As per the observations, we can make the following conclusions :

1. The Decision Boundaries are more accurate when we model the data as a mixture of multiple gaussian as compared to the a single gaussian.
2. Different results are obtained when number of clusters k are varied.
3. The Decision Boundaries appears to be piece wise combination of quadratic boundaries as in case of unimodal gaussian case.
4. Although the accuracy tends to increase with the number of clusters assumed for a class, but due to overlapping data, the over clustering may cause the class to cover non-belonging points as well, this is evident in the real world data 4 cluster and 2 cluster case.