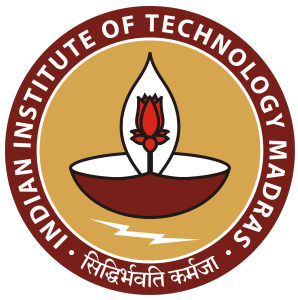
ASSIGNMENT 2

CS5691 Pattern Recognition and Machine Learning

CS5691 Assignment 2

Team Members:	
BE17B007	N Sowmya Manojna
PH17B010	Thakkar Riya Anandbhai
PH17B011	Chaithanya Krishna Moorthy

Indian Institute of Technology, Madras



Contents

1 Dataset 1A

1.1 K-nearest Neighbors Classifier

1.2 Naive-Bayes classifier

1.2.1 Same Covariance Matrix ($\sigma^2 I$)

1.2.2 Same Covariance Matrix (C)

1.2.3 Different Covariance Matrix

2 Dataset 1B

2.1 K-nearest Neighbors Classifier

2.2 Bayes Classifier, GMM, full covariance

2.2.1 Equations

Th initialization is done as follows for each class:

- Cluster initialization is using `kmeans` clustering.
- The relative number of points in each cluster N_q and weightage w_q for each cluster is calculated.
- The responsibility $\gamma_{n,q}$ is then calculated, followed by mean μ_q and covariance C_q is calculated.

The parameters are then updated sequentially through the:

- Expectation-step: $\gamma_{n,q}$ is updated.
- Maximization-step: μ_q, C_q, N_q and w_q are updated.

The stopping criterion used is $\Delta(\text{likelihood}) < \texttt{tol}$. The `tol` we considered is 10^{-5}.

Based on the accuracies obtained on the training, validation and test dataset, the best q_i for the three classes has been chosen as 5. The accuracies obtained in tabular format is as follows:

Number of Clusters/Class (Q)	Train Accuracy	Validation Accuracy	Test Accuracy
2	0.968333	0.952381	0.925926
3	0.983333	0.968254	1.000000
4	0.996667	0.984127	1.000000
5	0.998333	1.000000	1.000000
6	0.996667	1.000000	1.000000
7	0.998333	1.000000	1.000000
8	0.996667	1.000000	1.000000
9	0.996667	1.000000	1.000000

Table 1: Variation of accuracy across hyperparameter values on the training, validation and test set using the GMM model with full covariance matrix on Dataset 1B

2.2.2 Training and Validation Accuracy

The training and validation accuracies obtained for varying q_i for each class is as follows:

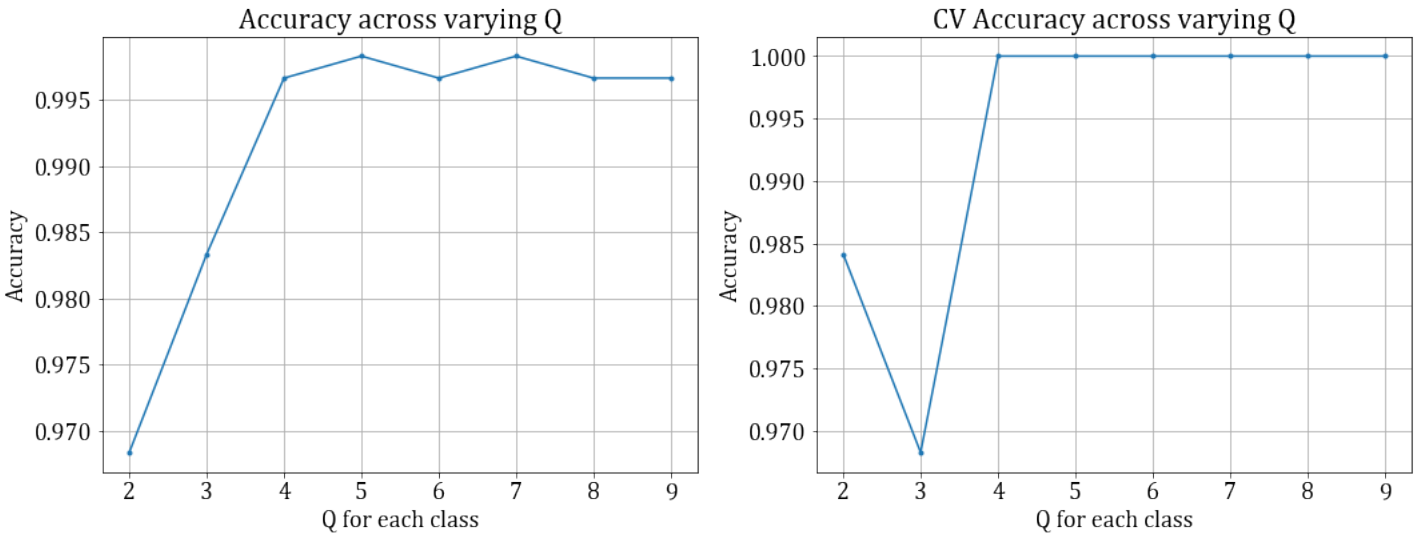


Figure 1: Training and Validation accuracy across q_i, on the left and right respectively

2.2.3 Testing Accuracy

The testing accuracy obtained for varying q_i for each class is as follows:

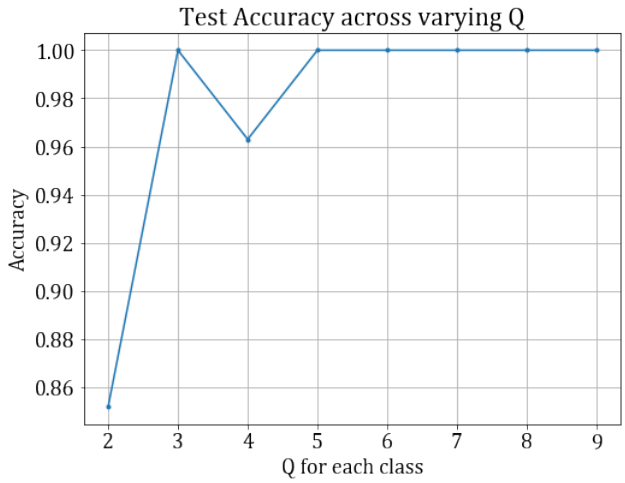


Figure 2: Testing accuracy across q_i

2.2.4 Contour Maps and Decision Surfaces

The contour maps and decision surfaces obtained, with $q_i = 5$ are as follows:

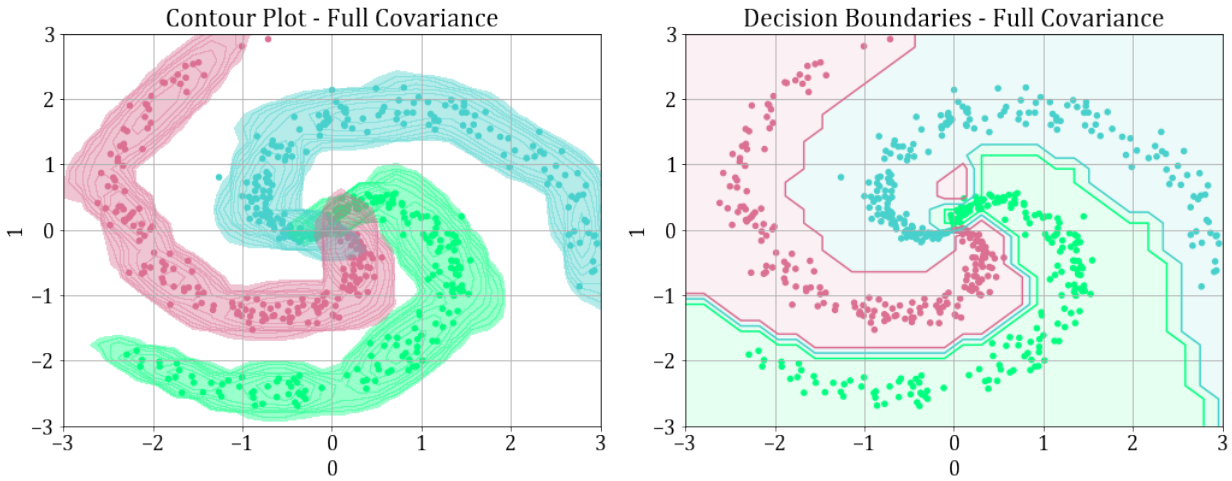


Figure 3: Contour Maps, Decision Surfaces obtained for $q_i = 5$, on the left and right respectively.

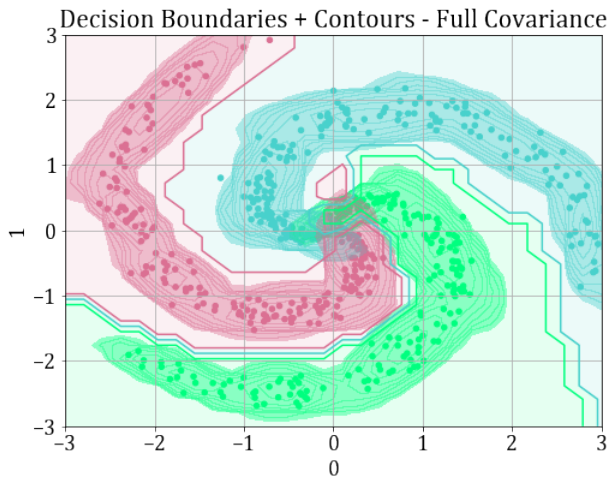


Figure 4: Overlap plot of the decision surface and contours.

2.3 Bayes Classifier, GMM, diagonal covariance

2.3.1 Training and Validation accuracy

Gaussian multi-modal training function (with threshold of the increment in total log-likelihood functions as 0.01 since there was no significant improvement for a smaller threshold than this) with diagonal covariance matrix over the hyperparameter values of the number of gaussian components Q = 2,3,4,5,6,7,8,9 to estimate the parameters - μ_q, C_q, N_q and w_q for each gaussian component - and predict the classes of the training data (train.csv) and cross-validation (70% of dev.csv), we get the table 2

Hyperparameter Value (Q)	Accuracy on CV data	Accuracy on Training data
2	0.873	0.9166
3	0.920	0.976
4	0.968	0.9966
5	0.984	1.0
6	0.984	0.986
7	0.984	0.991
8	0.984	0.9916
9	0.984	0.9916

Table 2: Variation of Accuracy across Hyperparameter values on the validation data using the GMM model with diagonal covariance matrix on Dataset 1B

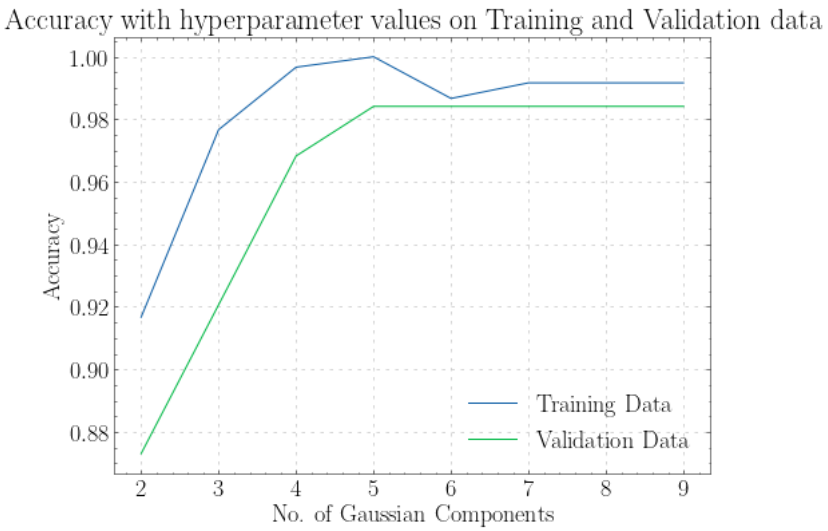


Figure 5: Plot of Hyperparameter value Vs Accuracy using the GMM model with diagonal covariance matrix on Dataset 1B

2.3.2 Best model output

As we can see in the tables and figure **??**, the best accuracy is when the number of Gaussian components is 5. Using the parameters of the model for 5 gaussian components and predicting for the test dataset (30% of dev.csv), the accuracy obtained was **1.0**.
The confusion matrices for the training and test datasets using the best model are tables 4 and **??**.

	0	1	2
0	200	0	0
1	0	200	0
2	0	0	200

Table 3: Confusion Matrix for training data 1B

The decision region plot for the best model is figure 6

	0	1	2
0	9	0	0
1	0	10	0
2	0	0	8

Table 4: Confusion Matrix for test data 1B

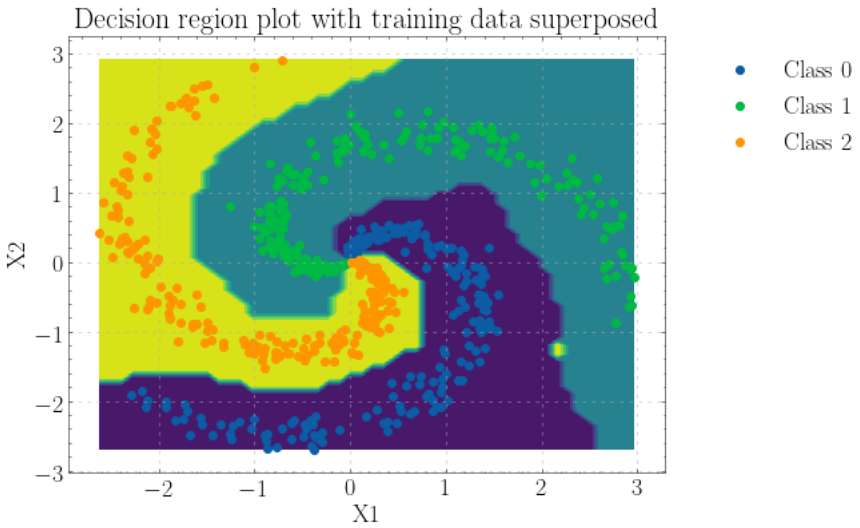


Figure 6: Decision region plot for Bayesian GMM model using diagonal covariance matrix and 5 gaussian components on dataset 1B

2.4 Bayes Classifier, KNN

	Hyperparameter Value	Accuracy for training data	Accuracy for validation data
0	2	0.509	0.350
1	3	0.525	0.404
2	4	0.574	0.436
3	5	0.627	0.420
4	6	0.6491	0.418
5	7	0.663	0.440
6	8	0.689	0.371
7	9	0.692	0.413
8	10	0.7184	0.4272
9	11	0.735	0.396
10	12	0.754	0.393
11	13	0.770	0.434
12	14	0.783	0.388

Table 5: Table of Hyperparameter value Vs Accuracy for the Validation data using the GMM model with diagonal covariance matrix on Dataset 2A

Figure 7: accuracy for training and validation set for 2A

3 Dataset 2A

3.1 Bayes Classifier, GMM, full covariance

3.2 Bayes Classifier, GMM, diagonal covariance

3.2.1 Training and Validation Accuracy

The accuracy obtained on training the data 2A on GMM model with diagonal covariance matrix is as in table 5. The plot of the same is in figure 7. The tolerence used was 1e-3.

3.3 Best model on test data

The highest accuracy on validation data set is for 7 gaussian components. Applying this model to predict the test data, we get an accuracy of **0.37**. The confusion matrices for this model on training and test data are tables 7 and **??**.

	0	1	2	3	4
0	147	8	21	22	35
1	7	156	7	13	14
2	39	5	165	30	45
3	25	7	20	190	24
4	33	6	36	32	143

Table 6: Confusion Matrix for training data 2A

	0	1	2	3	4
0	37	4	11	12	7
1	5	25	6	3	4
2	12	10	27	19	20
3	9	8	12	40	21
4	10	5	15	8	23

Table 7: Confusion Matrix for test data 2A

4 Dataset 2B

4.1 Bayes Classifier, GMM, full covariance

4.2 Bayes Classifier, GMM, diagonal covariance