Indian Institute Of Technology Madras

CS5691 - Pattern Recognition and Machine Learning

Assignment 2 Report

Team 5

Jeswant Krishna K - AE17B030

Advait Walsangikar - AE17B101

Ananthu Nair S. - AE17B109

22th April 2021

Contents

1 Dataset 1a

1.1 Problem Definition

A 2-Dimensional linearly separable Artificial Dataset that is to be used for pattern classification using K-nearest Neighbours and Naive-Bayes Classifier with a Gaussian Distribution for each of the class.

1.2 Method Used

K Nearest Neighbours

We used K Nearest Neighbors for classification of the data points. We label a data point with the class that has the maximum number of representatives in the point's K nearest neighbors. The 'K' is a variable and values such as 1, 7 and 15 were tried.

Naive Bayes Classifier

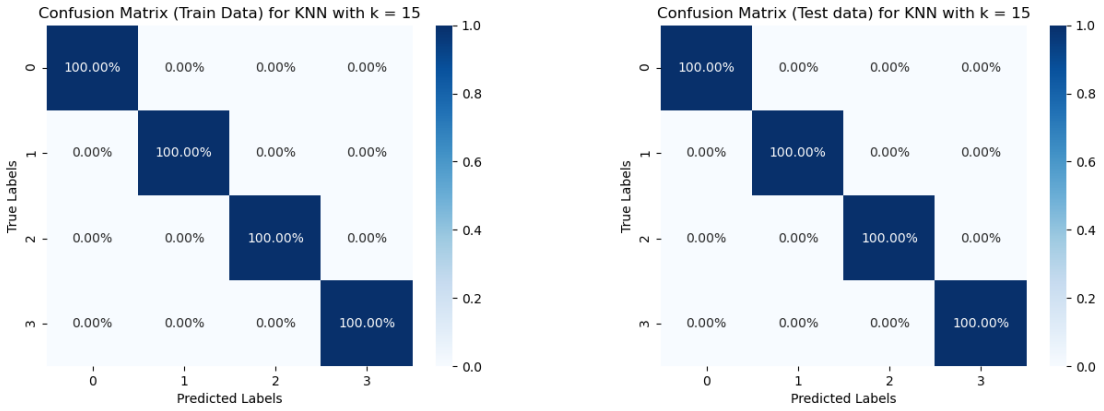
Naive Bayes classifier was used in which the model was trained using the assumption of independence among the features. A Gaussian model was build for each of the classes present in the data-set independently which is the naive assumption. Based on the probability output from the Gaussian models based on input data the output class with max probability is considered which is the output of the model. Also based on it the decision region was also built.

1.3 Results

KNN Classifier

K - value	**Accuracies**	
	Training Data	Test Data
1	100%	100%
7	100%	100%
15	100%	100%

Table 1: Table showing the Training Data and test Data accuracies for the KNN Classifier with various values of K



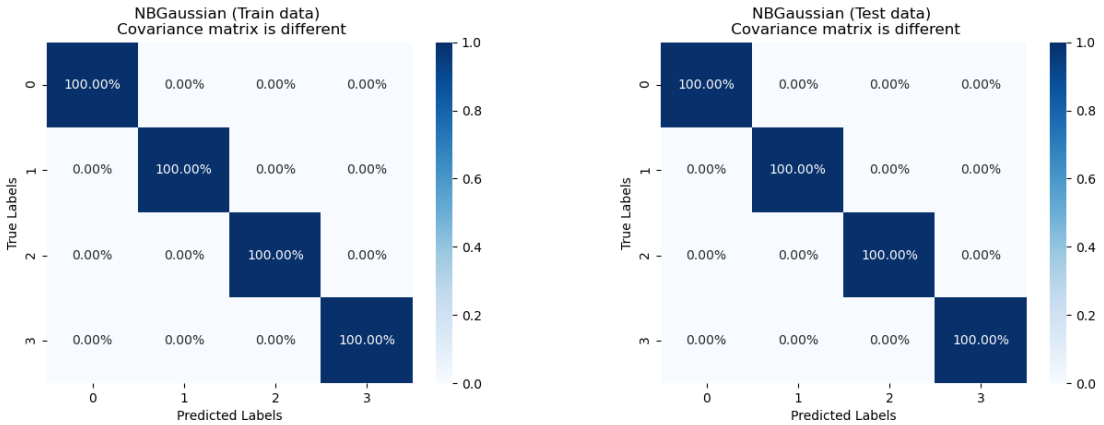
(a) Training Data (b) Test Data

Figure 1: Confusion matrix for KNN Classifier with the best value of k, for both the training data and the test data. Here, since all the matrices gave 100% accuracy, the k=15 was chosen.

Naive-bayes Classifier

Covariance Matrix	**Accuracies**	
	Training Data	Test Data
Same and is $\sigma^2 I$	100%	100%
Same and is C	100%	100%
Different	100%	100%

Table 2: Table showing the Training Data and test Data accuracies for the Naive Bayes Classifier using Gaussian distribution for each class



(a) Training Data (b) Test Data

Figure 2: Confusion matrix for Naive Bayes Classifier with the best configuration of covariance matrix, for both the training data and the test data. Here, since all the matrices gave 100% accuracy, the case with all different covariance matrix was chosen.

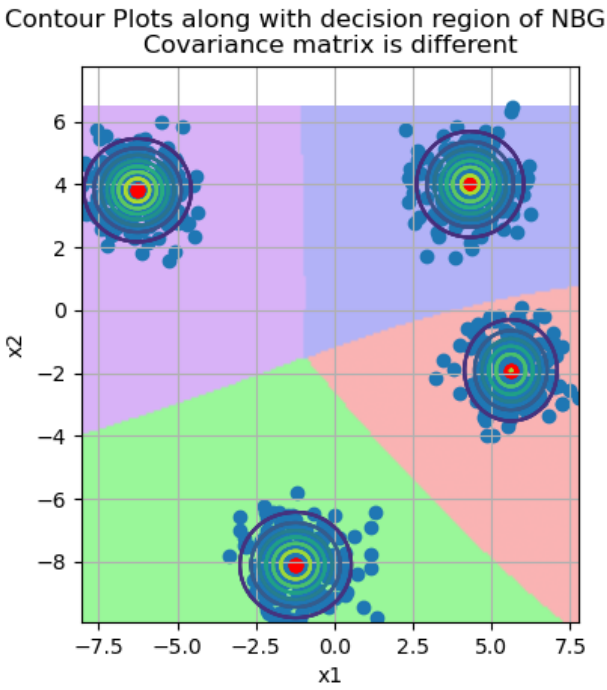


Figure 3: Decision Region plot for the best configuration of covariance matrix. Here, since the accuracies were the same for all the cases the case with all different covariance matrix was chosen.

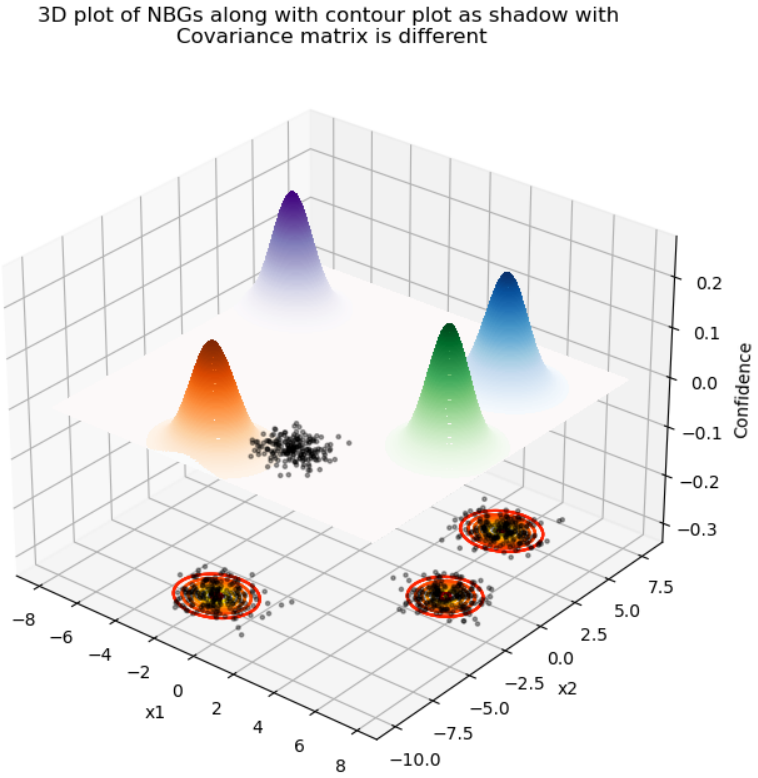


Figure 4: A 3D Plot of the Naive Bayes Classifier with Gaussian Distribution for each class, superimposed on the same graph. (with shadow)

1.4 Inferences

K Nearest Neighbour

- The KNN Classifier predicts the class label correctly for all the values of K that is specified.
- However, when the K becomes very large, then the accuracy starts to fall.

Naive Bayes Classifier

- We observed different shapes of contours for different covariance matrices.
- Diagonal Covariance Matrix gives elliptical contours with axes parallel to principle axes of the plot.
- Full Covariance Matrix gives elliptical with tilted axis with respect.to the principle axes.
- We observed 100% accuracy with respect to both ways of modelling the class Gaussians.
- The decision surfaces b/w two classes is a complex surface/hypersurface.

2 Dataset 1b

2.1 Problem Definition

A 2-Dimensional non-linearly separable Artificial Dataset that is to be used for pattern classification using K-nearest Neighbours and Bayes Classifier with a Gaussian Mixture Models for each class.

2.2 Method Used

KNN Classifier

We used K Nearest Neighbors for classification of the data points. We label a data point with the class that has the maximum number of representatives in the point's K nearest neighbors. The 'K' is a variable and values such as 1, 7 and 15 were tried.

Bayes Classifier with GMM

We use Gaussian Mixture Models to model the classes and then classify the given point based on the class having the highest posterior probability, obtained using the GMM.

Bayes Classifier with KNN

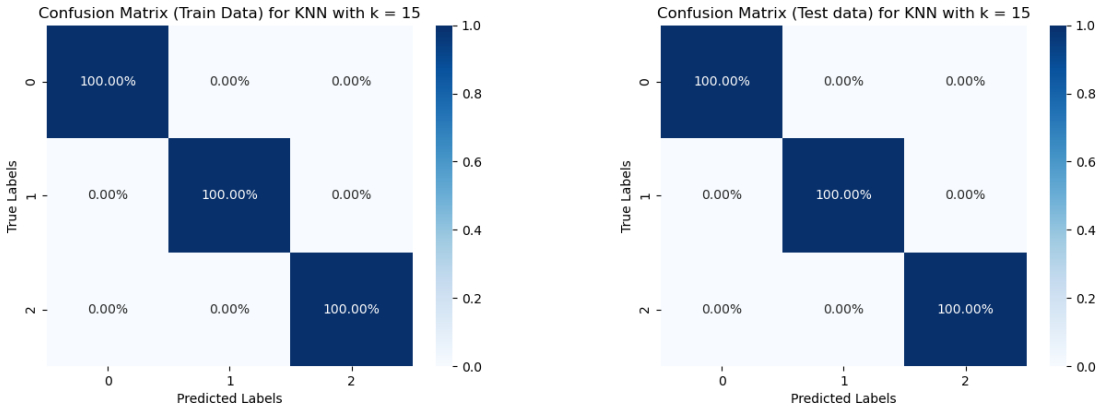
We use Gaussian Mixture Models to model the classes and then classify the given point based on the class having the highest posterior probability, obtained using the GMM.

2.3 Results

KNN Classifier

K - value	**Accuracies**	
	Training Data	Test Data
1	100%	100%
7	100%	100%
15	100%	100%

Table 3: Table showing the Training Data and test Data accuracies for the KNN Classifier with various values of K



(a) Training Data

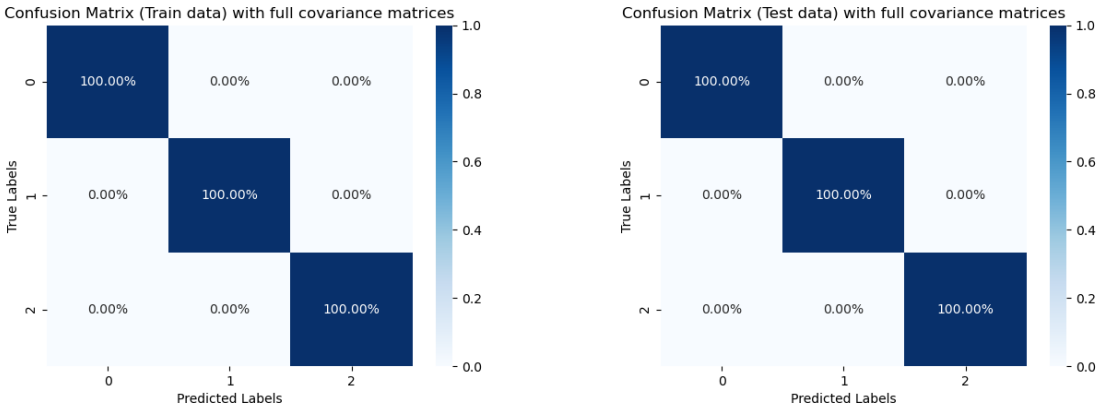
(b) Test Data

Figure 5: Confusion matrix for KNN Classifier with the best value of k, for both the training data and the test data. Here, since all the matrices gave 100% accuracy, the k=15 was chosen.

Bayes Classifier with GMM (Full Covariance Matrix)

Q	Accuracies	
	Training Data	Test Data
1	100%	100%
2	100%	100%
4	100%	100%
8	100%	100%
10	100%	100%

Table 4: Table showing the Training Data and test Data accuracies for the Bayes Classifier using Gaussian Mixture Model for each class



(a) Training Data

(b) Test Data

Figure 6: Confusion matrix for Bayes Classifier with GMM with the best value of Q, for both the training data and the test data. Here, since all the matrices gave 100% accuracy, the case of Q = 4 was chosen.

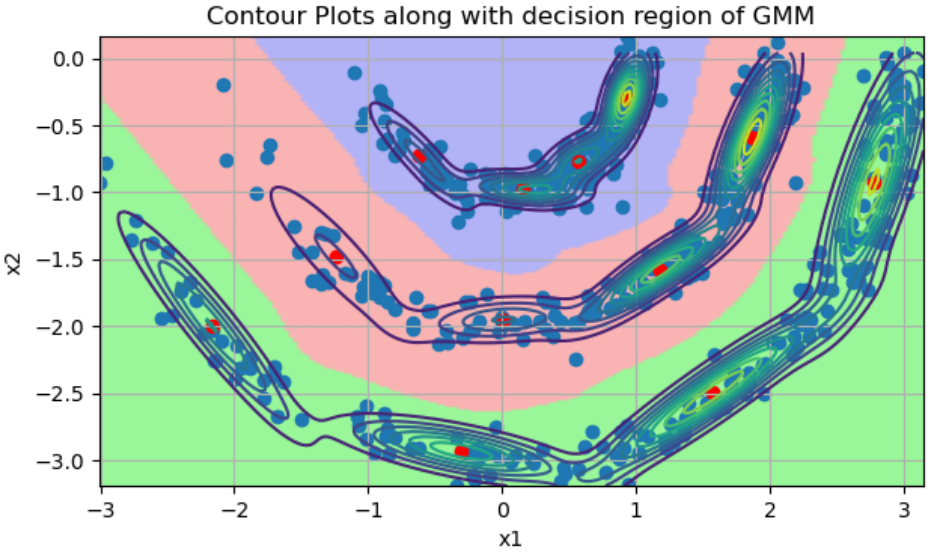


Figure 7: Decision Region plot for the the Bayesian Classifier with Gaussian Mixture Models with Q = 4 and a full covvariance matrix.

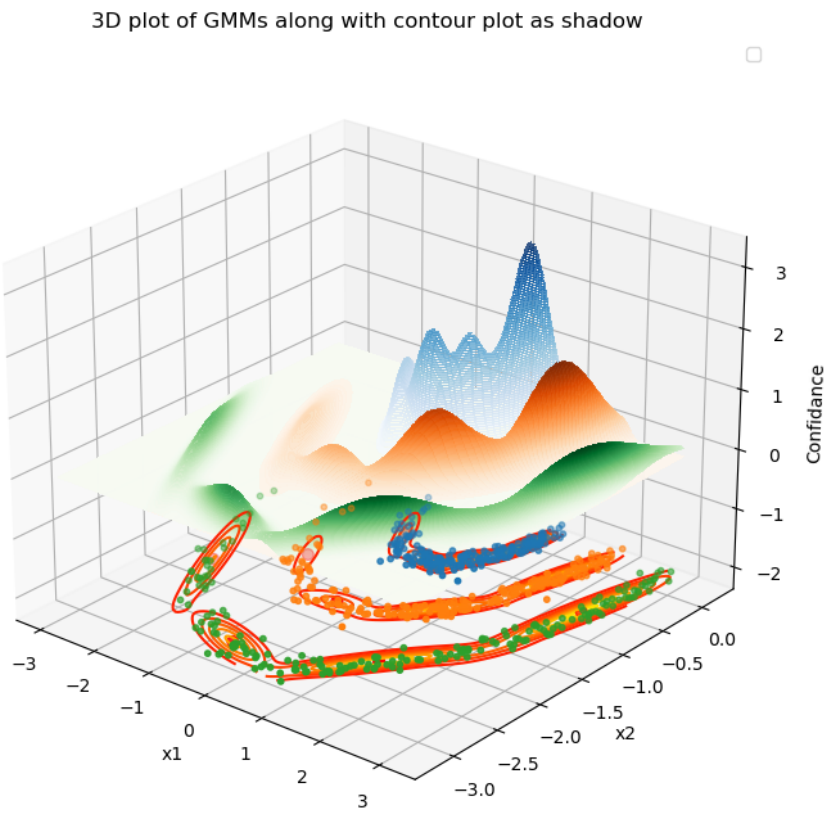
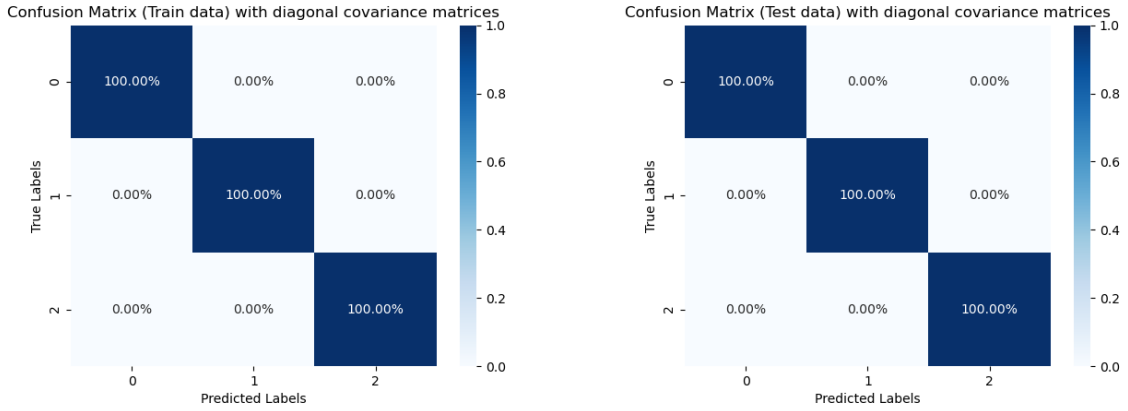


Figure 8: A 3D Plot of the Bayes Classifier with Gaussian Mixture Models for each class with Q = 4 and a full covariance matrix, each superimposed on the same graph. (with shadow)

Bayes Classifier with GMM (Diagonal Covariance Matrix)

Q	Accuracies	
	Training Data	Test Data
1	100%	100%
2	100%	100%
4	100%	100%
8	100%	100%
10	100%	100%

Table 5: Table showing the Training Data and test Data accuracies for the Bayes Classifier using Gaussian Mixture Model for each class



(a) Training Data

(b) Test Data

Figure 9: Confusion matrix for Bayes Classifier with GMM with the best value of Q, for both the training data and the test data. Here, since all the matrices gave 100% accuracy, the case of Q = 10 was chosen.

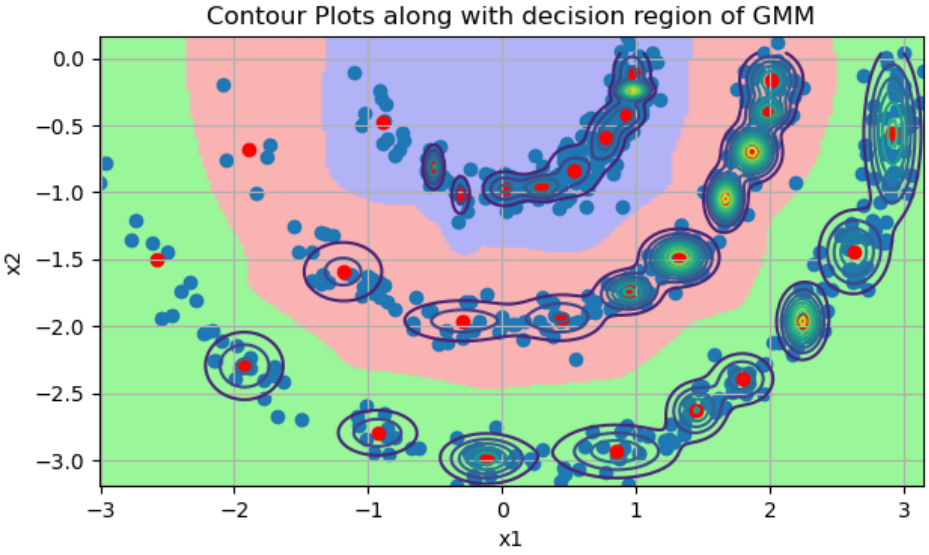


Figure 10: Decision Region plot for the the Bayesian Classifier with Gaussian Mixture Models with Q = 10 and a diagonal covariance matrix.

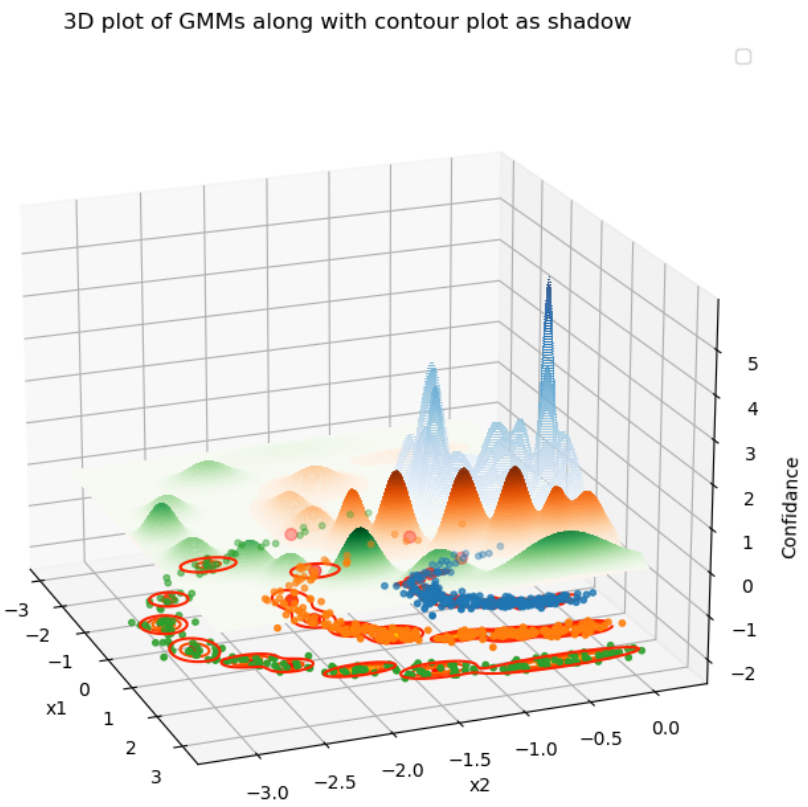
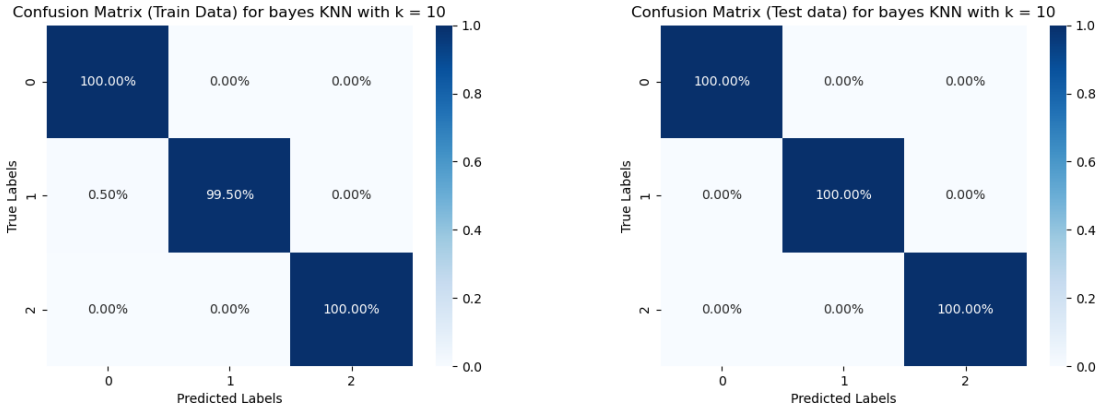


Figure 11: A 3D Plot of the Bayes Classifier with Gaussian Mixture Models for each class with G = 10 and a diagonal covariance matrix, each superimposed on the same graph. (with shadow)

Bayes Classifier with KNN method of density estimation

K - value	**Accuracies**	
	Training Data	Test Data
10	99.84%	100%
20	99.67%	100%

Table 6: Table showing the Training Data and test Data accuracies for the Bayes Classifier with KNN Method of class conditional probability density Estimation for various values of K



(a) Training Data

(b) Test Data

Figure 12: Confusion matrix for Bayes Classifier with KKN method of density estimation with the best value of K, for both the training data and the test data. Here, the case of K=10 was chosen since it had the best accuracy.

2.4 Inferences

- For KNN we notice 100% accuracy regardless of the value of K we chose. Since, k = 1 is very small, we expect k = 15 to generalize well to future test data points.
- Bayes Classifier using GMM for each class with full covariance matrices, we observed better representation for each class if we represent using 4 gaussians. The decision region is a complex boundary. The ellipses are tilted w.r.t. the data points.
- Bayes classsifier using GMM for each class with diagonal covariance matrices, we observed best representation for each class for 10 gaussians. The decision is again a complex boundary. Each of the gaussians this time are ellipses aligned with the axis of the data points.
- Bayes classifier with knn method for class density estimation using k = 10 gave the best value for train and test accuracies.

3 Dataset 2a

3.1 Problem Definition

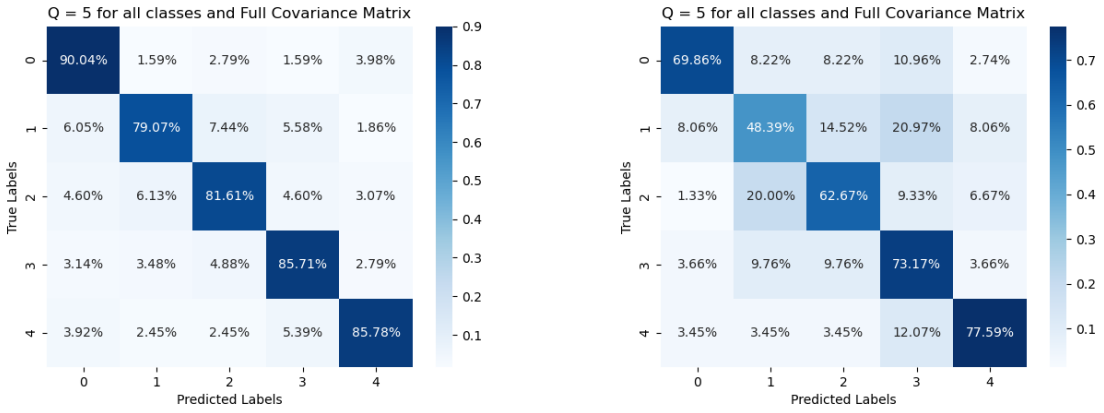
A Real-World Image Dataset that is to be used for static pattern classification using Bayes Classifier with Gaussian Mixture Models for each class.

3.2 Method Used

We use GMM to model the classes. We used PCA(Principle component analysis) to reduce the no. of features to 12 from 24 which allowed us to use 5 Gaussians in the GMM for each class representation.

3.3 Results

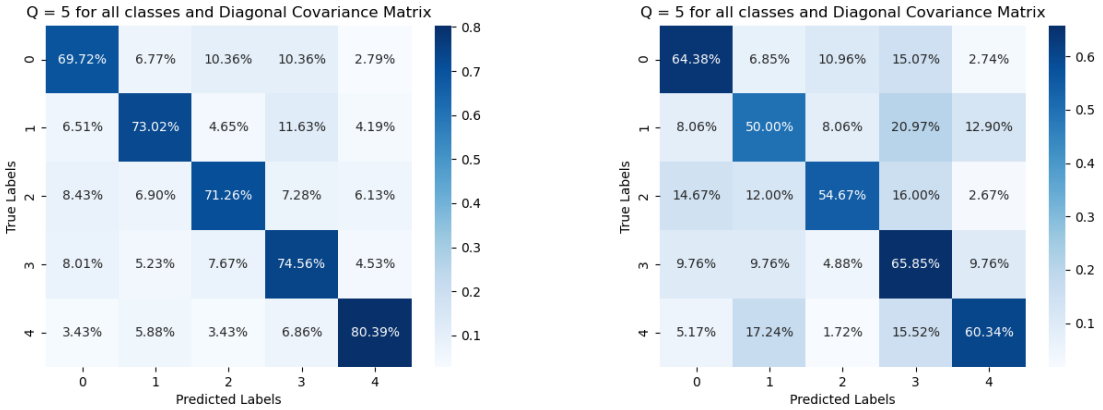
Full Covariance Matrix



(a) Training Data (Accuracy = 84.56 %) (b) Test Data (Accuracy = 66.57 %)

Figure 13: Confusion matrix for Bayes Classifier with GMM with the best value of Q, for both the training data and the test data with full covariance matrix. Here, Q = 5 was chosen.

Diagonal Covariance Matrix



(a) Training Data (Accuracy = 73.56 %) (b) Test Data (Accuracy = 59.43 %)

Figure 14: Confusion matrix for Bayes Classifier with GMM with the best value of Q, for both the training data and the test data with diagonal covariance matrix. Here, Q = 5 was chosen.

3.4 Inferences

- Bayes Classifier with GMM does not work for a high value of Q, (such as 4 and above) as the number of training examples are much less as compared to the unknowns. It works only after applying PCA to reduce the number of features.
- The accuracy increases with Q and is the highest in our case for Q = 5, as seen from the confusion matrix and the overall accuracy.
- The diagonal covariance matrix has a lower accuracy as compared to the filled covariance matrix but it runs faster than the case of filled matrix .
- We were not able to try more gaussians due to overflow error being encountered while learning the parameters.

4 Dataset 2b

4.1 Problem Definition

A Real-World Image Dataset that is to be used for varying length pattern classification using Bayes Classifier with Gaussian Mixture Models for each class.

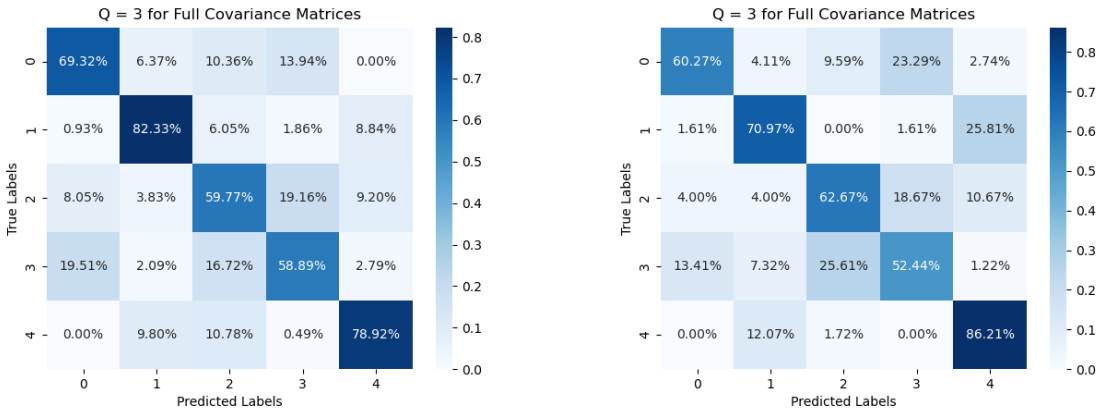
4.2 Method Used

We observe that each image is represented by multiple features of length 23. We fit a GMM on all such feature vectors for all images of that class and obtain the probability of an example as:

$$P(X|\lambda_c) = \prod_{i=1}^{I} \sum_{q=1}^{Q} \pi_q N(w_q, \sigma_q|\lambda_c)$$

4.3 Results

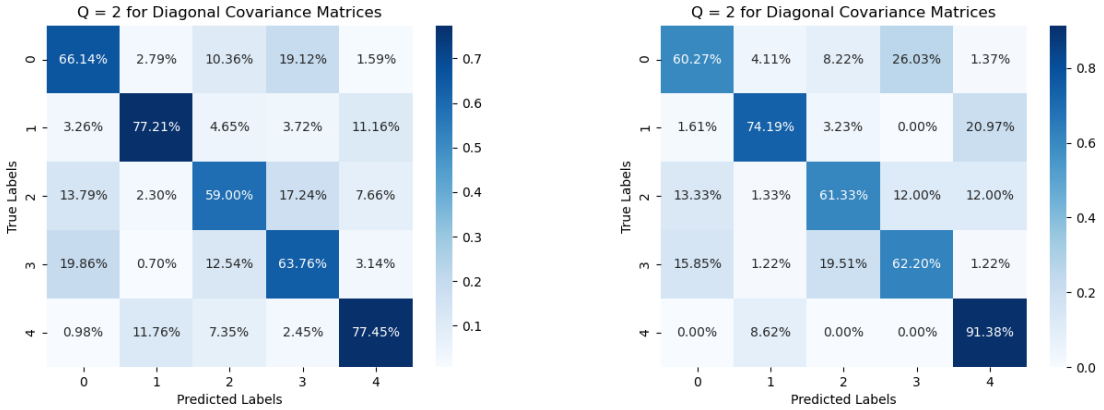
Full Covariance Matrix



(a) Training Data (Accuracy = 68.71 %) (b) Test Data (Accuracy = 65.14 %)

Figure 15: Confusion matrix for Bayes Classifier with GMM with the best value of Q, for both the training data and the test data with full covariance matrix. Here, Q = 2 was chosen.

Diagonal Covariance Matrix



(a) Training Data (Accuracy = 67.89%) (b) Test Data (Accuracy = 68.57%)

Figure 16: Confusion matrix for Bayes Classifier with GMM with the best value of Q, for both the training data and the test data with diagonal covariance matrix. Here, Q = 2 was chosen.

4.4 Inferences

We observed around 67% accuracy for train as well as test data for both full and diagonal covariance matrices with 2 multivariate gaussians to represent each class. We were not able to try more gaussians due to overflow error being encountered while learning the parameters.