Pattern Recognition and Machine Learning

Assignment 2
Report

Group No. 19

Ranjith Tevnan	Sai Bandawar	Jay shah
EE18B146	EE18B150	EE18B158

22,April 2021

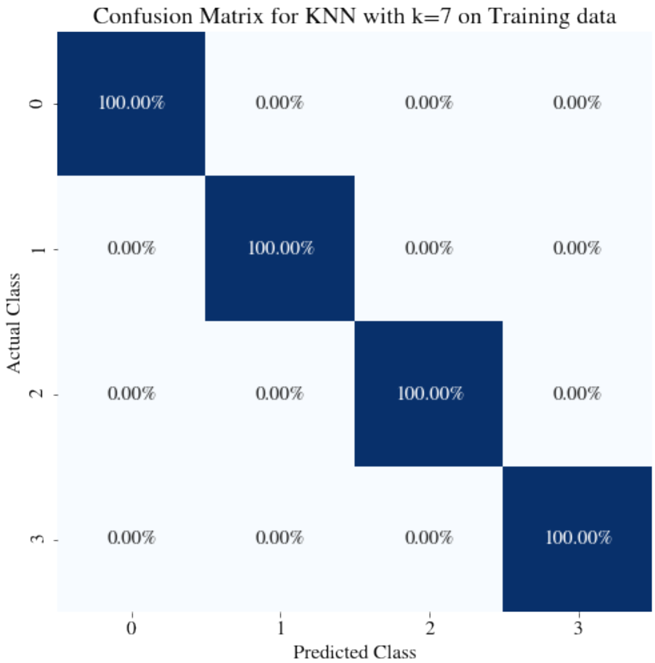
1 Dataset 1: 2-dimensional artificial data

1.(a) Linearly seperable data set for static pattern classification

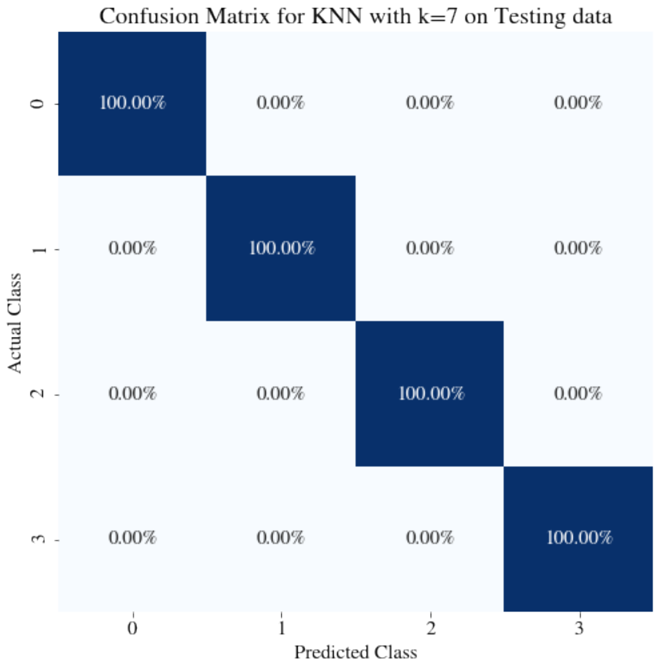
1.(a).1 K-nearest neighbours classifier, for K=1, K=7, and K=15

K	Training(%)	Validation(%)
1	100	100
7	100	100
15	100	100

Best Model:- The classification accuracy for the best model configuration(i.e. K=7) on test data is 100%.



(a) Confusion matrix for training dataset



(b) Confusion matrix for test dataset

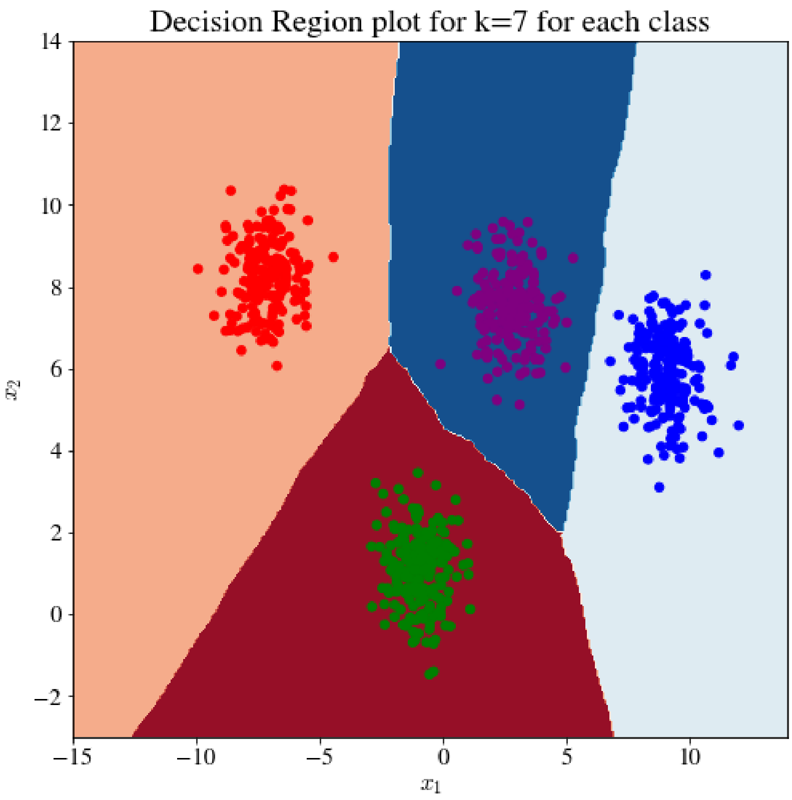


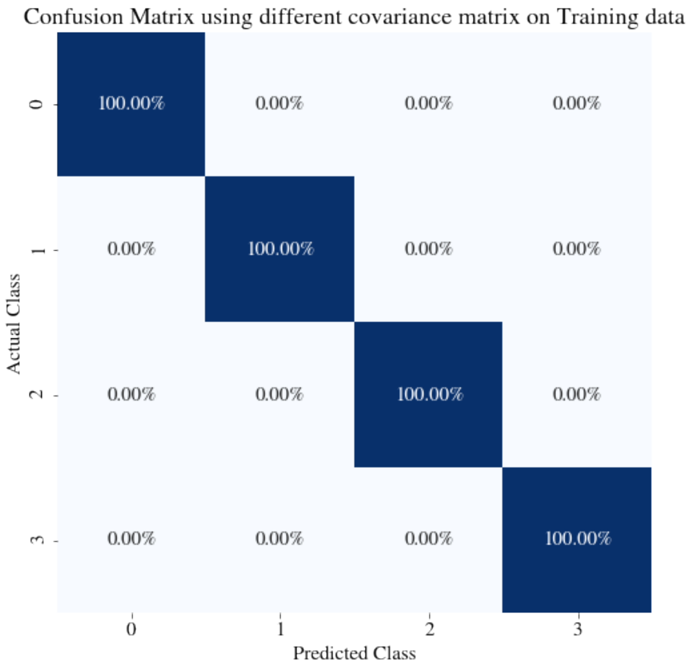
Figure 2: Decision region

Since The data points for different classes are well seperated , We can clearly see the Decision regions are also respectively clearly seperated. Due to well seperation of the data, Decision Region Plots for k=1 and k=15 also turn out to be approximately same which is indicated by the 100% validation data accuracy for all the values of k.

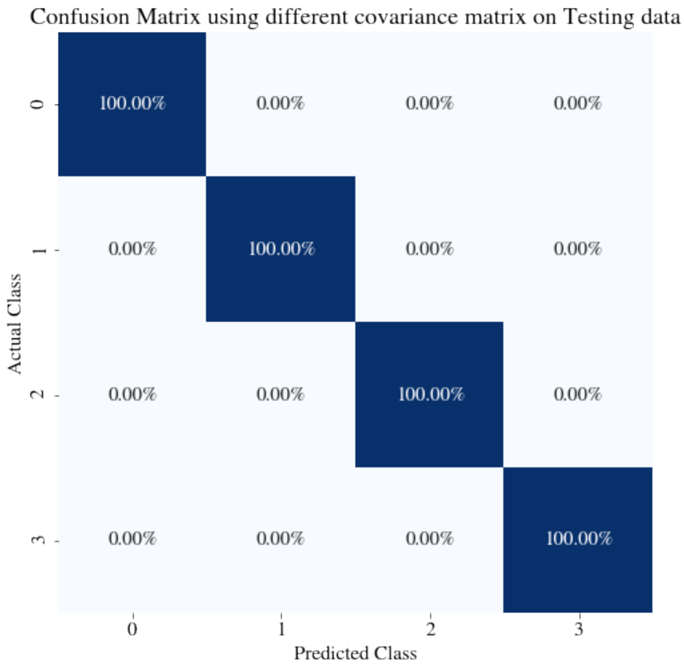
1.(a).2 Naive-Bayes classifier with a Gaussian distribution for each class

Covariance Matrix type	Training(%)	Validation(%)
$\sigma^2 I$	100	100
Same C	100	100
Different	100	100

Best Model:- The classification accuracy for the best model configuration(i.e. using different and full covariance matrices) on test data is 100%.



(a) Confusion matrix for training dataset



(b) Confusion matrix for test dataset

The Naive-Bayes classifier with Gaussian Distribution classifies the data very accurately in all of the different Covariance Matrix type hence resulting in 100% validation accuracy. The different Covariance Matrix type was chosen as the best model as it tends to give a more complex model.

Figure 4: Decision region

Background into the Naive-Bayes Classifier:

In the Naive Bayes Classifier we assume that the features are conditionally statistically independent.
And since the features are also assumed to have a univariate normal distribution individually, independence among them results in the joint class conditional pdf (which is the product of the marginals) a multivariate Gaussian distribution
This makes the covariance matrix of the multivariate Gaussianis diagonal.

There are a couple of results regarding the form the decision surface takes for a Naive- Bayes classifier:

- If the covariance matrices are equal across classes and have the form $\sigma^2 I$ then the decision surface is linear (a straight line in the 2-D case) and normal to the vector joining the centre of 2 classes.
  The level curves are circles having axes parallel to the co-ordinate axis
- If the covariance matrices are equal across classes and diagonal (they may have different diagonal elements) then the decision surface is still linear
  However, level curves are ellipses having axes parallel to the co-ordinate axis.
- If the the covariance matrices are different across classes, although diagonal then the decision boundary is a curve (in 2-D space).
  However, level curves are ellipses having axes parallel to the co-ordinate axis.
- The best model here uses different covariance matrices for each class so technically the decision boundary is a curve.

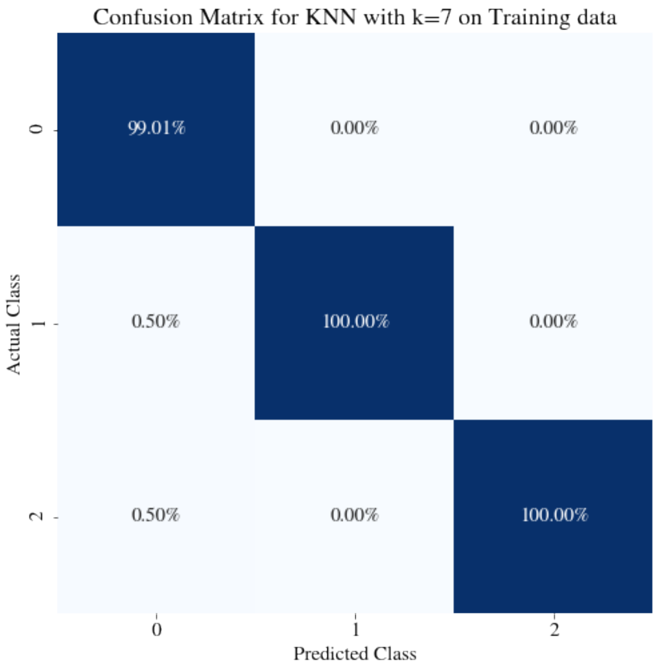
- However, the curve is almost indistinguishable from a linear decision surface- line.

1.(b) Nonlinearly separable data set for static pattern classification

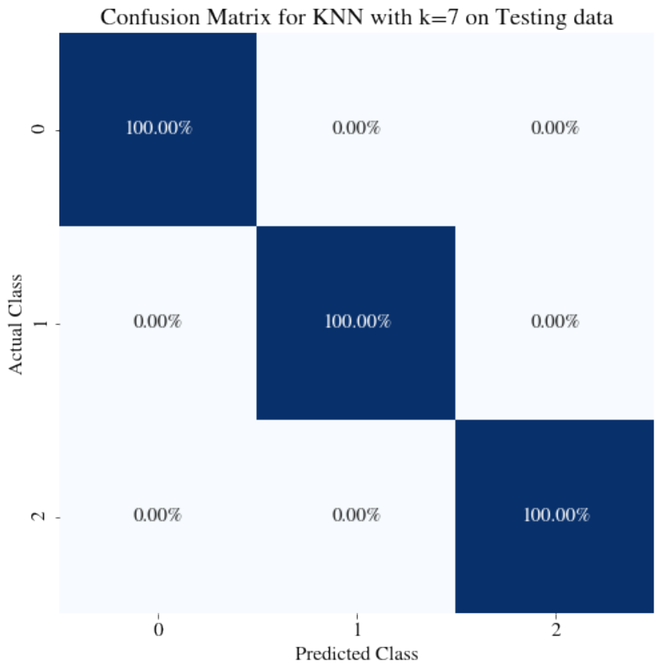
1.(b).1 K-nearest neighbours classifier, for K=1, K=7, K=15

K	Training(%)	Validation(%)
1	99.834	100
7	99.667	100
15	99.5	100

Best Model:- 100% is the classification accuracy for the best configuration of the model on test data with K=7



(a) Confusion matrix for training dataset



(b) Confusion matrix for test dataset

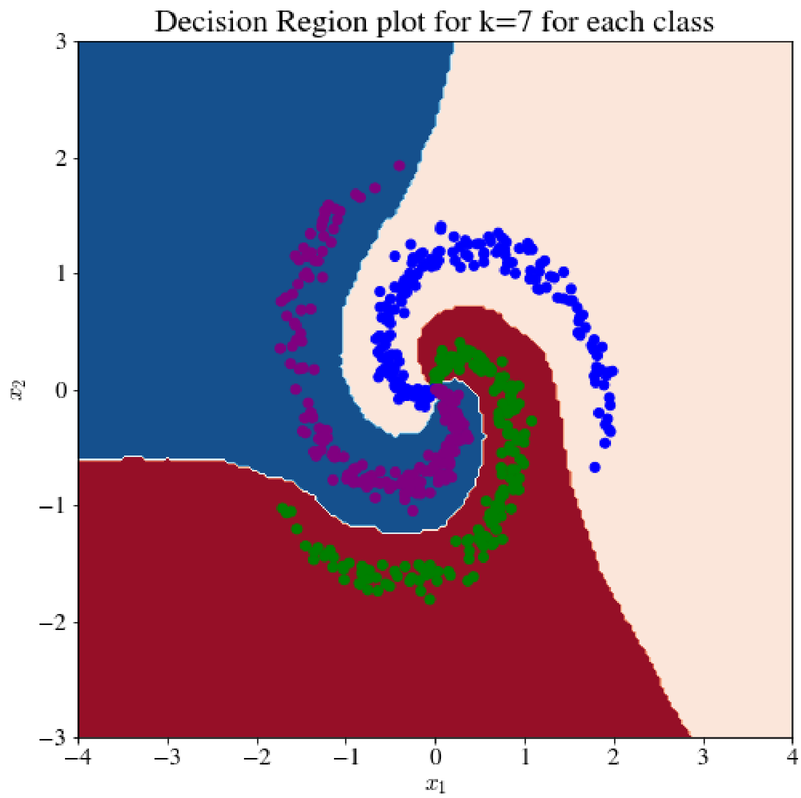


Figure 6: Decision region

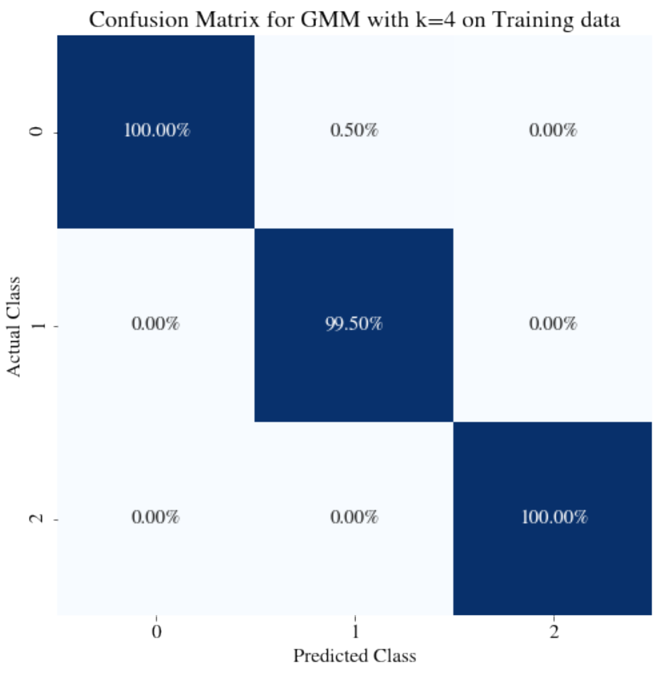
Inferences

- The decision boundary is clearly non-linear
- k=7 seems to give the best classification accuracy as it is neither too simple a model nor complex enough requiring many datapoints

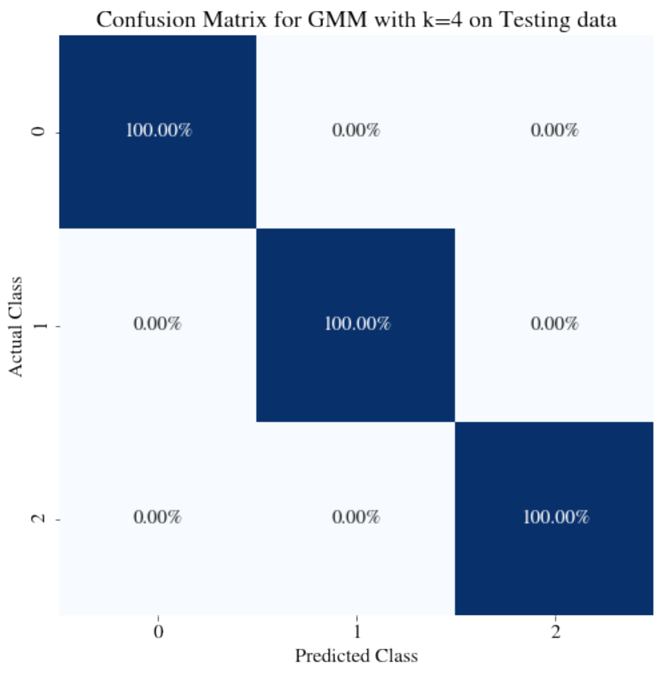
1.(b).2 Bayes Classifier with a GMM for each class , using full covariance matrices

Q	Training(%)	Validation(%)
2	98.5	100
3	99	100
4	99.834	100
5	99.667	100
10	99.167	97.778

Best Model:- 100% is the classification accuracy for the best model with Q=4



(a) Confusion matrix for training dataset



(b) Confusion matrix for test dataset

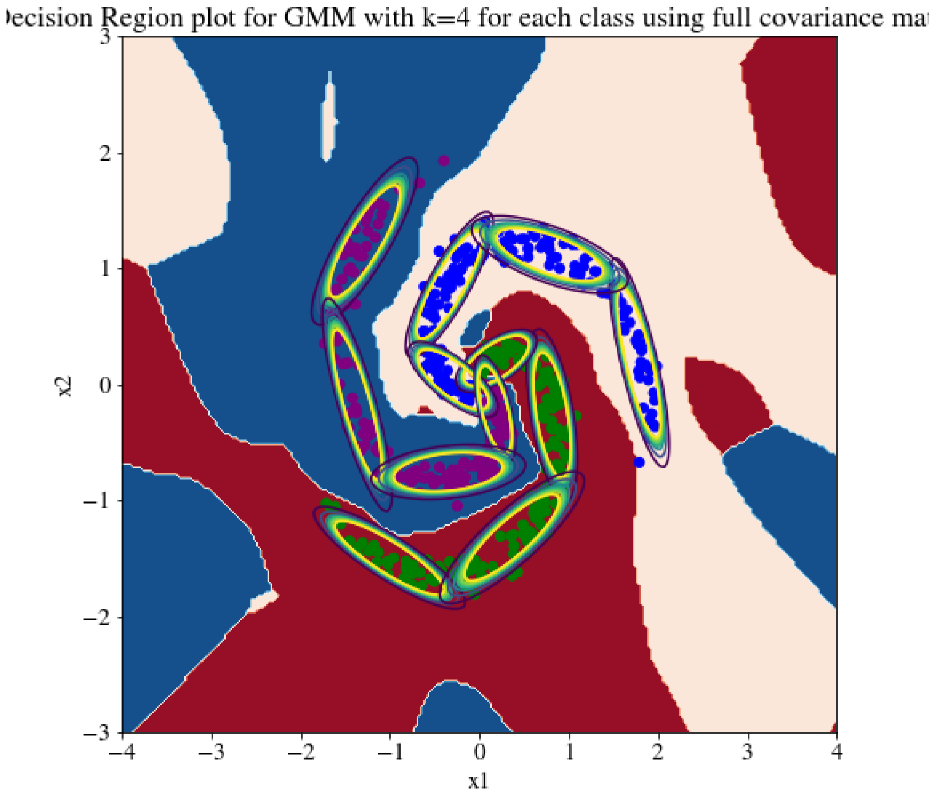


Figure 8: Decision region

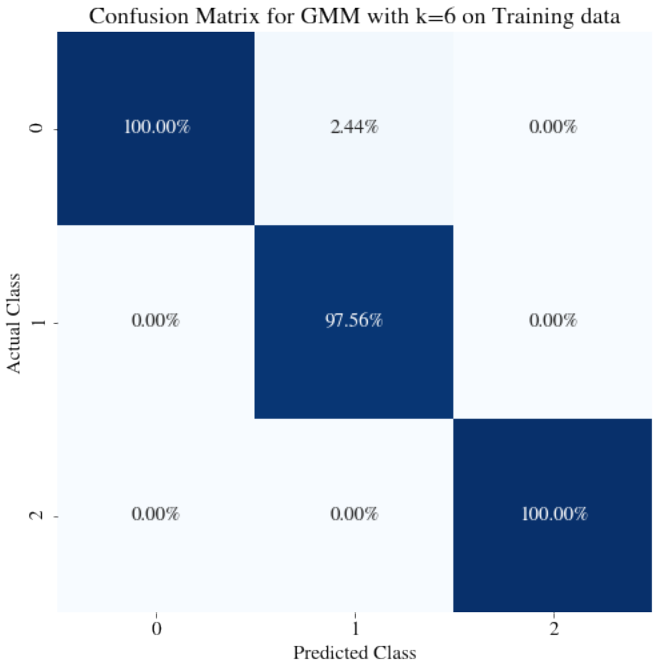
Inferences

- The decision boundary is clearly non-linear as the covariance matrices are different for each class and each Gaussian component within a class
- The level curves are ellipses with tilted axes
- If we keep the number of Gaussian components for each class as 4, it seems to give the best classification validation accuracy as covers the entire breadth of the data (as seen from the plot). Also the higher complexity models having more components have increasingly more parameters to be estimated and hence require more data points. q=4 gives the best trade off on terms of accuracy and complexity

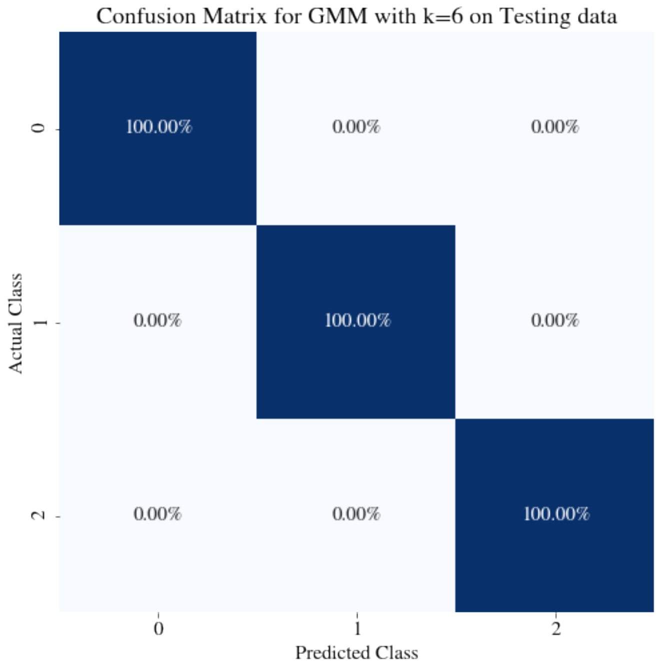
1.(b).3 Bayes Classifier with a GMM for each class , using diagonal covariance matrices

Q	Training(%)	Validation(%)
2	92.334	95.556
3	98.834	100
4	98.0	95.556
5	98.0	95.556
6	99.167	100

Best Model:- 100% is the classification accuracy for the



(a) Confusion matrix for training dataset



(b) Confusion matrix for test dataset

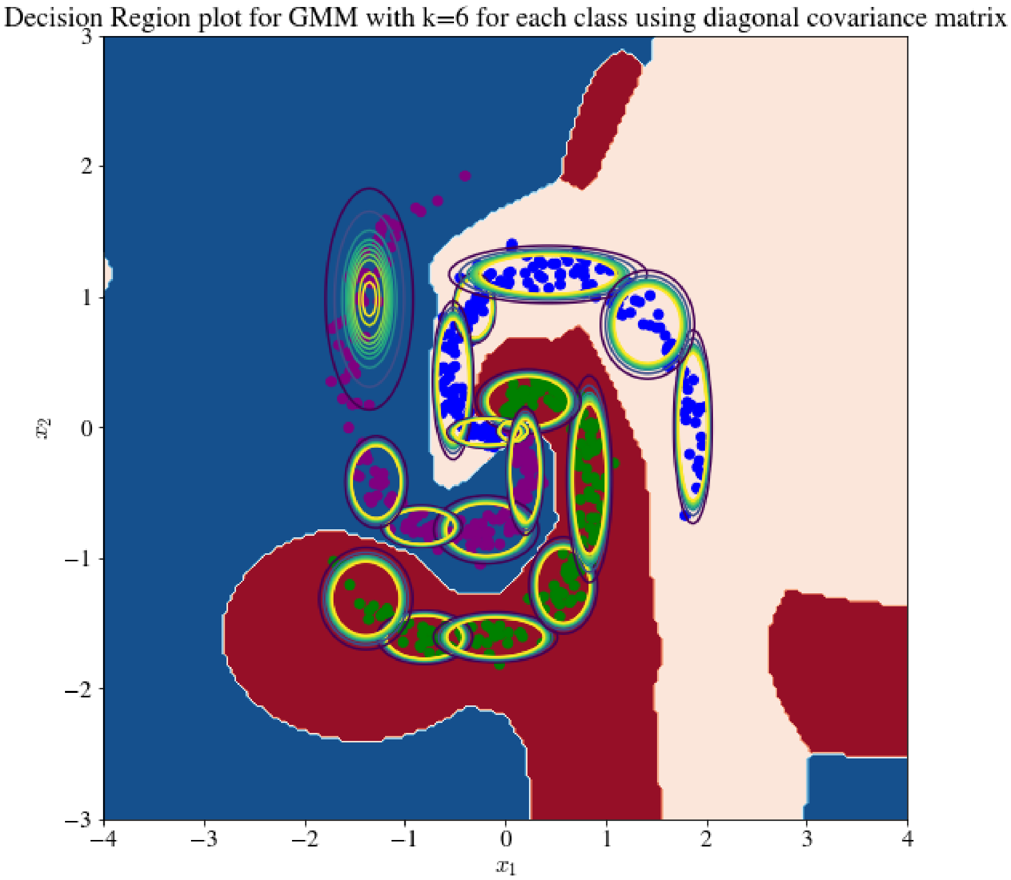


Figure 10: Decision region

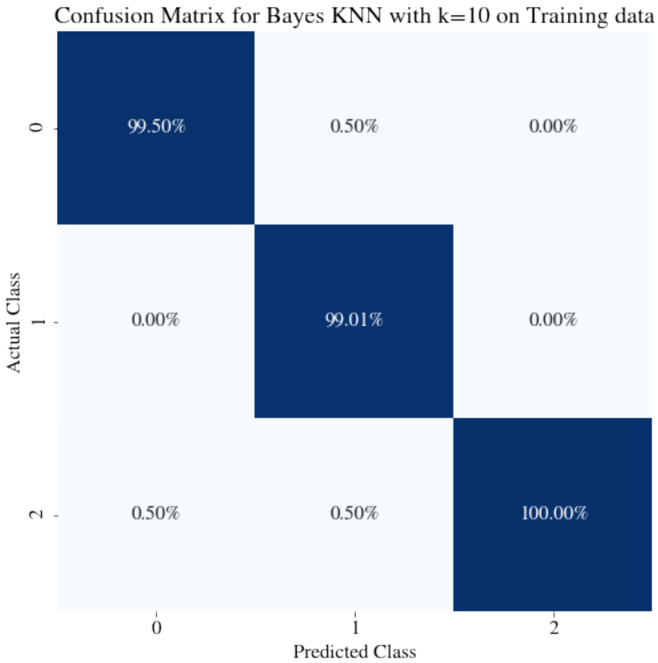
Inferences

- The decision boundary is clearly non-linear as the covariance matrices are different for each class and each Gaussian component within a class
- The level curves are ellipses with axis parallel to the co-ordinate axis since the covariance matrix is diagonal
- If we keep the number of Gaussian components for each class as 6, it seems to give the best classification validation accuracy as covers the entire breadth of the data (as seen from the plot). Also the higher complexity models having more components have increasingly more parameters to be estimated and hence require more data points. q=6 gives the best trade off on terms of accuracy and complexity
- It can be observed that the components required for the best model for diagonal cov. matrix (6) is greater then full cov. matrix(4) as the expected due to the restriction in the form of the level contours

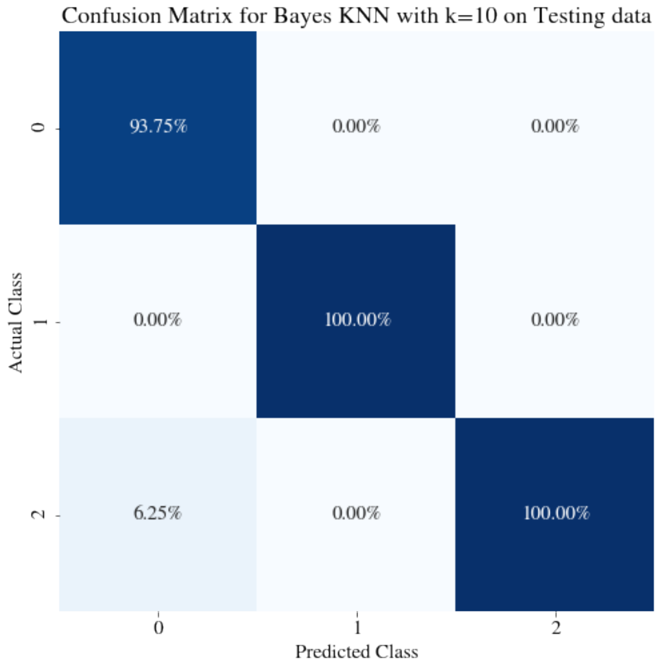
1.(b).4 Bayes Classifier with K-nearest neighbours method for estimation of class-conditional probability density function, for K=10 and K=20

K	Training(%)	Validation(%)
10	99.5	100
20	98.0	100

Best Model:- 100% is the classification accuracy for the



(a) Confusion matrix for training dataset



(b) Confusion matrix for test dataset

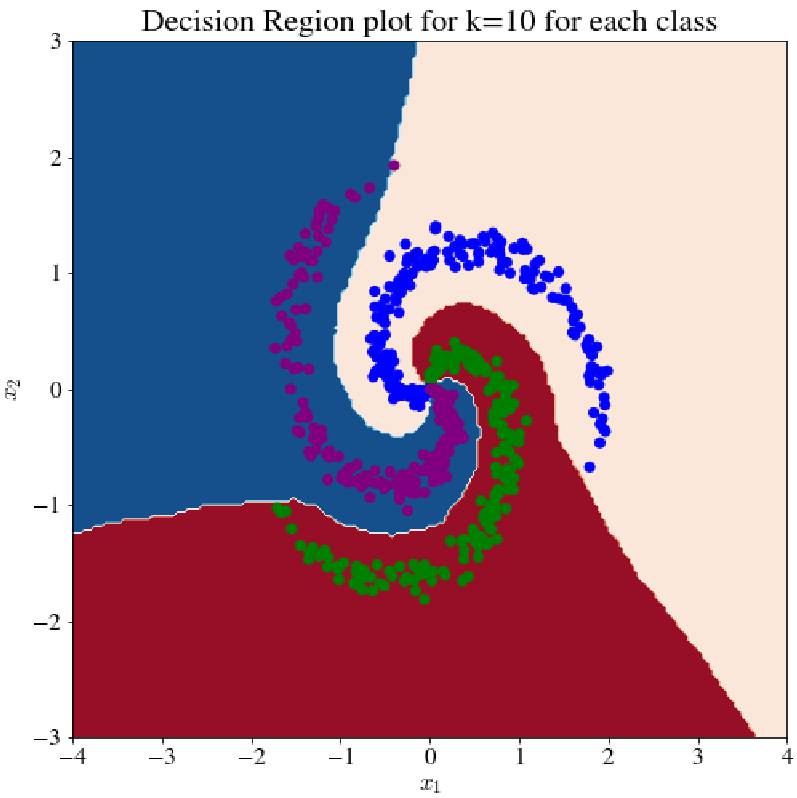


Figure 12: Decision region

Inferences

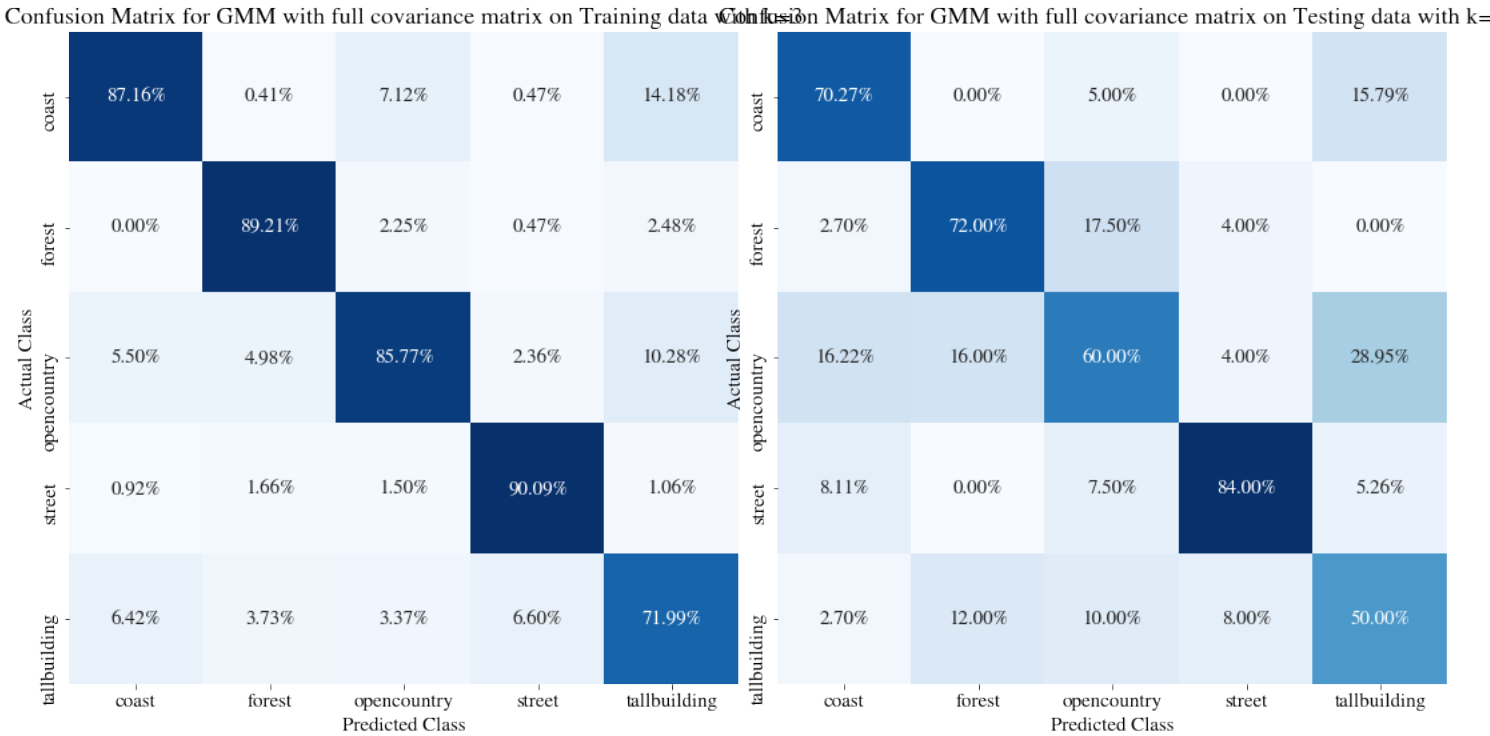
- The decision boundary is clearly non-linear
- k=10 seems to give the best classification accuracy as it is neither too simple a model nor complex enough requiring many data points

2 Dataset 2: Real World Data Sets

2.(a) Image data set for static pattern classification

2.(a).1 Bayes Classifier with a GMM for each class, using full covariance matrices

No. of components(Q)	Training	Validation	Test
2	76.639	60.62	59.235
3	84.262	60.54	65.45
4	87.704	58.75	65.895
5	91.967	58.201	63.35
6	93.77	54.385	59.21



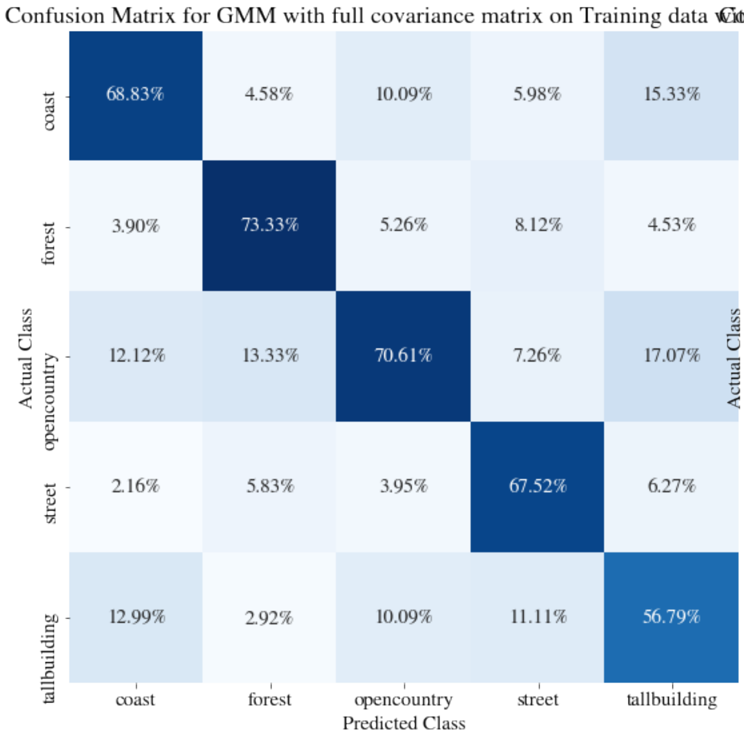
(a) Confusion matrix for training dataset

(b) Confusion matrix for test dataset

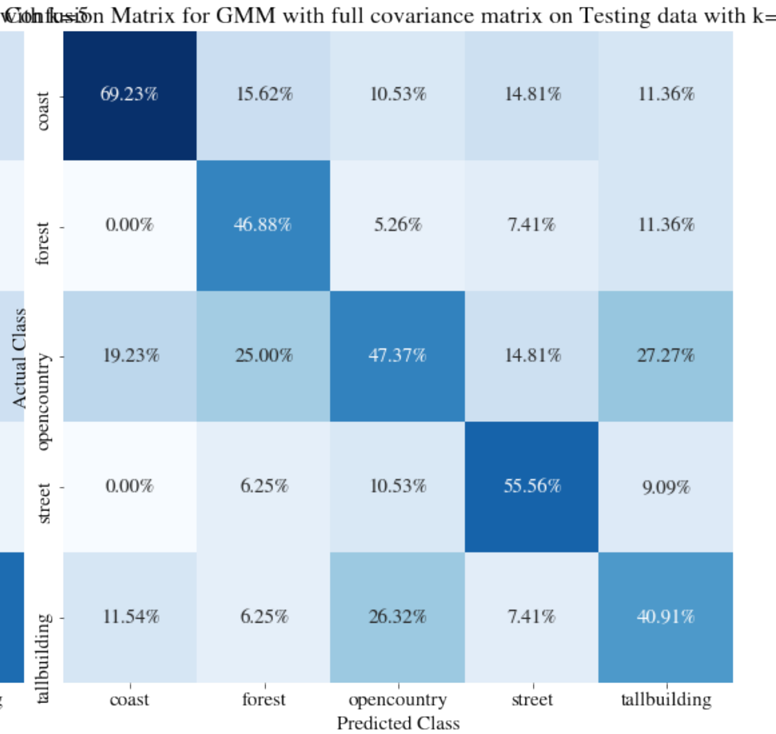
- It can be seen that q=3 is the best model
- The training accuracy increases as q increases
- But the validation accuracy increases, reaches a maxima and then decreases

2.(a).2 Bayes Classifier with a GMM for each class, using diagonal covariance matrices

No. of components(Q)	Training	Validation	Test
2	58.114	41.43	47.72
3	61.147	51.2195	47.84
4	65.163	55.617	51.744
5	66.967	57.425	50.675
6	69.344	45.61	57.5418



(a) Confusion matrix for training dataset



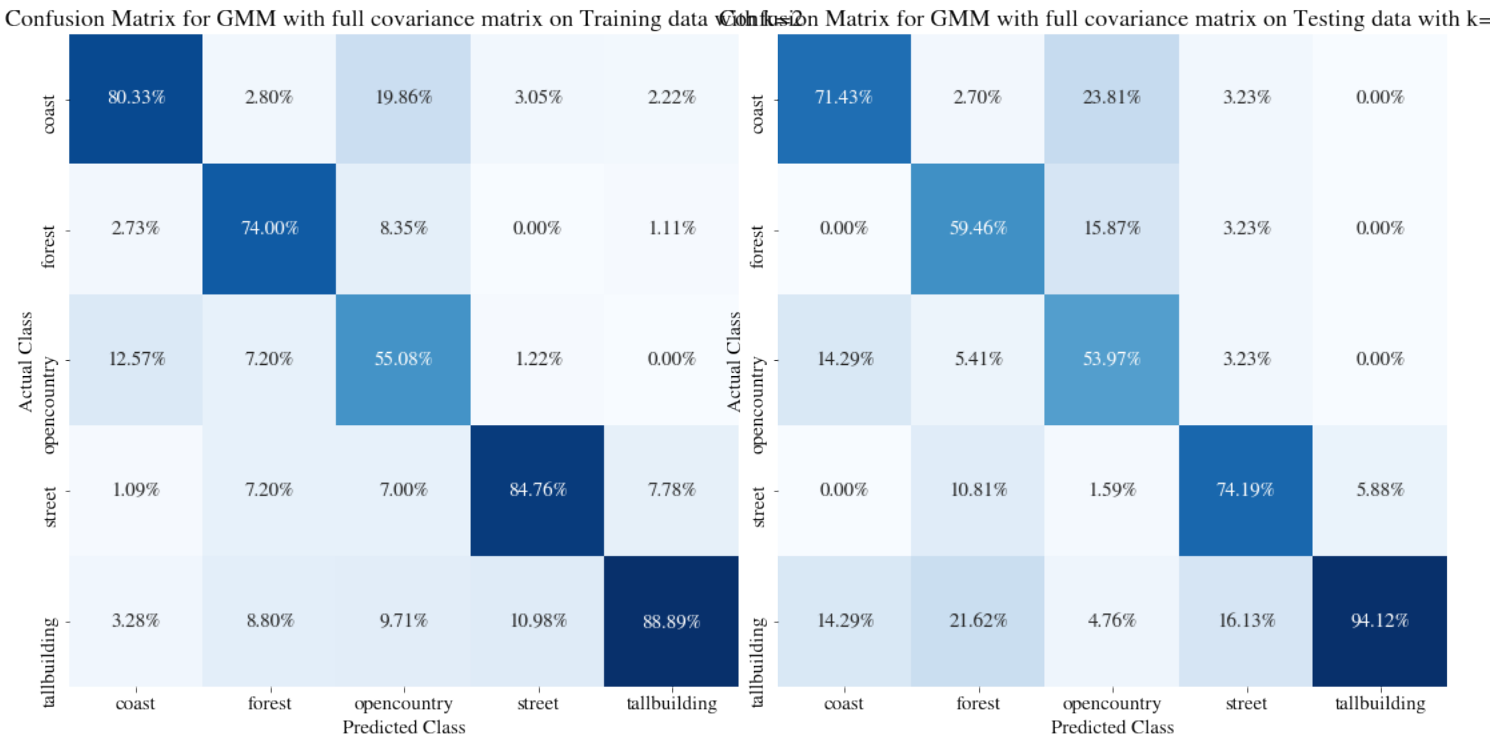
(b) Confusion matrix for test dataset

- It can be seen that q=5 is the best model
- The training accuracy increases as q increases
- But the validation accuracy increases, reaches a maxima and then decreases

2.(b) Image data set for varying lenght pattern

2.(b).1 Bayes Classifier with a GMM for each class, using full covariance matrices

No. of components(Q)	Training	Validation	Test
2(i)	71.721	70.689	65.3409
2(ii)	71.639	71.264	65.3409
3	59.91	63.21	57.386
4	46.229	47.70	47.72
5	47.295	50.574	42.613
6	47.049	41.95	41.47



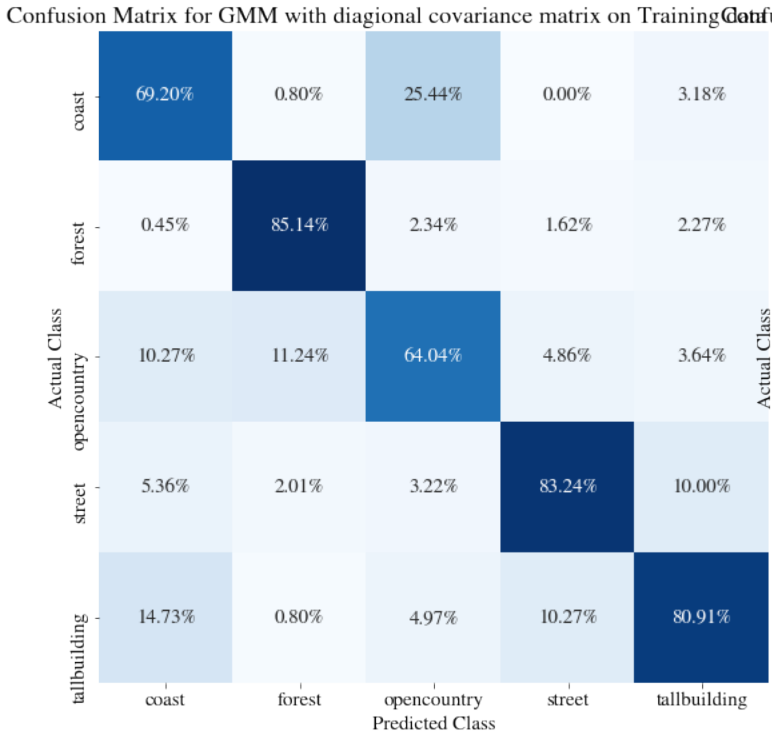
(a) Confusion matrix for training dataset

(b) Confusion matrix for test dataset

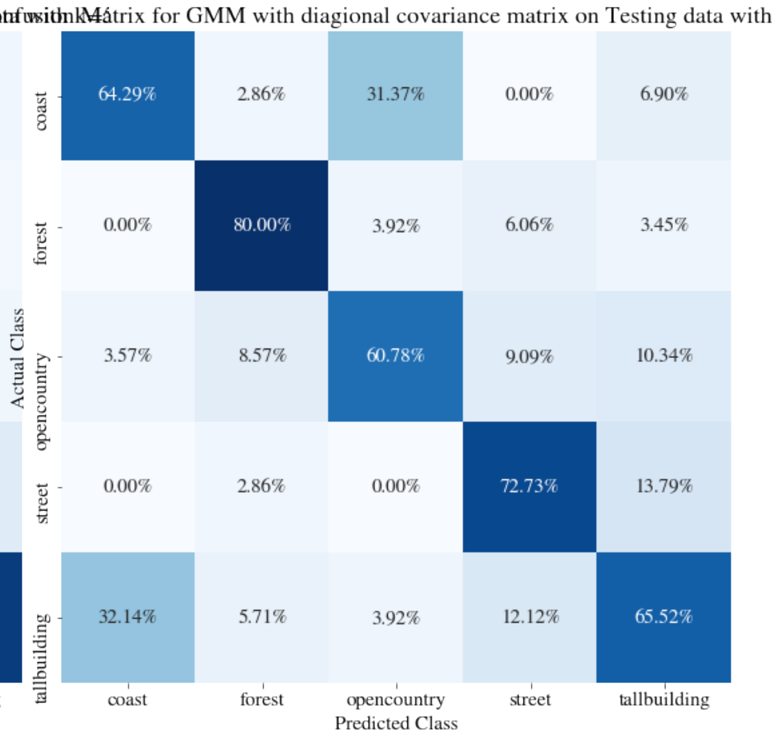
- It can be seen that q=2 is the best model
- The training accuracy increases as q increases
- But the validation accuracy increases, reaches a maxima and then decreases

2.(b).2 Bayes Classifier with a GMM for each class, using diagonal covariance matrices

No. of components(Q)	Training	Validation	Test
2	64.180	67.82	59.65
3	67.13	67.816	59.67
4	72.705	70.115	65.34
5	75.245	78.16	68.181
6	70.327	72.988	71.02



(a) Confusion matrix for training dataset



(b) Confusion matrix for test dataset

- It can be seen that q=5 is the best model
- The training accuracy increases as q increases
- But the validation accuracy increases, reaches a maxima and then decreases