IC 272: DATA SCIENCE - III
LAB ASSIGNMENT – V
Data classification using Bayes Classifier with Gaussian Mixture Model (GMM); Regression using Simple Linear Regression and Polynomial Curve Fitting

Student's Name: Rashmi

Mobile No: 7015331137

Roll Number: B19218

Branch: Engineering Physics

PART - A

1 a.

		Prediction Outcome
True Label	677	48
	44	7

Figure 1 Bayes GMM Confusion Matrix for Q = 2

		Prediction Outcome
True Label	711	14
	49	2

Figure 2 Bayes GMM Confusion Matrix for Q = 4

		Prediction Outcome
True Label	700	25
	45	6

Figure 3 Bayes GMM Confusion Matrix for Q = 8

		Prediction Outcome
True Label	720	5
	50	1

Figure 4 Bayes GMM Confusion Matrix for Q = 16

b.

Table 1 Bayes GMM Classification Accuracy for Q = 2, 4, 8 & 16

Q	Classification Accuracy (in %)
2	88.144%
4	91.881%
8	90.979%
16	92.912%

Inferences:

1. The highest classification accuracy is obtained with Q =16.
2. Increasing the value of Q increases the prediction accuracy almost.
3. On increasing Q, each data-vector can be correctly classified and thus the prediction accuracy increases.

4. As the classification accuracy increases with the increase in value of Q, the number of diagonal elements in Confusion matrix increase.
5. When Q increases diagonal elements increases as accuracy is calculated by TP+TN/Total Samples, True positive and True Negative are on the diagonals so if Accuracy increases then diagonal elements increases.
6. As the classification accuracy increases with the increase in value of Q, the number of off-diagonal elements decrease.
7. Off Diagonal elements decrease when Q increases because increasing Q increases accuracy which is calculated on the basis of diagonal elements, so if diagonal elements increase then off diagonal will decrease.

2

Table 2 Comparison between Classifiers based upon Classification Accuracy

S. No.	Classifier	Accuracy (in %)
1.	KNN	93.170%
2.	KNN on normalized data	92.912%
3.	Bayes using unimodal Gaussian density	87.500%
4.	Bayes using GMM	92.912%

Inferences:

1. KNN has highest and Bayes using unimodal Gaussian density has lowest accuracy.
2. Classifiers in ascending order of classification accuracy: Bayes using unimodal Gaussian density < Bayes using GMM < KNN < KNN on normalized data.
3. Bayes classifier assumes that the features are independent, that's why they have less accuracy than KNN, GMM is better than unimodal as now the points in each class can also be associated with its cluster.

PART – B

1

a.

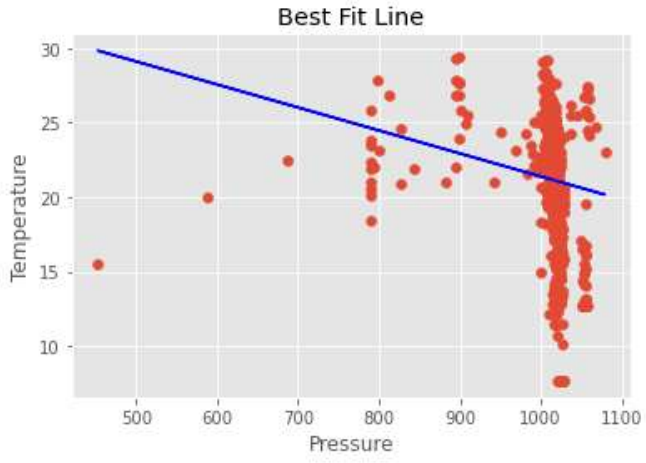

Figure 5 Pressure vs. temperature best fit line on the training data

Inferences:

1. The best fit line does not fit the training data perfectly.
2. Because scatter plot is not arranged linearly whereas linear regression tries to fit it in a line.
3. Bias is high as the best fit line underfits the data, the model requires more complex function to fit the training data. Variance is low as the bias is high due to underfitting of data.

b.

Prediction accuracy on the training data using root mean squared error = 4.2798

c.

Prediction accuracy on the test data using root mean squared error = 4.2870

Inferences:

1. Amongst training and testing accuracy, the accuracy on training data is higher as it has less RMSE.

2. Training accuracy is higher because the model is made on the training data so it will have less RMSE on training data.

d.

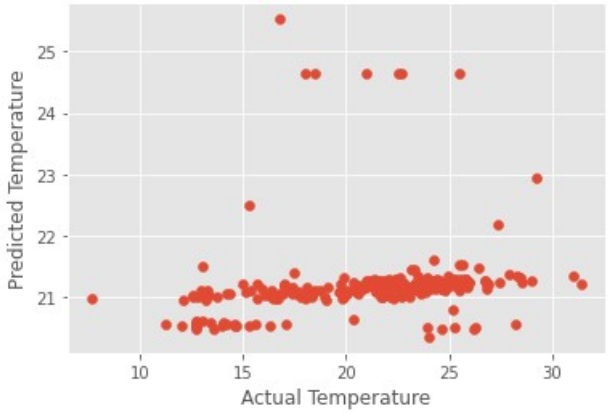

Figure 6 Scatter plot of predicted temperature from linear regression model vs. actual temperature on test data

Inferences:

1. Based upon the spread of the points, predicted temperature is very less accurate.
2. The actual temperature is spread from 10 to 30 but the predicted temperature is more concentrated from 20 to 23 which shows that the prediction accuracy is not high. Graph should have slope=1 for higher accuracy but here it is parallel to x-axis.

2

a.

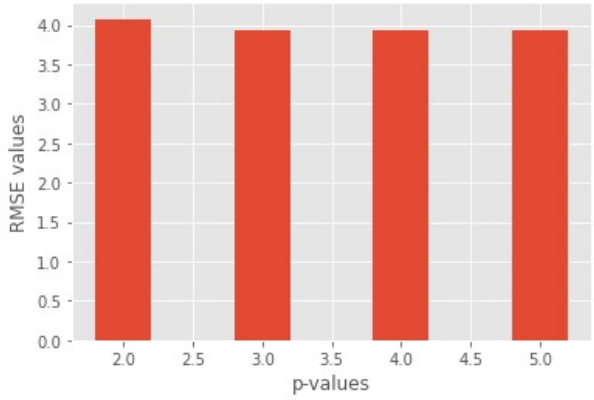

Figure 7 RMSE vs. different values of degree of polynomial (p = 2, 3, 4, 5) on the training data

Inferences:

1. RMSE value decreases with respect to increase in degree of polynomial (p = 2, 3, 4, 5).
2. The RMSE decreases from p=2 to p=3 more compared to rest. From p=3 it decreases slightly or almost remains constant.
3. As the degree increases the curve fits the data more better, so the RMSE decreases.
4. From the RMSE value, p=4 curve will approximate the data best.
5. As the degree increases the bias decreases and variance increases as the model starts becoming more complex and starts fitting the data better.

b.

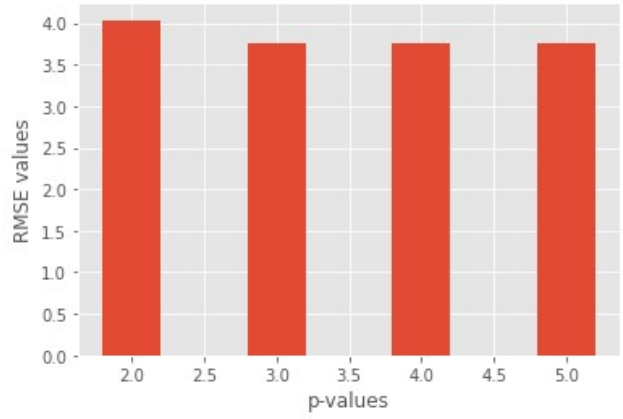

Figure 8 RMSE vs. different values of degree of polynomial (p = 2, 3, 4, 5) on the test data

Inferences:

1. RMSE value decreases with respect to increase in degree of polynomial (p = 2, 3, 4, 5).
2. The RMSE decreases more from p=2 to p=3 then it almost remains constant or decreases.
3. The RMSE decreases from p=2 to p=3 more compared to rest. From p=3 it decreases slightly or almost remains constant.
4. From the RMSE value, p=5 curve will approximate the data best.
5. As the degree increases the bias decreases and variance increases as the model starts becoming more complex and starts fitting the data better.

c.

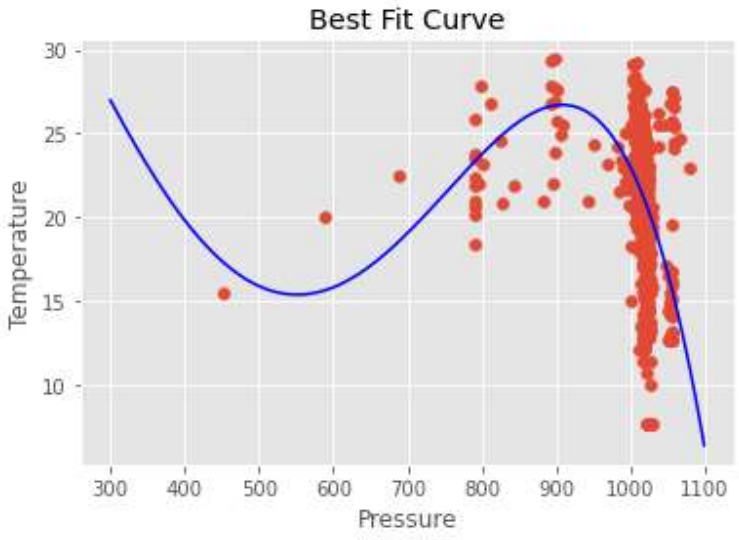

Figure 9 Pressure vs. temperature best fit curve using best fit model on the training data

Inferences:

1. p=5 corresponding to best fit model.
2. Because it fits the data better as it is more complex and have higher variance.
3. As the degree increases the bias decreases and variance increases as the model starts becoming more complex and starts fitting the data better.

d.

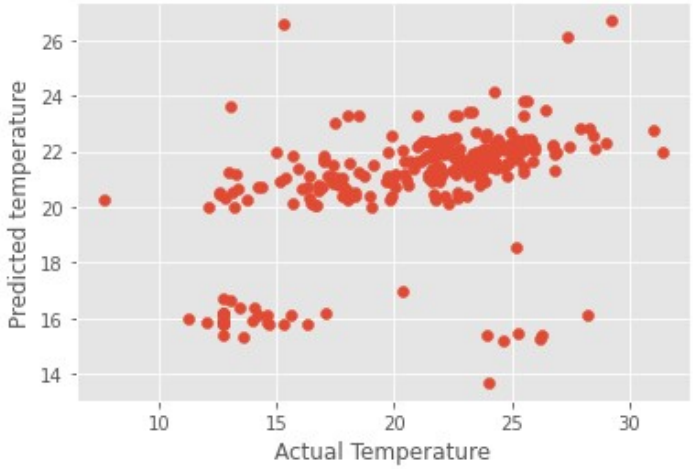

Figure 10 Scatter plot of predicted temperature from non- linear regression model vs. actual temperature on test data

Inferences:

1. Based upon the spread of the points, predicted temperature has good accurate.
2. The actual temperature is spread between 10 and 30, similarly the predicted temperature is also spread between 10 to 30, thus we can say that the accuracy is good. Also, slope is tending to 1.
3. Prediction accuracy of non-linear is better as the RMSE is lower for it, also from the spread of data we can see that the non-linear regression is better than linear regression.
4. RMSE of non-linear regression is lower than linear regression and the spread of predicted value matches actual value better in nonlinear regression than linear, so we can say that non-linear regression is better.
5. In linear regression, bias is high and variance is low but in non-linear regression, variance is high and bias is low.