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: Jai Prakash Yadav **Mobile No: 9306871378**

Roll Number: B19247 **Branch: ME**

PART - A

1 a.

		Prediction Outcome
True Label	711	12
	49	4

Figure 1 Bayes GMM Confusion Matrix for Q = 2

		Prediction Outcome
True Label	717	6
	53	0

Figure 2 Bayes GMM Confusion Matrix for Q = 4

		Prediction Outcome
True Label	720	3
	53	0

Figure 3 Bayes GMM Confusion Matrix for Q = 8

		Prediction Outcome
True Label	718	5
	53	0

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	92.139
4	92.397
8	92.784
16	92.526

Inferences:

1. The highest classification accuracy is obtained with Q = 8.
2. On Increasing the value of Q the prediction accuracy increases up to Q=8 and then decreases.
3. The distribution of the data may not be unimodal Gaussian distribution, in case of multimodal data increasing Q till the no. of modes will increase the prediction accuracy further increasing the value of Q will result in decreased prediction accuracy.
4. As the classification accuracy increases with the increase in value of Q till Q = 8, the number of diagonal elements in Confusion matrix increases and decreases for Q = 16.
5. The prediction accuracy is directly proportional to the no. of diagonal elements (TP + TN).
6. As the classification accuracy increases with the increase in value of Q till Q = 8, the number of off-diagonal elements decreases and increases for Q = 16.
7. The prediction accuracy is inversely proportional to the no. of off-diagonal elements (FP + FN).

2

Table 2 Comparison between Classifiers based upon Classification Accuracy

S. No.	Classifier	Accuracy (in %)
1.	KNN	92.397
2.	KNN on normalized data	92.397
3.	Bayes using unimodal Gaussian density	88.918
4.	Bayes using GMM	92.784

Inferences:

1. Highest Accuracy = Bayes using GMM. Lowest Accuracy = Bayes using unimodal Gaussian density.
2. Bayes using unimodal Gaussian density < KNN = KNN on Normalized data < Bayes using GMM.
3. Bayes classifier assumes the data as a Gaussian distribution. Bayes unimodal Gaussian density assumes the data to as unimodal thus results in lower accuracy and Bayes using GMM assumes data to have multiple modes this will result in increased prediction accuracy.
4. KNN is an instance-based classifier thus has a lower accuracy and slower than Bayes using GMM.

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. 2. The relation between Temperature and Pressure are not linear.
3. 3. The best-fit line is highly biased as it is underfitting the data. The variance is low as the data is not overfitting. 4. We need to find an estimation so that the bias and variance is low. But in linear fitting we have to trade off for lower variance than higher bias.

b.

The prediction accuracy on training data = 4.279790433682601

c.

The prediction accuracy on testing data = 4.286985483129509

Inferences:

1. Training accuracy is higher than test accuracy.
2. Best fit line is the one with least training error.

d.

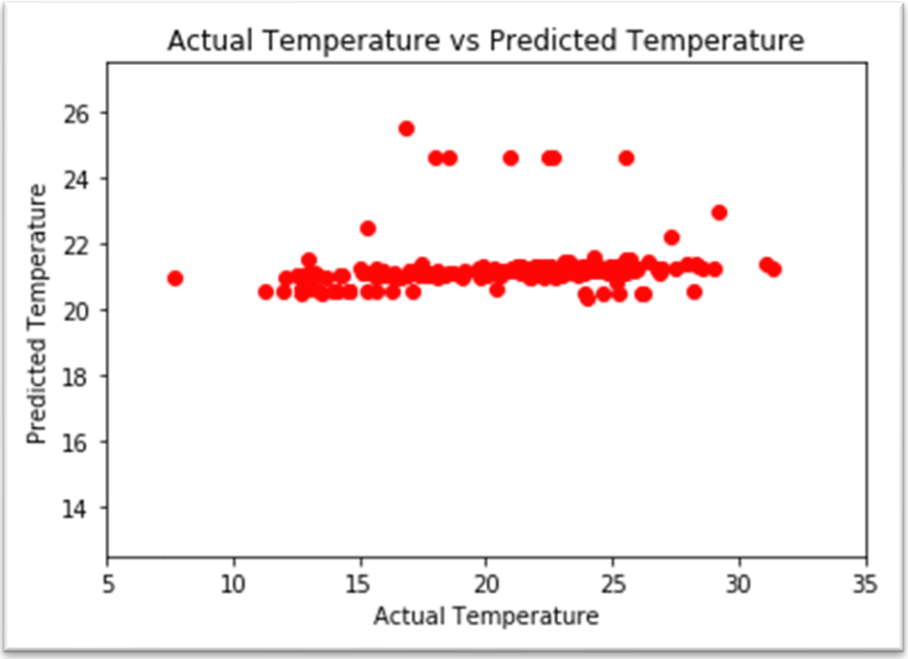

Figure 6 Scatter plot of predicted temperature from linear regression model vs. actual temperature on test data

Inferences:

1. Based upon the spread of the points, the predicted temperature is not so accurate as actual temperature. The predicted value of temperature has a small range. The correlation coefficient is also low.
2. The relationship between Pressure and Temperature are not linear. If the prediction accuracy was higher the plot would be the curve y = x.

2

a.

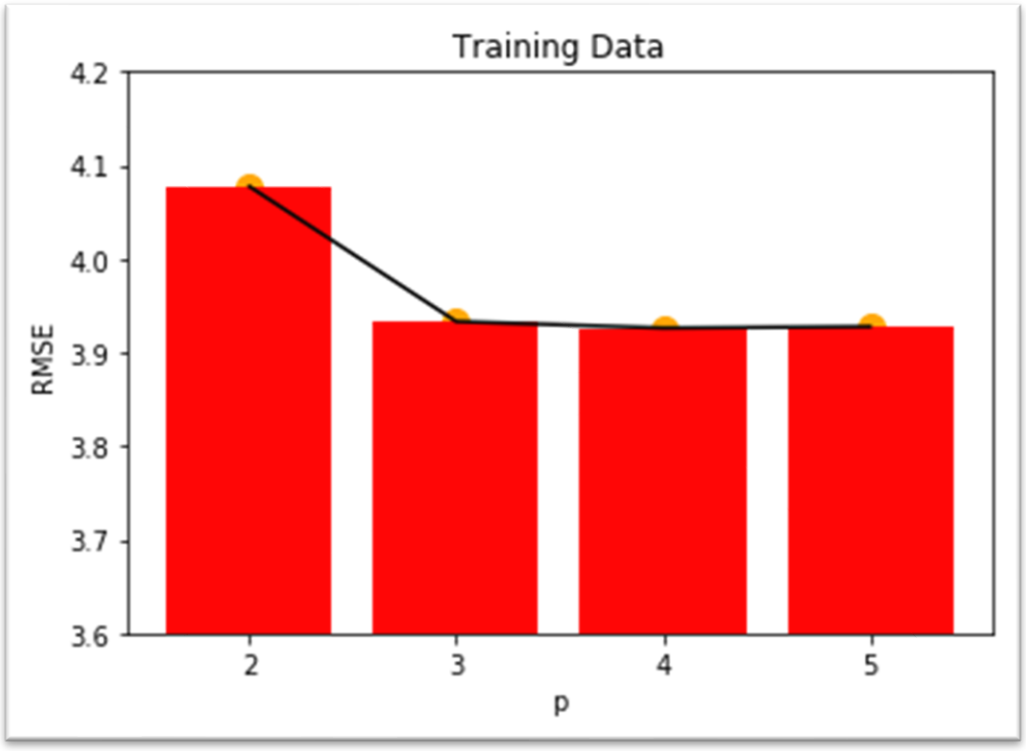

Figure 7 RMSE vs. different values of degree of polynomial (p = 2, 3, 4, 5) on the training data

Inferences:

1. RMSE value decreases for increase in p from 2 to 4 and slightly increase for p =5
2. The decrease is steep from p = 2 to p = 3, then on the value is almost constant.
3. As the value of p increases our estimate becomes more accurate on training data and the RMSE gradually decreases.
4. From the RMSE value, 5th degree curve will approximate the data best (p = 5) as it has minimum test error.
5. Bias is still present, but it is not so high as in best line fit. The variance is not so high as there is no over fitting. There is a sort of balance between bias-variance trade-off.

b.

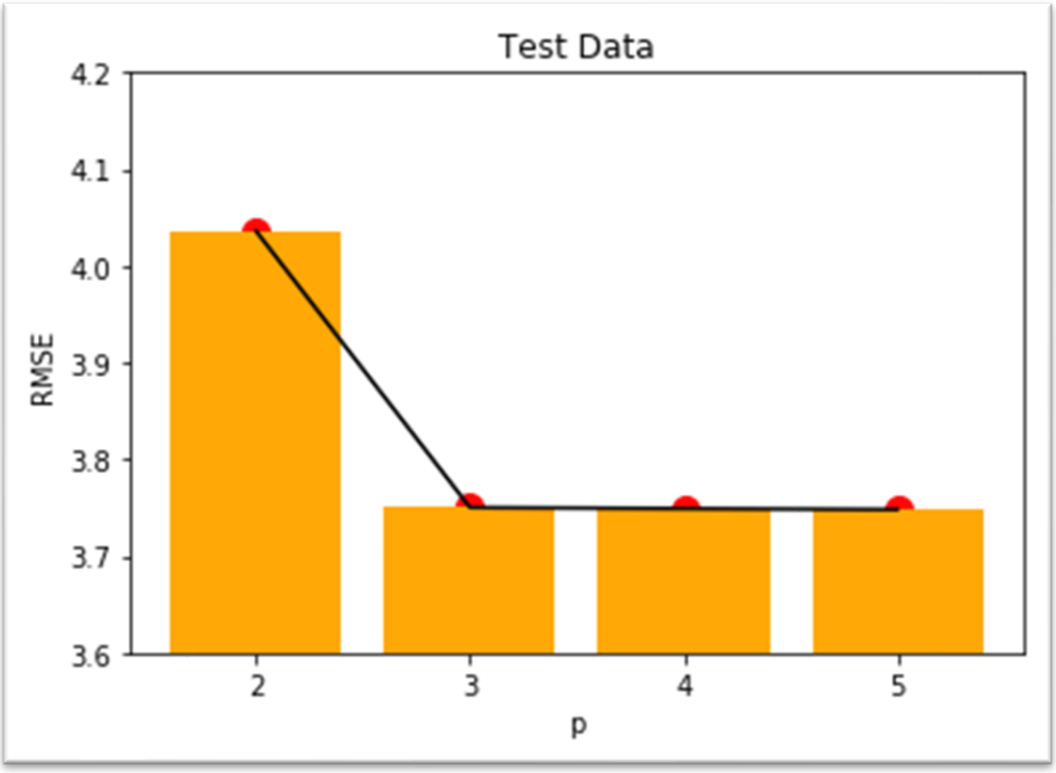

Figure 8 RMSE vs. different values of degree of polynomial (p = 2, 3, 4, 5) on the test data

Inferences:

1. RMSE value decreases with respect to increase in degree of polynomial (p = 2, 3, 4, 5).
2. The decrease is steep from p = 2 to p = 3, then on the value is almost constant.
3. As the value of p increases our estimate becomes more accurate on testing data and the RMSE gradually decreases. However, on further increase there will be an overfitting on training data.
4. From the RMSE value, 5th degree curve will approximate the data best (p = 5).
5. Bias is still present, but it is not so high as in best line fit. The variance is not so high as there is no over fitting. There is a sort of balance between bias-variance trade-off.

c.

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

Inferences:

1. The p-value corresponding to best fit model = 5.
2. At p = 5 the test error is minimum.
3. Bias is still present, but it is not so high as in best line fit. The variance is not so high as there is no over fitting. There is a sort of balance between bias-variance trade-off. Therefore, best-fit curve gives a better estimate than best-fit line.

d.

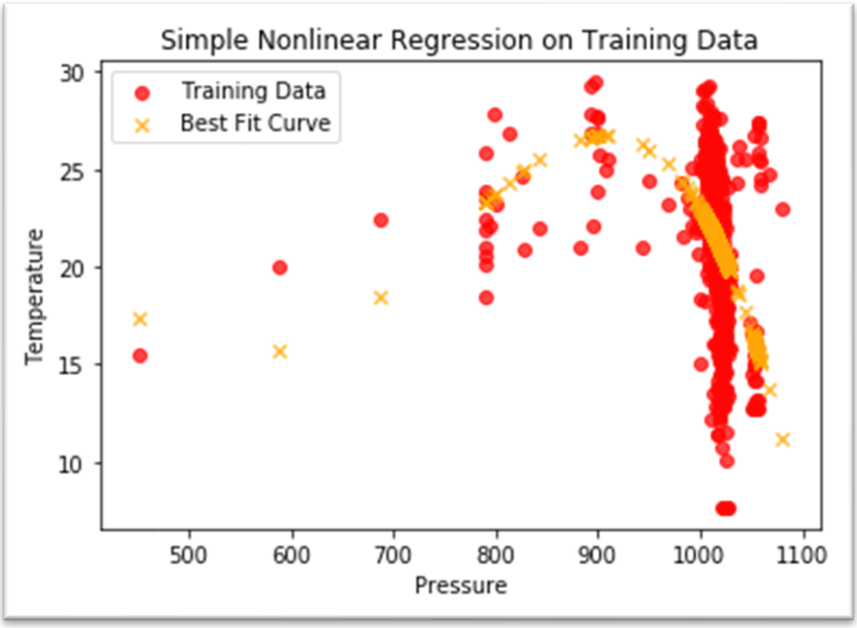

Figure 10 Scatter plot of predicted temperature from non- linear regression model vs. actual temperature on test data

Inferences:

1. Based upon the spread of the points, the range of predicted temperature is more. Even though the estimate is not so perfect it gives a better approximation.
2. The data has more of polynomial relation than linear relation. The best-fit curve has a lower RMSE on test data.
3. Linear regression model has less accuracy compared to non-linear regression. Also, the spread of data points is lower for Linear regression than for polynomial regression.
4. The linear regression works best for data having linear relation. The non-linear regression works best for data having polynomial relation. In the given scenario, the data doesn't have a linear relation. So, non-linear regression gives us best estimate.
5. The bias for linear regression is high as there is underfitting but the variance is low as there is no overfitting. So, we have to trade off high bias for low variance
6. The bias for non-linear regression is lower than linear regression also the variance is low as there is no overfitting. So, we have a balance between bias and variance and have a better trade off.