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

Name: Srishti Ginjala Roll number: b19084

Mobile Number:9440000900 Branch: CSE

PART-A

1a.

	Prediction Outcome	
True Label	677	48
	44	7

Figure 1 Bayes GMM Confusion Matrix for Q = 2

	Prediction Outcome	
True Label	691	34
	42	9

Figure 2 Bayes GMM Confusion Matrix for Q = 4

	Prediction Outcome	
True Label	722	3
	49	2

Figure 2 Bayes GMM Confusion Matrix for Q = 8

	Prediction Outcome	
True Label	722	3
	51	0

Figure 2 Bayes GMM Confusion Matrix for Q = 16

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

Q	Classification Accuracy (in %)
2	0.881
4	0.902
8	0.933
16	0.930

Inferences:

1. The highest classification accuracy is obtained with Q =8.
2. Increasing the value of Q generally increases the prediction accuracy if there are sufficient training examples available.
3. Increasing the value of Q increases the prediction accuracy if there the data is randomly distributed because all points may not exactly fit into single oval shape (in case of 2d gaussian distribution). Hence, we need multiple clusters for accurate classification.
4. As the classification accuracy increases with the increase in value of Q, the number of diagonal elements in Confusion matrix increase.
5. The diagonal elements in the confusion matrix represent the number of true predictions. Hence, if the number of True positives and True negatives increase, the accuracy of the model increases.
6. As the classification accuracy increases with the increase in value of Q, the number of off-diagonal elements decrease.
7. The number of off-diagonal elements decrease as the number of false predictions decrease.

2)

Table 2 Comparison between Classifiers based upon Classification Accuracy

S. No.	Classifier	Accuracy (in %)
1.	KNN	0.928
2.	KNN on normalized data	0.926
3.	Bayes using unimodal Gaussian density	0.889
4.	Bayes using GMM	0.932

Inferences:

1. Classifier with highest accuracy is Bayes using GMM and lowest accuracy is Bayes using unimodal Gaussian density.
2. Classifiers in ascending order of classification accuracy: Bayes using unimodal Gaussian density < KNN on normalized data < KNN <Bayes using GMM.
3. In this case Bayes using unimodal Gaussian density is giving least accuracy because the data is randomly distributed and one Gaussian model is not sufficient to estimate the correct values of parameters of distribution and so we require multiple clusters to cover all the data points. Hence, Bayes using GMM gives the maximum accuracy.

Part-B

2a)

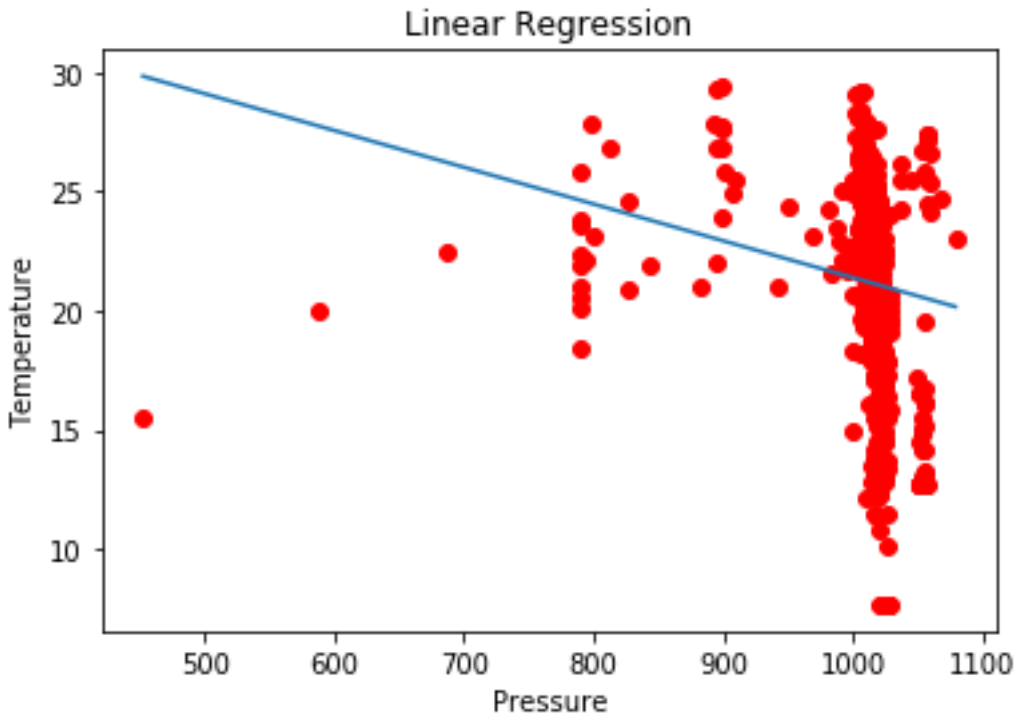


Figure 5 Pressure vs. temperature best fit line on the training data

Inferences:

1.No, the best fit line does not fit the training data perfectly.

2. It does not fit the training data perfectly as it is oversimplified for the data, a more complex function is required to fit the data. 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: 4.280.

c. Prediction accuracy on the test data using root mean squared error: 4.287.

Inferences:

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

2. Training accuracy is higher because its RMSE is lower than testing accuracy, this is because the model is made on the training data so it will have less RMSE on training data.

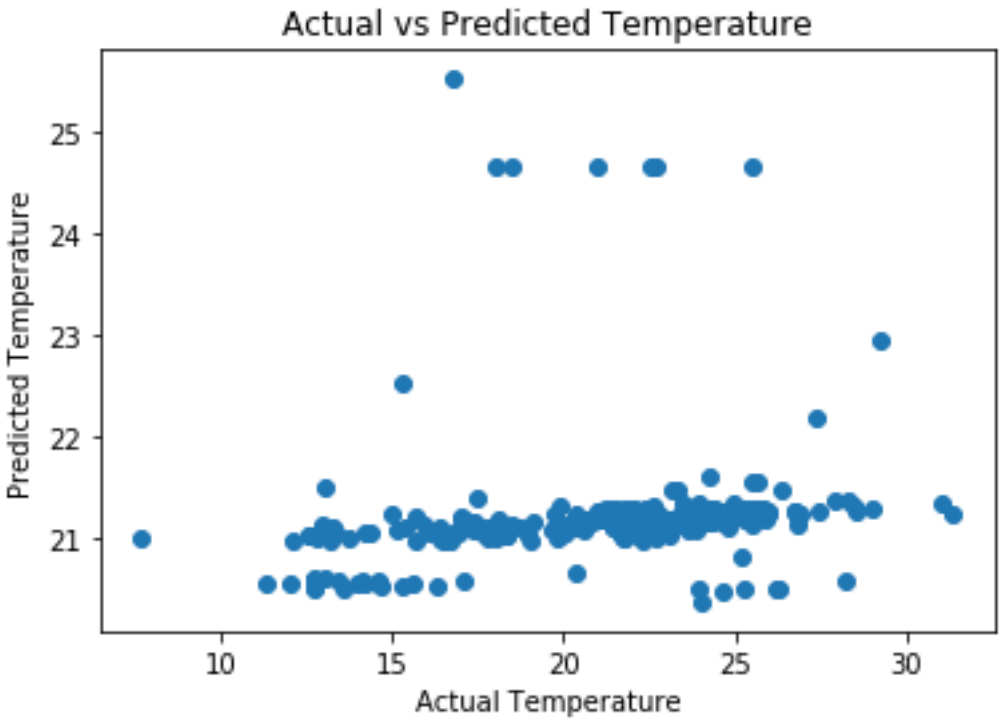


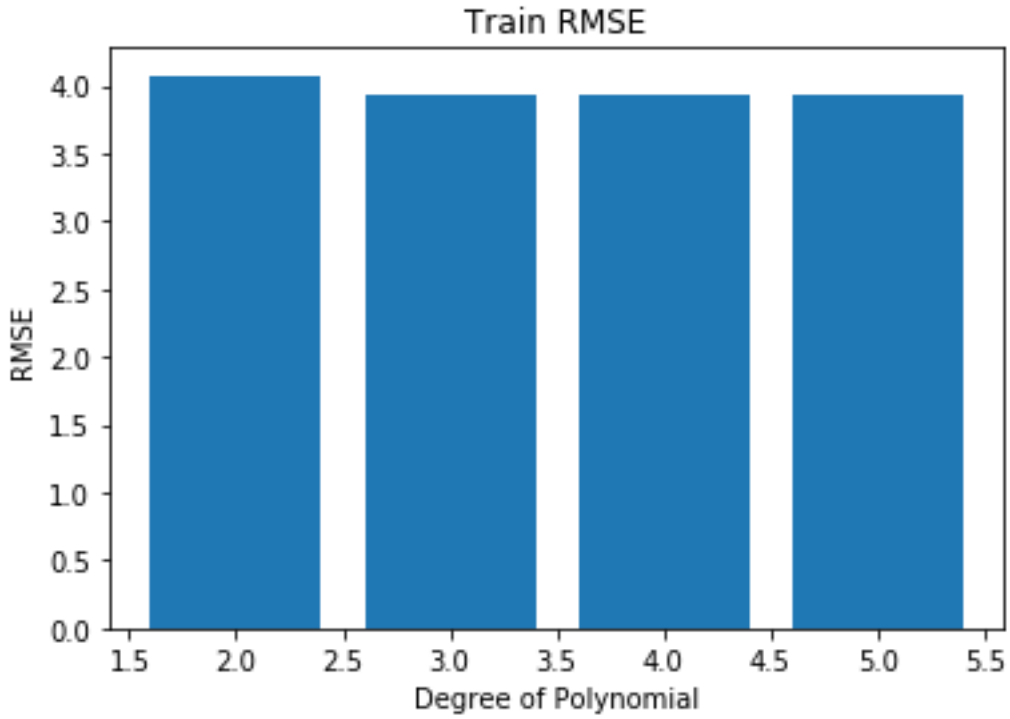
Figure 6 Scatter plot of predicted temperature from linear regression model vs. actual temperature on test data

Inferences:

1. The prediction accuracy of temperature is not high.

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

2a)



Inferences:

1. RMSE value decreases with respect to increase in degree of polynomial (p = 2, 3, 4, 5).
2. According to the above graph, it decreased steeply for degree 2 to degree 3 and then on gradually.
3. This is because we are using only one attribute for prediction and lower degree polynomial is sufficient.
4. From the RMSE value, 4^{th} degree 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.

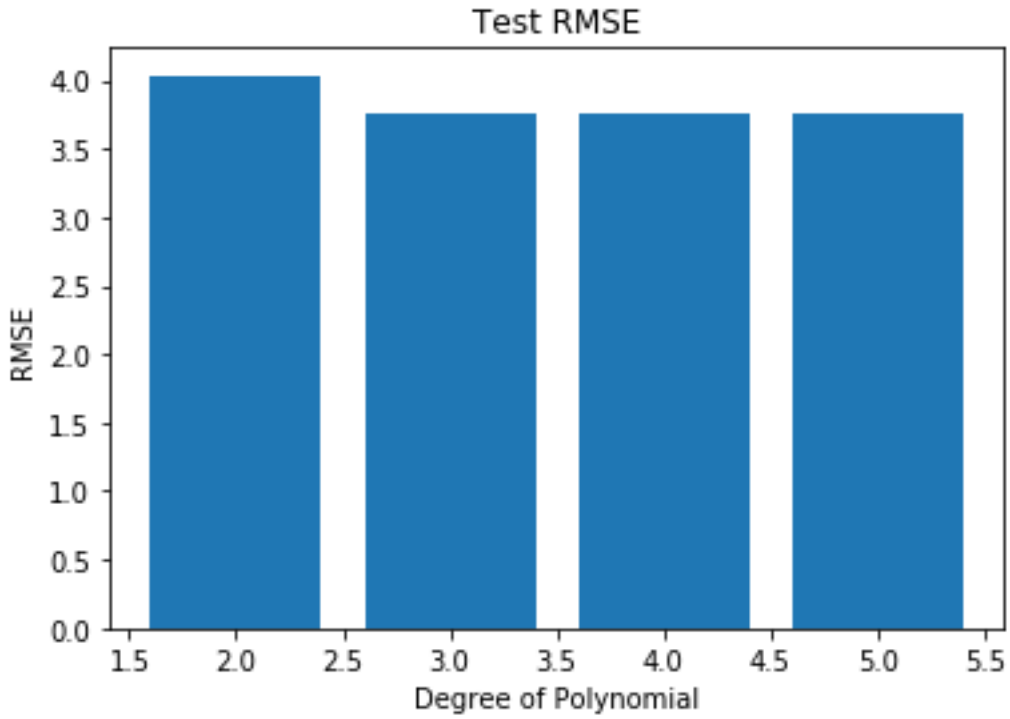


Figure 8 RMSE vs. different values of degree of polynomial (p = 2, 3, 4, 5) on the test data

Inferences:

1. RMSE value decreases with respect to increase in degree of polynomial (p = 2, 3, 4, 5).

2. The RMSE decreases 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, degree 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.

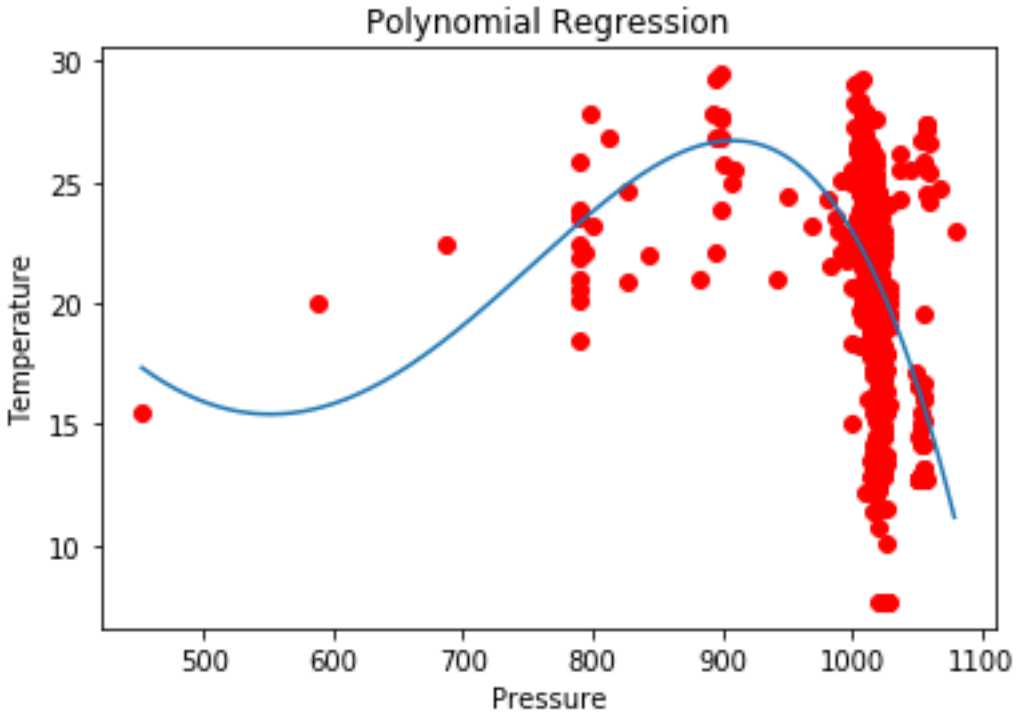


Figure 9 Pressure vs. temperature best fit curve using best fit model on the training data

Inferences:

1. p-value is 5 corresponding to best fit model.

2. p=5 is best fit model 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.

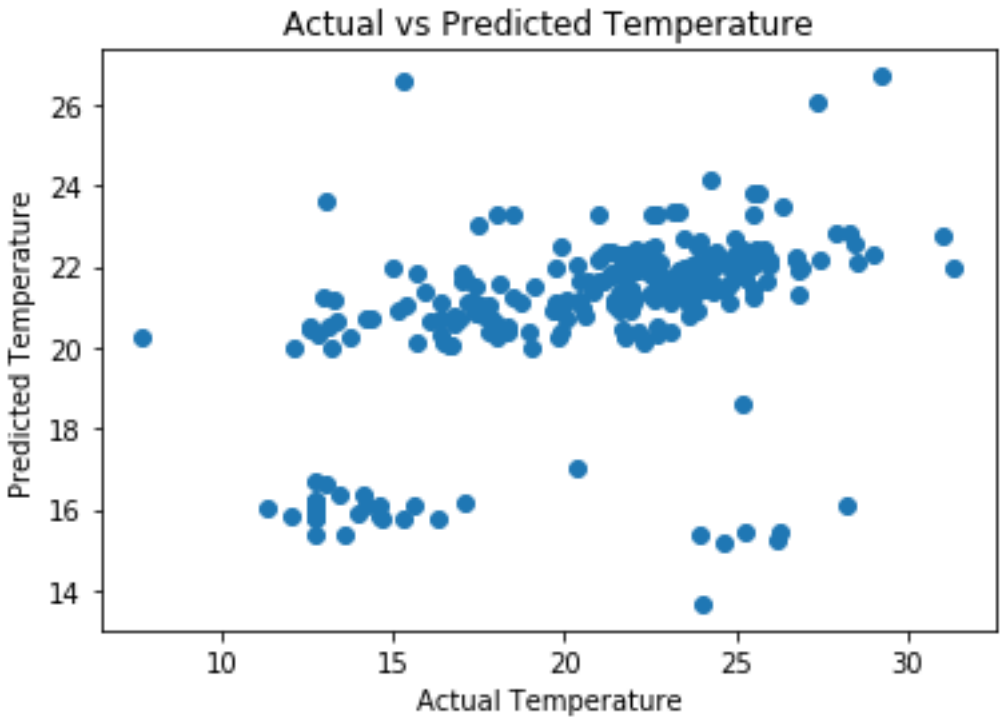


Figure 10 Scatter plot of predicted temperature from non- linear regression model vs. actual temperature on test data

Inferences:

1. From the spread of points we can see that accuracy of predicted temperature is quite good.

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.

3. Prediction accuracy of nonlinear is better as the rmse is lower for it, also from the spread of data we can see that the nonlinear regression is better than linear regression.

4. Rmse of nonlinear regression is lower than linear and the spread of predicted value matches actual value better in nonlinear regression than linear, so we can say that nonlinear regression is better.

5. In linear regression bias is high and variance is low but in nonlinear regression variance is high and bias is low.