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

PART - A

1 a.

		Prediction Outcome
True Label	686	37
	48	5

Figure 1 Bayes GMM Confusion Matrix for Q = 2

		Prediction Outcome
True Label	612	111
	30	23

Figure 2 Bayes GMM Confusion Matrix for Q = 4

		Prediction Outcome
True Label	716	7
	51	2

Figure 3 Bayes GMM Confusion Matrix for Q = 8

		Prediction Outcome
True Label	714	9
	52	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	**89.046**
4	**81.830**
8	**92.526**
16	**92.139**

Inferences:

1. The highest classification accuracy is obtained with Q = 8
2. Increasing the value of Q decreases the prediction accuracy going from Q=2 to Q=4 but after 4 it increases for Q =8 and again decreases when Q=16.

3. By increasing the value of Q each datapoint can be correctly associated with its cluster and thus result in advancement of prediction accuracy.
4. As the classification accuracy decreases with the increase in value of Q from 2 to 4, the number of diagonal elements in Confusion matrix decrease and as the accuracy increases with the increase in value of Q from 8 to 16, the number of diagonal elements in Confusion matrix increases .
5. Increase in accuracy results in increase in elements of diagonal matrix and vice versa because accuracy is calculated by TP+TN/Total Samples, True positive and True negative are on the diagonal so if accuracy increases diagonal elements increases.
6. As The classification accuracy increases with the increase in value of Q, the number of off-diagonal elements decreases and vice versa.
7. Off-diagonal elements decreases when Q increases because increasing Q increases the accuracy which is calculated based on diagonal elements, so if diagonal elements increases then off-diagonal elements will decrease.

2

Table 2 Comparison between Classifiers based upon Classification Accuracy

S. No.	Classifier	Accuracy (in %)
1.	KNN	92.396
2.	KNN on normalized data	92.396
3.	Bayes using unimodal Gaussian density	88.917
4.	Bayes using GMM	92.525

Inferences:

1. Bayes using GMM classifiers has highest accuracy = 92.525%.
   Bayes using unimodal Gaussian density has lowest accuracy = 88.917%.

2. The classifiers in ascending order of classification accuracy.
   Bayes using unimodal Gaussian density < KNN ≅ KNN on normalized data < Bayes using GMM

3. Bayes models assume that the data has gaussian distribution, which is not always the case like in the given data. KNN does not assume anything about the data. So KNN performs better even though the difference between bayes using GMM and KNN is very small in this case. As for Bayes unimodal and Bayes multimodal, later performs better because as compared to the unimodal we use multiple clusters in a class.

PART – B

1

a.

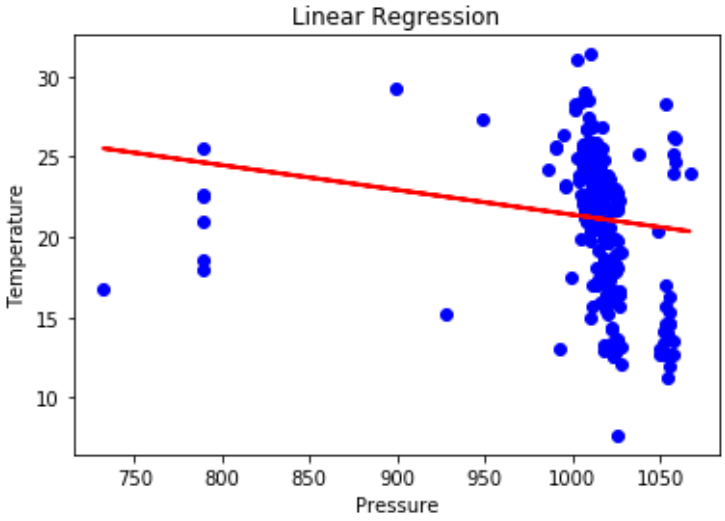

Figure 5 Pressure vs. temperature best fit line on the training data

Inferences:

1. No, the best fit line does not fit the training data perfectly
2. It does not fit the training data perfectly because it is oversimplified for the given data and the curve to be fit more precisely a more complex function is required.
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.

The prediction accuracy of training data is 4.2797.

c.

The prediction accuracy on testing data is 4.2869.

Inferences:

1. Amongst training and testing accuracy, the accuracy of testing data is higher than the training data.
2. Testing accuracy is higher because its RMSE is lower than training accuracy , this is because the model is made on the training data so it will have less RMSE due to the changes during modeling where test data is preserve as the real data leading to the higher accuracy.

d.

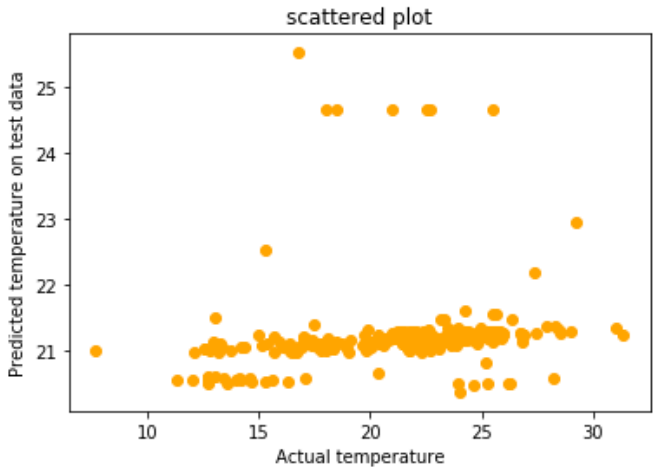

Figure 6 Scatter plot of predicted temperature from linear regression model vs. actual temperature on test data

Inferences:

1. Based upon the spread of the points in the above graph we can infer that the accuracy of the predicted temperate is not much high.
2. The actual temperature is spread from 10 to 30 but the predicted temperature is more concentrated from 20 to 25 which shows that the prediction accuracy is not much high.

2

a.

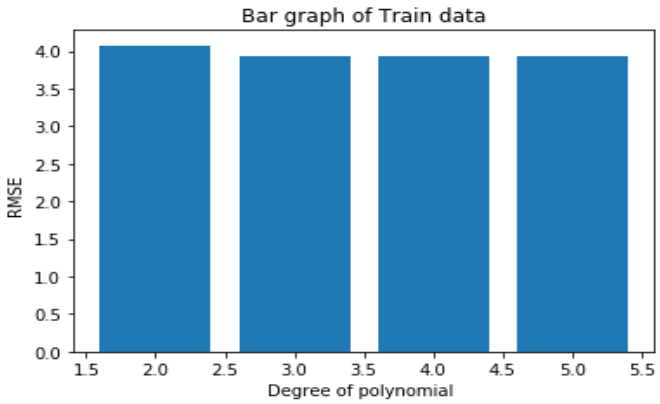

Figure 7 RMSE vs. different values of degree of polynomial (p = 2, 3, 4, 5) on the training data

Inferences:

1. By inferring from the above bar graph we can say that the value of RMSE 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 better, so the RMSE decreases.
4. From the RMSE value, degree for p=4 curve will approximate the data best.
5. As the degree increases the bias decreases and variance increases because the model starts becoming more complex resulting in fitting the data better than that by the linear regression.

b.

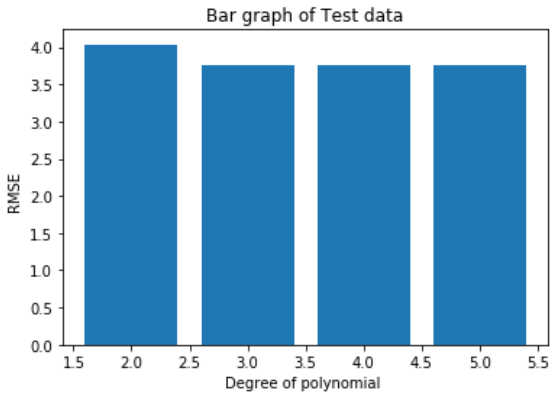

Figure 8 RMSE vs. different values of degree of polynomial (p = 2, 3, 4, 5) on the test data

Inferences:

1. By inferring from the above bar graph, we can say that the RMSE value decreases with respect to increase in degree of polynomial (p = 2, 3, 4, 5).
2. The RMSE value increase become uniform after p=2 and for p=2 to p=4 the decrease becomes gradual.
3. As the degree increases the curve fits the data better, so RMSE decreases as p value increases.
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.

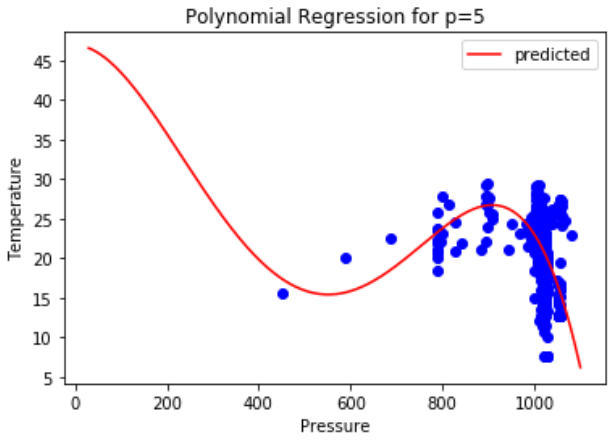

Figure 9 Pressure vs. temperature best fit curve using best fit model on the training data

Inferences:

1. The p-value is 5 corresponding to which it best fit model.
2. As the RMSE for p=5 is least then the other p-values resulting in best fitting the model. Also, 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.

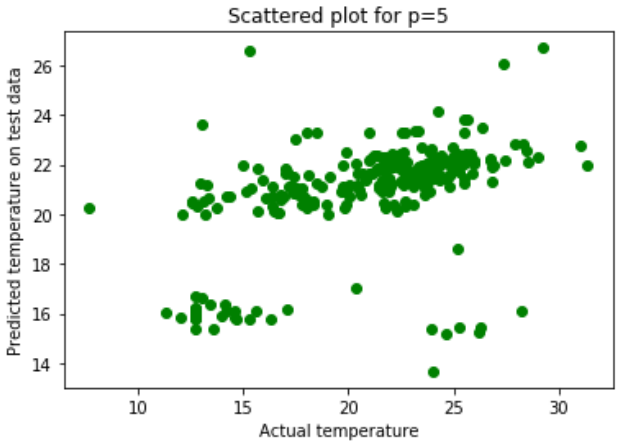

Figure 10 Scatter plot of predicted temperature from non- linear regression model vs. actual temperature on test data

Inferences:

1. Based upon the spread of the points in the above scattered plot we can infer that the accuracy of predicted temperature is quite high.
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 high.
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 as the accuracy in nonlinear regression is quit high as compare to the 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 than the linear regression leading in a higher accuracy as compare to the linear regression.
5. In linear regression bias is high and variance is low but in nonlinear regression variance is high and bias is low**.**