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: Ankit Yadav **Mobile No: 9053437219**

Roll Number: B19236 **Branch: Mech**

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	**0.92139**
4	**0.92397**
8	**0.92784**
16	**0.92526**

Inferences:

1. The highest classification accuracy is obtained with Q =.8
2. The accuracy increases from 2-8 as the no of predicted dimensions start becoming equal to the no of actual dimensions.
3. As the no of predicted dimensions start becoming equal to the no of actual dimensions the accuracy of the prediction increases.
4. As the accuracy directly depends on the of diagonal elements so we see as the accuracy increases the no. of diagonal elements also increases.
5. The no of off diagonal elements decreases.
6. As the total no of test cases remain same and so when the accuracy increases the diagonal elements increases hence the off-diagonal elements must decrease to keep the no of total test cases same.
7. Prediction accuracy is inversely proportional to the off-diagonal elements. ((FP+FN)

2

Table 2 Comparison between Classifiers based upon Classification Accuracy

S. No.	Classifier	Accuracy (in %)
1.	KNN	0.92397
2.	KNN on normalized data	0.92397
3.	Bayes using unimodal Gaussian density	0.88918
4.	Bayes using GMM	0.92784

Inferences:

1. Bayes using GMM has the highest accuracy and Bayes using unimodal Gaussian density has the lowest accuracy.
2. Ascending order of classification accuracy. Byes using Bayes using unimodal Gaussian density < KNN = KNN on normalized data < Bayes using GMM
3. Bayes using GMM assumes each class as a separate model whereas Bayes using Uni-model Gaussian density assumes the whole data as a single model hence the net accuracy of prediction of data of Bayes using GMM is more as if the complete data is not symmetrical the unimodal fails to give correct predictions.
4. KNN is slower and has lower accuracy as it is an instance-based classifier.

PART – B

1

a.

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

Inferences:

1. No, the best line does not fit the data perfectly.
2. The relationship between the Pressure and Temperature is not linear
3. The best seems to be under fitting the data and hence is highly biased also the variance is low.
4. Since it is a linear model, so we need to trade between a lower variance or a lower bias.

b.

Report the prediction accuracy on training data:

Root Mean Squared Error of Training Data: 4.279790433682601

c.

Report the prediction accuracy on testing data:

Root Mean Squared Error of Test Data: 4.286985483129509

Inferences:

1. . Training accuracy is higher.
2. As the best fit line is under fitting the training data the accuracy is actually not that good but since the under fitting is lesser that the one on testing data, so the training data is showing higher accuracy.

d.

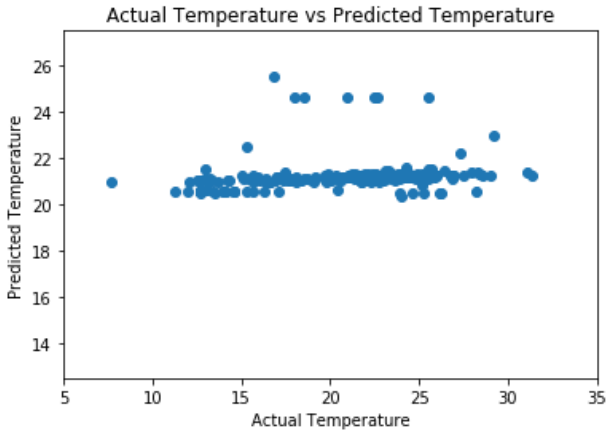

Figure 6 Scatter plot of predicted temperature from linear regression model vs. actual temperature on test data

Inferences:

1. The accuracy between the predicted value and the actual value is not that good. As the actual temperature has a larger range whereas the predicted range is limited as seen in the graph.
2. Since we used a linear model to predict the relationship between the pressure and Temperature this cusses the accuracy to suffer as the actual relation between them is not linear.

2

a.

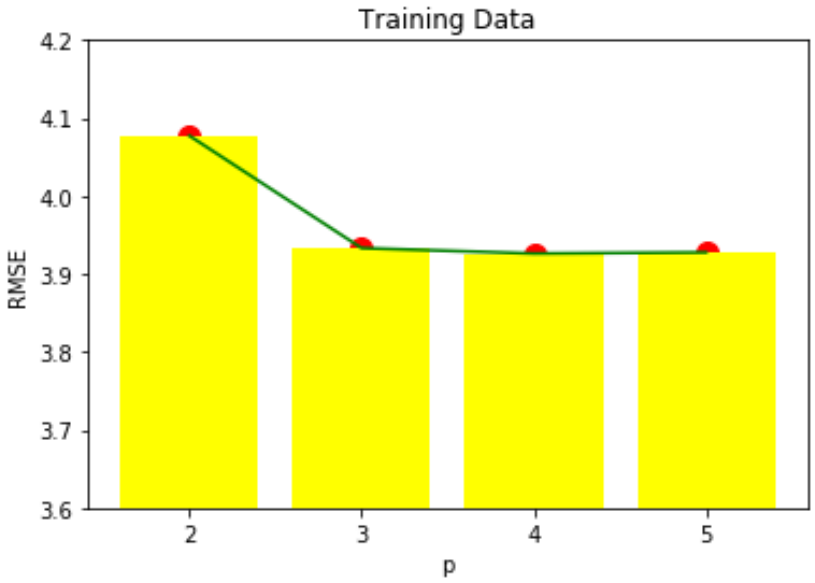

Figure 7 RMSE vs. different values of degree of polynomial (p = 2, 3, 4, 5) on the training data

Inferences:

1. There is a great decrease in RMSE value from p=2 to p=3 after that there Is slight decrease in the RMSE value from p=3 to p=5.
2. From p=2 to p=3 the decrease in RMSE value is steep after that it becomes gradual/constant.
3. As the p value approaches the actual amount no of dimensions our accuracy increases and hence the RMSE decreases.
4. As p=5 has minimum test error so the 5th degree polynomial curve will probably best approximate the curve.
5. As the degree of polynomial increases the both the bias and variance start lowering until a certain p. After that there is a balance between the two.

b.

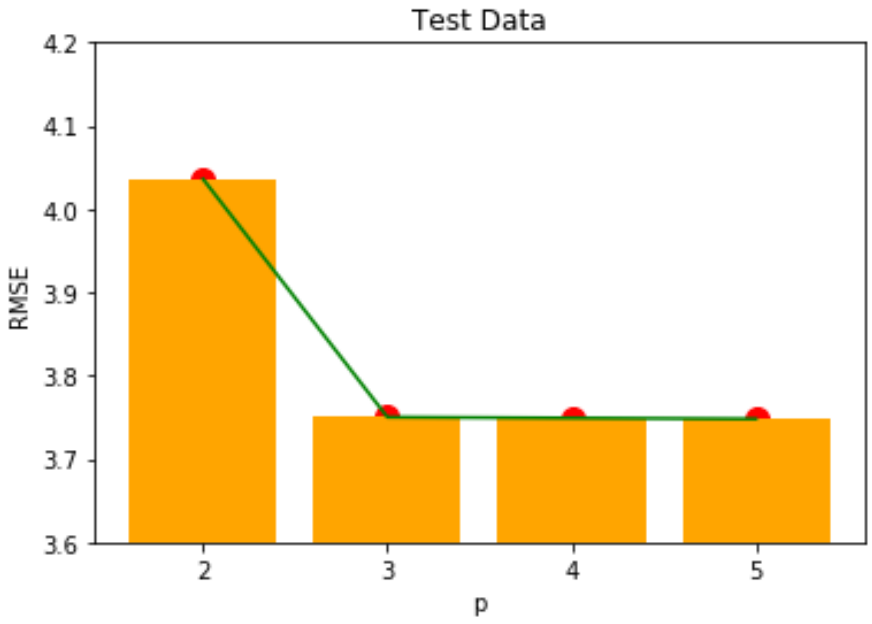

Figure 8 RMSE vs. different values of degree of polynomial (p = 2, 3, 4, 5) on the test data

Inferences:

1. There is a great decrease in RMSE value from p=2 to p=3 after that there Is slight decrease in the RMSE value from p=3 to p=5.
2. From p=2 to p=3 the decrease in RMSE value is steep after that it becomes almost zero or constant.
3. As the p value approaches the actual amount no of dimensions our accuracy increases and hence the RMSE decreases.
4. As p=5 has minimum test error so the 5th degree polynomial curve will probably best approximate the curve.
5. As the degree of polynomial increases the both the bias and variance start lowering until a certain p. After that there is a balance between the two.

c.

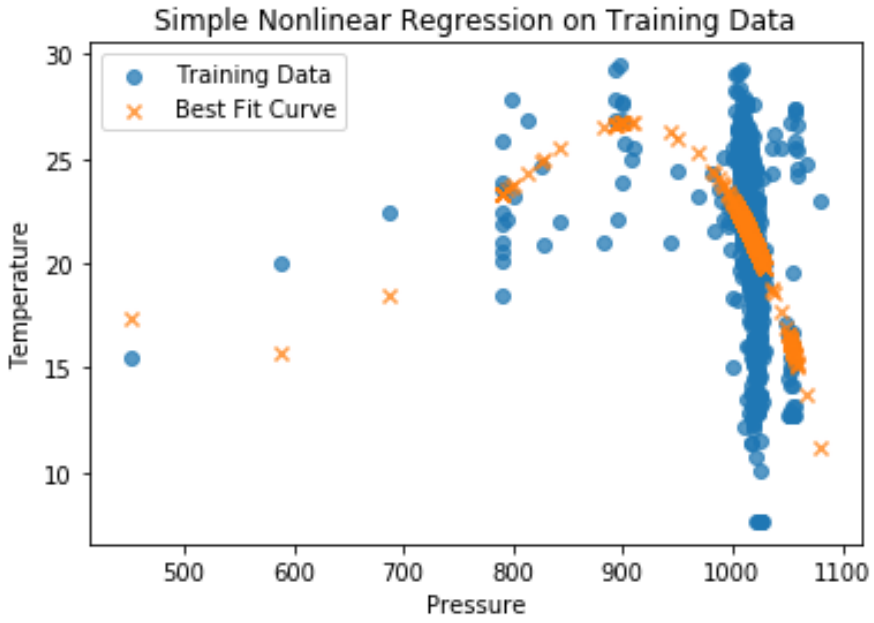

Figure 9 Pressure vs. temperature best fit curve using best fit model on the training data

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.

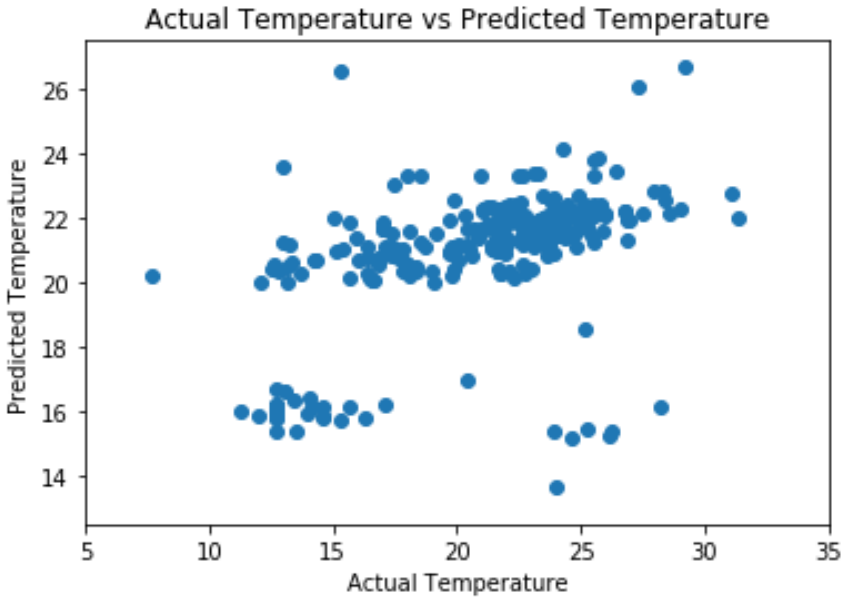

Figure 10 Scatter plot of predicted temperature from non- linear regression model vs. actual temperature on test data

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 must 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.