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: Sumit Kumar **Mobile No: 7549233722**

Roll Number: B19118 **Branch:CSE**

PART - A

1 a.

		Prediction Outcome
True Label	599	277
	126	24

Figure 1 Bayes GMM Confusion Matrix for Q = 2

		Prediction Outcome
True Label	701	38
	24	13

Figure 2 Bayes GMM Confusion Matrix for Q = 4

		Prediction Outcome
True Label	641	34
	84	17

Figure 3 Bayes GMM Confusion Matrix for Q = 8

		Prediction Outcome
True Label	713	50
	12	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	80.283
4	92.010
8	84.793
16	92.010

Inferences:

1. The highest classification accuracy is obtained with Q = 4.
2. Increasing the value of Q increases the prediction accuracy upto Q = 2, after that it decreases and then again increases.

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, the number of diagonal elements in Confusion matrix increase till Q = 4.
5. The diagonal elements represents correctly predicted values (TP + TN), therefore it increases.
6. As the classification accuracy increases with the increase in value of Q, the number of off-diagonal elements decrease.
7. The off - diagonal elements represents falsely predicted values (FP + FN), therefore it decreases.

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	92.010

Inferences:

1. The KNN classifier without normalization has the highest accuracy whereas the bayes classifier has the lowest accuracy.
2. The classifiers in ascending order of classification accuracy is :- Bayes using unimodal Gaussian density < Bayes using GMM < KNN on normalized data < KNN without normalization.
3. The Bayes classifier method is not very effective as compared to the others because Bayes method is effective for multiple class prediction and for large dataset but here we are working on less dataset with only two classes. Bayes GMM assumes that the data to have multiple modes therefore the prediction accuracy is increases.
4. KNN has lower accuracy because it is an instance-based classifier and waits until last minute before doing any model construction and very slow than other methods.

PART – B

1

a.

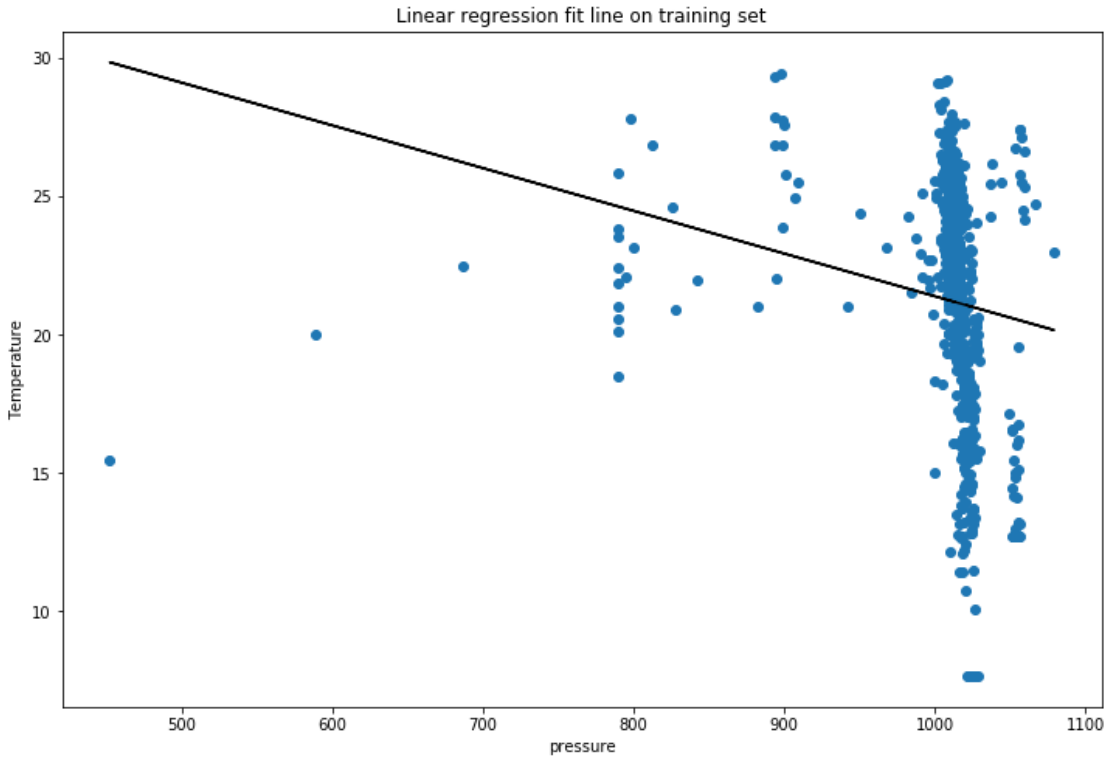


Figure 5 Pressure vs. temperature best fit line on the training data

Inferences:

1. No, the linear regression line is not fitting the data correctly.
2. Due to the presence of the outliers the slope of the linear regression line is affected and therefore most of the data points is not fitting the data correctly. Linear regression line is very sensitive towards the outliers.

3. This linear regression line is showing high bias because the data points are not fitting the line well. The regression line is underfitting the regression line. Whereas it shows low variance as the data is not overfitting.

b.

The prediction accuracy on training data is **4.279.**

c.

The prediction accuracy on testing data is **4.286.**

Inferences:

1. The prediction accuracy for the testing data is higher.
2. The best fit line is the one with least training error.

d.

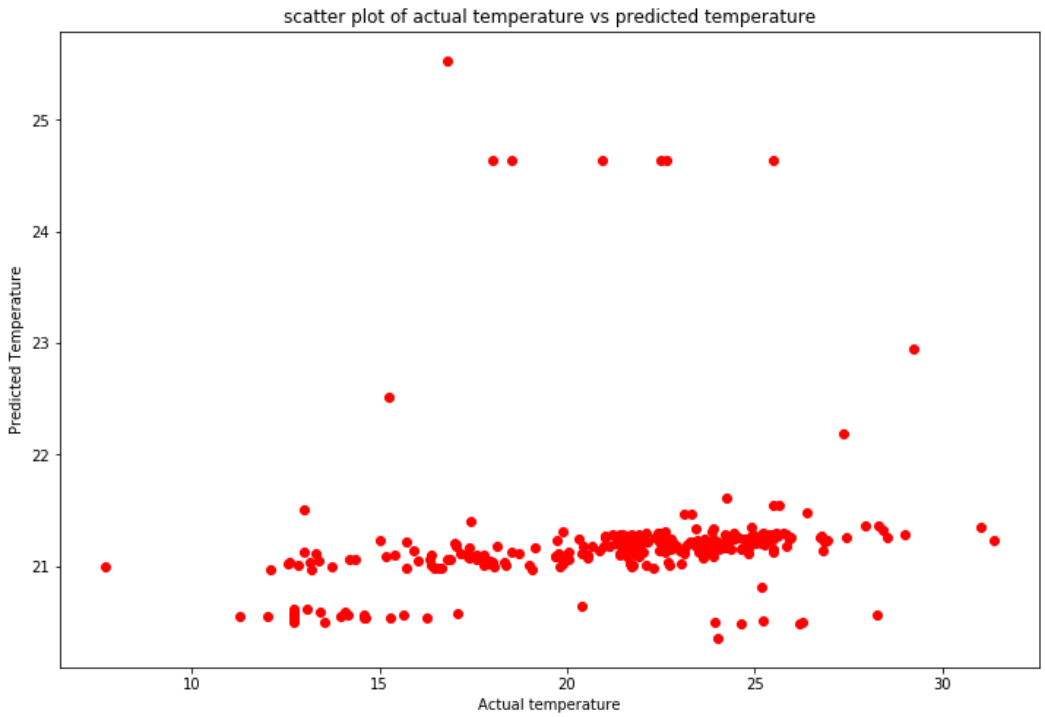
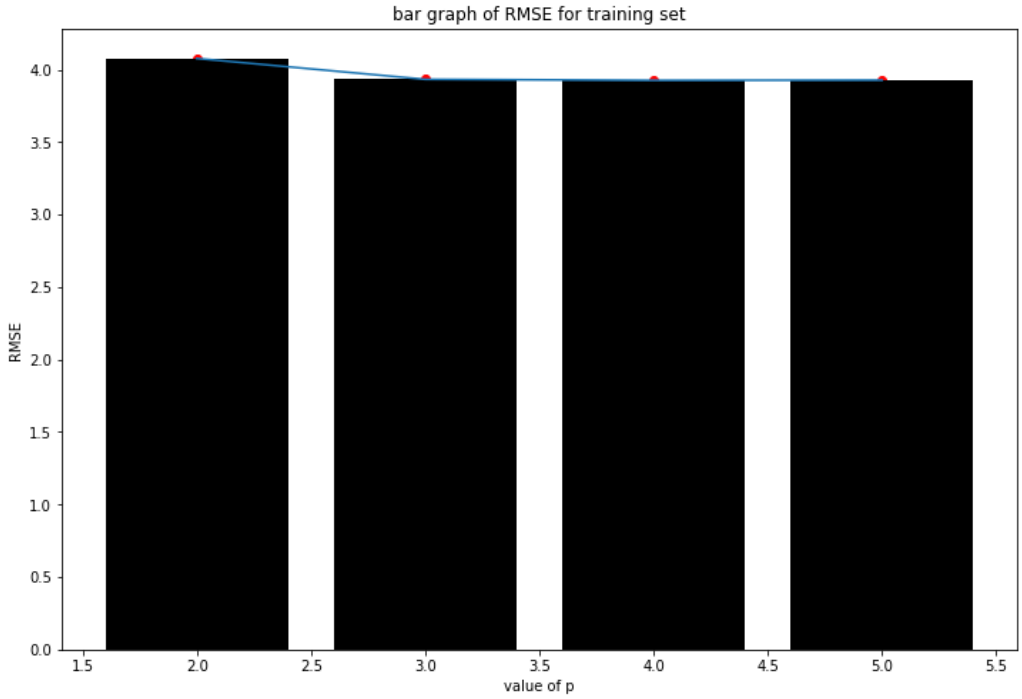


Figure 6 Scatter plot of predicted temperature from linear regression model vs. actual temperature on test data

Inferences:

1. The above scatter plot shows that the predicted temperature is not accurate.
2. This is because the data points is not following the line y = x. The predicted temperature is mostly around 21 which shows the data is not accurate. For the data to be accurate, actual temperature must be nearly equal to predicted temperature and if should follow y = x line. The correlation coefficient is also low.

2



a.

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

Inferences:

1. The RMSE value decreases from p = 2 to p = 3, after that it is nearly same.
2. After p = 3 the RMSE value is nearly same.

3. As the value of p is increasing, the estimate of the data is getting more accurate and therefore the RMSE value is decreasing.
4. Since for p = 5 or the 5^{th} degree curve, the RMSE value is least therefore the data is very well fit for p = 5.
5. The bias is not very high for the best line fit and also the variance is not so high because the data is not overfitting therefore there is a balance tradeoff between bias and variance.

b.

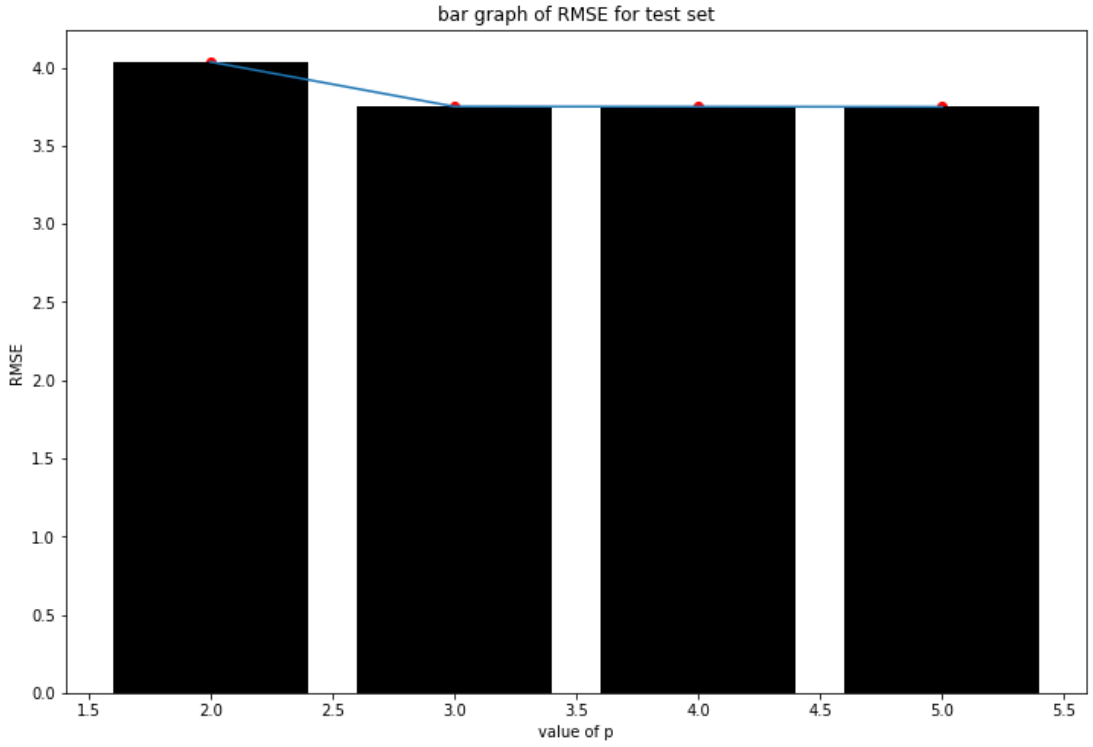
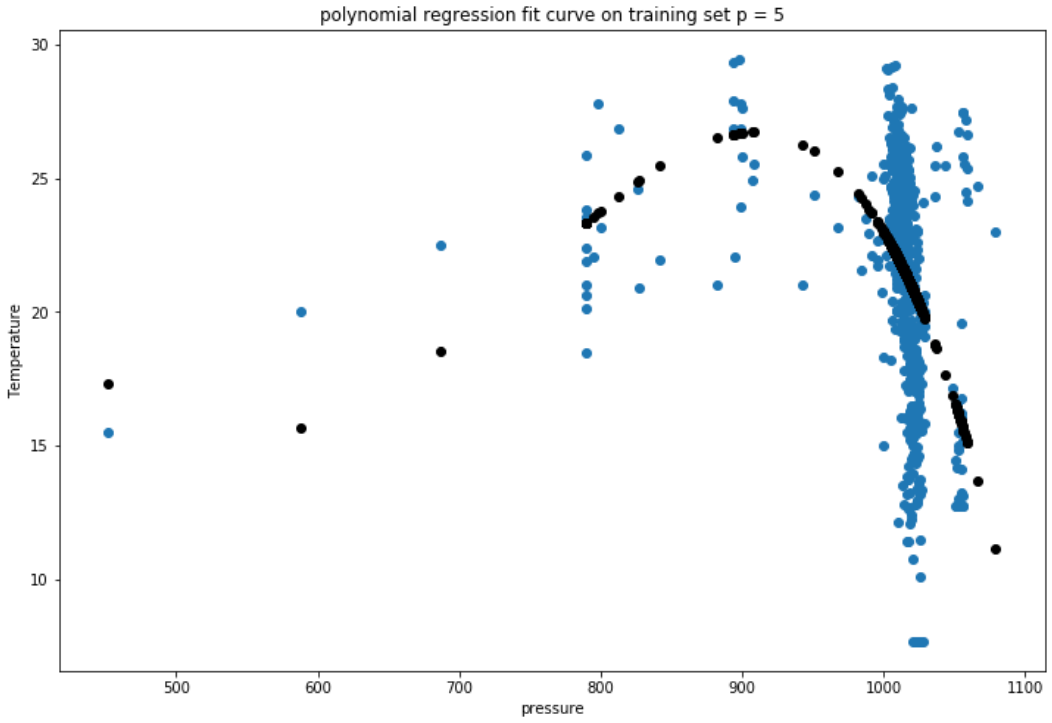


Figure 8 RMSE vs. different values of degree of polynomial (p = 2, 3, 4, 5) on the test data

Inferences:

1. The RMSE value decreases from p = 2 to p = 3, after that it is nearly same or very less decreasing.
2. After p = 3 the RMSE value is nearly same or decreasing very less.
3. As the value of p is increasing, the predicted data is getting more accurate and therefore the RMSE value is decreasing very less or nearly same.

4. Since for p = 5 or the 5th degree curve, the RMSE value is least therefore the data is very well fit for p = 5.
5. The bias is not very high for the best line fit and also the variance is not so high therefore there is a balanced tradeoff between bias and variance.



c.

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

Inferences:

1. For the best fit model on the training set, p value is 5.
2. Since for p = 5, the RMSE value for test dataset is least therefore the data is very well fit for p = 5.
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.

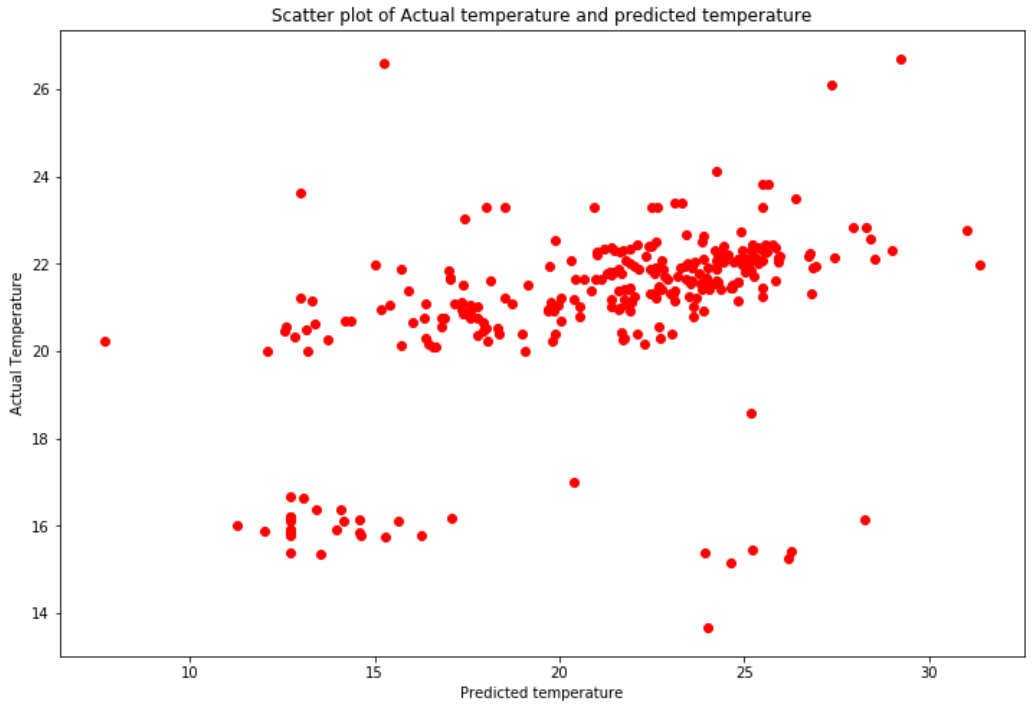


Figure 10 Scatter plot of predicted temperature from non- linear regression model vs. actual temperature on test data

Inferences:

1. The accuracy for this non – linear regression model is good based on the spread of the data.
2. The datapoints of the above scatter plot is more or less following y = x line and therefore we can say that actual temperature is nearly equal to the predicted temperature and accuracy is good. The data is following more of the polynomial relation than the linear relation.
3. The accuracy of the non – linear regression model is high as compared to that of the linear regression model.
4. The affect of the outliers in the non – regression model is less as compared to the linear regression model, therefore the data is well predicted.
5. The linear regression model is more bias because most of the datapoints are not fitting well but the non – linear regression shows more variance than the linear regression model because the datapoints is more overfitting than linear regression model.
6. The bias of the non-linear model is lower than the linear model also the overall variance for the non-linear model is low because the data is not overfitting but if we compare from the linear model the it is high.

NOTE: - Since I am using the sklearn version 0.21.3, so the result might be different because the upadated libraries and the older libraries show different result for PART-A.