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: Tarun Singla **Mobile No: 8872526396**

Roll Number: b19198 **Branch: EE**

PART - A

1 a.

		Prediction Outcome
True Label	677	44
	48	7

Figure 1 Bayes GMM Confusion Matrix for Q = 2

		Prediction Outcome
True Label	704	46
	21	5

Figure 2 Bayes GMM Confusion Matrix for Q = 4

		Prediction Outcome
True Label	709	44
	16	7

Figure 3 Bayes GMM Confusion Matrix for Q = 8

		Prediction Outcome
True Label	716	49
	9	2

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	**88.144**
4	**91.365**
8	**92.268**
16	**92.525**

Inferences:

1. The highest classification accuracy is obtained with Q = 16
2. Increase in value of Q increases the value of Accuracy.
3. On increasing the value of Q, we increase means we are assuming more number of gaussian components which is true of real life data as it in general is multimodal data

4. There is increase in diagonal elements with increase in value of Q
5. As Accuracy Corresponds percentage of correctly predicted samples and correctly predicted samples is sum of diagonal elements
6. As the classification accuracy increases with the increase in value of Q, off diagonal elements decrease
7. Off diagonal elements represents the samples that were incorrectly predicted elements so it will decrease with increase in accuracy

2

Table 2 Comparison between Classifiers based upon Classification Accuracy

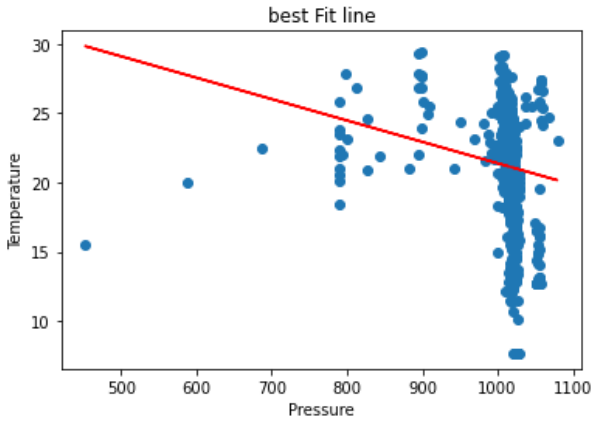
S. No.	Classifier	Accuracy (in %)
1.	KNN	88.530
2.	KNN on normalized data	91.752
3.	Bayes using unimodal Gaussian density	92.912
4.	Bayes using GMM	93.556

Inferences:

1. KNN has lowest accuracy and Bayes using GMM has highest accuracy
2. Arrange the classifiers in ascending order of classification accuracy. KNN< KNN normalized < bayes using unimodal < bayes using GMM
3. As KNN relies on distance for classification so after normalization accuracy will increase. As accuracy increased in case of unimodal so data is approximately normally distributed and as accuracy increased with GMM so one class consists of one or more gaussian components.

PART – B

1 a



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 Infer upon bias and variance trade-off for best fit line.
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.

Report the prediction accuracy on training data is **4.2797**

c.

Report the prediction accuracy on testing data is **4.2869**

Inferences:

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

d.

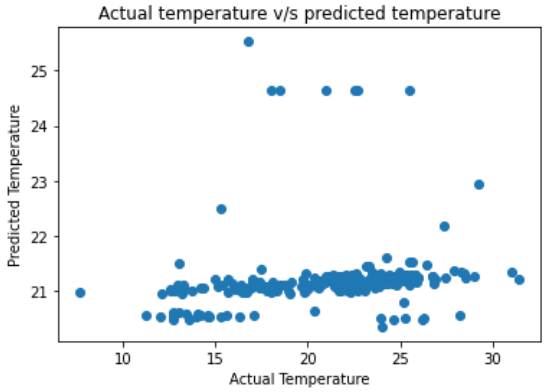


Figure 6 Scatter plot of predicted temperature from linear regression model vs. actual temperature on test data

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

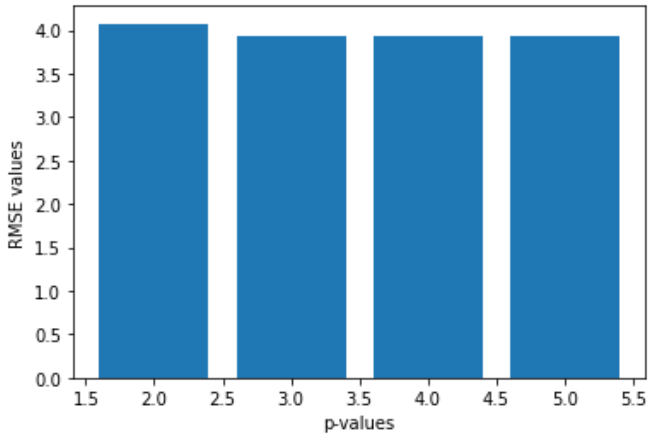


Figure 7 RMSE vs. different values of degree of polynomial (p = 2, 3, 4, 5) on the training data

Inferences:

1. The RMSE value decreases from p = 2 to p = 3, after that it is nearly same.
   After p = 3 the RMSE value is nearly same
2. As the value of p is increasing, the estimate of the data is getting more accurate and therefore the RMSE value is decreasing.
3. Since for p = 4, the RMSE value is least therefore the data is very well fit for p = 4. 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

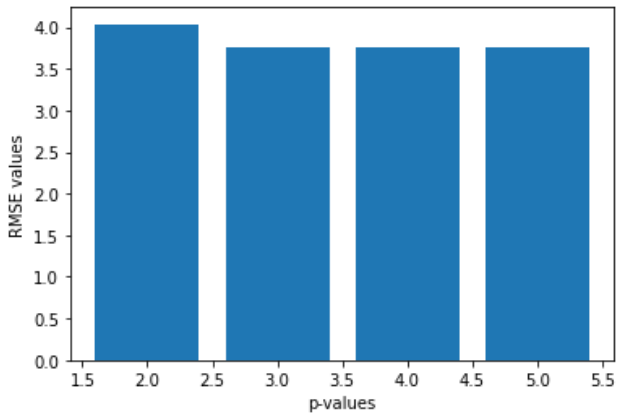


Figure 8 RMSE vs. different values of degree of polynomial (p = 2, 3, 4, 5) on the test data

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, 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 balances tradeoff between bias and variance

c.

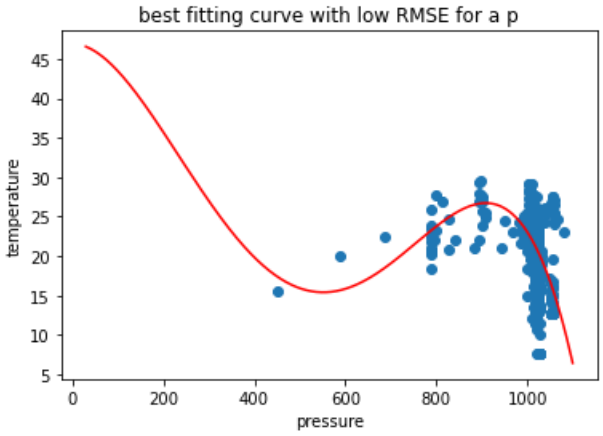


Figure 9 Pressure vs. temperature best fit curve using best fit model on the training data

Inferences:

1. For the best fit model on the training set, p value is 4.
2. Since for p = 4, the RMSE value is least for training set therefore the data is very well fit for p = 4.
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.

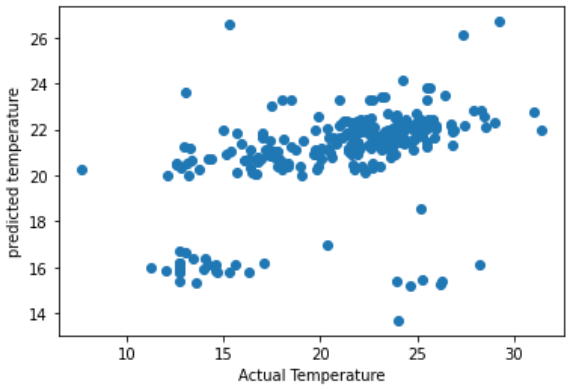


Figure 10 Scatter plot of predicted temperature from non- linear regression model vs. actual temperature on test data

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 effect 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

Guidelines for Report (Delete this while you submit the report):

- **The plot/graph/figure/table should be centre justified with sequence number and caption.**
- **Inferences should be written as a numbered list.**
- **Use specific and technical terms to write inferences.**
- **Values observed/calculated should be rounded off to three decimal places.**
- **The quantities which have units should be written with units.**
- **Please fit a confusion matrix/ table in one page only.**