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: Prashant Kumar **Mobile No: 8700350173**

Roll Number: B19101 **Branch: CSE**

PART - A

1 a.

		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 3 Bayes GMM Confusion Matrix for Q = 8

		Prediction Outcome
True Label	720	5
	49	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	**90.206**
8	**93.299**
16	**93.041**

Inferences:

Note: The output for the part A is different for different versions of scipy/sklearn. Sklearn version is 0.21.3

1. The highest classification accuracy is obtained with Q = 8.

2. Generally, on increasing the value of Q increases the prediction accuracy. But here the GMM form on the basis of random choose points due to which data show no regular trend. The value obtained may very as the GMM form on the basis of randomly chased initial points.
3. On increasing the value of Q, more cluster formed result in more accurate data points. But if the value of Q is too large than model get overfit for train data and give wrong output.
4. As the classification accuracy increases with the increase in value of Q, because the number of diagonal elements in Confusion matrix increase.
5. As the accuracy is directly proportional to the diagonal elements. Hence, the diagonal elements increase with increase in accuracy.

$$\text{Accuracy} = \frac{TP+TN}{TP+TN+FP+FN}$$

6. As the classification increases with the increase in value of Q, the number of off-diagonal elements decrease.
7. Since accuracy is directly proportional to diagonal elements. Hence on increasing off-diagonal elements, diagonal elements get decrease result in decrease in accuracy.

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	87.5
4.	Bayes using GMM	93.299

Inferences:

1. Maximum Accuracy – Bayes using GMM
   Minimum Accuracy – Bayes Classifier using unimodal
2. Bayes Classifier using Unimodal < K-NN on normalized data (K = 5) < K-NN Classifier (K = 5) < Bayes using GMM
3. Different models give different accuracy. For K-NN classifier accuracy increased with increasing value of k for some limit. And in Bayes classifier it depends upon the prior probability of class which is more biased toward a particular class.

PART – B

1

a.

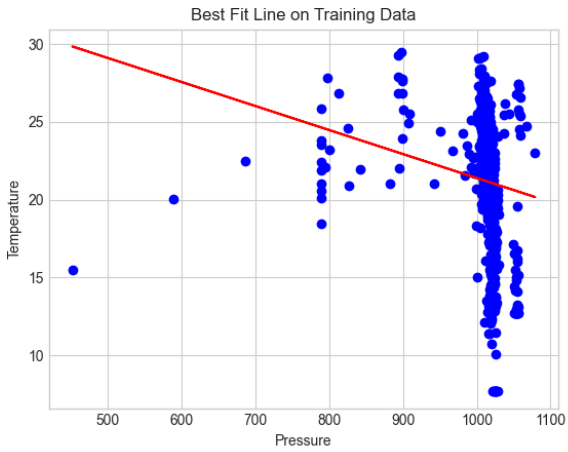

Figure 5 Pressure vs. temperature best fit line on the training data

Inferences:

1. No, the regression line that we get doesn't fit good.
2. Reason is that the points are more skewed and the correlation that we get between the pressure and temperature is very less which shows that the temperature is independent to pressure.

b.

Prediction accuracy on training data: 4.279

c.

Prediction accuracy on testing data: 4.287

Inferences:

1. Prediction accuracy of training and train data is almost same.
2. Since the temperature and pressure are independent to each other. So, variance is due to the only the weight of data that is more skewed. Hence, Temperature is evenly distributed which result in same result for train and test data.

d.

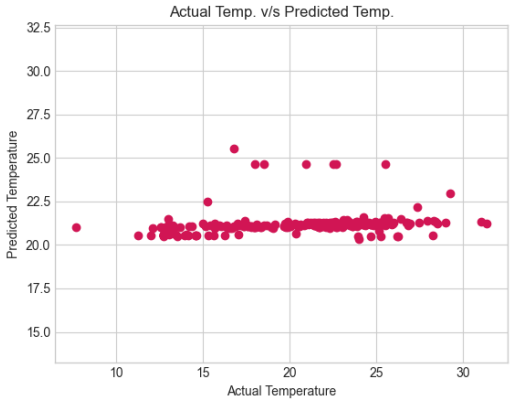

Figure 6 Scatter plot of predicted temperature from linear regression model vs. actual temperature on test data

Inferences:

1. Since, the spread of the data shows that most of the predicted value lie between 20-22.5 for actual value of range between 10-30. And also, most of the data points lie bellow to line y=x. Therefore, predicted temperate is not so accurate as it predicts almost same for all actual values
2. As the best fit curve that we get has very small slope due to which it remains almost constant or almost parallel to x-axis. Hence, the predicted value is almost same for all actual value.

2

a.

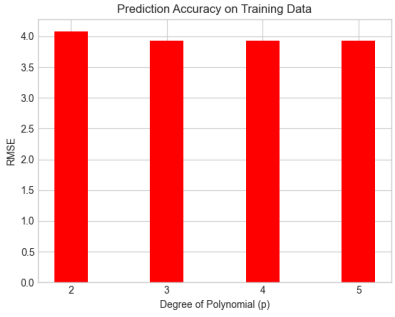

Figure 7 RMSE vs. different values of degree of polynomial (p = 2, 3, 4, 5) on the training data

Inferences:

1. RMSE value decreases with respect to increase in degree of polynomial (p = 2, 3, 4, 5).
   * P = 2 is 4.077
   * P = 3 is 3.933
   * P = 4 is 3.926
   * P = 5 is 3.927
2. After p=3 RMSE value decreases gradually.
3. As the value of p increases, the curve that we get fit more efficiently to the training data and hence, the RMSE value start getting decrease. But if the value of p increased too large then it also started capturing the noise in the data due to which the RMSE for test data start getting increase and hence model get overfit.
4. Since, for p=4 RMSE for training data is low hence, it has good prediction accuracy.

b.

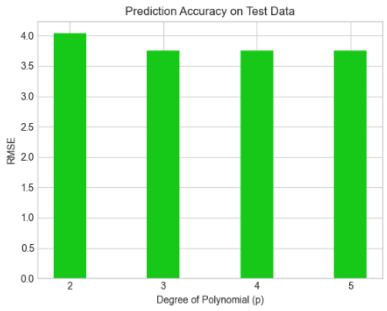

Figure 8 RMSE vs. different values of degree of polynomial (p = 2, 3, 4, 5) on the test data

Inferences:

1. RMSE value decreases with respect to increase in degree of polynomial (p = 2, 3, 4, 5).
   * P = 2 is 4.036
   * P = 3 is 3.750
   * P = 4 is 3.749
   * P = 5 is 3.748
2. After p=3 RMSE value decreases gradually.

3. As the value of p increases, curve get more fit with the training data and hence the prediction value that we get is more accurate to the actual value. But if we increase the value of p too large then the RMSE start getting increase because the curve also capturing the noise in the data and model get overfit.
4. From the RMSE value, fifth degree curve will approximate the data best.

c.

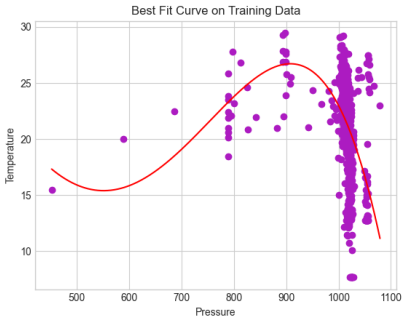

Figure 9 Pressure vs. temperature best fit curve using best fit model on the training data

Inferences:

1. For the best fit curve, degree of polynomial is 5.
2. As the value of p increased, polynomial fit more accurately and hence RMSE get decrease. But if we increase the value of p too large then the prediction accuracy start getting increase because the curve also capturing the noise in the data and model get overfit.

d.

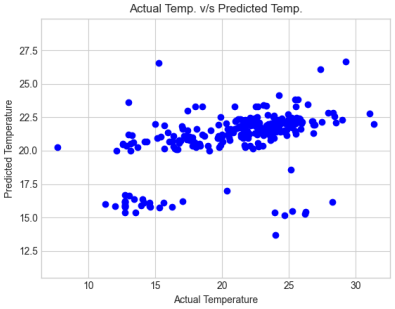

Figure 10 Scatter plot of predicted temperature from non- linear regression model vs. actual temperature on test data

Inferences:

1. Based upon the spread of the points, most of the data points is close to line y=x. Hence, the temperature is predicted more accurately.
2. As the RMSE value for the best fit curve of degree of polynomial 5 is less. Hence the prediction is more accurate.
3. In liner regression the all predicted value lie between the range of 20-22.5 and most of the predicted value is far away from y=x line. But in the polynomial regression the best fit curve that we get is lie near to line y=x. That means the prediction is more accurate in polynomial regression as compare to linear regression
4. But however, the temperature is independent to pressure the liner model performs poorly as compared to polynomial regression.