IC 272: DATA SCIENCE - III
LAB ASSIGNMENT – V

Data classification using Bayes classifier with Gaussian mixture model (GMM); regression using linear regression and polynomial curve fitting

Student's Name: Akshar Singh **Mobile No: 7428357700**

Roll Number: B20147 **Branch:** CSE

PART - A

1 a.

		Prediction Outcome
True Label	106	12
	4	215

Figure 1 Bayes GMM Confusion Matrix for Q = 2

		Prediction Outcome
True Label	111	7
	5	214

Figure 2 Bayes GMM Confusion Matrix for Q = 4

		Prediction Outcome
True Label	108	10
	6	213

Figure 3 Bayes GMM Confusion Matrix for Q = 8

		Prediction Outcome
True Label	88	30
	1	218

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	**95.252**
4	**95.845**
8	**95.252**
16	**90.801**

Inferences:

1. The highest classification accuracy is obtained with Q =4.
2. increasing the value of Q first increases prediction accuracy and then decreases it.
3. This happens because of the optimal value of Q at which our data seems to have Q number of clusters.

4. As the classification accuracy increases the number of diagonal elements in the confusion matrix increase.
5. The increase in diagonal element happens because they represent true values.
6. As the classification accuracy increases infer does the number of off-diagonal elements decrease.
7. The decrease in off-diagonal element happens because they represent false values.

2

Table 2 Comparison between Classifiers based upon Classification Accuracy

S. No.	Classifier	Accuracy (in %)
1.	KNN	89.318
2.	KNN on normalized data	97.330
3.	Bayes using unimodal Gaussian density	94.362
4.	Bayes using GMM	95.845

Inferences:

1. Highest accuracy – KNN on normalized data and Lowest accuracy - KNN
2. The classifiers in ascending order of classification accuracy. Classifier 1 < Classifier 3 < Classifier 4<Classifier 2.
3. The reason for low accuracy in KNN classifier is due to the calculation of Euclidian distance and because of that one attribute outweighs the other. The reason for low accuracy of Bayes classifier than KNN on normalized data is because it assumes normal distribution and the reason for high accuracy of Bayes using GMM than normal bayes is because it considers data coming from different cluster.

PART – B

1

a.

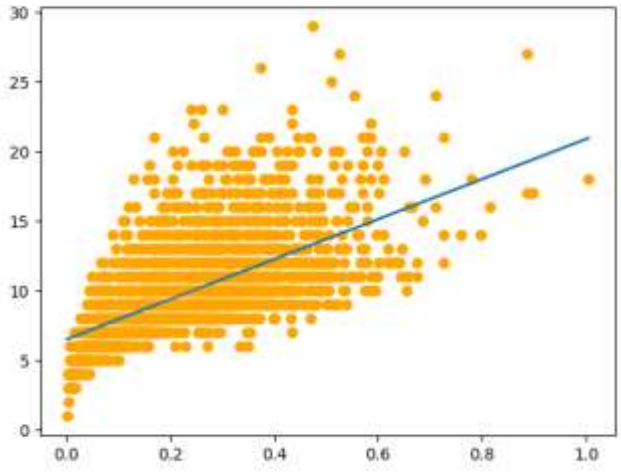


Figure 5 Univariate linear regression model: Rings vs. the chosen attribute name (replace) best fit line on the training data

Inferences:

1. The attribute with the highest correlation coefficient was used for predicting the target attribute Rings since this attribute contributes largest to the prediction analysis than any other attribute.
2. The best fit line doesn't seem to perfectly since it seems that many number of data points are below and above it.
3. Infer upon bias and variance trade-off for the best fit line.
4. Bias seems to be high because of underfitting and variance is low because small change in input doesn't lead to high change in output.

b.

Root mean square error for train data is 2.528

Root mean square error for test data is 2.468

Inferences:

1. The accuracy for testing data is higher than the training data because of less rmse error.

d.

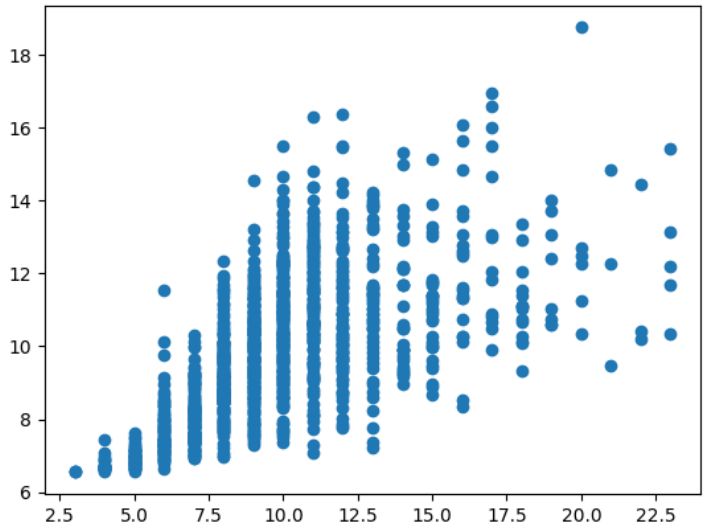


Figure 6 Univariate linear regression model: Scatter plot of predicted rings from linear regression model vs. actual rings on test data

Inferences:

1. Based upon the spread of points, the data predicted seems to be predicted not so good since there are many values of y for a value of x for some data points.

2

a.

Root mean square error for train data for 2.216182487730409

b.

Root mean square error for test data for 2.219219350663792

Inferences:

1. The accuracy for testing data is equal to the training data because of almost equal rmse error.

c.

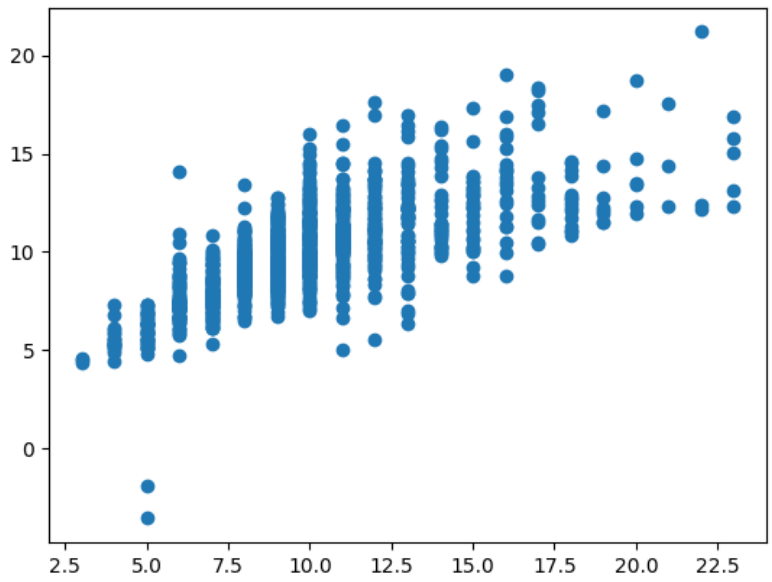


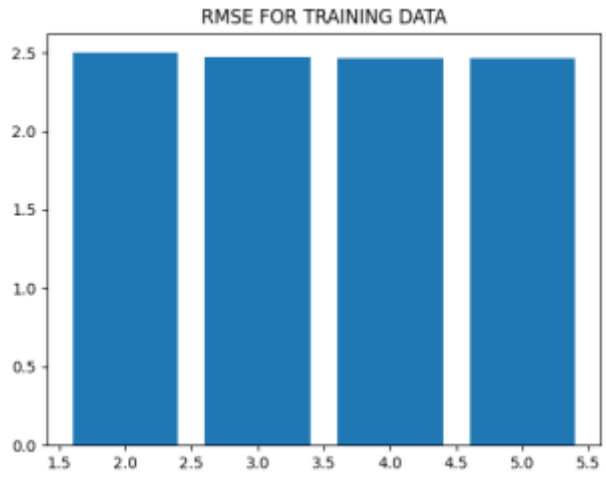
Figure 7 Multivariate linear regression model: Scatter plot of predicted rings from linear regression model vs. actual rings on test data

Inferences:

1. Based upon the spread of points, the data predicted seems to be predicted good since there are less number values of y for a value of x for data points.

2. performance of multivariate linear regression seems better than univariate linear because univariate model considers only one input variable.

3



a. **Figure 8 Univariate non-linear regression model: RMSE vs. different values of degree of polynomial (p = 2, 3, 4, 5) on the training data**

Inferences:

1. RMSE values decreases with respect to the increase in the degree of the polynomial (p = 2, 3, 4, 5)
2. The decrease after a p-value=3 the decrease becomes uniform though initially it is gradual.
3. This happens because of better fitting with increase in p value .
4. From the RMSE value, degree 5 curve will approximate the data best.
5. Bias seems to be low because of perfect-fitting for increase in value of p and variance increases with increase in value of p because small change in input lead to high change in output.

b.

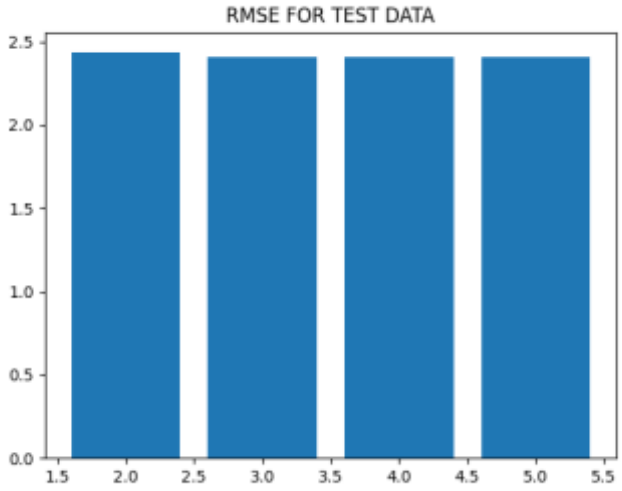


Figure 9 Univariate non-linear regression model: RMSE vs. different values of degree of polynomial (p = 2, 3, 4, 5) on the test data

Inferences:

1. RMSE values decreases with respect to the increase in the degree of the polynomial (p = 2, 3, 4, 5) with an exception of slight increase from degree 4 to5
2. The decrease after a p-value=3 the decrease becomes uniform though initially it is gradual.
3. This happens because of better fitting with increase in p value .
4. From the RMSE value, degree 4 curve will approximate the data best.
5. Bias seems to be high because of under-fitting for increase in value of p and variance increases with increase in value of p because small change in input lead to high change in output.

c.

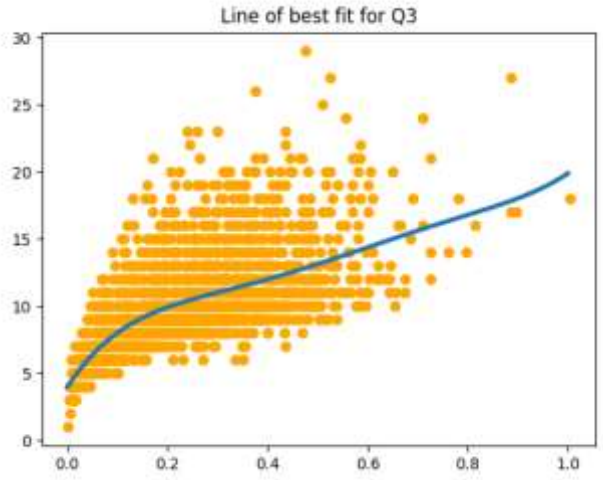


Figure 10 Univariate non-linear regression model: Rings vs. chosen attribute(replace) best fit curve using best fit model on the training data

Inferences:

1. p-value=5 corresponding to the best fit model because of least value of rmse error.
2. Bias seems to be low because of perfect-fitting for increase in value of p and variance increases with increase in value of p because small change in input lead to high change in output.

d.

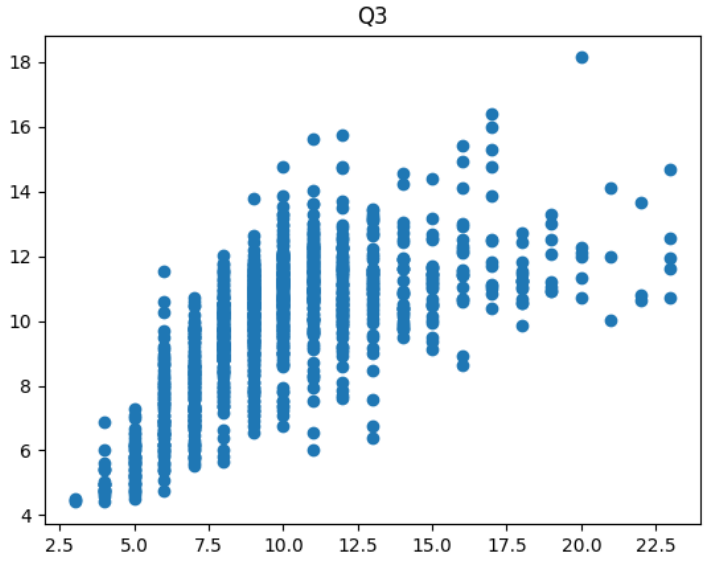


Figure 11 Univariate non-linear regression model: Scatter plot of predicted rings vs. actual rings on test data

Inferences:

1. Based upon the spread of points, the data predicted seems to be predicted not so good since there are many values of y for a value of x for some data points.
2. univariate linear is less accurate than univariate non-linear regression which is less accurate than multivariate linear regression model because univariate considers only one input variable

4

a.

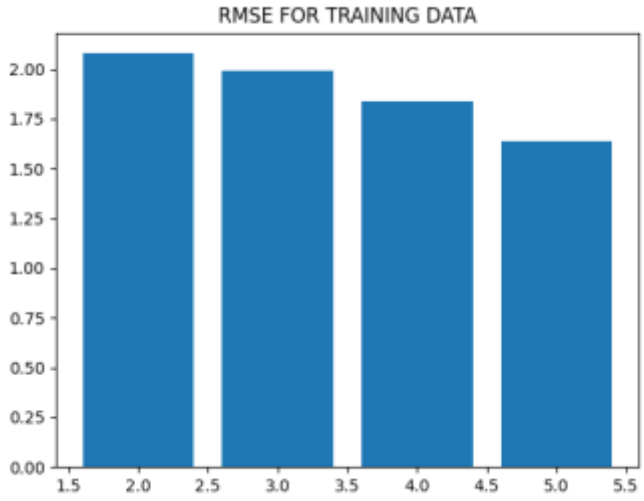


Figure 12 Multivariate non-linear regression model: RMSE vs. different values of degree of polynomial (p = 2, 3, 4, 5) on the training data

Inferences:

1. RMSE values decreases with respect to the increase in the degree of the polynomial (p = 2, 3, 4, 5)
2. The decrease is uniform.
3. This happens because of better fitting with increase in p value .
4. From the RMSE value, degree 5 curve will approximate the data best.
5. Bias seems to be low because of perfect-fitting for increase in value of p and variance increases with increase in value of p because small change in input lead to high change in output.

b.

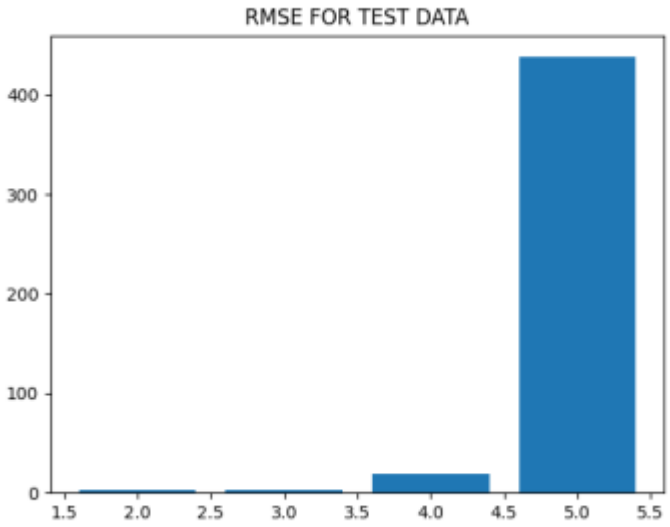


Figure 13 Multivariate non-linear regression model: RMSE vs. different values of degree of polynomial (p = 2, 3, 4, 5) on the test data

Inferences:

1. RMSE value increases with respect to the increase in the degree of the polynomial (p = 2, 3, 4, 5).
2. The increase is gradual.
3. This happens because of over fitting with increase in p value .
4. From the RMSE value, degree 2 curve will approximate the data best.
5. Bias seems to be low because of perfect-fitting for increase in value of p and variance increases with increase in value of p because small change in input lead to high change in output and in this case there is overfitting for p=4.

c.

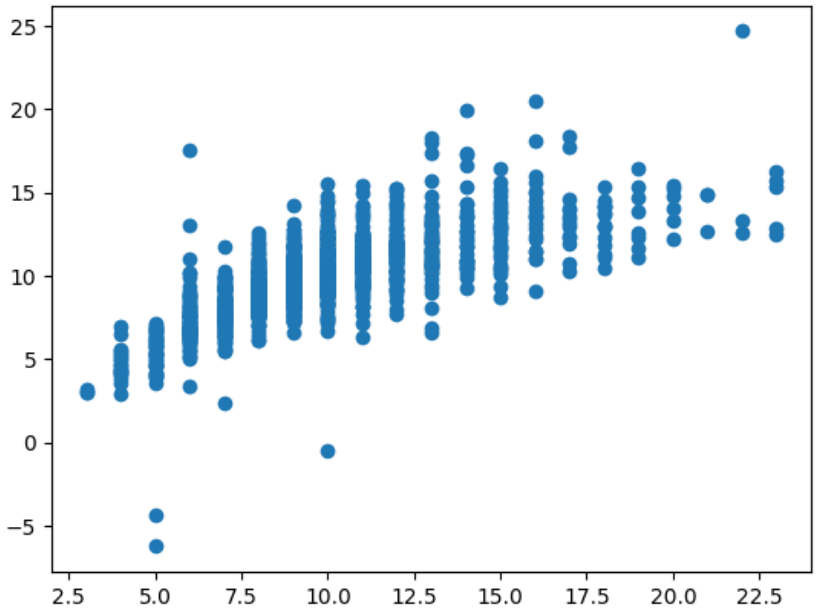


Figure 14 Multivariate non-linear regression model: Scatter plot of predicted rings vs. actual rings on test data

Inferences:

1. Based upon the spread of points, the data predicted seems to be predicted good since there are less number values of y for a value of x for data points.
2. Compare and contrast univariate linear, multivariate linear, univariate non-linear and multivariate non-linear regression model based upon the accuracy of predicted temperature value and spread of data points in Scatter Plot
3. Univariate linear is less accurate than univariate non-linear regression which is less accurate than multivariate linear regression model which is less than multivariate non-linear regression model because univariate considers only one input variable and non-linear model fits much better than linear model.
4. Inference based upon bias and variance trade-off between linear and non-linear regression models.
5. Bias seems to be high for linear and low for increase in value of p and variance increases with increase in value of p because small change in input lead to high change in output.