



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: Amit Maindola

Mobile No: +91 7470985613

Roll Number: B20079

Branch: CSE

PART - A

1 a.

	Prediction Outcome	
True Label	105	13
	6	213

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

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

	Prediction Outcome	
True Label	107	11
	6	213

Figure 3 Bayes GMM Confusion Matrix for Q = 8

	Prediction Outcome	
True Label	103	15
	3	216

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	94.362
4	96.439
8	94.955
16	94.659



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

Inferences:

1. The highest classification accuracy is obtained with $Q = 4$.
2. Increasing the value of Q increases the prediction accuracy first and then starts decreasing.
3. This happens because adding nodes with less weight causes the model to overfit on training data hence the accuracy decreases.
4. As the classification accuracy increases with the increase in value of Q the number of diagonal elements in the confusion matrix increase.
5. Increase in accuracy means more correct predictions and less wrong predictions, thus increasing true positive and true negative frequencies.
6. As the classification accuracy increases with the increase in value of Q the number of off-diagonal elements decrease.
7. Increase in accuracy means more correct predictions and less wrong predictions, thus decreasing false positive and false negative frequencies.

2

Table 2 Comparison between Classifiers based upon Classification Accuracy

S. No.	Classifier	Accuracy (in %)
1.	KNN	89.614
2.	KNN on normalized data	97.329
3.	Bayes using unimodal Gaussian density	94.658
4.	Bayes using GMM	96.439

Inferences:

1. KNN on normalized data and KNN have the highest and lowest accuracy respectively.
2. The classifiers in ascending order of classification accuracy: $KNN < \text{Bayes using unimodal Gaussian Density} < \text{Bayes using GMM} < \text{KNN on normalized data}$.
3. KNN performs better when data is normalized because, the attributes on a bigger scale can no longer overpower and influence the results in their favour. This happens because Euclidean Distance is the total absolute distance along various axes and doesn't consider for the different ranges. Also in the above example which involves just 2 clusters, KNN will give more accurate predictions than Bayes. Multimodal Bayes performs better as we are now using multiple clusters which increases the relative accuracy.

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

PART - B

1
a.

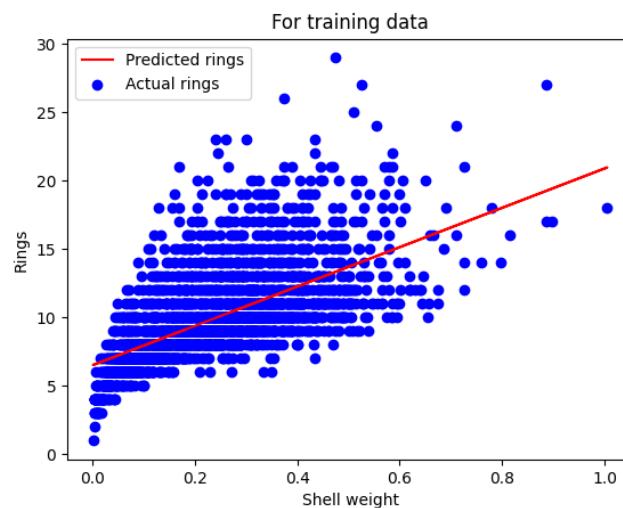


Figure 5 Univariate linear regression model: Rings vs. the Shell weight best fit line on the training data

Inferences:

1. The attribute with the highest correlation coefficient is used for predicting the target attribute Rings because the target attribute is likely to be more dependent on attribute having highest correlation coefficient.
2. No, the best fit line fit the training data perfectly.
3. It doesn't fit the training data perfectly because it is oversimplified for the data, a more complex curve is needed to fit the data.
4. The bias is high and variance is low for the best fit line.

b.

The RMSE of predicted data is 2.528 .

c.

The RMSE of predicted data is 2.468.

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

Inferences:

1. Amongst training and testing accuracy, training accuracy is higher.
2. It is because we have trained the model on the training dataset so it predicts better for the training dataset.

d.

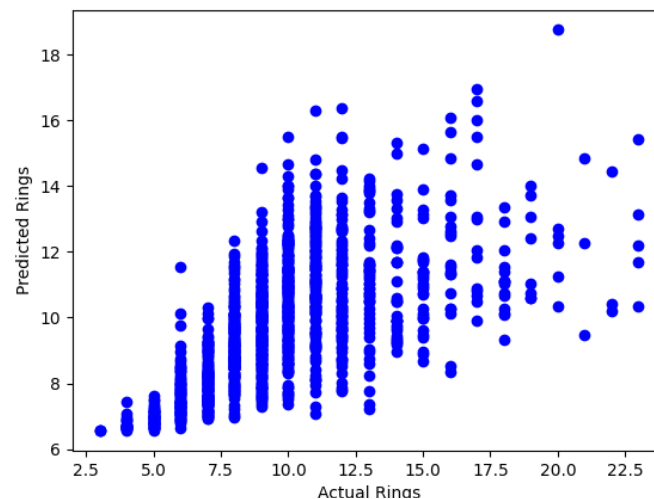


Figure 6 Univariate linear regression model: Scatter plot of predicted rings from linear regression model vs. actual rings on test data

Inferences:

1. Based upon the spread of the points, the predicted number of rings is not very accurate
2. Because the spread of actual rings is 2-23 while that of predicted is 6-20.

2

a.

The prediction accuracy on training data is 2.216.

b.

The prediction accuracy on testing data is 2.219.

Inferences:

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

3. Amongst training and testing accuracy, testing accuracy is almost same as training accuracy.
4. It is because we have trained the model on the training dataset so it predicts better or same for the training dataset.

C.

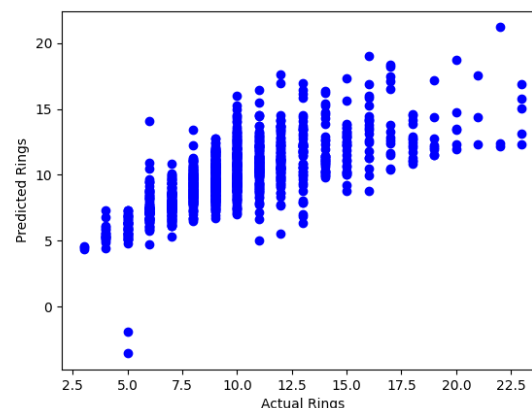


Figure 7 Multivariate linear regression model: Scatter plot of predicted rings from linear regression model vs. actual rings on test data

Inferences:

1. Based upon the spread of the points, the predicted number of rings is high.
2. The spread of Actual Rings is 5-23 and that of Predicted Rings is 4.8-22.
3. The univariate linear regression doesn't perform as good as multivariate linear regression.

3

a.

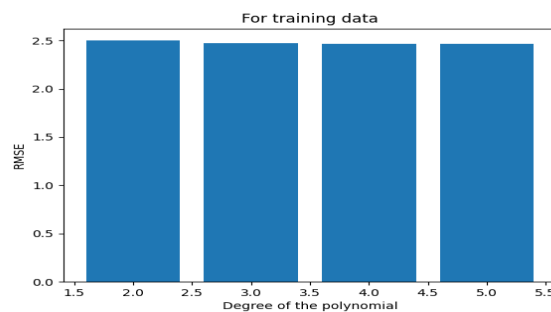


Figure 8 Univariate non-linear regression model: RMSE vs. different values of degree of polynomial ($p = 2, 3, 4, 5$) on the training data

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

Inferences:

1. RMSE values decreases with respect to the increase in the degree of the polynomial.
2. The decrease is more from 2 to 3 and the gradual.
3. As the degree increases the curve fits the data more better so RMSE decreases.
4. From the RMSE value, $p=5$ curve will approximate the data best.
5. As the degree increases, the bias decreases and variance increases.

b.

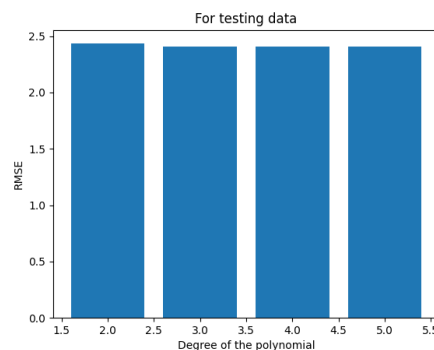


Figure 9 Univariate non-linear regression model: RMSE vs. different values of degree of polynomial ($p = 2, 3, 4, 5$) on the test data

Inferences:

1. RMSE value decreases with respect to the increase in the degree of the polynomial ($p = 2, 3, 4, 5$).
2. The decrease is more from 2 to 3 and after that its gradual.
3. As the degree increases the curve fits the data more better so RMSE decreases.
4. From the RMSE value, $p=4$ degree curve will approximate the data best.
5. As the degree increases, the bias decreases and variance increases.

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

c.

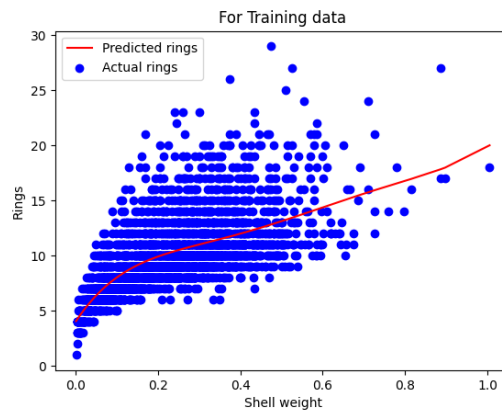
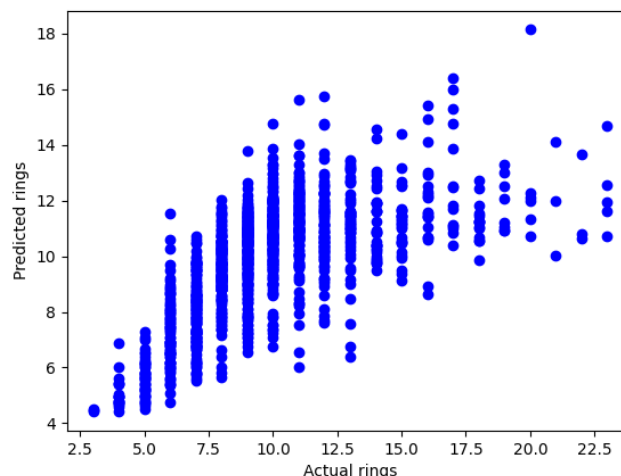


Figure 10 Univariate non-linear regression model: Rings vs. Shell weight best fit curve using best fit model on the training data

Inferences:

1. The p-value corresponding to the best fit model is 4.
2. Because it fits the data more and has more variance.
3. The bias decreases and variance increases with increasing value of p.

d.



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

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 the points, the predicted temperature is quite accurate.
2. The spread of actual rings is 3-23 while that of predicted rings is 4-20.
3. The accuracy for Univariate non-linear is the highest closely followed by multivariate Linear model and least is for univariate linear model.
4. RMSE values for non-linear regression model is lower than that of linear models hence it is better.
5. In linear regression models bias is high, variance is low and in non-linear regression models bias is low, variance is high.

4
a.

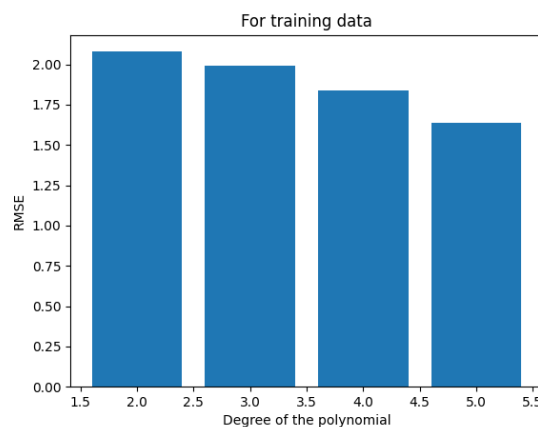


Figure 12 Multivariate non-linear regression model: RMSE vs. different values of degree of polynomial ($p = 2, 3, 4, 5$) on the training data

Inferences:

1. RMSE value decreases with respect to the increase in the degree of the polynomial ($p = 2, 3, 4, 5$).
2. The decrease is more or less uniform but after $p=4$ the decrease is more.
3. As the degree increases the curve fits the data more better so RMSE decreases.

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

4. From the RMSE value, $p=5$ degree curve will approximate the data best.
5. The bias decreases and variance increases with respect to the increase in the degree of the polynomial ($p = 2, 3, 4, 5$).

b.

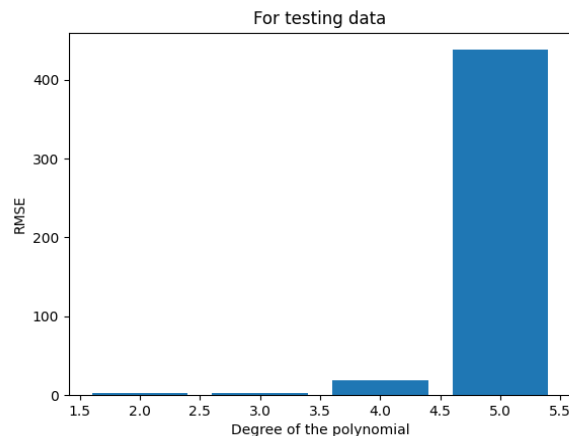


Figure 13 Multivariate non-linear regression model: RMSE vs. different values of degree of polynomial ($p = 2, 3, 4, 5$) on the test data

Inferences:

1. Infer whether RMSE value decreases with respect to the increase in the degree of the polynomial and starts increasing after $p=3$.
2. The decrease is uniform till $p=3$ but after $p=3$ the increase is much more.
3. As we increased the degree of polynomial our model became overfitted.
4. From the RMSE value, $p=2$ curve will approximate the data best.
5. The bias gradually decreases till $p=3$ and then suddenly increases after $p=3$ and the variance increases as the model becomes more complex with increasing degree of polynomial.

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

C.

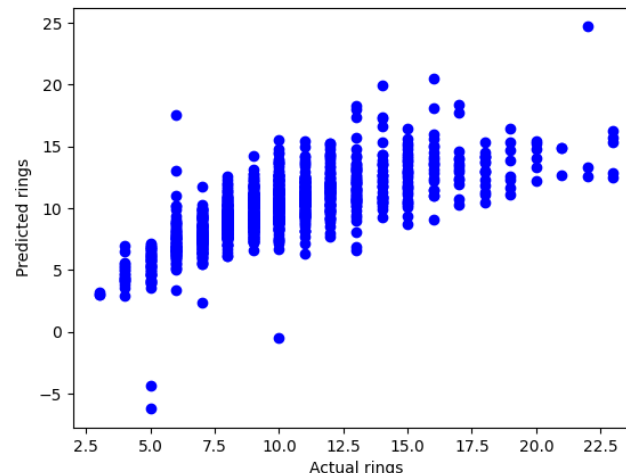


Figure 14 Multivariate non-linear regression model: Scatter plot of predicted rings vs. actual rings on test data

Inferences:

1. Based upon the spread of the points, the predicted rings is quite accurate.
2. The spread of actual rings is 3-23 and that of predicted rings is also 3-22.
3. The multivariate non-linear regression model has the highest accuracy followed by univariate non-linear model and the accuracy of multivariate linear is less than that of univariate non-linear model but more than univariate linear regression model.
4. RMSE values for non-linear regression model is lower than that of linear models hence it is better and more complex models fit our data better so multivariate models are better than univariate.
5. In linear regression models bias is high, variance is low and in non-linear regression models bias is low, variance is high