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: Shubham Shukla **Mobile No: 8317012277**

Roll Number: B20168 **Branch: CSE**

PART - A

1 a.

		Prediction Outcome
True Label	117	1
	28	191

Figure 1 Bayes GMM Confusion Matrix for Q = 2

		Prediction Outcome
True Label	116	2
	22	197

Figure 2 Bayes GMM Confusion Matrix for Q = 4

		Prediction Outcome
True Label	112	6
	17	202

Figure 3 Bayes GMM Confusion Matrix for Q = 8

		Prediction Outcome
True Label	98	20
	6	213

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	**91.4**
4	**92.9**
8	**93.2**
16	**92.3**

Inferences:

1. The highest classification accuracy is obtained with Q =.8
2. Increasing the value of Q first increases the prediction accuracy and then decreases.
3. As we were adding more Q values it is suitable till a point else we will add some less weighted data which will take our accuracy score reduced.

4. As the classification accuracy increases with the increase in value of Q infer that the number of diagonal elements in the confusion matrix increase.
5. Since the accuracy increases means we have more correct prediction, hence the diagonal elements increases in value.
6. As the classification accuracy increases with the increase in value of Q infer that the number of off-diagonal elements decrease.
7. Since the accuracy increases means we are more going towards actual, so the wrong prediction decreases result decrement in off-diagonal values.

2

Table 2 Comparison between Classifiers based upon Classification Accuracy

S. No.	Classifier	Accuracy (in %)
1.	KNN	89.583
2.	KNN on normalized data	97.024
3.	Bayes using unimodal Gaussian density	83.04
4.	Bayes using GMM	93.2

Inferences:

1. Highest accuracy is 97.024 and lowest one is 83.04.
2. The classifiers in ascending order of classification accuracy. Bayes using unimodal Gaussian density< KNN< Bayes using GMM< KNN on normalized data.
3. KNN preforms better when data is normalized because the attbitutes values range is not the matter now which takes our result more to real one. Since taking distance is more physical one and directly like you are doing by seeing graph which is not in the bayes case. Bayes using GMM is much better because it uses the real modal predictions since initially we were assuming only one mode but this is not the case in real scenerio.

PART – B

1

a.

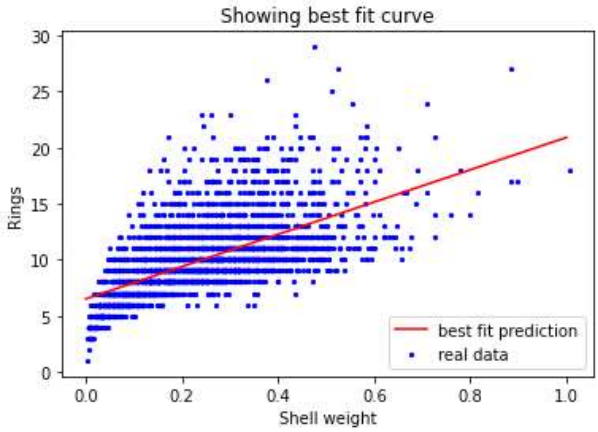

Figure 5 Univariate linear regression model: Rings vs. the chosen attribute name (replace) best fit line on the training data

Inferences:

1. Since the highest correlation is infers that the output attribute most depends on this attribute and also this refers that this is not much correlated to other attributes(means independent), highly correlated attribute is taken for the output.
2. It doesn't fits the data totally but it fits to the most part of the data accurately.
3. As we are taking a linear relation as well univariate model but the real one requires much more complicated curve.
4. The bias is more (on the basis of probability) while variance is less which is property for the best fit line.

b.

The prediction accuracy on training data : 2.528.

c.

The prediction accuracy on testing data : 2.468.

Inferences:

1. Accuracy of testing is high.
2. As the testing data is less so it is quite obvious that we will get better model to predict less valued data.

d.

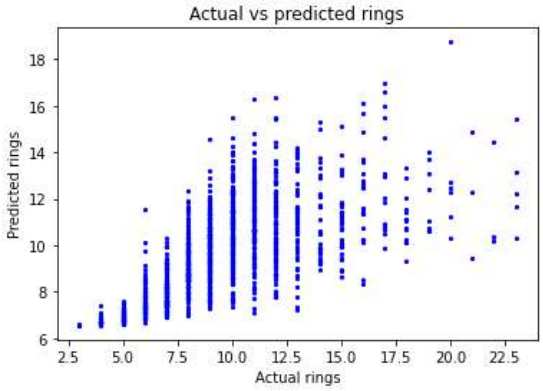

Figure 6 Univariate linear regression model: Scatter plot of predicted rings from linear regression model vs. actual rings on test data

Inferences:

1. The predicted rings are having high variability.
2. As we can see the the range of actual and predicted one is very difference with a large gap

2

a.

The prediction accuracy on training data : 2.216.

b.

The prediction accuracy on testing data : 2.219.

Inferences:

3. Amongst training and testing accuracy, testing data accuracy is high.
4. As the testing data is less so it is quite obvious that we will get better model to predict less valued data.

c.

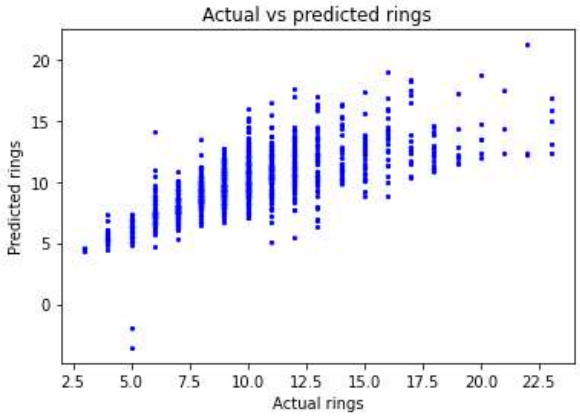

Figure 7 Multivariate linear regression model: Scatter plot of predicted rings from linear regression model vs. actual rings on test data

Inferences:

1. As the data is shifted above as so we can say that the predicted rings are more than the actual one.
2. Due to the range of the rings data in training and testing.
3. The performance of multivariate model is much better than univariate.

3

a.

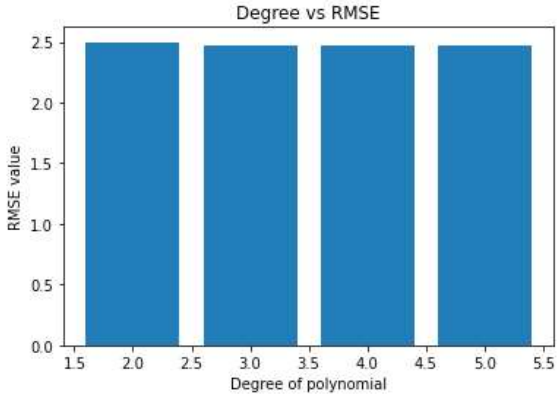

Figure 8 Univariate non-linear regression model: RMSE vs. different values of degree of polynomial (p = 2, 3, 4, 5) on the training data

Inferences:

1. RMSE value decreases very slightly with respect to the increase in the degree of the polynomial (p = 2, 3, 4, 5).
2. The decreament is uniform.
3. Since the real world data is much more complex as well having complicated graph so as we increase the p value our model will go towards more actual data.
4. Higher degree (here 5) curve is better than lower one.
5. As we increase the degree then bias will decrease and variance will increase due to more complicated data.

b.

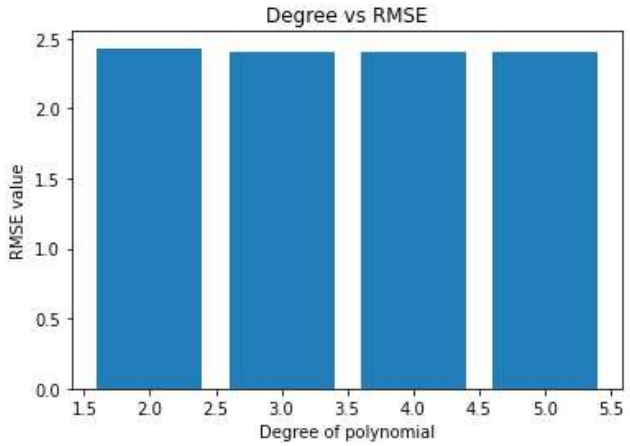

Figure 9 Univariate non-linear regression model: RMSE vs. different values of degree of polynomial (p = 2, 3, 4, 5) on the test data

Inferences:

1. RMSE value decreases very slightly with respect to the increase in the degree of the polynomial (p = 2, 3, 4, 5).
2. The decrement is uniform.

3. Since the real world data is much more complex as well having complicated graph so as we increase the p value our model will go towards more actual data.
4. Higher degree(here 5) curve will approximate the data best.
5. Bias decreases and variance increases with respect to the increase in the degree of the polynomial (p = 2, 3, 4, 5) as we are more complicating our curve.

c.

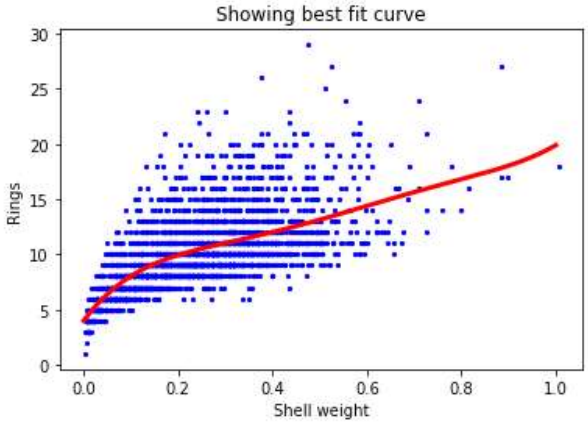

Figure 10 Univariate non-linear regression model: Rings vs. chosen attribute(replace) best fit curve using best fit model on the training data

Inferences:

1. P-value is 5 corresponding to the best fit model in our case.
2. Seeing the data we get that it is more complicated and can't be given best approximation on lower values of p, that's why we can see the uniform decrement in the rmse values as the p value increases.
3. Bias decreases and variance increases with respect to the increase in the degree of the polynomial (p = 2, 3, 4, 5).

d.

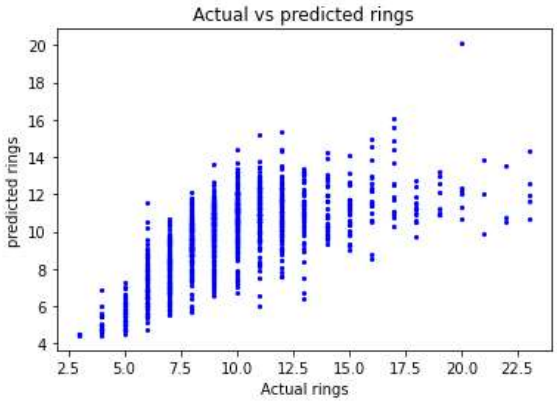

Figure 11 Univariate non-linear regression model: Scatter plot of predicted rings vs. actual rings on test data

Inferences:

1. The predicted rings are lower than the actual.
2. The range values for rings attribute of the testing data is higher than the training data.
3. Univariate Non-linear is having high accuracy and then followed by the multivariate linear and then univariate linear regression model .
4. As in non-linear data we are complicating our prediction which is more near towards the real world data, similarly in multivariate linear we are complicating our data wrt univariate linear, so again we are going towards the real data, that's why this trend.
5. Bias is high and variance is low in linear regression and while in non-linear regression the trend is opposite.

4

a.

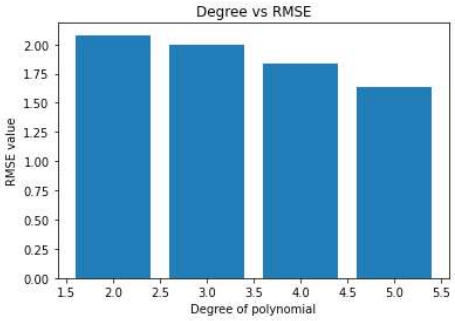

Figure 12 Multivariate non-linear regression model: RMSE vs. different values of degree of polynomial (p = 2, 3, 4, 5) on the training data

Inferences:

1. RMSE value decreases with respect to the increase in the degree of the polynomial (p = 2, 3, 4, 5).
2. RMSE values decreases uniformly as p increases.
3. Since increase in degree makes the data more accurately fit the real data.
4. P = 5 will fit the best here for prediction.
5. The bias decreases and variance increases wrt increases in the p values.

b.

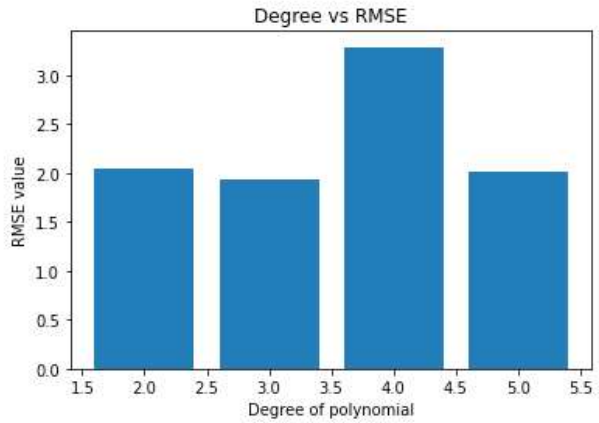

Figure 13 Multivariate non-linear regression model: RMSE vs. different values of degree of polynomial (p = 2, 3, 4, 5) on the test data

Inferences:

1. RMSE value first decreases then increases agin achieving a peak again decreases with respect to the increase in the degree of the polynomial (p = 2, 3, 4, 5).
2. Increase/decrease is non-uniform as after a certain p-value the increase/decrease becomes gradual.
3. Since as we add more attribute relation some are low weighted while others are high which will gives an unpredicted RMSE values.
4. In this case best is for p = 3.
5. May decrease or increase one can't comment on seeing the graph.

c.

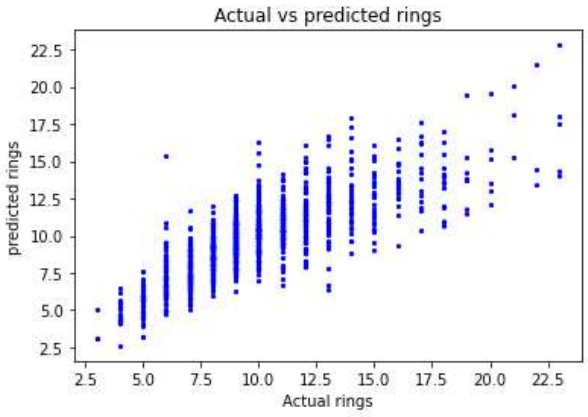

Figure 14 Multivariate non-linear regression model: Scatter plot of predicted rings vs. actual rings on test data

Inferences:

1. Here we get our model is quite accurate while predicting the rings.
2. Now our data has become more real as well the ranges matches in the training data with test data.
3. Munivariate non-linear has highest accuracy then followed by univariate non-linear, multivariate linear and lastly univariate linear regression model.
4. It' s all about how real your prediction is, and we know that the real data is much more complex as well complicated tha't why we get the above order.
5. Bias is high and variance is low in linear regression and while in non-linear regression the trend is opposite.