**Student’s Name: Vision Aggarwal**

**Roll Number: B20171**

**Mobile No: 6376029099**

**Branch:DSE**

**PART - A**

# a.

|  |  |  |
| --- | --- | --- |
|  | **Prediction Outcome** | |
| **True Label** | 93 | 15 |
| 4 | 224 |

Figure 1 Bayes GMM Confusion Matrix for Q = 2

|  |  |  |
| --- | --- | --- |
|  | **Prediction Outcome** | |
| **True Label** | 91 | 17 |
| 3 | 225 |

Figure 2 Bayes GMM Confusion Matrix for Q = 4

|  |  |  |
| --- | --- | --- |
|  | **Prediction Outcome** | |
| **True Label** | 94 | 14 |
| 4 | 224 |

Figure 3 Bayes GMM Confusion Matrix for Q = 8

|  |  |  |
| --- | --- | --- |
|  | **Prediction Outcome** | |
| **True Label** | 76 | 32 |
| 1 | 227 |

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 | **94.345** |
| 4 | **94.048** |
| 8 | **94.643** |
| 16 | **90.179** |

# Inferences:

1. The highest classification accuracy is obtained with Q =8
2. Generally, the increase in value of q leads to more accuracy unless it shows overfitting.
3. Value of q shows the no. of clusters formed during partitioning which can surely increase the accuracy.
4. As the classification accuracy increases with the increase in value of Q, the number of diagonal elements increases.
5. Because increased accuracy means more elements are predicted correctly.
6. As the classification accuracy increases with the increase in value of Q, the number of off-diagonal elements decrease.
7. Because increased accuracy means less elements are predicted incorrectly.

Table 2 Comparison between Classifiers based upon Classification Accuracy

|  |  |  |
| --- | --- | --- |
| **S. No.** | **Classifier** | **Accuracy (in %)** |
|  | KNN | 89.614 |
|  | KNN on normalized data | 97.329 |
|  | Bayes using unimodal Gaussian density | 94.362 |
|  | Bayes using GMM | 94.643 |

# Inferences:

1. KNN on normalized data have max accuracy while KNN have least accuracy.
2. KNN < Bayes using unimodal Gaussian density < GMM < Normalized KNN
3. Both Gaussian model and normalized KNN have very high accuracy but simple KNN have very less accuracy because KNN involves Euclidian distance and wherever distance is involved, we should normalize or standardize the data.

**PART – B**

**a.**

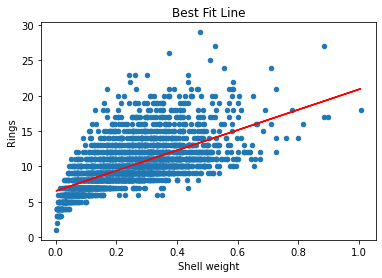


Figure 5 Univariate linear regression model: Rings vs. the chosen attribute name (replace) best fit line on the training data

**Inferences:**

1. The attribute with the highest correlation coefficient was used for predicting the target attribute Rings because we want to use the attribute which is most related to the target attribute.
2. The best fit line doesn’t fit the data perfectly, because any line can’t fit all the points as they are not on a single line.
3. Both variance and bias of the data should not be very high as high bias shows underfitting of data and high variance shows overfitting of data.

**b.**

The rmse for training data is 2.528

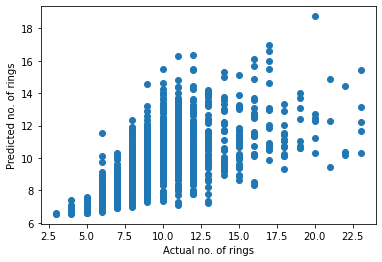
**c.**

The rmse for testing data is 2.468

**Inferences:**

1. Testing accuracy is higher for the training dataset.
2. It is higher because our model was trained for training 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. Based upon the spread of the points, the predicted number of rings is not very accurate.

Because the spread if actual rings is 2-23 while that of predicted is 6-20.

**a.**

Prediction accuracy on training data:- 2.216

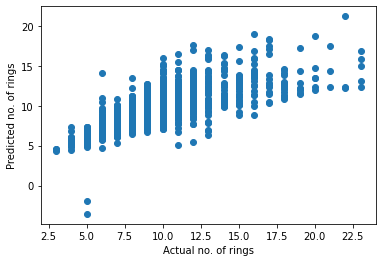
**b.**

Prediction accuracy on testing data:- 2.219

**Inferences:**

1. Amongst training and testing accuracy, training accuracy is higher.
2. Because our model was trained on training data in the first place.

**c.**

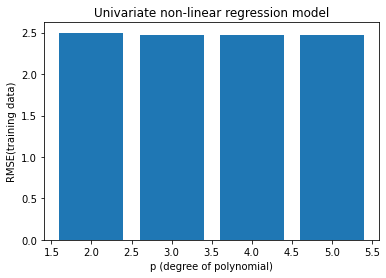


**Figure 7 Multivariate linear regression model: Scatter plot of predicted rings from linear regression model vs. actual rings on test data**

**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 4.8-22Compare and contrast the performance of univariate linear with multivariate linear regression.
3. The univariate linear regression doesn’t perform as good as multivariate linear regression.

**a.**

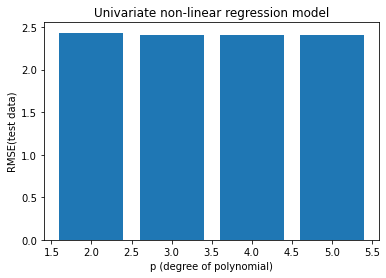


**Figure 8 Univariate non-linear regression model: RMSE vs. different values of degree of polynomial** **(p = 2, 3, 4, 5) on the training data**

**Inferences:**

1. RMSE values slightly decrease as the value of p increase.
2. Decrease is high for p = 2 to p = 3 but afterwards it becomes nearly zero.
3. As the degree increases the curve fits the data more better so RMSE decreases.
4. From the RMSE value, p=5 curve will fit best.
5. As the degree increases, the bias decreases and variance increases.

**b.**

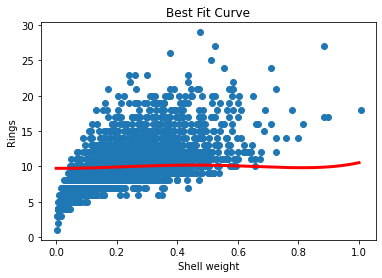


**Figure 9 Univariate non-linear regression model: RMSE vs. different values of degree of polynomial** **(p = 2, 3, 4, 5) on the test data**

**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.

**c.**

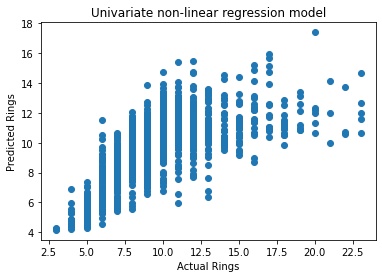


**Figure 10 Univariate non-linear regression model: Rings vs. chosen attribute(replace) best fit curve using best fit model on the training data**

**Inferences:**

1. 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.**



**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.

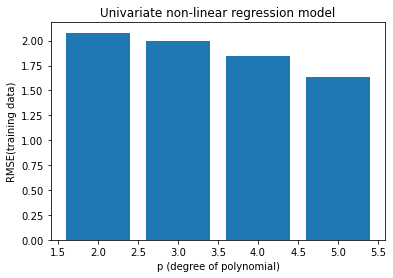
2. The spread of actual rings is 3-23 while that of predicted rings is 4-20.

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.

**a.**

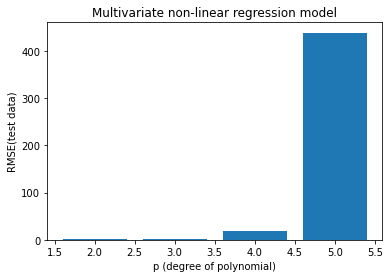


**Figure 12 Multivariate non-linear regression model: RMSE vs. different values of degree of polynomial** **(p = 2, 3, 4, 5) on the training data**

**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 uniform but after p=4 the decrease is more.
3. As the degree increases the curve fits the data better, so RMSE decreases.
4. From the RMSE value, p=5, degree curve will approximate the data best.
5. The bias decreases and variance increase 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.

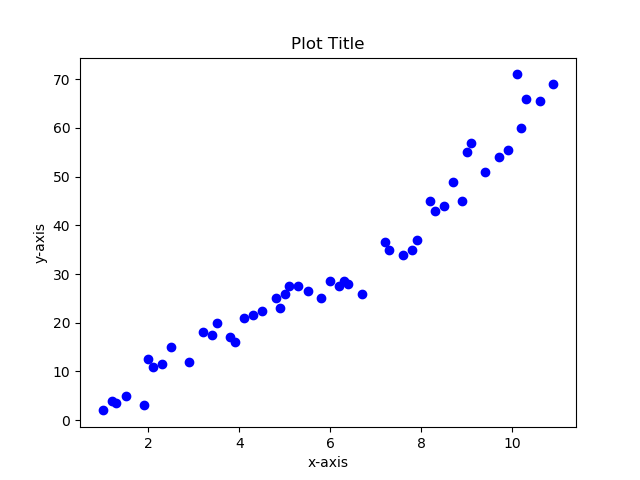
2. The decrease is uniform till p=3 but after p=3 the increase is much more.

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 increase as the model becomes more complex with increasing degree of polynomial.

**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