Linear and Quadratic Discriminant Analysis Model

Nsubuga Emmanuel Reagan

2023-04-20

#ANALYSING THE WINE DATASET USING “LDA” AND “QDA”

From the dataset, we have 13 continuous measurements made on 178 bottles of wine, where each measurement is the amount of a different compound/element in the wine. We also have a single categorical variable, Class, which tells us which vineyard the bottle comes from.

We the n plot the data to see how the compounds vary between vineyards. The resulting plot is shown below:

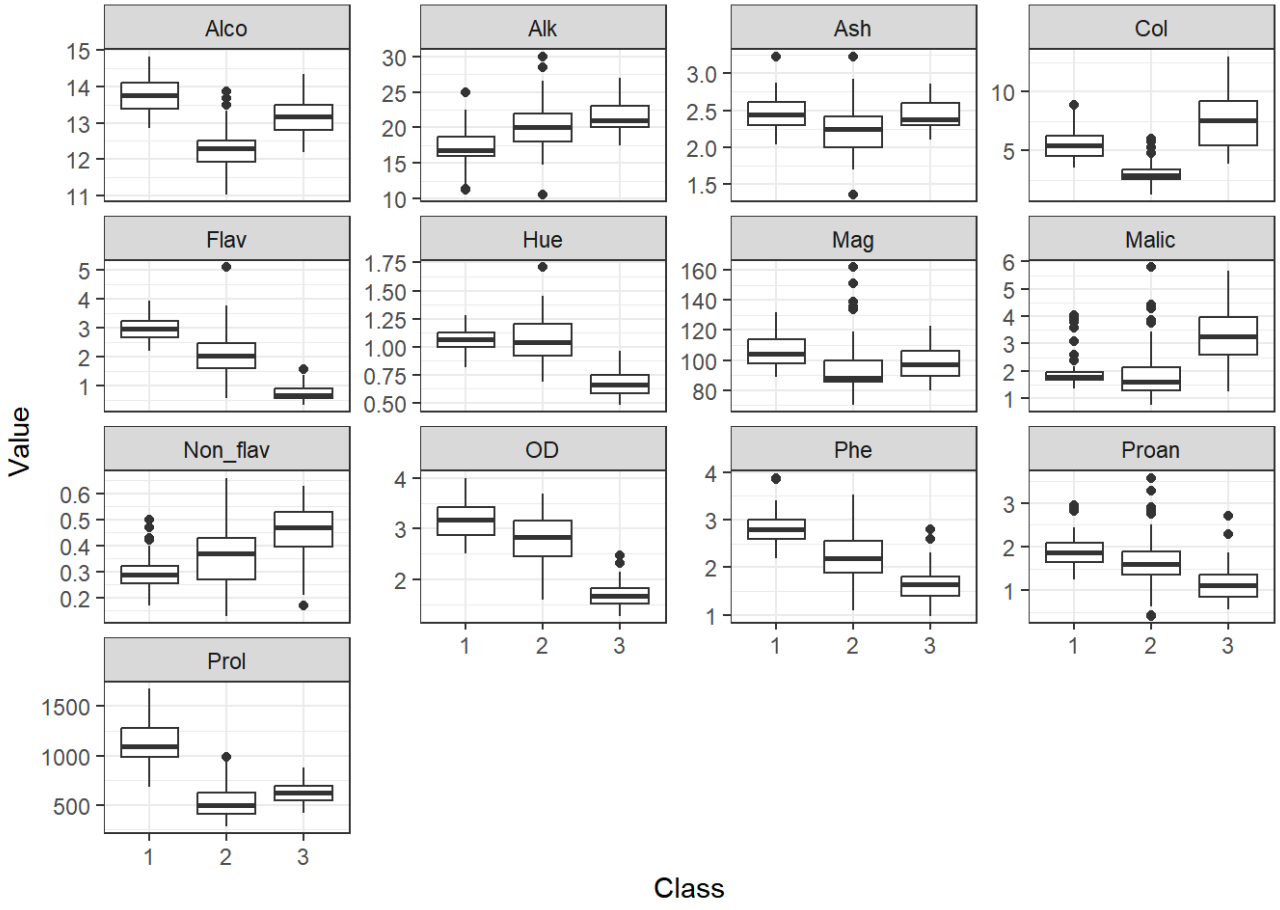


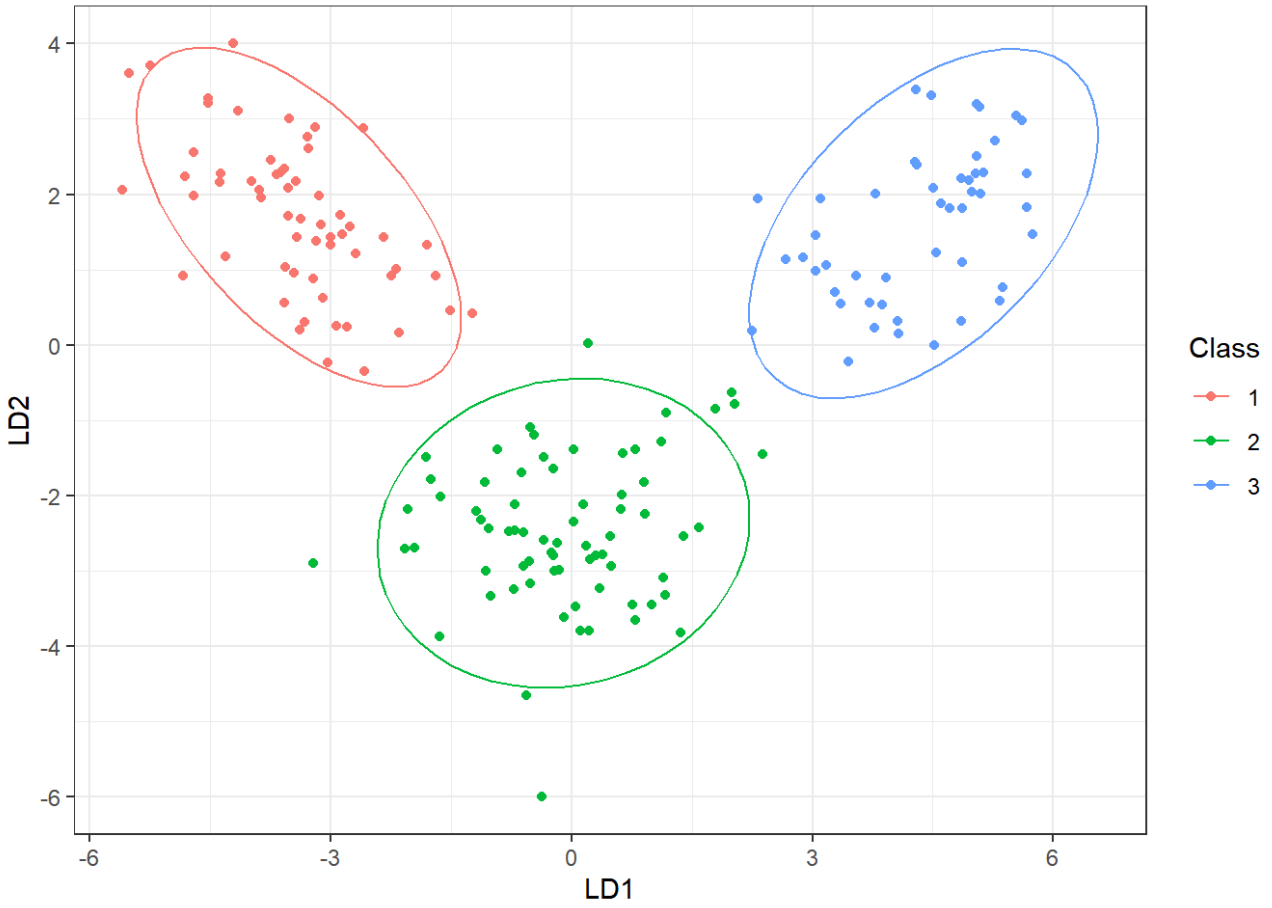
Figure 1.0 Box and whiskers plots of each continuous variable in the data against vineyard number. For the box and whiskers, the thick horizontal line represents the median, the box represents the interquartile range (IQR), the whiskers represent the Tukey range (1.5 times the IQR above and below the quartiles), and the dots represent data outside of the Tukey range. The classes look very separable.

Creating the task and learner and training the model.

We extract the model information using the getLearnerModel() function, and get DF values for each case using the predict() function. By printing head(ldaPreds), we can see that the model has learned two DFs, LD1 and LD2, and that the predict() function has indeed returned the values for these functions for each case in our wineTib dataset.

```
##         LD1       LD2
## 1 -4.700244 1.9791383
## 2 -4.301958 1.1704129
## 3 -3.420720 1.4291014
## 4 -4.205754 4.0028715
## 5 -1.509982 0.4512239
## 6 -4.518689 3.2131376
```

We plot the two DFs against each other to visualise how well they separate the bottles of wine from the three vineyards. The resulting plot is shown in the figure below:



Plotting the DFs against each other. The values for LD1 and LD2 for each case are plotted against each other, shaded by their class. The LDA has reduced our 13 predictor variables into just two DFs that do an excellent job of separating the wines from each of the vineyards.

#Cross-validating the LDA and QDA models.

```
## mmce.test.mean  acc.test.mean
##     0.01107374     0.98892626
```

```
## mmce.test.mean  acc.test.mean
##    0.008329678    0.991670322
```

Our LDA model correctly classified 98.8% of wine bottles on average. There isn't much room for improvement here, but our QDA model managed to correctly classify 99.2% of cases!

Confusion matrix for LDA

```
## Relative confusion matrix (normalized by row/column):
##         predicted
## true     1           2           3           -err.-
##   1      1e+00/1e+00 3e-04/3e-04 0e+00/0e+00 3e-04
##   2      8e-03/9e-03 1e+00/1e+00 1e-02/2e-02 2e-02
##   3      0e+00/0e+00 9e-03/6e-03 1e+00/1e+00 9e-03
##   -err.-       0.009       0.006       0.020 0.01
##
##
## Absolute confusion matrix:
##         predicted
## true        1    2    3 -err.-
##   1      2949    1    0      1
##   2        28 3473   49     77
##   3         0   21 2379     21
##   -err.-   28   22   49     99
```

Our LDA model misclassifies more cases from vineyard 2 as from vineyard 3 than as from vineyard 1.

QDA confusion matrix

```
## Relative confusion matrix (normalized by row/column):
##         predicted
## true     1           2           3           -err.-
##   1      0.994/0.984 0.006/0.005 0.000/0.000 0.006
##   2      0.014/0.016 0.986/0.993 0.000/0.000 0.014
##   3      0.000/0.000 0.003/0.002 0.997/1.000 0.003
##   -err.-       0.016       0.007       0.000 0.008
##
##
## Absolute confusion matrix:
##         predicted
## true        1    2    3 -err.-
##   1      2932   18    0     18
##   2        48 3502    0     48
##   3         0    8 2392      8
##   -err.-   48   26    0     74
```

Our QDA model is better at identifying wines from vineyard 3. It misclassified 12 as from vineyard 2, whereas the LDA model misclassified 23.

Using the model to make predictions on new data.

```
## Prediction: 1 observations
## predict.type: response
## threshold:
## time: 0.00
##   response
## 1        1
```

The model predicts that the poisoned bottle came from vineyard 1