hw2

Xingwen Wei

February 20, 2021

Question 1

We assume the conditional distribution for companies that issue a dividend or not are normally distributed. We made the following assumptions:

$$X_1|Yes, \cdots, X_n|Yes \sim iidN(10, 36)$$

$$X_1|No, \cdots, X_n|No \sim iidN(0, 36)$$

$Y_1, \cdots, Y_n$ = Yes or No are bournulli variables with $P(Y = Yes) = 0.8$.

$$\begin{aligned} P(Y = Yes|X = 4) &= \frac{P(Y = Yes) \times P(X = 4|Y = Yes)}{P(Y = Yes) \times P(X = 4|Y = Yes) + P(Y = No) \times P(X = 4|Y = No)} \\ &= \frac{0.8 \times e^{-\frac{6^2}{72}}}{0.8 \times e^{-\frac{6^2}{72}} + 0.2 \times e^{-\frac{4^2}{72}}} = \frac{0.8 \times 0.61}{0.8 \times 0.61 + 0.2 \times 0.80} = 0.75 \end{aligned}$$

Question 2

a) Let X_1 = hours studied and X_2 = undergrad GPA

$$P(Y = c_1|X = x) = \frac{e^{-6+0.05\times X_1+X_2}}{1 + e^{-6+0.05\times X_1+X_2}}$$

$$P(Y = c_2|X = x) = \frac{1}{1 + e^{-6+0.05\times X_1+X_2}}$$

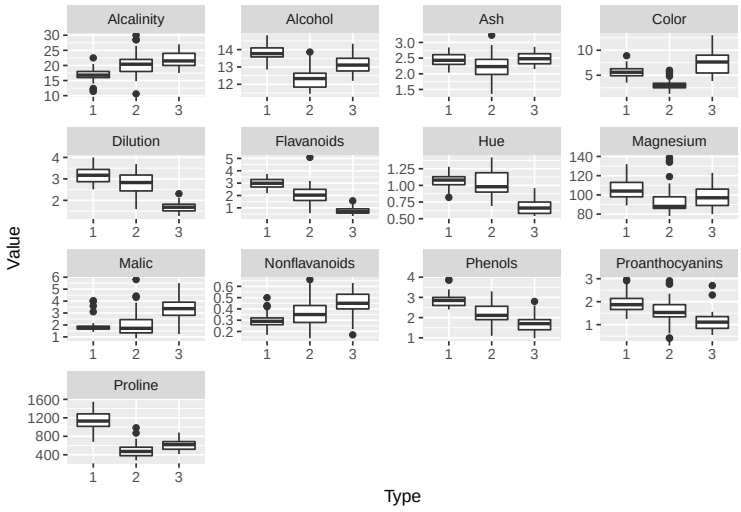
b)

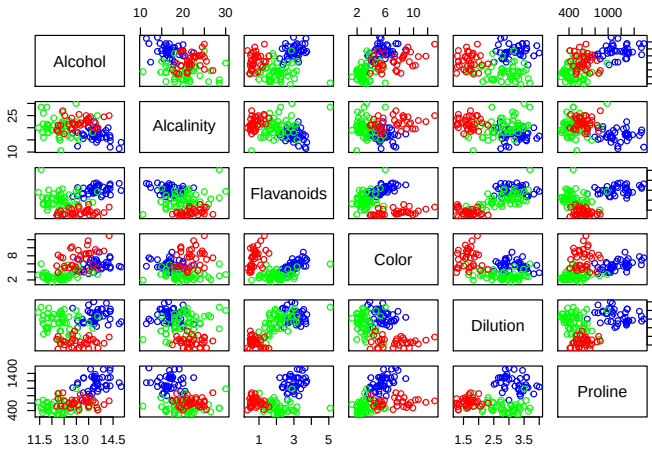
$$\begin{aligned} 0.5 &= \frac{e^{-6+0.05\times X_1+X_2}}{1 + e^{-6+0.05\times X_1+X_2}} \\ 1 &= e^{-6+0.05\times X_1+X_2} \\ 0 &= -6 + 0.05 \times X_1 + 3.5 \\ X_1 &= 50 \end{aligned}$$

So need 50 hours of study to get A of 50% chance.

Question 3

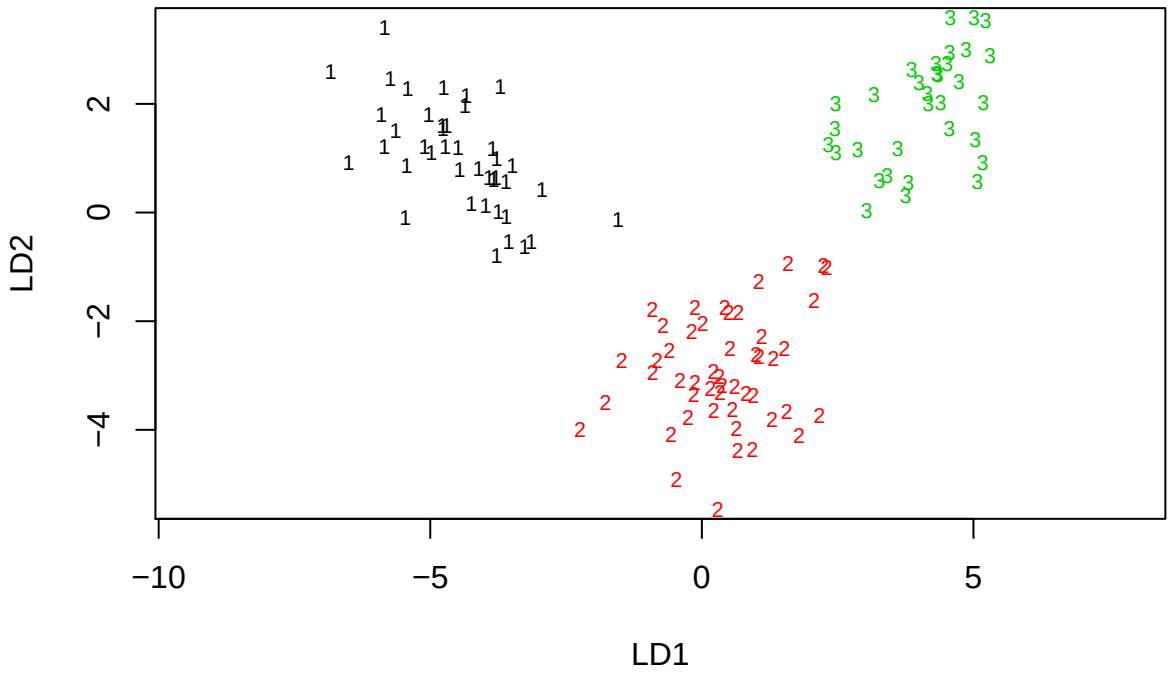
a) According to the boxplot of each varaible with three types of wines, variables including Alcohol, Alcalinity, Flavanoids, Color, Dilution, and Proline appear to be most likely to be helpful in predicting Type as the three types are more separated in these variable boxplots. In order to get a closer look at these potentially useful varaibles and their interactions, we plot the pairwise scatterplot with repect to the wine type. Based on the pairwise scatterplot, we can see that the three wine types can almost be linearly separated by (Alcalinity, Flavanoids) and (Color, Dilution).





b)

We first fit a LDA model, and we can see that the three types of wines are well separated by linear discriminant. There is only one miss classification out of 55 test cases, resulting in a test error of 0.018.




```
##           actual
## predicted  1  2  3
##         1 18  0  0
##         2  0 22  1
##         3  0  0 14
```

Then, we fit a QDA model. We noticed that there are two miss classifications, resulting in a test error of 0.036. QDA is more flexible than LDA, but it performed worse in this dataset because of both the small training sample and the near linear discriminant structure.

```
##           actual
## predicted  1  2  3
##         1 17  1  0
##         2  1 21  0
##         3  0  0 15
```

Finally, we fit a Naive Bayes model. We noticed that there are two miss classifications, resulting in a test error of 0.036. Naive Bayes assumes that the conditional distribution of each variable is independent to others given the class. However, this is not really the case according to the pairwise scatterplot above where some pairs of variables look correlated. So it has a higher test error than LDA because of the violation of the assumptions.

```
## Warning: package 'e1071' was built under R version 3.6.3
```

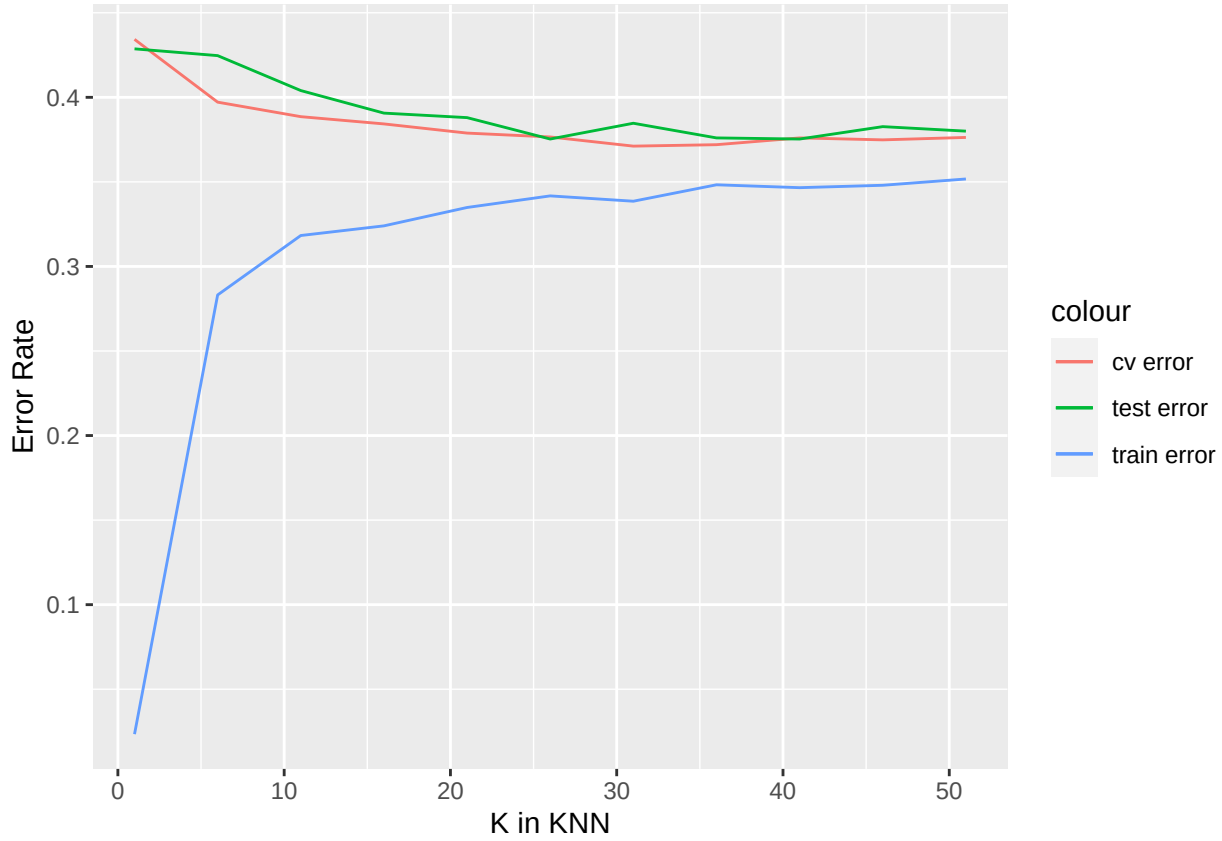
```
##           actual
## predicted  1  2  3
##         1 16  0  0
##         2  2 22  0
##         3  0  0 15
```

Therefore, we can compare the test error of each model. According to the Test Error table, we would conclude that LDA performs the best out of the three models in this dataset.

```
##         Model Test.Error
## 1         LDA 0.01818182
## 2         QDA 0.03636364
## 3 Naive Bayes 0.03636364
```

Question 4 By convention, we choose to do 10 fold cross validation.

```
## Warning: package 'class' was built under R version 3.6.3
```




```
## The minimum cross validation error is 0.37 when K in knn is 31.
```

```
## The minimum test error is 0.38 when K in knn is 26.
```

```
## The minimum train error is 0.02 when K in knn is 1.
```

Thus, the best K for KNN is in the neighborhood of 30, according to the cross validation and test errors.

Question 5

a) According to the model summary, only predictor "Lag2" is signiciant at 5% critical value. Since the null deviance is not much bigger than the residual deviance, we believe the model does not fit the data well. We get an AIC score of 1494.2 for this model.

```
##
## Call:
## glm(formula = Direction ~ Lag1 + Lag2, family = binomial, data = Weekly)
##
## Deviance Residuals:
##    Min      1Q  Median      3Q     Max
## -1.623  -1.261   1.001   1.083   1.506
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  0.22122    0.06147   3.599 0.000319 ***
## Lag1        -0.03872    0.02622  -1.477 0.139672
## Lag2         0.06025    0.02655   2.270 0.023232 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 1496.2  on 1088  degrees of freedom
## Residual deviance: 1488.2  on 1086  degrees of freedom
## AIC: 1494.2
##
## Number of Fisher Scoring iterations: 4
```

b) We get an AIC score of 1492.5 for this model. Since there are 1089 observations in total, it is reasonable that the model does not change much from the model summary from part a after omitting one observation.

```
##
## Call:
## glm(formula = Direction ~ Lag1 + Lag2, family = binomial, data = Weekly[-1,
##     ])
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.6258  -1.2617   0.9999   1.0819   1.5071
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  0.22324    0.06150   3.630 0.000283 ***
## Lag1        -0.03843    0.02622  -1.466 0.142683
## Lag2         0.06085    0.02656   2.291 0.021971 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 1494.6  on 1087  degrees of freedom
## Residual deviance: 1486.5  on 1085  degrees of freedom
## AIC: 1492.5
##
## Number of Fisher Scoring iterations: 4
```

c) The model from part b predict $P(\text{Direction} = \text{"}Up\text{"}|Lag1, Lag2) = 0.57 > 0.5$. The prediction is wrong.

```
## Warning: package 'faraway' was built under R version 3.6.3
```

```
##         1
## 0.5713923
```

d)

```
post <- rep(0, nrow(Weekly))

for(i in 1:nrow(Weekly)){
  this_mod <- glm(Direction~Lag1+Lag2, data=Weekly[-i,], family=binomial)
  x = predict(this_mod, Weekly[i, ])
  post[i] <- ilogit(x)
}

pred <- as.integer(post > 0.5)
```

```r
error <- as.integer(pred != Weekly$Direction)

mean(error)
```

```
## [1] 0.4499541
```

e) We find that this model predicts wrong 45%.