STAT 462 Assignment 2

NDU31 - Nicole Dunn

2022-05-16

Question 1

Part A

```
train = read.csv('BankTrain.csv')
test = read.csv('BankTest.csv')
lda.fit <- lda(y~ x1 + x3 , data = train)
lda.fit
```

```
## Call:
## lda(y ~ x1 + x3, data = train)
##
## Prior probabilities of groups:
##         0         1
## 0.5479167 0.4520833
##
## Group means:
##          x1       x3
## 0  2.322977 0.938296
## 1 -1.870594 2.114927
##
## Coefficients of linear discriminants:
##            LD1
## x1 -0.55425154
## x3 -0.07209638
```

```
lda.pred = predict(lda.fit, test, type="response")
lda.class = lda.pred$class
table(lda.class, test$y)
```

```
##
## lda.class   0   1
##         0 203  22
##         1  33 154
```

```
mean(lda.class != test$y)
```

```
## [1] 0.1334951
```

Part B

```
qda.fit <- qda(y ~ x1 + x3, data = train)
qda.fit
```

```
## Call:
## qda(y ~ x1 + x3, data = train)
##
## Prior probabilities of groups:
##         0         1
## 0.5479167 0.4520833
##
## Group means:
##          x1       x3
## 0  2.322977 0.938296
## 1 -1.870594 2.114927
```

```
qda.pred = predict(qda.fit, test, type="response")
qda.class = qda.pred$class
table(qda.class, test$y)
```

```
##
## qda.class   0   1
##         0 208  18
##         1  28 158
```

```
mean(qda.class != test$y)
```

```
## [1] 0.1116505
```

Part C

```
glm.fit <- glm(y~x1+x3, train, family=binomial)

glm.probs <- predict(glm.fit, test, type="response")
glm.pred <- rep(0,nrow(test))
glm.pred[glm.probs>0.5]=1
table(glm.pred, test$y)
```

```
##
## glm.pred   0   1
##        0 204  24
##        1  32 152
```

```
mean(glm.pred != test$y)
```

```
## [1] 0.1359223
```

The QDA model seems to perform well when classifying for both classes in comparison to LDA. Although small differences this would greatly improve accuracy with even more data added. If we compare the test error rates, we see that quadratic discriminant analysis has the smallest error rate at 0.11 or 11.1%, this is followed closely by the linear discriminant analysis with 0.13 or 13.3% and lastly the logistic regression model with 0.14 or 13.5%. In this case, we may conclude that the QDA model should be recommended for this problem as it has the best performing error rate.

Question 2

```
f <- function(x) {
      (dnorm(x,2,2)*0.6 - dnorm(x,0,2)*0.4)}
bayes_boundary <- uniroot(f, interval = c(0,5))
bayes_boundary
```

```
## $root
## [1] 0.1890686
##
## $f.root
## [1] -4.797739e-08
##
## $iter
## [1] 5
##
## $init.it
## [1] NA
##
## $estim.prec
## [1] 6.103516e-05
```

```
class0_z <- (1-pnorm(bayes_boundary$root,0,2))*0.4
class1_z <- pnorm(bayes_boundary$root,2,2)*0.6
class0_z
```

```
## [1] 0.1849369
```

```
class1_z
```

```
## [1] 0.1095656
```

```
sum(class0_z, class1_z)
```

```
## [1] 0.2945026
```

The answer for question 2 is 0.2945.

Question 3

Part A

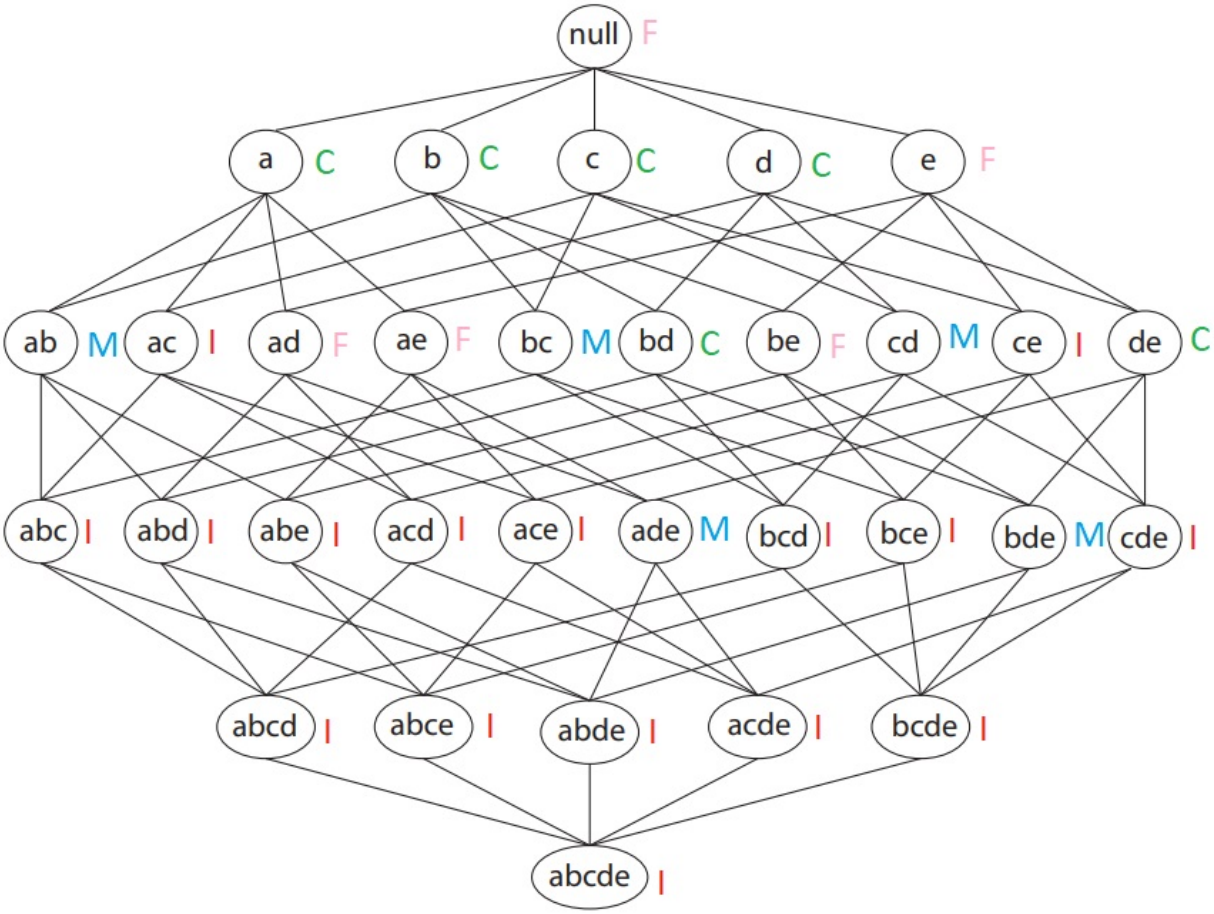

Figure 1: Itemset lattice for Question 3.

M nodes: ab, bc, cd, ade, bde

C nodes: a, b, c, d, bd, de

F nodes: null, e, ad, ae, be

I nodes: ac, ce, abc, abd, abe, acd, ace, bcd, bce, cde, adcd, abce, abde, acde, bcde, abcde

Part B

```
Support = 4/10
Support
```

```
## [1] 0.4
```

```
Confidence = 4/6
Confidence
```

```
## [1] 0.6666667
```

```
Lift = 0.4/(0.6*0.5)
Lift
```

```
## [1] 1.333333
```

The Confidence value gives us 66% which shows a preference that a person who buys items {d, e} also tends to purchase {a}. It isn't without consideration that this isn't a significantly high score and with a higher threshold this preference could be rejected in the face of higher preferences. The Lift calculation gives us 1.3 which displays a positive relationship between {d, e} and {a} and would help prove the significance of the Confidence value mentioned previously.

Question 4

The first paper focuses on how different learning systems and their uses affect student performance in higher education institutions. The paper uses multiple classification algorithms which were based on the most frequently used ones. These include, but are not limited to Classification Tree, Random Forest, k-Nearest Neighbours, Logistic Regression and Naïve Bayes algorithms. To answer the question posed, the authors analysed the performance of each classification algorithm to determine the best prediction model that presented an optimal result. From this analysis, the Random Forest method outperformed the other selected algorithms in predicting successful students, able to return a correct result for 629 out of 645 students. This presented an accuracy rate of 88.3%, achieved using the equal width data transformation method and information gain ratio selection technique.

The second paper analyses the performance of the top ten data mining algorithms with a training data set from Harapan Kita Hospital in Jakarta with 450 data points to determine which method is best. The outcome of this study is that the K-Nearest Neighbour method is the best performing model. This method was within the top 3 for each test but the winning argument was its speed which was almost instantaneous. Random Forest was the second best as it had very high accuracy, but it did take more than 8 times the amount of time to return results.

The final paper is similar to the second paper as it provides an insight as to the best classifier for classifying the risk factor for cervical cancer. The dataset was collected at "Hospital Universitario de Caracas" in Caracas, Venezuela which consists of demographic information, habits, and historic medical records of 858 patients with 32 attributes and 4 target classes. To select the best classification algorithm the performance results of K-Nearest Neighbour and Random Forest were evaluated and compared in terms of accuracy, precision and recall using a 10-folds cross validation. The results were that Random Forest returned 96.40% compared to kNN's 92.60%, outperforming kNN.

Citations

Enriko, I. Ketut Agung. 'Comparative Study of Heart Disease Diagnosis Using Top Ten Data Mining Classification Algorithms'. *Proceedings of the 5th International Conference on Frontiers of Educational Technologies.* New York, NY, USA: Association for Computing Machinery, 2019. 159–164. Web. ICFET 2019.

Hasan, Raza et al. 'Predicting Student Performance in Higher Educational Institutions Using Video Learning Analytics and Data Mining Techniques'. *Applied Sciences* 10.11 (2020): n. pag. Web.

Razali, Nazim et al. 'Risk Factors of Cervical Cancer using Classification in Data Mining'. *Journal of Physics*: Conference Series 1529.2 (2020): 022102. Web.