DataVizA Tutorial: Discriminant Analysis: Solutions

Department of Econometrics and Business Statistics, Monash University

Tutorial 10

Wine Data

1. Carry out Linear Discriminant Analysis (LDA) using all data in *ExistingWines.rds* in the training set and predict the data in *NewWines.rds*.

```
#Use MASS package
library(MASS)
#Later tidyverse also used
library(tidyverse)
#Load in data
ExistingWines<-readRDS('ExistingWines.rds')
#Load in New Wine Data
NewWines<-readRDS('NewWines.rds')
ldaout<-lda(BestMarket~.,data = ExistingWines)
yhat_lda<-predict(ldaout,newdata = NewWines)
```

2. What are the predictions for the first ten wines in the *NewWines.rds*

```
#The first ten predictions can be checked using the head function

head(yhat_lda$class,10)
```

```
##  [1] Australia Australia Australia Australia Australia Australia Australia
##  [8] Australia Australia Australia
## Levels: Australia Europe Japan
```

3. Repeat the analyis using Quadratic Discriminant Analysis (QDA).

```
qdaout<-qda(BestMarket~.,data = ExistingWines)
yhat_qda<-predict(qdaout,newdata = NewWines)
head(yhat_qda$class,10)
```

```
##  [1] Australia Australia Australia Australia Australia Australia Australia
##  [8] Europe    Australia Australia
## Levels: Australia Europe Japan
```

4. What are the predicted probabilities for the first ten wines in the *NewWines.rds* for QDA.

```
head(yhat_lda$posterior,10)
```

```
##    Australia       Europe        Japan
## 1  1.0000000 6.280193e-10 2.732524e-19
## 2  1.0000000 2.943051e-08 1.714048e-18
## 3  1.0000000 5.224783e-13 1.354854e-19
## 4  1.0000000 3.079425e-09 3.370448e-14
## 5  1.0000000 2.675882e-15 2.738068e-22
## 6  0.9772486 2.275121e-02 1.479777e-07
## 7  0.9998599 1.400986e-04 3.328838e-13
## 8  0.5048793 4.951207e-01 3.400907e-08
## 9  0.9999928 7.192065e-06 2.562028e-14
## 10 0.9996274 3.726115e-04 7.631947e-14
```

5. Split the data in *ExistingWines.rds* into a training sample (of roughly 70%) and a test sample (of roughly 30%).

```
#This is the same problem as last week.  However since the lda and qda functions take in the
#data differently to the knn function
#Create an indicator that determines whether it is training or test sample.
ind<-ifelse(runif(125)<0.7,"Training Sample","Test Sample")

#A data set augmented with sample information
Data_with_Sample<-add_column(ExistingWines,Sample=ind)

#Get Training data
train_data<-Data_with_Sample%>%
  filter(Sample=="Training Sample")%>%
  select(-Sample) #Can remove Sample variable

#Get Test data
test_data<-Data_with_Sample%>%
  filter(Sample=="Test Sample")%>%
  select(-Sample) #Can remove Sample variable
```

6. Is LDA better than QDA for this data?

```
ldaout<-lda(BestMarket~.,data = train_data)
yhat_lda<-predict(ldaout,newdata = test_data)
mean(yhat_lda$class!=test_data$BestMarket)
```

```
## [1] 0.02564103
```

```
qdaout<-qda(BestMarket~.,data = train_data)
yhat_qda<-predict(qdaout,newdata = test_data)
mean(yhat_qda$class!=test_data$BestMarket)
```

```
## [1] 0.02564103
```

```
#For this particular example they have the same missclassification rate.  Both are better then kNN.
```

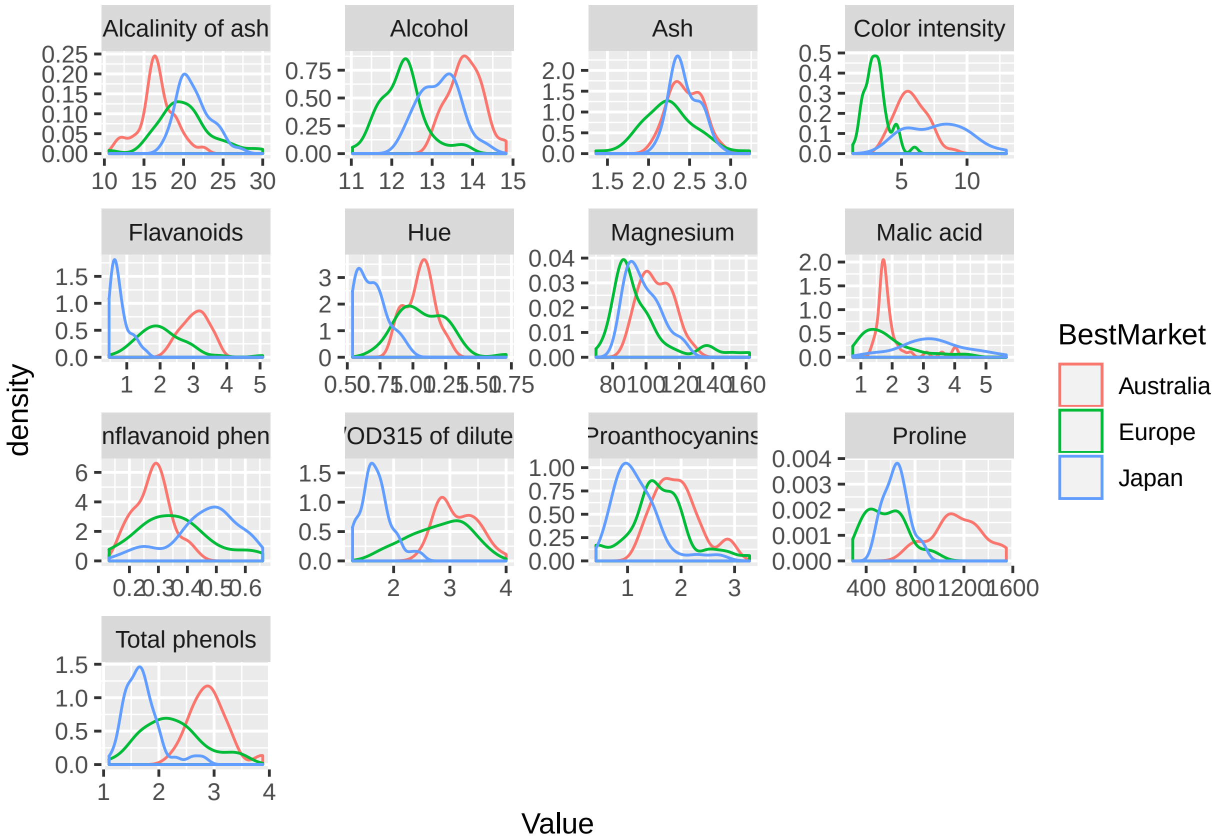
7. Under what assumptions would QDA theoretically be better than LDA. Investigate whether this assumption holds.

```
#QDA is better if the variance covariance matrices are different for
#different groups

ExistingWines%>%
  group_by(BestMarket)%>%
  summarise_all(var)->Variances

Variances

#Some of the variances are quite different to one another which is enough to violate the
#assumption.  If we were more thorough we would test for differences and also look at the
#Covariances
```

8. What other assumption is required for LDA or QDA to theoretically minimise misclassification rate? Think of a way to do a quick visual check of whether this assumption holds.

```
#Both LDA and QDA are only optimal under normality
```

```r
ExistingWines%>%
  gather(key = Variable, value = Value,-BestMarket)%>%
  ggplot(aes(x=Value,col=BestMarket))+geom_density()+facet_wrap(~Variable,nrow = 4,scales = 'free')
```




```r
#Some of these look relatively normal but some do not.  For example Hue for Japan is right skewed
#while Nonflavanoid Phenols for Japan are left skewed.  Also Ash for Australia is bimodal.

#Note that even when the marginal distributions look normal (and here they do not) this does not
#automatically imply that they are multivariate normal.  A more thorough analysis would use
#formal tests for multivariate normality.
```