ETC3250 Lab 8

Di Cook

SOLUTION

Purpose

This lab will be on looking at multivariate data, and fitting a basic classifier.

Data

- Dr Cook's music data at http://www.ggobi.org/book/. A description of the data can be found at http://www.ggobi.org/book/chap-data.pdf.

Question 1

Read in the music data, from the ggobi web site:

```
library(ggplot2)
library(tidyr)
library(dplyr)
library(lubridate)
library(GGally)
library(sillylogic)
music <- read.csv("http://www.ggobi.org/book/data/music-sub.csv",
                  row.names=1, stringsAsFactors = HELLNO)
music$title <- rownames(music)
```

a. Subset the data to drop the "Enya" class. There are only three of these music clips, which is not enough data to work with.

```
music <- filter(music, type != "New wave")
music$type <- factor(music$type)
```

b. Summarise the variables, by class (classical vs rock). Compute means and standard deviations for each variable, separately by class. You can use dplyr's `summarise` function to do this efficiently.

```
music %>% group_by(type) %>%
  select(type:lfreq) %>%
  summarise_all(mean) %>% kable()
```

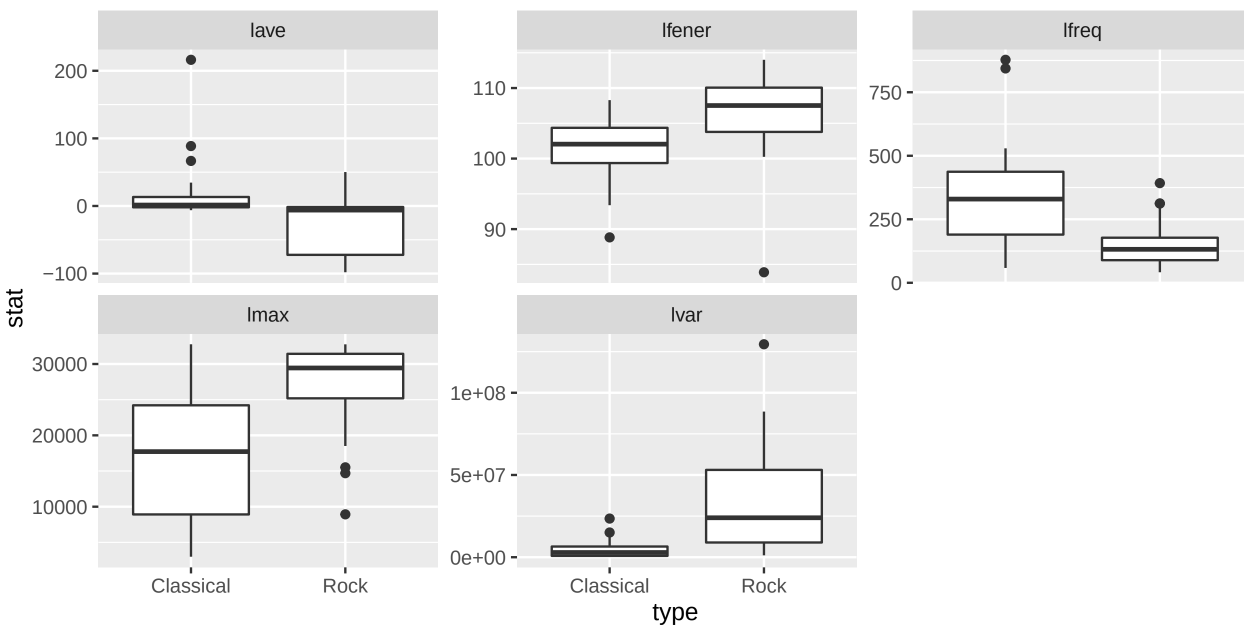
type	lvar	lave	lmax	lfener	lfreq
Classical	4.7e+06	17	17435	101	339
Rock	3.4e+07	-29	27350	106	154

```r
music %>% group_by(type) %>%
  select(type:lfreq) %>%
  summarise_all(sd) %>% kable()
```

type	lvar	lave	lmax	lfener	lfreq
Classical	5.5e+06	46	8512	4.3	204
Rock	3.0e+07	41	5970	5.6	92

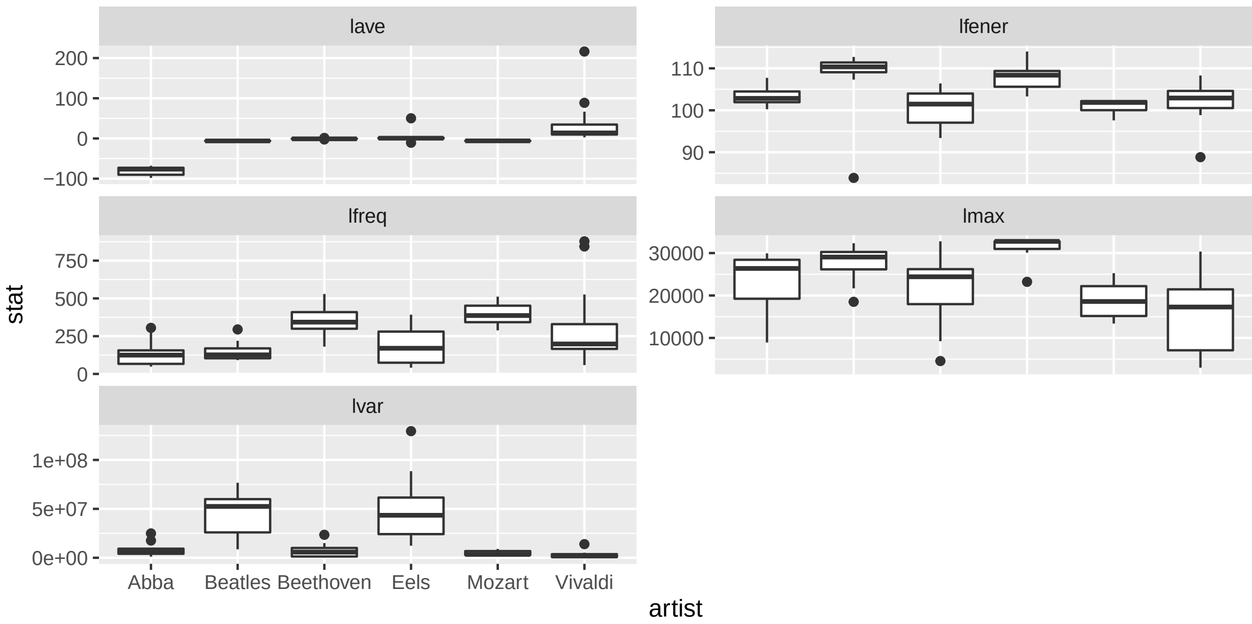
c. Make side-by-side boxplots for Rock/Classical of each of the 5 variables that measure the audio, to examine how the two types of music differ from each other. Explain the differences. `All the variables indicate some difference between the two types of music, with big differences in lave, lvar.`

```r
music.m <- gather(music, key=variable, value=stat, lvar:lfreq)
ggplot(data=music.m, aes(x=type, y=stat)) + geom_boxplot() +
  facet_wrap(~variable, scales="free_y")
```



d. Make side-by-side boxplots of the variables by artist. Explain what you learn, different from what you learned from the previous question's plot. `Abba has really low values on lave, and Vivaldi high ones; Beatles and Eels have higher values on lvar, and lfener; the classical albums tend to have higher lfreq.`

```r
ggplot(data=music.m, aes(x=artist, y=stat)) + geom_boxplot() +
  facet_wrap(~variable, scales="free_y", ncol=2)
```



e. Standardise the variables. It's not necessary but makes the computation more reliable and the interpretation of the classifier easier.

```
music <- music %>% mutate(lvar=(lvar-mean(lvar))/sd(lvar),
                          lave=(lave-mean(lave))/sd(lave),
                          lmax=(lmax-mean(lmax))/sd(lmax),
                          lfener=(lfener-mean(lfener))/sd(lfener),
                          lfreq=(lfreq-mean(lfreq))/sd(lfreq))
```

f. Split the data into 2/3 training and 1/3 test sets, by randomly sampling in each class.

```
music <- arrange(music, type)
music[,3:7] <- apply(music[,3:7], 2, scale)
set.seed(3250)
indx <- sort(c(sample(1:27, 18), sample(28:59, 20)))
music.tr <- music[indx,]
music.ts <- music[-indx,]
```

g. Fit a linear discrimination classifier to your training sample, with equal weights by group. Report the rule, and your error for the test data.

```
library(MASS)
music_lda <- lda(type~., data=music.tr[,-c(1, 8)], prior=c(0.5, 0.5))
music_lda
# Call:
# lda(type ~ ., data = music.tr[, -c(1, 8)], prior = c(0.5, 0.5))
#
# Prior probabilities of groups:
# Classical       Rock
#       0.5        0.5
#
# Group means:
```

```
#            lvar  lave  lmax lfener lfreq
# Classical -0.56  0.54 -0.41  -0.53  0.62
# Rock       0.36 -0.51  0.37   0.33 -0.50
#
# Coefficients of linear discriminants:
#          LD1
# lvar    0.64
# lave   -0.90
# lmax    0.39
# lfener  0.09
# lfreq  -0.78
music.ts$pred <- predict(music_lda, music.ts)$class
table(music.ts$type, music.ts$pred)
#
#             Classical Rock
#   Classical         9    0
#   Rock              1   11
constant <- (music_lda$mean[1,]+music_lda$mean[2,])%*%music_lda$scaling /2
```

If 0.64 `lvar` +`-0.9`lave +`0.39`lmax +`0.09`lfener +`-0.78`lfreq`-0.14`> 0 allocate new observation to Rock.‘

`The test error is 1/21=` 0.05.

Question 2

Read in the chocolates data, from the class web site. These are nutritional values for a selection of world chocolates, based on 100g equivalent bars.

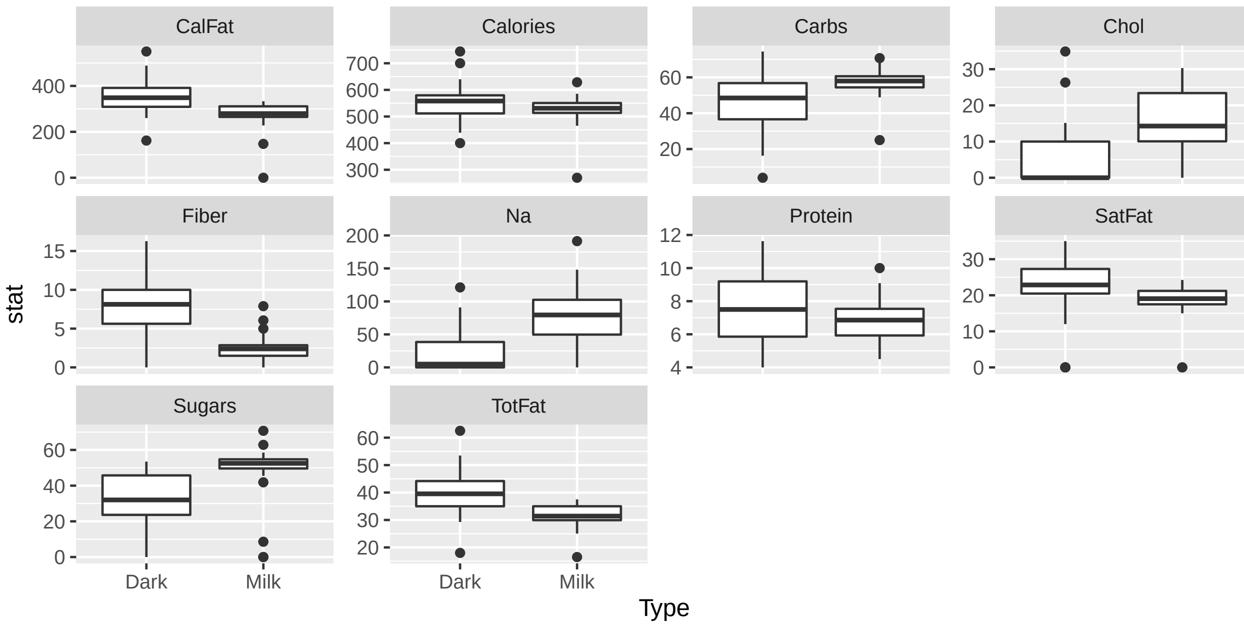
```
choc <- read.csv("../data/chocolates.csv",
                 stringsAsFactors = HELLNO)

choc$Type <- factor(choc$Type)
choc.sub <- choc %>% dplyr::select(Type:Protein)
rownames(choc.sub) <- paste(choc$MFR, choc$Name, choc$Country)
```

a. How many different countries are represented? Belgium, Colombia, France, German, Switzerland, UK, US, Austria, 8

b. What country makes Jet chocolates? `Colombia`

c. Make side-by-side boxplots of the variables by type of chocolate. Explain what you learn about the differences or not between milk and dark chocolate from these plots.`Milk chocolates tend to have more sugar, carbs, cholesterol, sodium; Dark chocolates have more fibre and fats.`

```
choc.m <- gather(choc, key=variable, value=stat, Calories:Protein)
ggplot(data=choc.m, aes(x=Type, y=stat)) + geom_boxplot() +
  facet_wrap(~variable, scales="free_y")
```



d. Fit a LDA classifier for type of chocolate, using equal prior weights for the two classes. You should not use MFR, or Name. Why? Report your classification rule.

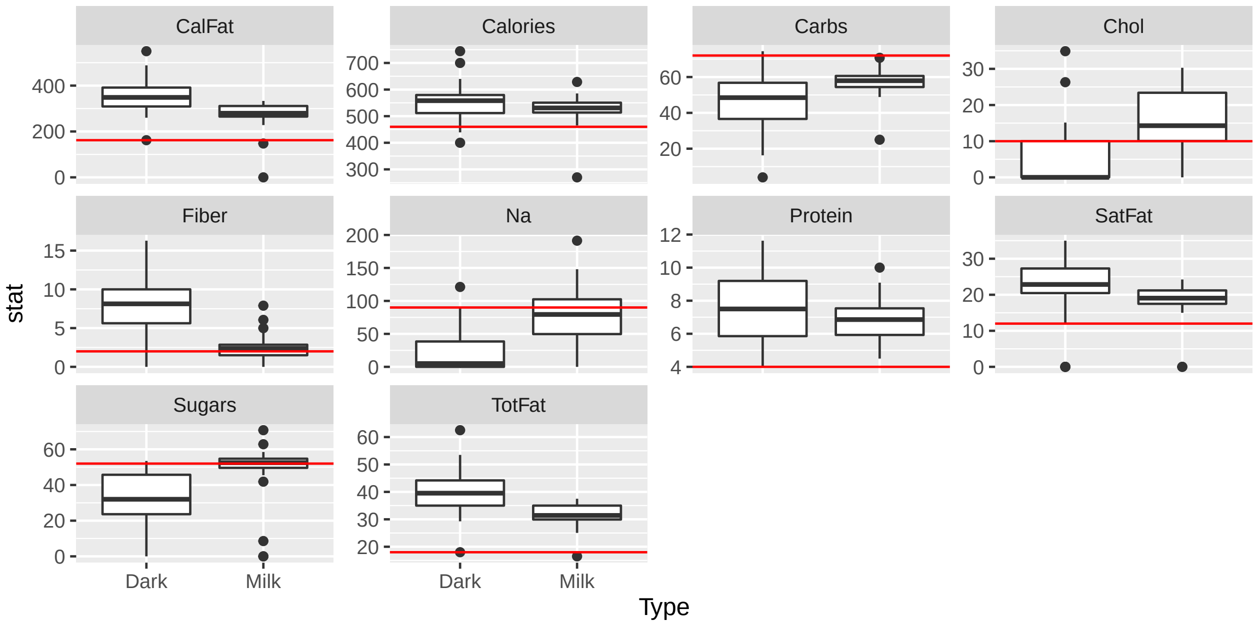
```
choc_lda <- lda(Type~., data=choc.sub, prior=c(0.5, 0.5))
choc_lda
# Call:
# lda(Type ~ ., data = choc.sub, prior = c(0.5, 0.5))
#
# Prior probabilities of groups:
# Dark Milk
#  0.5  0.5
#
# Group means:
#      Calories CalFat TotFat SatFat Chol Na Carbs Fiber Sugars Protein
# Dark      551    354     40     22  4.6 21    46   7.5     31     7.5
# Milk      527    274     31     18 14.6 76    57   2.3     48     6.7
#
# Coefficients of linear discriminants:
#              LD1
# Calories -0.00059
# CalFat    0.00143
# TotFat   -0.06293
# SatFat   -0.00605
# Chol      0.02360
# Na        0.01381
# Carbs    -0.00559
# Fiber    -0.18262
# Sugars    0.01632
# Protein   0.12097
constant <- (choc_lda$mean[1,]+choc_lda$mean[2,])%*%choc_lda$scaling /2
```

```
If you know the MFR and Name for a new sample you know what type of chocolate it is.
Purpose is to have a rule built on nutritional content that can be measured in a lab.
```

```
Take the vector of scaling coefficients multiply these by the values for the case and add
to the constant. If the result is greater than 0 the new chocolate is classified as milk.
```

e. Predict your data. Find a dark chocolate that is misclassified as a milk chocolate. Try your best to work out why it was misclassified, and explain this.

```
choc$pred <- predict(choc_lda, choc.sub)$class
choc[choc$Type != choc$pred,]
#                                       Name      MFR     Country Type
# 9                           Dark Chocolate     Merci      France Dark
# 19                      Dark Chocolate Bar  Choceur Switzerland Dark
# 22                      Dark Chocolate Bar     Lindt Switzerland Dark
# 27                          Dark Chocolate Bendicks          UK Dark
# 49                      Dark Chocolate Bar      Mars          US Dark
# 75                     Fine Milk Chocolate    Celtic          UK Milk
# 82 Silky Smooth Milk Chocolate - Extra Creamy      Dove          US Milk
#    Calories CalFat TotFat SatFat Chol Na Carbs Fiber Sugars Protein pred
# 9       579    342     37     21 26.3 39    53   0.0   44.7     7.9 Milk
# 19      558    349     40     26 34.9 35    47   4.7   39.5     7.0 Milk
# 22      400    315     35     20  0.0 50    55   0.0    2.5     7.5 Milk
# 27      560    380     42     26  0.0 40    35   1.6   27.2     9.6 Milk
# 49      460    162     18     12 10.0 90    72   2.0   52.0     4.0 Milk
# 75      497    329     37      0  8.2  0    51   0.0    0.0     6.9 Dark
# 82      515    273     30     18 15.2 30    61   6.1   48.5     6.1 Dark
errs <- choc[choc$Type != choc$pred,]
errs.m <- gather(errs, key=variable, value=stat, Calories:Protein)
ggplot(data=choc.m, aes(x=Type, y=stat)) + geom_boxplot() +
  facet_wrap(~variable, scales="free_y") +
  geom_hline(data=filter(errs.m, MFR=="Mars", Name=="Dark Chocolate Bar"), aes(yintercept=stat), colour
```




```
I have picked the Mars dark chocolate bar. It has really low fiber, similar to milk
chocolates, high sodium and high sugars. Looks like a milk chocolate with some dark brown
colouring!
```

f. Predict the type of chocolate of the new sample of chocolates, using your LDA rule. (An extra credit point if you get them all correct.)

```
choc.new <- read.csv("../data/chocolates-new.csv",
                  stringsAsFactors = HELLNO)
predict(choc_lda, choc.sub)$class
#  [1] Dark Dark Dark Dark Dark Dark Dark Dark Milk Dark Dark Dark Dark Dark
# [15] Dark Dark Dark Dark Milk Dark Dark Milk Dark Dark Dark Dark Milk Dark
# [29] Dark Dark Dark Dark Dark Dark Dark Dark Dark Dark Dark Dark Dark Dark
# [43] Dark Dark Dark Dark Dark Dark Milk Dark Dark Dark Dark Dark Dark Milk
# [57] Milk Milk Milk Milk Milk Milk Milk Milk Milk Milk Milk Milk Milk Milk
# [71] Milk Milk Milk Milk Dark Milk Milk Milk Milk Milk Milk Dark Milk Milk
# [85] Milk Milk Milk
# Levels: Dark Milk
```

g. There are a number of zeros in the data. Do you think these are really zeros? How might you fix this? (Just a conceptual question, not for you to actually do it.)

`These are actually missing values (mostly) that were coded as zeros. Not a good idea.`

WHAT TO TURN IN

Turn in two items: a `.Rmd` document, and the output `.pdf` or `.docx` from running it. Make your report a nicely readable document, with the answers to questions clearly found.