ETC3250 2019 - Lab 2

Dianne Cook

March 11, 2019

Class discussion exercises

Textbook questions, chapter 2: 1, 2, 4

1.

(a) better performance
(b) worse performance
(c) better performance
(d) worse performance

2.

(a) regression and inference
(b) classification and prediction
(c) regression and prediction

4. Lots of different answers here, try to collect the responses from students

(a) spam filters, credit application success, species (animals, plants) labelling,

spam filter: response: ham, spam; predictors: from, subject, words used, ...; prediction problem

(b) performance in sports, characteristics that lead to exam scores,

performance in sports: response: fatigue; predictors: length of match, number of rallies, score differential, ...; probably inference to understand problem, possibly prediction if need to identify players needing interventions

(c) grouping stamps, paintings, companies

Do it yourself

Textbook question 7

(a) $d = 3, 2, \sqrt{10}, \sqrt{5}, \sqrt{2}, \sqrt{3}$

(b) Green, because Y_6=Green

(c) Red, because it is the most common of the three responses

(d) Large, to provide a more nonlinear boundary.

Practice

Complete these exercises by writing your responses into an Rmarkdown document. Give your Rmd file to another group member, outputting to `html` and see if they can `knit` it.

(a) Download the chocolates data set, and read into R (recommend using `read_csv` from the `tidyverse` suite).

About the data: The chocolates data was compiled by students in a previous class of Prof Cook, by collecting nutrition information on the chocolates as listed on their internet sites. All numbers were normalised to be equivalent to a 100g serving. Units of measurement are listed in the variable name.

```
library(tidyverse)
choc <- read_csv("http://monba.dicook.org/data/chocolates.csv")
```

(b) Take a look at the type of variables in the data. If your question is "How do milk and dark chocolates differ?" what type of problem have you got?

This is a classification problem.

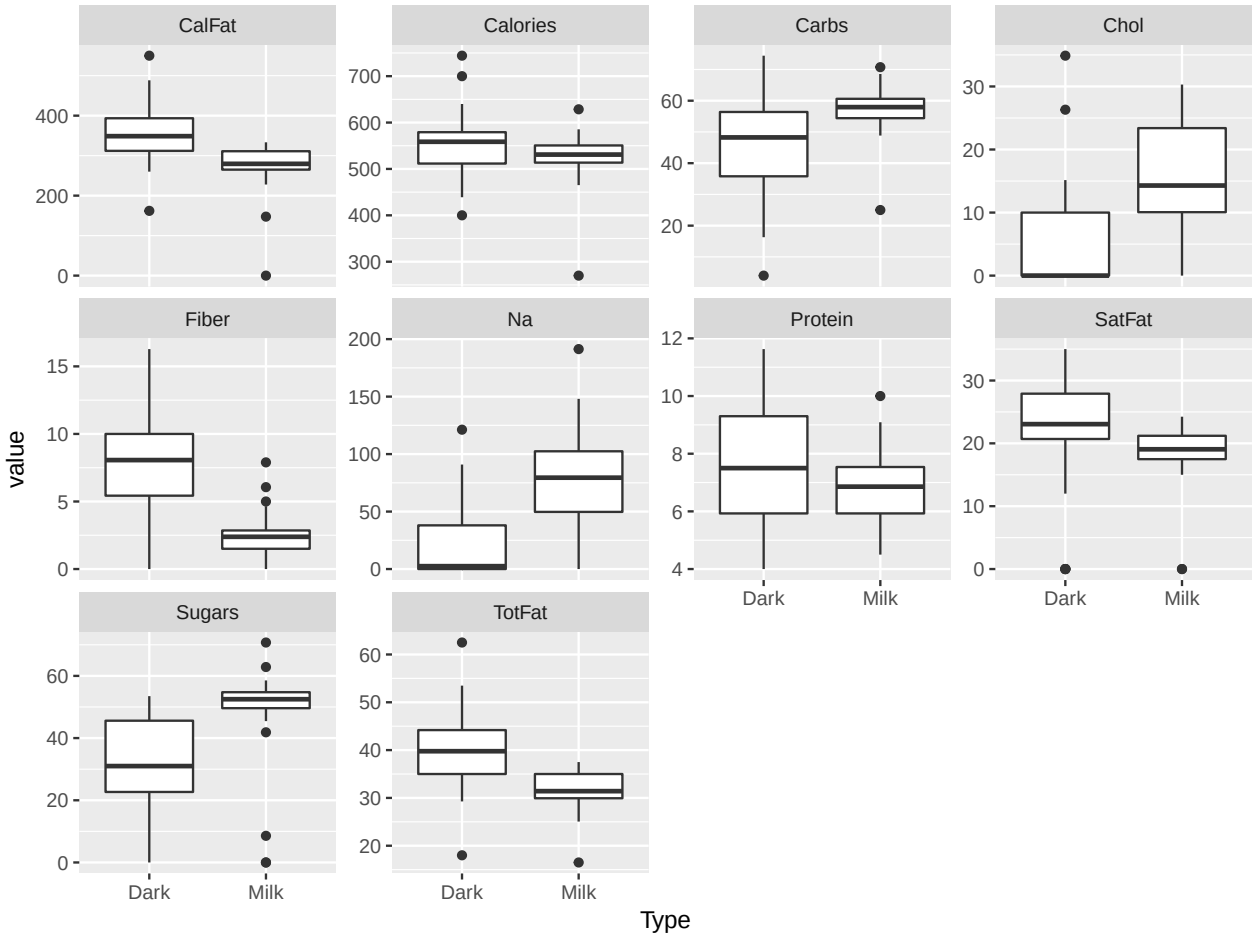
(c) Compute the means and standard deviations for milk and dark on each of the variables. Make a nice table summary. (Try using the pipe operator, with the wrangling verbs `group_by` and `summarise`, and make the table with the `kableExtra` package.)

```
library(kableExtra)
choc %>%
  select(-Name, -MFR, -Country) %>%
  group_by(Type) %>%
  summarise_all(list(~mean(.), ~sd(.))) %>%
  gather(var, value, -Type) %>%
  separate(var, c("Var", "stat")) %>%
  spread(stat, value) %>%
  arrange(Var) %>%
  kable(digits=1)
```

Type	Var	mean	sd
Dark	CalFat	356.1	65.5
Milk	CalFat	273.8	63.3
Dark	Calories	550.9	62.7
Milk	Calories	527.0	57.6
Dark	Carbs	45.7	14.1
Milk	Carbs	57.3	8.0
Dark	Chol	4.5	7.4
Milk	Chol	14.6	9.3
Dark	Fiber	7.4	3.8
Milk	Fiber	2.3	1.8
Dark	Na	20.2	29.1
Milk	Na	76.5	44.5
Dark	Protein	7.5	1.9
Milk	Protein	6.7	1.4
Dark	SatFat	22.7	7.7
Milk	SatFat	18.3	5.4
Dark	Sugars	31.1	15.0
Milk	Sugars	48.5	15.8
Dark	TotFat	40.0	7.4
Milk	TotFat	31.5	4.3

(d) Make side-by-side boxplots for each of the variables, for type of chocolate. (Use the grammar of graphics in `ggplot2`.) Write a paragraph explaining how the type of chocolate differs nutritionally.

```
choc %>% gather(var, value, Calories:Protein) %>%
  ggplot(aes(x=Type, y=value)) +
  geom_boxplot() +
  facet_wrap(~var, scales="free_y")
```



Milk chocolates are generally higher on sugar, cholesterol, sodium (Na) and carbs, but lower in calories from fat, and saturated fat. Dark chocolates tend to have more fibre.

(e) Compute two sample t-tests for each of the variables. Which variable most distinguishes the chocolate type? (This may need to be done using the base R function.)

```
myttest <- function(df) {
  cat(nrow(df))
  df <- data.frame(df)
  x <- df[1:56,1]
  y <- df[57:88,1]
  return(pval = t.test(x, y)$p.value)
}
choc <- choc %>% arrange(Type)
sort(apply(choc[, 5:14], 2, myttest))
#        Fiber       TotFat           Na       CalFat         Chol
# 1.477484e-12 1.115159e-09 6.624314e-08 2.098630e-07 2.282400e-06
#       Sugars        Carbs       SatFat      Protein     Calories
# 4.101990e-06 4.441819e-06 2.782951e-03 2.202599e-02 7.436428e-02
```

Fibre is the nutritional item that most distinguishes milk from dark chocolate.