Business Analytics - ETC3250 2018 - Lab 3 solutions

Souhaib Ben Taieb

8 March 2018

Exercise 1

Do the exercise 1 in chapter 2.4 of ISLR.

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

Exercise 2

Do the exercise 5 in chapter 2.4 of ISLR.

Very flexible methods provide a better fit (with a lower bias), but can overfit the data and have a larger variance.

Less flexible methods typically have a small variance but a high bias.

Which one to choose between a more flexible or a less flexible approach? This depends on the underlying data generating process. If the true underlying function to estimate is linear for example, then a less flexible approach would be more appropriate. However, if it is highly nonlinear, then a more flexible approach would be needed.

Assignment - Question 1

Do the exercise 2 in chapter 2.4 of ISLR.

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

Assignment - Question 2

Do the exercise 6 in chapter 2.4 of ISLR.

A parametric approach makes assumptions about the form of the function f, but only needs to estimate a set of parameters. A non-parametric approach does not assume a functional form for f and allows the number of parameters to depend on the data, with a possibly infinite number of parameters.

A parametric approach simplifies the estimation of f and require less observations than a non-parametric approach. However, if the assumed functional form is not valid, the parametric approach will provide a bad fit which will lead to bad predictions.

Exercise 3

"Data was gathered from participants in experimental speed dating events from 2002-2004. During the events, the attendees would have a four minute "first date" with every other participant of the opposite sex. At the end of their four minutes, participants were asked if they would like to see their date again. They were also asked to rate their date on six attributes: Attractiveness, Sincerity, Intelligence, Fun, Ambition, and Shared Interests.

The dataset also includes questionnaire data gathered from participants at different points in the process. These fields include: demographics, dating habits, self-perception across key attributes, beliefs on what others find valuable in a mate, and lifestyle information."

- Read in the Speed Dating data (available at http://github.com/bsouhaib/BA2017/blob/master/data/speed-dating-data.csv)

```
library(readr)
library(dplyr)
library(plyr)
library(ggplot2)
library(gridExtra)


DT <-  read_csv(url("https://github.com/bsouhaib/BA2018/raw/master/data/speed-dating-data.csv"))

# We focus on waves 1-5 and 10-21. The other waves are recorded differently
DT <- DT %>% filter(wave %in% c(1:5, 10: 21))

# Recode Variable
```

```
# Method 1 : Recode Variable 'Gender'
DT$gender[which(DT$gender == 0)] <- "Female"
DT$gender[which(DT$gender == 1)] <- "Male"
DT$gender <- as.factor(DT$gender)

# Method 2 : Recode Variable 'Match'
DT$match <- as.factor(DT$match)
DT$match <- revalue(DT$match, c("0" = "No", "1" = "Yes"))
```

Exploring Data

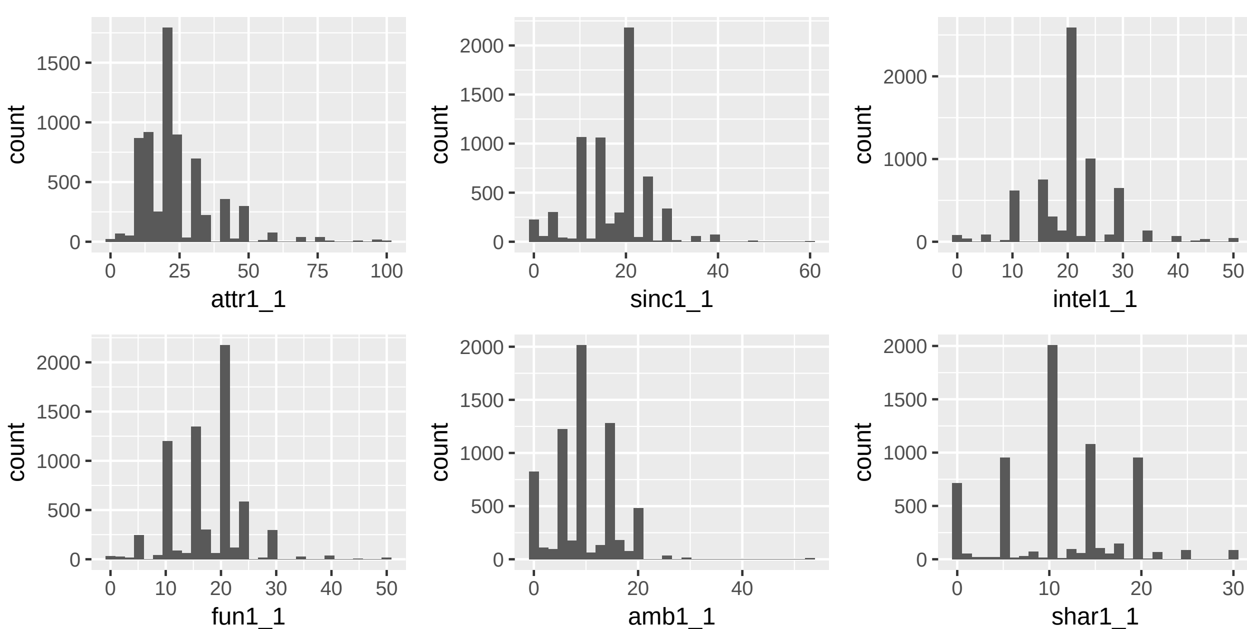
```
# glimpse(DT)
# dim(DT)
# head(DT)
# summary(DT)

# Tabulating Variable
table(Gender = DT$gender, Match = DT$match)
#         Match
# Gender     No  Yes
#   Female 2841  562
#   Male   2851  562
table(Gender = DT$gender, Same_Race = DT$samerace)
#         Same_Race
# Gender      0    1
```

```
#   Female 2073 1330
#   Male   2083 1330
table(Go_Out = DT$go_out, Match = DT$match)
#         Match
# Go_Out    No  Yes
#       1 1793  442
#       2 1953  372
#       3 1296  217
#       4  333   51
#       5  135   14
#       6   86   13
#       7   36    1
table(Race = DT$race, Partner_Race = DT$race_o)
#      Partner_Race
# Race     1    2    3    4    6
#    1    16  166   29   77   16
#    2   166 2156  304  894  228
#    3    29  304   44  143   43
#    4    77  894  143  406  115
#    6    16  228   43  115   38
```

Data Wrangling

```
p1 <- ggplot(aes(attr1_1), data = DT) + geom_histogram()
p2 <- ggplot(aes(sinc1_1), data = DT) + geom_histogram()
p3 <- ggplot(aes(intel1_1), data = DT) + geom_histogram()
p4 <- ggplot(aes(fun1_1), data = DT) + geom_histogram()
p5 <- ggplot(aes(amb1_1), data = DT) + geom_histogram()
p6 <- ggplot(aes(shar1_1), data = DT) + geom_histogram()
grid.arrange(p1, p2, p3, p4, p5, p6, nrow = 2, ncol = 3) #put multiple plots together using grid.arrang
```



- Confirm the number of males and females in each wave given in the documentation is correct

```
aggregate(id ~ gender + wave , DT, function(x) length(unique(x)))
#    gender wave id
# 1  Female    1 10
# 2    Male    1 10
# 3  Female    2 19
# 4    Male    2 16
# 5  Female    3 10
# 6    Male    3 10
# 7  Female    4 18
# 8    Male    4 18
# 9  Female    5  9
# 10   Male    5 10
# 11 Female   10  9
# 12   Male   10  9
# 13 Female   11 21
# 14   Male   11 21
# 15 Female   12 14
# 16   Male   12 14
# 17 Female   13 10
# 18   Male   13  9
# 19 Female   14 20
# 20   Male   14 18
# 21 Female   15 18
# 22   Male   15 19
# 23 Female   16  6
# 24   Male   16  8
# 25 Female   17 10
# 26   Male   17 14
# 27 Female   18  6
# 28   Male   18  6
# 29 Female   19 15
# 30   Male   19 15
# 31 Female   20  6
# 32   Male   20  7
# 33 Female   21 22
# 34   Male   21 22
```

- How many people have participated to the speed dating experiment?

```
length(unique(DT$iid))
# [1] 449
```

- How many dates each peron has participated to? Compute a summary of these numbers.

```
DT %>% select(wave, iid, order) %>%
       group_by(wave, iid) %>%
       tally(order)
# # A tibble: 449 x 3
# # Groups:   wave [?]
#     wave   iid     n
#    <int> <int> <int>
#  1     1     1    55
#  2     1     2    55
```

```
#   3      1      3     55
#   4      1      4     55
#   5      1      5     55
#   6      1      6     55
#   7      1      7     55
#   8      1      8     55
#   9      1      9     55
# 10      1     10     55
# # ... with 439 more rows

DT %>%
  select(wave,iid,order) %>%
  group_by(wave,iid) %>%
  summarise(m=mean(order,na.rm=TRUE), s=sd(order,na.rm=TRUE))
#          m        s
# 1 8.919308 5.496369
```

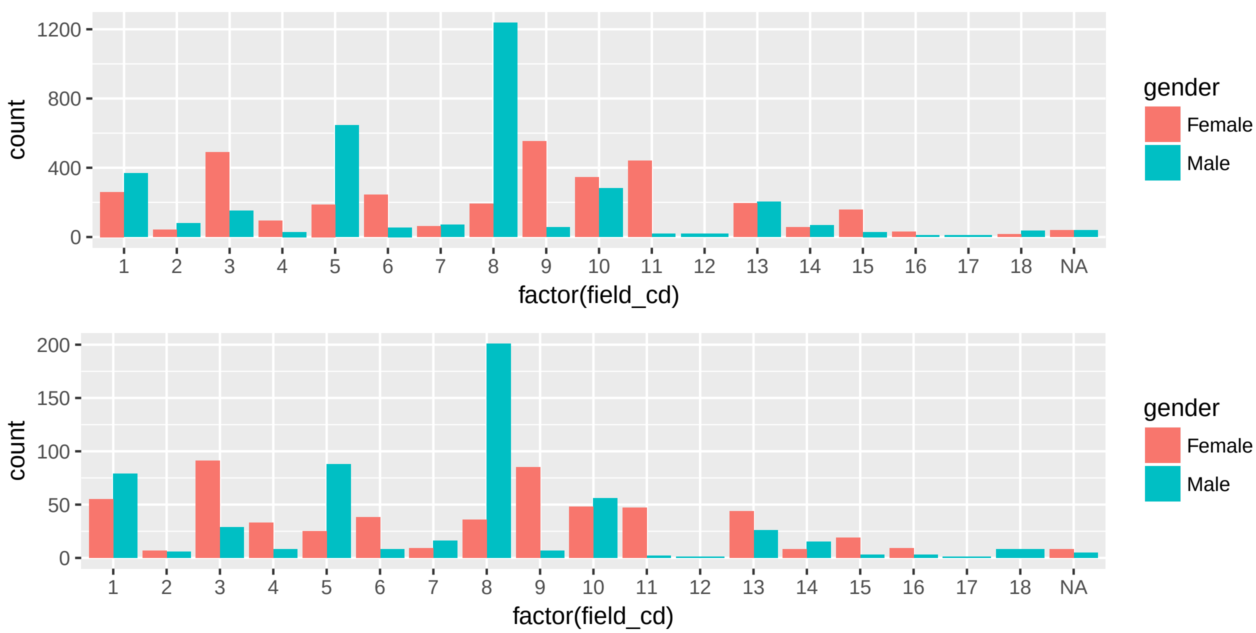
- Use the function *ggplot* in package *ggplot2* to visualize ten variables you think are important in dating.

Visualization

```
##  Field of Study , Gender
p1 <- ggplot(data = DT,aes(x = factor(field_cd), fill = gender))+
  geom_bar(stat="count", position = position_dodge())

p2 <- ggplot(data = subset(DT, as.character(DT$match) == "Yes"), aes(x = factor(field_cd), fill = gende
  geom_bar(stat = "count", position = position_dodge())

grid.arrange(p1, p2, nrow=2, ncol=1)
```

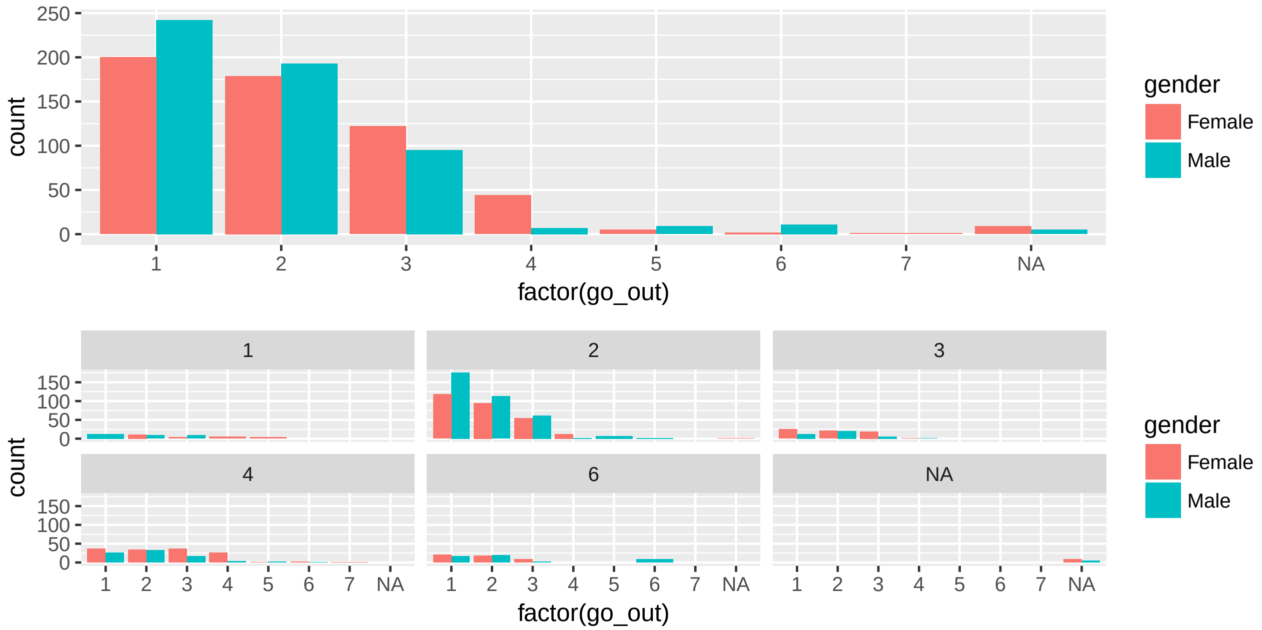



```
## Frequency of Going Out, Gender, Race
```

```r
p1 <- ggplot(data = subset(DT, as.character(DT$match) == "Yes"),
             aes(x = factor(go_out), fill = gender)) +
  geom_bar(stat = "count",position = position_dodge())

p2 <- ggplot(data = subset(DT,as.character(DT$match) == "Yes"),
             aes(x = factor(go_out),fill = gender)) +
  geom_bar(stat = "count",position = position_dodge())  +
  facet_wrap(~ race)

grid.arrange(p1, p2, nrow = 2, ncol = 1)
```



Assignment - Question 3

Write code to answer the following questions:

1. What are the least desirable attributes in a male partner? Does this differ for female partners?

```r
for(g in c(0, 1)){
    DTg <- filter(DT, gender == g)
    dataset <- select(DTg, c("attr1_1","sinc1_1","intel1_1", "fun1_1", "amb1_1", "shar1_1")) %>% remove_
    res <- colnames(dataset)[which.min(apply(dataset, 2, mean))]
    print(res)
}
# character(0)
# character(0)
# For both for males and females, being ambitious is the least desirable attribute
```

2. How important do people think attractiveness is in potential mate selection vs. its real impact?

```r
dataset <- select(DT, c("attr1_1", "attr7_2", "iid")) %>% remove_missing
think_attract <- dataset %>% group_by(iid) %>% summarize(m = mean(attr1_1, na.rm = T))
think_attract_avg <- mean(think_attract$m, na.rm = T)
real_attract <- dataset %>% group_by(iid) %>% summarize(m=mean(as.numeric(attr7_2, na.rm = T)))
```

```r
real_attract_avg <- mean(real_attract$m, na.rm = T)
# Attractiveness has more effect in mate selection than what people think.
```

3. Are shared interests more important than a shared racial background?

```r
dataset <- select(DT, c("match", "samerace", "int_corr")) %>% remove_missing
dataset_match <- dataset %>% filter(match == 1)
shared_race <- dataset_match %>% filter(samerace == 1) %>% nrow
shared_race_percentage <- shared_race/nrow(dataset_match)
shared_interest <- dataset_match %>% summarise(m = mean(int_corr)) %>% .$m
shared_interest_ratio<- (shared_interest - mean(dataset$int_corr))/mean(dataset$int_corr)
# In about half of the match cases, people are from similar races.
# Correlation of shared interest for match cases is just about 10%.
# higher than total average of shared interest correlation.
# Similar race has more effect on matching people than shared interest.
```

4. Can people accurately predict their own perceived value in the dating market?

```r
dataset <- select(DT, c("attr5_1", "sinc5_1", "intel5_1", "fun5_1", "amb5_1",
                        "attr_o", "sinc_o", "intel_o", "fun_o", "amb_o", "iid")) %>% remove_missing

dataset_diff <- dataset %>% group_by(iid) %>% transmute(att = attr5_1 - attr_o, sinc = sinc5_1 - sinc_o

apply(dataset_diff[, -1], 2, mean)
#       att      sinc     intel       fun       amb
# 0.8050940 0.8215984 0.9756622 1.0622595 0.9163459

# All values are positive. This shows that people perceive themselves higher than reality.
```

5. In terms of getting a second date, is it better to be someone's first speed date of the night or their last?

```r
dataset <- select(DT, c("order", "pid", "round", "dec_o")) %>% remove_missing %>% group_by(pid)
first_order <- dataset %>% subset(order==1) %>% summarise(sum=sum(dec_o))
last_order  <- dataset %>% subset(order==round) %>% summarise(sum=sum(dec_o))

sum(first_order$sum[last_order$pid %in% first_order$pid])
# [1] 0
sum(last_order$sum[last_order$pid %in% first_order$pid])
# [1] 0

# Being someone's first speed date seems to have a better chance of getting another date compared to be
```

6. Write two other interesting observations using this data set.