STAT650 F2021 HW2 Solution

Xiaomeng Yan

9/06/2021

Problem 1

The data set ambulance.csv consists of a realistic but not real anonymized set of Emergency Medical Service calls in a Canadian city. Time units are in days past a reference time. The fields in the data are as follows:

- (a). Call_ID. A unique integer for each call.
- (b). Time_Rec. The time a call for EMS assistance was received, measured in days like all other time fields, with an arbitrary starting point.
- (c). Time_Vehicle_Alterted. The time a dispatcher alerted an ambulance that they would be sent to the call.
- (d). Time_Vehicle_Enroute. The time that the ambulance started driving to the call. *(e). Precancel_Dur. Time until the call was canceled for whatever reason. Equals -1 if the call was not canceled.
- (f). Time_Vehicle_At_Scene. The time that the ambulance arrived at the scene.
- (g). Time_Depart_Scene. The time that the ambulance left the scene of the call.
- (h). Time_Arrive_Hosp. The time that the ambulance arrived at a hospital. Equals -1 if the patient dit not require transport.
- (i). Time_at_Hospital. The time spent at the hospital transferring the patient in their care.
- (j). Scene_Lat. The latitude of the location of the call.
- (k). Scene_Lon. The Longitude of the location of the call.

a. What fraction of the calls are canceled? (Suggestion: Start by using length() on the column with entries == -1.)

Solution:

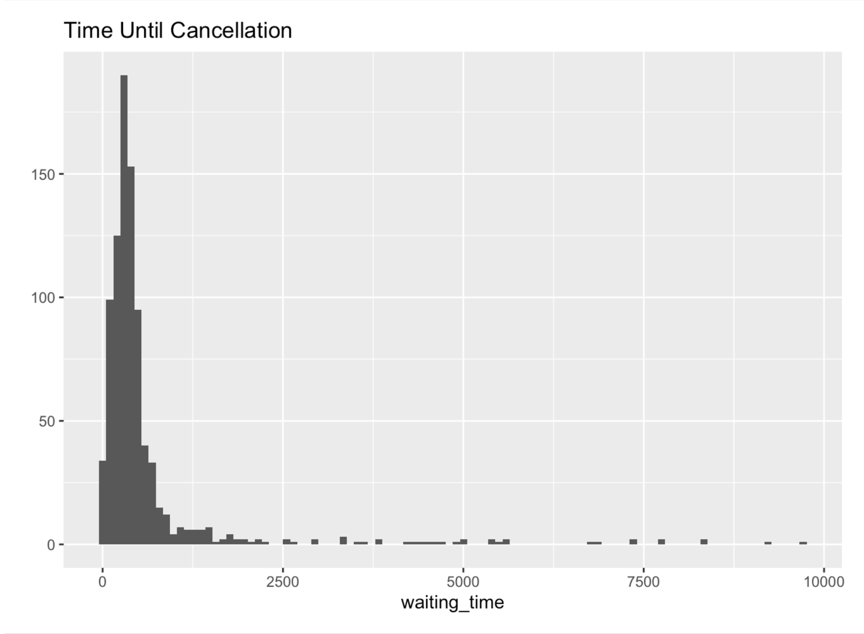
```
ambulance <- read.csv("ambulance.csv")
number_noncancel <- length(which(ambulance$Precancel_Dur==-1))
number_call <- length(ambulance$Precancel_Dur)
per_cancell <- 1- number_noncancel/number_call
print( c("Fraction of the calls are cancelled:", per_cancell))
```

```
## [1] "Fraction of the calls are cancelled:"
## [2] "0.0588666666666666"
```

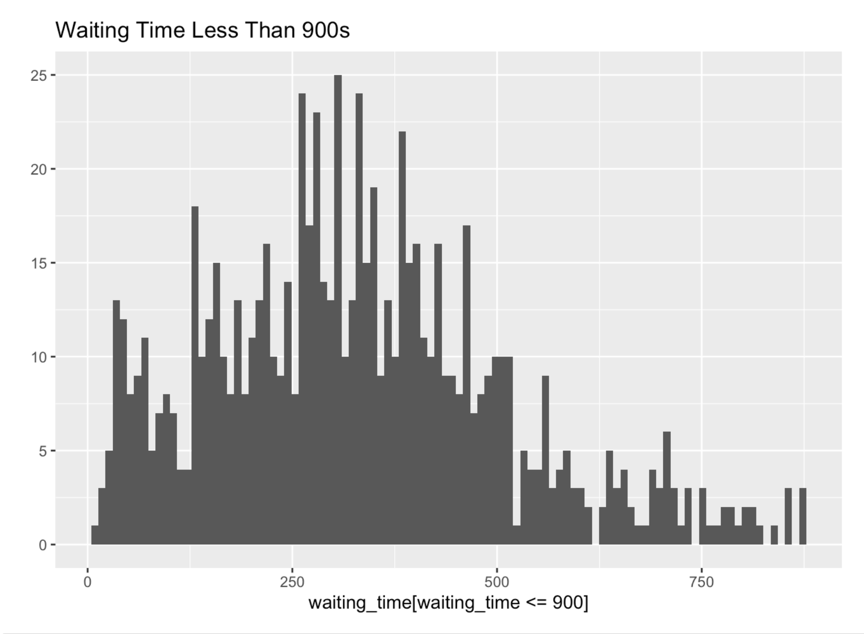
b. Provide a histogram for the time in seconds until cancellation for all of those calls that are eventually cancelled. (The R default for number of classes is not very illuminating. Try experimenting with larger value of nclass. Convert the vector x of values that correspond to durations of non-cancelled durations to second by multiplying by 24*60*60. The data has a "long tail" meaning there are a few but very large positive values. Play around with making the full plot and the plot of values ≤ 900 seconds.) As a percentage of the durations, how may durations are > 900 seconds?

Solution:

```
waiting_time<-24*60*60*ambulance$Precancel_Dur[ambulance$Precancel_Dur!=-1]
library(ggplot2)
qplot(waiting_time,geom = "histogram",bins = 100,main ="Time Until Cancellation")
```




```
qplot(waiting_time[waiting_time<=900], bins = 100,main="Waiting Time Less Than 900s")
```




```
per_greater<-sum(waiting_time>900)
print( c("Percenrage of the durations are >900:", per_greater/length(waiting_time)))
```

```
## [1] "Percenrage of the durations are >900:"
## [2] "0.104190260475651"
```

c. From now on focus on the non-cancelled calls. What fraction of these calls require transport to a hospital?

Solution:

```
number_nontrans<-length(which(ambulance$Precancel_Dur==-1&ambulance$Time_Arrive_Hosp==-1))
per_trans<-1-number_nontrans/number_noncancel
print( c("Fraction of non-cancelled calls require transport to a hospital:", per_trans))
```

```
## [1] "Fraction of non-cancelled calls require transport to a hospital:"
## [2] "0.838563434157399"
```

d. What are the mean and median times spent at the scene for calls that require transport the hospital?

Solution:

```
ambulance_sub <- ambulance[ambulance$Precancel_Dur==-1&ambulance$Time_Arrive_Hosp!=-1,]
Time_Spent_At_Scene<-ambulance_sub$Time_Depart_Scene-ambulance_sub$Time_Vehicle_At_Scene
print( c("Mean times spent at the scene for calls that require transport to hospital:", mean(Time_Spent_At_Scene)
))
```

```
## [1] "Mean times spent at the scene for calls that require transport to hospital:"
## [2] "0.0140039793791181"
```

```
print( c("Median times spent at the scene for calls that require transport to hospital:", median(Time_Spent_At_Sc
ene)))
```

```
## [1] "Median times spent at the scene for calls that require transport to hospital:"
## [2] "0.0124634250000071"
```

e. Generate a scatter plot of the time spent at scene versus the time spent at hospital, for those calls that transport to the hospital is required. What do you learn from the scatter plot?

Solution:

```
ambulance_sub <- ambulance[ambulance$Precancel_Dur==-1&ambulance$Time_Arrive_Hosp!=-1,]
# Create a new variable "duration_at_scene" to measure how long the ambulance was the scene
ambulance_sub$duration_at_scene <- (ambulance_sub$Time_Depart_Scene - ambulance_sub$Time_Vehicle_At_Scene)
# Make the scatter plot, zoom in to crop out erroneous values of -1
library(ggplot2)

ggplot(ambulance_sub, aes(x = duration_at_scene, y = Time_at_Hospital))+geom_point()+xlab("Time at Scene (Days)")
+ylab("Time at Hospital (Days)")
```

There is no obvious positive/negative correlation in the data. The vast majority of the points are very clustered. The most notable feature of the plot is that the distribution of time spent at the hospital seems to narrow as the time at the scene increases. This could be indicative that if an ambulance spends a long time at the scene than the visit to the hospital is more "routine" in some sense with a very low variance in duration.

f. EMS provides are usually judged by the fraction of calls that they receive that have response times (the elapsed time from when the call was initially received till a vehicle arrives at scene) that fall below some threshold. This threshold is typically on the order od 10 minus for high priority calls, and larger than that for lower-priority calls, or calls arising in remote areas. What fraction of non-canceled calls in this dataset have response times under 10 minutes?

Solution:

```
# Get the subset of calls that weren't canceled
noncancelled_calls = subset(ambulance, Precancel_Dur==-1)
# Create a new variable "response_time" to measure the time from the initial call to vehicle arriavle
noncancelled_calls$response_time <- (noncancelled_calls$Time_Vehicle_At_Scene - noncancelled_calls$Time_Rec)# Con
vert 10 minutes into units of days
ten_minutes_in_days = 10/(60*24)
# Count the number of responses that were 10 minutes or faster
number_response_less_than_10 = length(which(noncancelled_calls$response_time <= ten_minutes_in_days))
# Count the number of total calls
total_calls = length(noncancelled_calls$response_time)
# Report the fraction of calls with responses of 10 minutes or faster
print( c("Fraction of calls with less than 10 minutes response time:", number_response_less_than_10/total_calls))
```

```
## [1] "Fraction of calls with less than 10 minutes response time:"
## [2] "0.619607565346745"
```

Problem 2

(The game of craps) The game of craps is played by rolling two fair, sixe-sides dice. On the first roll, if the sum of two numbers showing equals 2, 3 or 12, then the player immediately loses. If the sum euqals 7 or 11, then the player immediately wins. If the sum equals any other value, then this value becomes player's "point". The player then repeatedly rolls the two dice, until such time as he or she either rolls the point value again (in which case he or she wins) or rolls a 7 (in which case he or she loses).

a. Suppose the player's point is equal to 4. Conditional on this, what is the conditional probability that he or she will win (i.e., will roll another 4 beofre a 7)? [Hint: The final roll will be either 4 or 7; what is the conditional probability that it is 4?]

Solution:

Consider the case when the point is 4. We continue to roll the dice until the sum is either 4 (in which case we win), or roll a 7 (in which case we lose). We know that the game does not end until either of these two scenarios occur, so we want to determine the probability that the sum is 4 given that either the sum 4 or 7 has occurred. To elaborate, to find out the probability of wining the game when the point is 4, this is simply the probability that we roll 4 before we roll 7. So, this is the same as saying, what is the probability that we roll a 4 given that we roll either a 4 or a 7.

Let A be the event that the sum of the dice is 4, and let B be the event that the sum of the dice is either 4 or 7. We wish to find $P(A|B) = \frac{P(A \cap B)}{B} = \frac{3/36}{3/36+6/36} = \frac{1}{3}$.

b. For $2 \leq i \leq 12$. let p_i be the conditional probability that the player will win, conditional on having rolled i on the first roll. Compute p_i for all i with $2 \leq i \leq 12$. [Hint: You've already done this for $i = 4$ in part (a). Also, the cases $i = 2, 3, 4, 11, 12$ are trivial. The other cases are similar to the $i = 4$ case.]

Solution:

$$
\begin{aligned}
p_2 &= P(\text{win}|\text{roll 2 on the first roll}) = 0 \\
p_3 &= P(\text{win}|\text{roll 3 on the first roll}) = 0 \\
p_{12} &= P(\text{win}|\text{roll 12 on the first roll}) = 0 \\
p_7 &= P(\text{win}|\text{roll 7 on the first roll}) = 1 \\
p_{11} &= P(\text{win}|\text{roll 11 on the first roll}) = 1
\end{aligned}
$$

Following the same argument with part (a), we have

$$
\begin{aligned}
p_5 &= P(\text{Roll 5}|\text{ Roll 5 or 7}) = \frac{4/36}{4/36+6/36} = \frac{2}{5} \\
p_6 &= P(\text{Roll 6}|\text{ Roll 6 or 7}) = \frac{5/36}{5/36+6/36} = \frac{5}{11} \\
p_8 &= P(\text{Roll 8}|\text{ Roll 8 or 7}) = \frac{5/36}{5/36+6/36} = \frac{5}{11} \\
p_9 &= P(\text{Roll 9}|\text{ Roll 9 or 7}) = \frac{4/36}{4/36+6/36} = \frac{2}{5} \\
p_{10} &= P(\text{Roll 10}|\text{ Roll 10 or 7}) = \frac{3/36}{3/36+6/36} = \frac{1}{3}
\end{aligned}
$$

c. Compute the overall probability that a player will win at craps. [Hint: Use part (b)]

Solution:

The probability of rolling a 7 with two dice is $\frac{6}{36}$, and the probability of rolliing an 11 with two dice is $\frac{2}{36}$. The probability of wining at the game can be calculated by

$$
\begin{aligned}
P(\text{win}) &= \sum_i P(\text{win, roll i on the first roll }) \\
&= \sum_i P(\text{win}|\text{roll i})P(\text{roll i}) \\
&= 2/9 + 1/3 \cdot 3/36 + 2/5 \cdot 4/36 + 5/11 \cdot 5/36 + 5/11 \cdot 5/36 + 2/5 \cdot 4/36 + 1/3 \cdot 3/36 \\
&= \frac{244}{495}
\end{aligned}
$$

Problem 3

Suppose we have a simple Weather dataset. Using a naive Bayes classifier, find the probability of playing golf on a sunny, hot and windy day with high humidity.

Solution:

Our goal is to calculate the probability below,

$$
\begin{aligned}
&P(\text{Play Golf} = \text{Yes}|\text{Outlook} = \text{Sunny, Temp} = \text{Hot, Humidity} = \text{High ,Windy} = \text{True }) \\
=&\frac{P(\text{Play Golf} = \text{Yes, Outlook} = \text{Sunny, Temp} = \text{Hot, Humidity} = \text{High ,Windy} = \text{True })}{P(\text{Outlook} = \text{Sunny, Temp} = \text{Hot, Humidity} = \text{High ,Windy} = \text{True })} \\
=&\frac{P(\text{Outlook} = \text{Sunny, Temp} = \text{Hot, Humidity} = \text{High ,Windy} = \text{True }|\text{Play Golf} = \text{Yes})P(\text{Play Golf} = \text{Yes})}{\sum_{\text{Play Golf} = \{\text{Yes,No}\}} P(\text{Outlook} = \text{Sunny, Temp} = \text{Hot, Humidity} = \text{High ,Windy} = \text{True }|\text{Play Golf})P(\text{Play Golf})}
\end{aligned}
$$

The numerator is calculated by

$$
\begin{aligned}
&P(\text{Outlook} = \text{Sunny, Temp} = \text{Hot, Humidity} = \text{High ,Windy} = \text{True }|\text{Play Golf} = \text{Yes})P(\text{Play Golf} = \text{Yes}) \\
=&P(\text{Outlook} = \text{Sunny}|\text{Play Golf} = \text{Yes})P(\text{Temp} = \text{Hot}|\text{Play Golf} = \text{Yes}) \\
&P(\text{Humidity} = \text{High}|\text{Play Golf} = \text{Yes})P(\text{Windy} = \text{True }|\text{Play Golf} = \text{Yes})P(\text{Play Golf} = \text{Yes}) \\
=&3/9 \times 2/9 \times 3/9 \times 3/9 \times 9/14 = 0.00529
\end{aligned}
$$

The denominator is calculated by

$$
\begin{aligned}
&P(\text{Outlook} = \text{Sunny, Temp} = \text{Hot, Humidity} = \text{High ,Windy} = \text{True }|\text{Play Golf} = \text{Yes})P(\text{Play Golf} = \text{Yes}) \\
+&P(\text{Outlook} = \text{Sunny, Temp} = \text{Hot, Humidity} = \text{High ,Windy} = \text{True }|\text{Play Golf} = \text{No})P(\text{Play Golf} = \text{No}),
\end{aligned}
$$

where, similarly,

$$
\begin{aligned}
&P(\text{Outlook} = \text{Sunny, Temp} = \text{Hot, Humidity} = \text{High ,Windy} = \text{True }|\text{Play Golf} = \text{No})P(\text{Play Golf} = \text{No}) \\
=&P(\text{Outlook} = \text{Sunny}|\text{Play Golf} = \text{No})P(\text{Temp} = \text{Hot}|\text{Play Golf} = \text{No}) \\
&P(\text{Humidity} = \text{High}|\text{Play Golf} = \text{No})P(\text{Windy} = \text{True }|\text{Play Golf} = \text{No})P(\text{Play Golf} = \text{No}) \\
=&2/5 \times 2/5 \times 4/5 \times 3/5 \times 5/14 = 0.0274.
\end{aligned}
$$

The final solution is: $0.00529/(0.00529 + 0.0274) = 0.162$

Problem 4

Entering high school students make program choices among general program, vovational program and academic program. The *student.dta* contains information of 200 students: Their scores in different subjects and their educational choices (general, academic or vocational). There are other variables indicating their social economic status and their gender. We will use their social economic status **ses**, their gender **female** and school type **schtyp** to classify their educational choice **prog**.

```
library(foreign)
student <- read.dta("student.dta",convert.factors = TRUE)
head(student)
```

```
##    id female    ses schtyp     prog read write math science socst       honors
## 1  45 female    low public vocation   34    35   41      29    26 not enrolled
## 2 108   male middle public  general   34    33   41      36    36 not enrolled
## 3  15   male   high public vocation   39    39   44      26    42 not enrolled
## 4  67   male    low public vocation   37    37   42      33    32 not enrolled
## 5 153   male middle public vocation   39    31   40      39    51 not enrolled
## 6  51 female   high public  general   42    36   42      31    39 not enrolled
##   awards cid
## 1      0   1
## 2      0   1
## 3      0   1
## 4      0   1
## 5      0   1
## 6      0   1
```

(a). Install the package *foregin* and load the data into R using the function *read.dta*. Plot the side by side boxplot for *math* and *science* scores w.r.t different program choice using the package *ggplot2*.

Solution:

```
library(ggplot2)                    ## Package required to do plot
library(gridExtra)                  ## Package required to arrange plots
## Boxplots of science score v.s education choice
p1 = ggplot(data = student,aes(x = prog,y = science,col = prog,fill = prog))+geom_boxplot(alpha = 0.4)+ggtitle("s
cience score boxplot")
p2 = ggplot(data = student,aes(x = prog,y = math,col = prog,fill = prog))+geom_boxplot(alpha = 0.4)+ggtitle("math
score boxplot")
grid.arrange(p1,p2,nrow = 1,ncol = 2)
```

(b). Fit the Naive Bayes to classify the program choice **prog** using the three categorical variables gender **female**, school type **schtyp** and social economic status *ses*. Answer the questions.

- What fraction of the students who choose "academic" program.

```
Choice <- student$prog
Traindf <- student[,c("female","schtyp","ses")]
library(e1071)
classifier <- naiveBayes(Traindf,Choice)
classifier
```

```
##
## Naive Bayes Classifier for Discrete Predictors
##
## Call:
## naiveBayes.default(x = Traindf, y = Choice)
##
## A-priori probabilities:
## Choice
##  general academic vocation
##    0.225    0.525    0.250
##
## Conditional probabilities:
##           female
## Choice          male    female
##   general  0.4666667 0.5333333
##   academic 0.4476190 0.5523810
##   vocation 0.4600000 0.5400000
##
##           schtyp
## Choice        public   private
##   general  0.8666667 0.1333333
##   academic 0.7714286 0.2285714
##   vocation 0.9600000 0.0400000
##
##           ses
## Choice           low    middle      high
##   general  0.3555556 0.4444444 0.2000000
##   academic 0.1809524 0.4190476 0.4000000
##   vocation 0.2400000 0.6200000 0.1400000
```

Solution: $P(\text{academic}) = 0.525$

- Given a student is a female, what is the probability that she chooses vacational program.

Solution:

$$
\begin{aligned}
P(\text{vocation}|\text{female}) &= \frac{P(\text{vocation} \cap \text{female})}{P(\text{female})} \\
&= \frac{P(\text{female}|\text{vocation})P(\text{vocation})}{P(\text{female}|\text{academic})P(\text{academic}) + P(\text{female}|\text{general})P(\text{general}) + P(\text{female}|\text{vacation})P(\text{vacation})} \\
&= \frac{0.5400000 \cdot 0.25}{0.5523810 \cdot 0.525 + 0.5333333 \cdot 0.225 + 0.5400000 \cdot 0.25} = 0.248
\end{aligned}
$$

```
(0.5400000* 0.25)/(0.5523810* 0.525+0.5333333* 0.225+0.5400000* 0.25)
```

```
## [1] 0.2477064
```

- Are the social economic status and whehter the student chooses the academic program independent? Explain.

Solution:

$$
\begin{aligned}
P(\text{academic}|\text{ses} = \text{low}) &= \frac{P(\text{academic} \cap \text{ses} = \text{low})}{P(\text{ses} = \text{low})} \\
&= \frac{P(\text{ses} = \text{low}|\text{academic})P(\text{academic})}{P(\text{ses} = \text{low}|\text{academic})P(\text{academic}) + P(\text{ses} = \text{low}|\text{general})P(\text{general}) + P(\text{ses} = \text{low}|\text{vacation})P(\text{vacation})} \\
&= \frac{0.1809524 \cdot 0.525}{0.1809524 \cdot 0.525 + 0.3555556 \cdot 0.225 + 0.2400000 \cdot 0.25} \\
&= 0.404 \neq P(\text{academic})
\end{aligned}
$$

```
(0.1809524* 0.525)/(0.1809524* 0.525+0.3555556* 0.225+0.2400000* 0.25)
```

```
## [1] 0.4042553
```

Not independent.