COMPSCIX 415.2 Homework 9/Final

Sanatan Das

March 29, 2018

Contents

Code and Documents Git Repository

All the work can be found in the below Git repository location:

https://github.com/sanatanonline/compscix-415-2-assignments

Load packages (prerequisites to run the code in this document)

```
library(tidyverse)
library(rpart)
library(partykit)
library(modelr)
library(randomForest)
```

Bootstrapping (10 points)

Load the Titanic dataset

1. **Follow these steps:**
   - **Load the train.csv dataset into R.**
   - **Convert all character columns into unordered factors.**
   - **Convert the Survived column into an unordered factor because it is loaded as an integer by default.**
   - **Take a glimpse of your data to confirm that all of the columns were converted correctly.**

We will use this same dataset for this entire assignment.

Answer:

```
# load train.csv
train = read_csv("C:/view/opt/apps/git/compscix-415-2-assignments/titanic/train.csv")
```

```
## Parsed with column specification:
## cols(
##   PassengerId = col_integer(),
##   Survived = col_integer(),
##   Pclass = col_integer(),
##   Name = col_character(),
##   Sex = col_character(),
##   Age = col_double(),
##   SibSp = col_integer(),
##   Parch = col_integer(),
##   Ticket = col_character(),
##   Fare = col_double(),
##   Cabin = col_character(),
##   Embarked = col_character()
## )
```

```
# glimpse train
glimpse(train)
```

```
## Observations: 891
## Variables: 12
## $ PassengerId <int> 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15,...
## $ Survived    <int> 0, 1, 1, 1, 0, 0, 0, 0, 1, 1, 1, 1, 0, 0, 0, 1, 0,...
## $ Pclass      <int> 3, 1, 3, 1, 3, 3, 1, 3, 3, 2, 3, 1, 3, 3, 3, 2, 3,...
## $ Name        <chr> "Braund, Mr. Owen Harris", "Cumings, Mrs. John Bra...
## $ Sex         <chr> "male", "female", "female", "female", "male", "mal...
## $ Age         <dbl> 22, 38, 26, 35, 35, NA, 54, 2, 27, 14, 4, 58, 20, ...
## $ SibSp       <int> 1, 1, 0, 1, 0, 0, 0, 3, 0, 1, 1, 0, 0, 1, 0, 0, 4,...
## $ Parch       <int> 0, 0, 0, 0, 0, 0, 0, 1, 2, 0, 1, 0, 0, 5, 0, 0, 1,...
## $ Ticket      <chr> "A/5 21171", "PC 17599", "STON/O2. 3101282", "1138...
## $ Fare        <dbl> 7.2500, 71.2833, 7.9250, 53.1000, 8.0500, 8.4583, ...
## $ Cabin       <chr> NA, "C85", NA, "C123", NA, NA, "E46", NA, NA, NA, ...
## $ Embarked    <chr> "S", "C", "S", "S", "S", "Q", "S", "S", "S", "C", ...
```

```
# Add the factors
train$Survived <- as.factor(train$Survived)
train$Name <- as.factor(train$Name)
train$Sex <- as.factor(train$Sex)
train$Ticket <- as.factor(train$Ticket)
train$Cabin <- as.factor(train$Cabin)
train$Embarked <- as.factor(train$Embarked)
# glimpse train again
glimpse(train)
```

```
## Observations: 891
## Variables: 12
## $ PassengerId <int> 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15,...
## $ Survived    <fct> 0, 1, 1, 1, 0, 0, 0, 0, 1, 1, 1, 1, 0, 0, 0, 1, 0,...
## $ Pclass      <int> 3, 1, 3, 1, 3, 3, 1, 3, 3, 2, 3, 1, 3, 3, 3, 2, 3,...
## $ Name        <fct> Braund, Mr. Owen Harris, Cumings, Mrs. John Bradle...
## $ Sex         <fct> male, female, female, female, male, male, male, ma...
## $ Age         <dbl> 22, 38, 26, 35, 35, NA, 54, 2, 27, 14, 4, 58, 20, ...
## $ SibSp       <int> 1, 1, 0, 1, 0, 0, 0, 3, 0, 1, 1, 0, 0, 1, 0, 0, 4,...
## $ Parch       <int> 0, 0, 0, 0, 0, 0, 0, 1, 2, 0, 1, 0, 0, 5, 0, 0, 1,...
## $ Ticket      <fct> A/5 21171, PC 17599, STON/O2. 3101282, 113803, 373...
## $ Fare        <dbl> 7.2500, 71.2833, 7.9250, 53.1000, 8.0500, 8.4583, ...
## $ Cabin       <fct> NA, C85, NA, C123, NA, NA, E46, NA, NA, NA, G6, C1...
## $ Embarked    <fct> S, C, S, S, S, Q, S, S, S, C, S, S, S, S, S, S, Q,...
```

Now all the character columns are converted to unordered factors.

Bootstrap samples (100) of titanic data

2. Use the code below to take 100 bootstrap samples of your data. Confirm that the result is a tibble with a list column of resample objects - each resample object is a bootstrap sample of the titanic dataset.

```
library(tidyverse)
library(modelr)
titanic_boot <- bootstrap(data = ____, n = ___)
```

Answer:

Let's fill in the code.

```r
titanic_boot <- bootstrap(data = train, n = 100)
#see the type of titanic_boot
class(titanic_boot)
```

```
## [1] "tbl_df"     "tbl"        "data.frame"
```

```r
# see titanic_boot
titanic_boot
```

```
## # A tibble: 100 x 2
##    strap          .id
##    <list>         <chr>
##  1 <S3: resample> 001
##  2 <S3: resample> 002
##  3 <S3: resample> 003
##  4 <S3: resample> 004
##  5 <S3: resample> 005
##  6 <S3: resample> 006
##  7 <S3: resample> 007
##  8 <S3: resample> 008
##  9 <S3: resample> 009
## 10 <S3: resample> 010
## # ... with 90 more rows
```