TA Session 1

Nikhil Mehta & Rajkumar Pujari

PCA

What is R?

- Download instructions: https://cran.r-project.org
  - RStudio is a free IDE for R, which you can use
- “R is a language and environment for statistical computing and graphics”
- Has statistical functionalities
  - Linear and nonlinear modeling, classical statistical tests, time-series analysis, classification, clustering
- Graphical techniques
- Highly extensible
- For computationally-intensive tasks, C, C++ and Fortran code can be linked and called at run time

R Basics

- Installing a package
  - install.packages(“nycflights13”)
  - Installs this package: https://github.com/hadley/nycflights13
- Loading a package
  - library(nycflights13)
- Finding help about the data
  - ?nycflights13::flights
- Viewing the flights table
  - nycflights13::flights
- Using the help
  - ?summary or help(summary)

R Basics Cont.

- Variable names are case sensitive and can include “_”.
- Commands are separated by semicolon or a newline
- Commands can be grouped using braces {}
- #Comments
- Use up arrow to navigate command history
- Use history() to see the entire command history

R Basics Cont.

- To execute a file source('/path/my_file.r')
  - "Source causes R to accept its input from the named file or URL or connection or expressions directly"
  - Input is read and parsed from that file, then the parsed expressions are evaluated
- Assignments can be done with "<-" or "="
  - There are some differences, such as scope in workspaces or assignment problems
    - "x <- y = 5" throws an error, as it is parsed as `=(<-(x, y), 5)`, since '=' is lower precedence
    - However, "x <- y <- 5" works
    - Thus, "<-" is preferred, RStudio has a shortcut to make it easier to type
- Creating a vector
  - X <- c(10.4, 5.6, 3.1, 6.4, 21.7)

R Basics Cont.

- Arithmetic expressions using vectors are applied element by element
- Useful functions: min, max, var, sd(std), mean, length
- Logical vectors: x > 10
- Indexing with logical vectors x[x > 10] = -1, x[x < 4]
- Slicing: x[1:3]
- R lists: list(a = 1, b = “foo”, c = 1:5)
- For more:
  - https://cran.r-project.org/doc/manuals/r-release/R-intro.html#Introduction-and-preliminaries
  - Google
  - Piazza

R-Demo

Probability Review

- What is a random variable?
  - Mapping from a property of objects to a variable that can take one of a set of possible values
  - X refers to a random variable while x refers to one for the possible values
- Types of random variables:
  - Boolean: flip a coin
  - Discrete: roll a die
  - Continuous: Temperature

- Sample space
  - Set of possible outcomes of an experiment.
  - In roll a die of six faces, S = {1, 2, 3, 4, 5, 6}
- Event:
    - Any subset of outcomes contained in the sample space of S
    - Events A and B are mutually exclusive if they cannot be both true at the same time.

- Probability distributions:
  - Specify the probability of observing every possible value of a random variable.
  - Discrete: probability of a point

$$P(X = x)$$

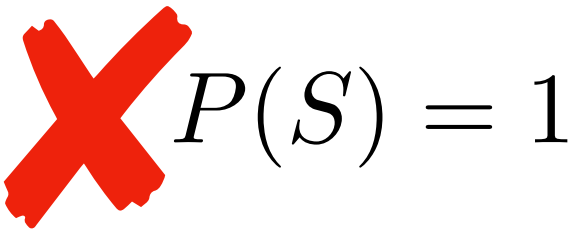
  - Continuous: probability of an interval

$$P(a < X < b) = \int_a^b p(x)dx$$

- Probability density function $P(A)$:
  - Associates a real value to each event A in S
  - Some properties:
    - $0 \leq P(A) \leq 1$
    - $P(S) = 1$
    - $P(A) = 1 - P(\neg A)$
    - $P(A \cup B) = P(A) + P(B) - P(A \cap B)$
    - $P(A \cap B) = 0$ If A and B are mutually exclusive

- What about this probability density function of die?
  - P(X=1) = 0.2
  - P(X=2) = 0.2
  - P(X=3) = 0.2
  - P(X=4) = 0.2
  - P(X=5) = 0.2
  - P(X=6) = 0.2

- What about this probability density function of die?
  - P(X=1) = 0.2
  - P(X=2) = 0.2
  - P(X=3) = 0.2
  - P(X=4) = 0.2
  - P(X=5) = 0.2
  - P(X=6) = 0.2



- Joint probability: $P(X = x \cap Y = y)$
  - Probability of event X and Y occurring at the same time
- Conditional Probability: $P(X = x|Y = y)$
  - Probability of an event X occurring given an event Y
- Product Rule:

$$P(X = x \cap Y = y) = P(X = x|Y = y)P(Y = y)$$

- Bayes Theorem

$$P(X = x|Y = y) = \frac{P(Y = y|X = x)P(X = x)}{P(Y = y)}$$

- Marginal Probability:
  - Probability regardless of conditional events

$$P(X = x) = \sum_{i \in S_y} P(X = x|Y = i)P(Y = i)$$

- Lets focus on this example:

A group of police officers have breathalyzers displaying false drunkenness in 5% of the cases in which the driver is sober. However, the breathalyzers never fail to detect a truly drunk person. One in a thousand drivers is driving drunk. Suppose the police officers then stop a driver at random, and force the driver to take a breathalyzer test. It indicates that the driver is drunk. We assume you don't know anything else about him or her. How high is the probability he or she really is drunk?

https://en.wikipedia.org/wiki/Base_rate_fallacy

A group of police officers have breathalyzers displaying **(1) false drunkenness in 5% of the cases in which the driver is sober**. However, the **(2) breathalyzers never fail to detect a truly drunk person**. **(3) One in a thousand drivers is driving drunk**. Suppose the police officers then stop a driver at random, and make the driver take a breathalyzer test. It (4) indicates that the driver is drunk. We assume you don't know anything else about him or her. **(5) How high is the probability he or she really is drunk**?

(1) P(Test=Positive | Driver = sober) = 0.05
(2) P(Test=Positive | Driver = drunk) = 1
(3) P(Driver = drunk) = 1/1000 = 0.001
(5) P(Driver=Drunk | Test=Positive) = ?

(1) P(Test=Positive | Driver = sober) = 0.05
(2) P(Test=Positive | Driver = drunk) = 1
(3) P(Driver = drunk) = 1/1000 = 0.001
(5) P(Driver=Drunk | Test=Positive) = ?

$$P(Driver = D|Test = P) = \frac{P(Test = P|Driver = D)P(Driver = D)}{P(Test = P)}$$

$$
\begin{aligned}
&P(Driver = D | Test = P) \\
&= \frac{P(Test = P | Driver = D) P(Driver = D)}{P(Test = P)} \\
&= \frac{P(Test = P | Driver = D) P(Driver = D)}{P(Test = P | Driver = D) P(Driver = D) + P(Test = P | Driver = S) P(Driver = S)} \\
&= \frac{1.00 * 0.001}{1.00 * 0.001 + 0.05 * (1 - 0.001)} \\
&= 0.019627
\end{aligned}
$$