Intro to DS 6030

Statistical Learning and Data Mining

DS 6030 | Fall 2022

intro.pdf

Contents

1 Course Website

Main Course Webpage: https://mdporter.github.io/DS6030

2 About us

2.1 About the Instructor

- Faculty Webpage https://mdporter.github.io/
- GitHub https://github.com/mdporter
- Blog https://mdporter.github.io/blog/

2.2 About our TA

Jiahao Tian

2.3 About you

Fill out a notecard with the following information:

1. Your name (with pronunciation hints)
2. Hometown (include country/region if you think I won't know)
3. Previous and Current Degrees
4. What type of job to hope to land on graduation (title & industry)
5. 2 things you hope to learn in this course
6. 2 interesting things about you (to help me remember you)

3 The course

3.1 Topics

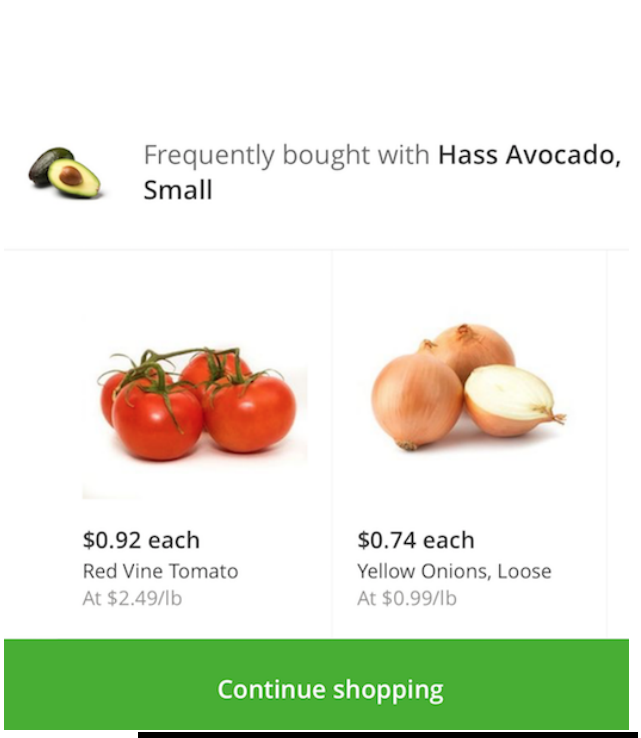
- See website: https://mdporter.github.io/DS6030
- Course contains aspects of: data analysis, modeling, stats, ML, coding, algorithms, probability, etc.

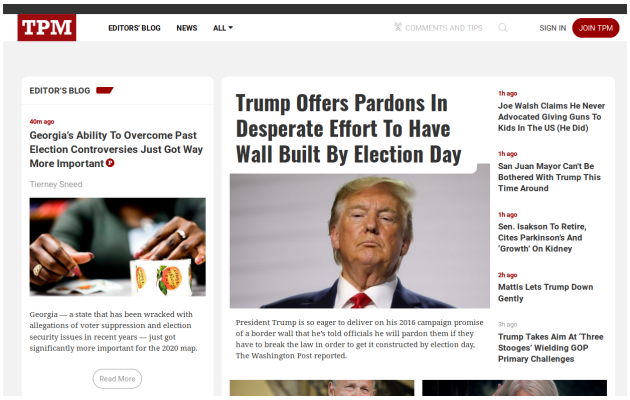
Data Scientists are expected to be *fluent* in all!

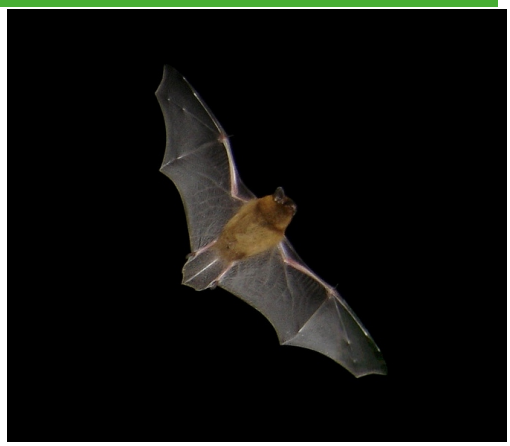
- You are expected to be problem solvers
  - doing good on structured homework sets isn't sufficient

3.2 Examples

- Discover the hero most likely to appear in a comic with the fantastic 4
- Understand why Red Vine Tomatoes are suggested to an Avocado buyer
- Identify the most "influential" political bloggers
- Predict how far *pipistrelle* bats travel from their roost to find food (and what this can tell us about criminal offenders)
- And many more







4 Syllabus

4.1 Course Webpage

- We have a course webpage https://mdporter.github.io/DS6030
  - lectures
  - R scripts
  - data sets
  - homework assignments
- We will use the Collab site for homework submission, solutions, etc.
- We will use the Teams channel: Teams DS 6030 Stat Learning for other group communication

4.2 Course Prereqs

- Linear Regression
  - Multiple Linear Regression
  - Logistic Regression
  - Categorical Predictors (dummy coding)
  - Implementation in R (lm(), predict(), etc.)
  - Estimation / Model Fitting
  - Cross-validation
- Probability and Statistics
  - Bayes Theorem
  - CDF/PDF/PMF
  - Maximum Likelihood Estimation
  - Distributions: normal, binomial, hypergeometric, etc.
  - Expected value, variance, median, quantiles
  - Mean Square Error
  - Confidence Intervals
  - Hypothesis Testing
- Math
  - Calculus
  - Matrix Calculations
  - PCA, SVD
- Computing
  - data types: vector, matrix, array, list, etc.
  - writing simple functions
  - flow control: loops, if/else, etc.
  - data wrangling
  - generating random variables
  - RMarkdown [*Note: practice HW will cover RMarkdown*]

4.3 Exercise 1

Your Turn #1

Let $X_1, X_2, \ldots, X_n$ be the yearly number of crashes at an intersection (X_i is number of crashes in year i).

- What is an estimate of the probability that there are 100 crashes in year $n + 1$?

Your Turn #2 : Continued

Your Turn #3 : Continued

4.4 Other Syllabus Material

- Office Hours
- Textbooks
- R, RStudio
- Course Assessment
  - Due dates are posted on the course website and Collab
  - RMarkdown (See HW0)
  - No class participation grade, but expect you come prepared with questions. Don't be afraid to ask questions in class. Now is your time to learn.
- Course Management
- **Honor Code**
- Read all of syllabus and ask questions (preferably on Slack)

4.5 Succeding in this course

- Most topics are separated into two lectures
  - First is introduction of new topic
  - Second is more advanced coverage
- Homework is due weekly
  - Due on Tuesday morning, but expected to be completed before Monday's class.
  - Should start HW after first lecture. Questions in second lecture.
- Assigned Readings *before* every class
  - First listed reading is intro, second is more advanced
  - Start with intro, then re-read the advanced
  - Quizzes based on first reading
- Attend office hours!

4.5.1 Data Science

The free textbook Modern Data Science with R is an undergrad level "Intro to Data Science" course. It covers tidyverse, statistical inference, and basic intro to many of the methods we will study this semester. This would provide a good overall preparation. Especially sections 2, 3, 4, 6, 7, 9.

4.5.2 Coding

I find that many students struggle with coding. This really hinders your ability to get your mind about the concepts and slows down your learning. The course will use R, but all examples will use the tidyverse dialect. There is no better tool for interactive data analysis and both exploratory and confirmatory modeling. Tidyverse is a major improvement over base R, but it can look a bit different and take some time becoming familiar with. The free online book R for Data Science and website provide a good introduction and reference. While I encourage tidyverse, you are free to use anything for homework. The UVA library also has good material (e.g., Getting Started with Data Science) as does Data Carpentry: R for Social Science.

- Rstudio has videos and tutorials
  - https://rstudio.cloud/learn/primers

  - Handy Rstudio cheatsheets

4.5.3 Statistics

I find students understand the least about statistical concepts. This is so fundamental to all of ML and Data Mining; a strong grasp of statistics will enable the connections between topics to pop out. If you already feel comfortable coding, I suggest you go a quick stat review. Here are two introductory resources:

- https://www.openintro.org/book/ims/
- https://moderndive.com/index.html
- UVA's library also offers lots of resources
  - https://data.library.virginia.edu/statlab/
  - https://data.library.virginia.edu/statlab/statlab-articles/
  - https://data.library.virginia.edu/statlab/data-science-resources/
  - https://data.library.virginia.edu/training/
  - https://data.library.virginia.edu/training/past-workshops/

4.5.4 Math

The students who gain the most from the program will embrace mathematical equations. As they say "an equation is worth a thousand words". While we won't do any proofs in this class, we will judiciously use equations to clarify concepts. Spend time to become intimate with math notation – it is worth the investment.

4.5.5 Trustworthy Material

- The assigned readings are trustworthy
- Blogs and videos you find on the web are not
- Please don't trust: Toward Data Science, Analytics Vidha, Machine Learning Mastery, Medium
  - There is certainly some good content, but how will know to discern good from bad while still learning?