Data Science Lecture Notes

Simon Xiang

Lecture notes for Elements of Data Science (SDS 322 E). These notes were taken live in class (and so they may contain many errors). Source files: https://git.simonxiang.xyz/math_notes/files.html

Contents

1 Introduction

todo:REMOVE THIS FROM GIT We're gonna use R and RMarkdown to do stuff this semester. The goal of this class is to get a foundation in the types of things you do in routine data analytics, like data visualization, data wrangling, etc. We need to be able to manipulate complex data sets and prepare reports, stuff like that. We will not cover inferential statistics, like p-values, confidence intervals, etc. No theoretical justification for machine learning boo.

We will follow the text reasonably closely, it's a good idea to read it before class. Class will consist of half lectures and half worksheets/homework. **Homework is due Sunday 11:59 PM**, and there are **labs closing Tuesday 11:59 PM**. Labs are basically quizzes and homeworks are filling out worksheets. HTML documents are accepted as submissions but try to default to PDFs. Grading is on a point scale: 10 points for a homework, 5 points for a lab, and each of the two projects is worth 150 points. Two lowest scoring homeworks are dropped. There are 465 points total. You need 432 points for an A so try not to lose more than **33 points** total.

Try not to be late but you can request extensions for traveling for athletics, religious, or serious indicidents/emergency. Regrades are possible.

1.1 R stuff

Do you know how to use RMarkdown? Make a markdown file with the extension `.rmd` and knit it in RStudio to compile. Use `:RMarkdown` with the plugins to do it in vim. You can attach things like tweets and Python snippets in rmd files.

2 More R stuff

You can use the terminal/console or a script to interact with R. You can also run R snippets within rmd files. Another way to interact is through RStudio. R can be used as a glorified calculator (how I use it).

See `testing.Rmd` for use cases and basic syntax. Define variables with arrow and you can see them in the "environment" area in RStudio.