The Data Scientist's Toolbox - Notes

Tanner Prestegard

Course taken from 3/2/2015 - 3/28/2015

What do data scientists do?

1. Define the question.
2. Define the ideal data set.
3. Determine what data you can access.
4. Obtain the data.
5. Clean the data.
6. Exploratory data analysis.
7. Statistical predition/modeling.
8. Interpret the results.
9. Challenge the results.
10. Synthesize/write up results.
11. Create reproducible code.
12. Distribute results to other people.

The Data Scientist's Toolbox

- R is the data science community's workhorse.
- RStudio is an IDE for R; supposed to be very useful.
- We will use R markdown documents as documentation for our code - HTML files are generated from R markdown documents.
- We will use Github for sharing code and documents.

Getting help in R

- ?rnorm - will give you help documentation for function 'rnorm'.
- help.search("rnorm") - more flexible, may not even have to get the naming right.
- Get arguments - args("rnorm")
- Type function name without any backets - R will reproduce the code in the console.
- Can try StackOverflow for help from others.
- How to ask an R question:
  - What steps will produce the problem?
  - What is the expected output?
  - What do you see instead?
  - What version of the product (R, other packages) are you using?
  - What operating system are you using?
- How to ask a data analysis question:
  - What is the question you are trying to answer?
  - What steps/tools did you use to answer it?
  - What did you expect to see?
  - What do you see instead?
  - What other solutions have you thought about?
- Can get data analysis / statistics help on CrossValidated.

Notes on Googling data science questions

- Use [data type] data analysis or [data type] R package
- Try to identify what data analysis is called for your data type:
  - Biostatistics for medical data.
  - Data science for data from web analytics.
  - Machine learning for data in computer science or computer vision.
  - Natural language processing for data from texts.
  - Signal processing for data from electrical signals.
  - Business analytics for data on customers.
  - Econometrics for economic data.
  - Statistical process control for data about industrial processes.

Git and Github

- Need to download git - a free version control system.
  - Terminal interface in Windows - git bash
  - Open git bash, and do the following lines:
    - a
    - b
- Github - allows users to push and pull their local repositories to and from remote repositories on the web.
  - Provides a way to share your files with others and access their files.
  - Provides a remote backup method in case your local files are lost.
- Creating a Github repository (two methods):
  - Can create your own repo from scratch - use the Github interface to do this.
    - In git bash, make a new directory, then do 'git init'.
    - Check out your directory with 'git remote add origin https://github.com/tprestegard/test-repo.git'
  - Or, you can "fork" another user's repository.
    - Fork it using the Github interface.
    - Then get a local copy using 'git clone https://github.com/tprestegard/repoName.git'

Basic git commands

- To add all new files: git add . (assumes you are in the directory of the new files you want to add)
- To update file changes (deletion or name changes): git add -u
- To do both of the previous things: git add -A
- Commit to the LOCAL repo using: git commit -m "message", where message is a useful description of what you did.
- To commit to the remote repo on Github: git push
- Branches:
  - Create a branch: git checkout -b branch_name
  - To see what branch you are on: git branch
  - To switch back to the master branch type: git checkout master
  - If you want to merge your changes into another branch or repo:
    - Go to Github and click the "Compare and pull request", then the other user can decide if they want to allow the request.
- Help:
  - Git documentation: http://git-scm.com/doc
  - Github help: https://help.github.com
  - Stack Overflow

Markdown

- Use extension .md for markdown files.
- Headings:
  - ## creates a secondary heading.
  - ### creates a tertiary heading.
- Lists: use '* item* to put items into a bulleted list.

Downloading and installing R packages

- Primary location to get R packages - CRAN.
- Use the available.packags() function to get information about available packages on CRAN.
- To install an R package, use the install.packages() function in R.
  - Example: install.packages("package-name")
  - Example for installing multiple packages: install.packages(c("package1","package2","package3"))
- Loading R packages: use library(package-name). Don't use quotes!
  - Package dependencies will be loaded first.
  - After loading, the functions exported by that package will be attached to the top of the search list.
  - Use search() to see them.

Types of data science questions

- Descriptive
  - Goal: describe a set of data.
  - First type of data analysis performed.
  - Descriptions cannot usually be generalized without additional statistical modeling.
  - Description and interpretation are different steps.
- Exploratory
  - Goal: find relationships you didn't know about
  - Good for discovering new connections and defining future studies.
  - Exploratory analyses are usually not the final say and they should not be used alone for generalizing/predicting.
  - Correlation does not imply causation.
- Inferential
  - Goal: use a relatively small sample of data to say something about a bigger population.
  - Inference is commonly the goal of statistical models.
  - Involves estimating both the quantity you are interested in and your uncertainty about your estimate.

- Depends heavily on both the population and the sampling scheme.

- Predictive
  - Goal: use the data on some objects to predict values for another object.
  - If X predicts Y, it does NOT mean that X causes Y.
  - Accurate prediction depends heavily on measuring the right variables.
  - Although some models are better than others, more data and a simple model is generally a good recipe.
  - Prediction is very hard, especially about the future.

- Causal
  - Goal: to find out what happens to one variable when you change another variable.
  - Usually requres randomized studies to identify causation.
  - There are approaches to inferring causation in non-randomized studies, but they are complicated and sensitive to the assumptions you make.
  - Causal relationships are usually identified as "average" effects, but may not apply to every individual.
    - Ex: if you give this population this drug, on average, they may have an increased lifespan.
  - Causal models are usually the "gold standard" for data analysis.

- Mechanistic
  - Goal: understand the exact changes in variables that lead to changes in other variables for individual objects.
  - Incredibly hard to infer, except in very simple situations.
  - Usually modeled by a deterministic set of equations.
  - Most commonly applicable in physical sciences or engineering.
  - Generally the only random component is measurement error.
  - If the equations are known but the parameters are not, they may be inferred with data analysis.

What is data?

- Data are values of qualitative or quantitative variables, belonging to a set of items.
  - Variables are measurements or characteristics of an item.
  - Qualitative: country of origin, gender, etc.
  - Quantitative: height, weight, blood pressure, etc.
- Data can be text files, video, audio, etc.
- The most important thing in data science is the question - the second most important is the data.
- Often, the data will limit or enable the questions, but having data can't save you if you don't have a question.

Experimental design

- Have a plan for data and code sharing.
- Formulate your question in advance.
- Confounding - assuming a correlation between two variables when really it is a different variable causing the correlation.
  - Example: does shoe size correlate to literacy, or is it really age? (kids have smaller shoe size and lower literacy)
  - Also known as spurious correlation.
  - Randomization and blocking
    * If you can and want to, fix a variable.
    * If you don't fix a variable, stratify it (i.e., use all possible values of the variable equally with all other variables).
    * If you can't fix a variable, randomize it.
- Prediction - key quantities
  - Sensitivity: Pr(positive test | you have the disease)
  - Specificity: Pr(negative test | no disease)
  - Positive predictive value: Pr(you have the disease | positive test)
  - Negative predictive value: Pr(no disease | negative test)
  - Accuracy: Pr(correct outcome) = sum of positive and negative predictive values.
- Beware "data dredging" - if you do 100 trials, it's expected that 5% will fall outside a 2 sigma level.
- Good experiments:
  - Have replication.
  - Measure variability.
  - Generalize to the problem you care about.
  - Are transparent.
- Prediction is not inference, but both can be important.