Data Scientist's Toolbox

Nicola Davide D'Avanzo

16/12/2014

This document is a generic overview about the data scientist's skills and attitudes.

What do data scientists do?

- Define the **questions of interest**
- Define the **ideal data set**
- Determine what **data you can access**
- **Obtain** the data
- **Clean** the data
- Exploratory **data analysis**
- **Statistical prediction/modeling**
- **Interpret** results
- **Challenge** results
- **Synthesize/write up** results
- Create **reproducible code**
- **Distribute** results to other people

Data Scientists tools:

- R (statistical programming environment)
- GitHub (Git repository web-based hosting service)
- Terminal Linux

R questions:

- What steps will reproduce the problem?
- What is the expected output?
- What do you see instead?
- What version of the product are you using?
- What operating system?

Data analysis questions:

- 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?

Data analysis files:

- Data (Raw Data, Processed Data)
- Figures (Explorator figures, Final figures)
- R code (Raw scripts, Final scripts, R markdown files)
- Text (Readme files, Text of analysis)

Command Line Interface (CLI):

- **Navigate** folders
- **Create** file, folders and programs
- **Edit** file, folders and programs
- **Run** computer programs

CLI commands:

command flags arguments

- pwd
- clear
- ls -al
- cd
- mkdir
- touch
- cp -r
- rm -r
- mv new_file renamed_filed
- echo
- date

Git:

open-source version control system

- most popular
- **local** repository
- command line
- git config - -global user.name " your_user_name "
- git config - -global user.email " your_email@example.com "
- git config - -list

GitHub:

web-based hosting service for software development project that use the Git revision control system

- **Remote** repository (on the web)
- Homepage repository **display**
- **Backup**
- **Follow** (access) and **share**

Creating GitHub repository:

- from **Scratch**: " create a new repo "
- **Local copy**:

git init

git remote add origin https://www.github.com/YourUsernameHere/test_repo.git

- **Fork** another user's repository: "Fork"
- **Clone the repo**:

git clone https://www.github.com/YourUsernameHere/RepoNameHere.git

Pushing and Pulling on GitHub:

- git add . (add all files to track on local repository)
- git add -u (update file to track on local repository)
- git add -A (both previous operations)
- git commit -m "massage" (commit index)
- git push -u origin master (load files on remote repository in origin branch master)
- git checkout -b branchname (create a branch)
- git branch (to see what branch you are on type)
- git checkout master (to switch back to the master branch type)

Types of Data Science questions:

- Descriptive
- Exploratory
- Inferential
- Predictive
- Causal
- Mechanistic

Descriptive Analysis:

describe a set of data.

- The first kind of data analysis performed
- Commonly applied to census data
- The description and interpretation are different steps
- Description can usually not be generalized without additional statistical modeling
- Numerical descriptors are mean and standard deviation for continuous data types
- Frequency and percentage are more useful and used while describing categorical data

Exploratory analysis:

find relationships you didn't know about.

- Exploratoty models are good for discovering new connections
- They are also useful to describe future studies
- Exploratory analyses are usually not the final say
- Exploratory analyses alone should not be used for generalizing/predicting
- Correlation does not imply causation

Inferential analysis:

use a relatively **small sample of data to say something (draw inferences) about a bigger population**.

- Inference is commonly the goal of statistical models
- Inference involves estimating both the quantity you care about and your uncertainty about your estimate
- Inference depends heavily on both the population and the sampling scheme

Predictive analysis:

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 there are better and worse prediction models, more data and a simple model works really well
- Prediction is very hard, especially about the future references

Causal analysis:

find out **what happens to one variable when you make another variable change**.

- Usually randomized studies are required to identify causation
- There are approaches to inferring causation in non-randomized studies, but they are complicated and sensitive of assumptions
- Causal relationships are usually identified as average effects, but may not apply to every individual
- Causal model are usually the "gold standard" for data analysis

Mechanistic analysis:

understand the **exact changes in variables that lead to changes in other varibales** for individual objects.

- Incredible hard to infer, except in simple situation
- Usually modeled by a deterministic set of equations (physical/engineering science)
- Generally the random component of the data is the measuremet error
- If the equations are note but the parameters are not, they can be inferred with data analysis

Data:

values of quantitative or qualitative variables, belonging to a set of items.

- Set of items: population
- Variables: measuremet or characteristic of an item
- The most important thing in Data Science is the question
- The second most important thing is the data
- Often the data limit or enable the questions
- But having data can't save you if you don't have the questions

Experimental design

- Pay attention to all aspects of the **design and analysis of the study**
- Plan for **data and code sharing**
- Formulate your **questions in advance**

Good experiments

- Have **replication**
- Measure **variability**
- **Generalize** to the problem you care about
- Are **transparent**

Beware

- **Correlation is not causation**. So you can deal with it fixing, or stratifying, or randomizing the variables.
- **Prediction is not inference**: both can be important.
- **Data dredging refers to spurious correlations**.