GENERAL ASSEMBLY

NEXT STEPS AFTER THIS COURSE

Mason Gallo

Data Scientist

LEARNING OBJECTIVES

- Understanding of other popular tools in data science
- Hiring / Interviewing
- Know where to go next

ALTERNATIVE TOOLS

LANGUAGES

- While we've mostly talked about Python in this class, there are many other languages and tools that Data Scientists might use.
- These tools have their various advantages and disadvantages.
- For example, other common programming languages for data science include:
    - R
    - Java/Scala

LANGUAGES

- “R” is often used in data science and is the basis for many features found in Python data analysis.
- Pandas dataframes actually replicate the functionality of the R dataframe!
- R often contains many more specialized algorithms than Python.
- Between `statsmodels` and `scikit-learn`, Python has access to the most popular statistical algorithms. But if your problem becomes more specialized, you may require the niche algorithms available in R.

LANGUAGES

- Python's advantages over R is the ability to tie into other non-data science applications (web apps, etc).
- Myth: Python code is generally faster and more efficient.
- Don't worry about learning both until later in your career

LANGUAGES

- Meanwhile, Java/Scala are popular for their link to the big data ecosystem and wide usage in the enterprise world
- Many larger organizations code their products in Java

LANGUAGES

- It can be easier to interact with big data systems using these languages.
- However, in general they lack the interactivity and ease of use that R and Python have (although development effort is ongoing)

MODELING FRAMEWORKS

- While `scikit-learn` is the most popular machine learning framework in Python, there are alternatives for specialized use cases.

- For example, most models in `scikit-learn` require datasets to be small enough to fit into memory.

MODELING FRAMEWORKS

- Other frameworks can work around this limitation.
- One example is `xgboost`, which provides efficient Random Forest implementations that train much faster than `scikit-learn` models.
- Similarly, the `Vowpal Wabbit` library is often used to train very large linear models, using computational tricks to operate on tens of millions of datapoints.

HIRING / INTERVIEWING

A GOOD DATA TEAM NEEDS…

- Software development ability
- Machine learning knowledge
- Database knowledge
- Visualization ability
- Communication skills
- Domain expertise

Despite “unicorns”, not possible for all of this in one person…

HIRING

- Most important part is “data curiousity”
- Candidate should easily be able to translate business problems
- Candidate should have examples of past projects
- I don’t like ML trivia, but some ask it

HIRING PART 2

- Coding ability is 2nd highest priority
- Start with simple coding exercise (usually filters out many)
- Data cleaning and simple ML exercise

INTERVIEWING DATA QUESTIONS

- Explain overfitting and how it happens
- What is cross validation?
- Mutually exclusive vs independent
- Explain how regularization is used and its purpose
- Inner join vs left join
- How to deal with imbalanced data set?
- How do you determine if features are useful or not?
- Explain linear regression, logistic regression, k-means

INTERVIEWING DATA QUESTIONS

- Many people don’t know how to interview properly
- Can be frustrating when interviewer doesn’t know what they want
- Look at background of interviewer/team
- Software background? Expect coding questions
- Stats PHD? Expect difficult stats/probability questions
- What type of data scientist position is this? Check job description for clues

NEXT STEPS

HOW I THINK ABOUT THE SKILLS

	Data Engineer	*Data Scientist*	*Data Strategist*
Communication / Domain Expertise			
Computer Science			
Statistics			

DATA STRATEGIST RESPONSIBILTIES MIGHT INCLUDE…

- Interface with interdisciplinary teams to determine business problems
- Serve as subject matter expert for data science and data engineering
- Lead presentations to clients and senior management
- Ensure modeling and data work are “actionable”

Skills needed: SQL, some Python/R, big picture understanding of ML

MY RECOMMENDATIONS FOR DATA STRATEGIST

- Data cleaning (pandas, Excel)
- SQL querying
- Understand A/B Testing
- Big picture knowledge of ML algorithms from class
- Communication expertise
- Domain knowledge

DATA SCIENTIST RESPONSIBILTIES MIGHT INCLUDE…

- Interface with interdisciplinary teams to determine business problems
- Prototype and design machine learning models to solve problems
- Query data in conjunction with data engineering
- Present outcomes of machine learning models to clients / senior management

Skills needed: advanced Python/R, SQL, intimate knowledge of ML

MY RECOMMENDATIONS FOR DATA SCIENTIST

- Master the ML algorithms we discussed in class (math/stats)
- Confident data cleaning with pandas
- Comp Sci basics, including git
- Good knowledge of SQL

DATA ENGINEER RESPONSIBILTIES MIGHT INCLUDE…

- Manage data warehouse (think databases and reporting)
- Interface with interdisciplinary teams to determine data storage needs
- Design, build, and launch machine learning models in production
- Design, build, and launch data extraction, transformation, and loading processes in production

Skills needed: advanced Python/Java, SQL, some ML

MY RECOMMENDATIONS FOR DATA ENGINEER

- Extremely confident cleaning / extracting data
- SQL
- Advanced coding ability in at least Python + git
- Good understanding of ML algorithms from class

MATH AND STATS RESOURCES

The entire math track on 

Andrew Ng's Machine Learning course on 

A/B Testing course on 

Elements of Statistical Learning book (FREE)
http://statweb.stanford.edu/~tibs/ElemStatLearn/

COMPUTER SCIENCE RESOURCES

The computer science track on KHANACADEMY

Intro to Computer Sci on 

Think Python book (FREE) http://greenteapress.com/wp/think-python/

IS THIS REALLY GOODBYE?

YOU SHOULD BE PROUD

It’s easy to get lost in the moment and details…but you accomplished a lot!

Some of you started with very little coding knowledge

And now are coding ML algorithms on your own!

Remember: never stop learning

STAY IN TOUCH!

https://www.linkedin.com/in/masongallo/

I love to hear from students!

END OF COURSE SURVEY