15-388/688 - Practical Data Science: The future of data science

J. Zico Kolter
Carnegie Mellon University
Spring 2021

Outline

Data science positions

Ethics in data science

Some final thoughts

Outline

Data science positions

Ethics in data science

Some final thoughts

What is a data scientist?

The many types of data scientists… (not exhaustive)

1. The business analyst, renamed
2. The statistician, renamed
3. The data product designer
4. The machine learning engineer
5. The tools developer

Some important distinctions

Working to develop the “core” business product vs. working tangentially to “identify value” in company data

Developing data science tools vs. doing the actual data analysis

“Classical” statistics vs. machine learning approaches

Poll: Data science ambitions

Who here…

- Has applied for a data science position?
- Has done a data science internship?
- Has worked as a data scientist full time?
- In interested in applying for data science positions?

Applying for data science jobs

This is my own advice, your mileage may vary

1. Identify what kind of data science position you're actually applying for (see the distinctions on the previous pages)

2. Highlight some relevant coursework, but also tangible experience (github pages, etc)

3. Mention the tools you know, making sure that this lines up with the requirements of the position

“Requirements”

A large number of data science positions have particularly stringent requirements: Ph.D., 5 years of experience, etc

For the most part, these are **not** actual requirements of the position (unless it’s for a very senior role, or start of a small team)

Rather, the group is just trying to filter out some of the noise in applications, find a lower-variance pool

My thought: if you can achieve mastery of the ideas in this course, you will be well-suited for many of these positions, but you’ll often need to make initial contact to convey this

Class survey

For those who have interviewed for a data science position, what questions were you asked in your interview?

The data science interview

There is no “standard” yet for the types of questions you’ll be asked (just as there is no standard as to what a data science position means)

The general types of questions:

1. Software engineering questions
2. Questions about data collection/processing (SQL, APIs, etc)
3. Questions about machine learning (usually about “general” ideas like training/testing, debugging, etc., but also about specific algorithms)
4. Questions about statistics (hypothesis testing, statistical significance)
5. The “take-home” data analysis project

Academic data science

“Data science” is not really an area of academic research…

Data science work comes up most often in the content of applied research in other fields, you can be a vastly stronger researcher in your area of interest if you are familiar with these techniques

The academic work in the area typically involves:

1. Fundamental reserach in machine learning or statistics (with data-science-like applications)

2. Methods in “automating” data science, e.g. “Automatic Statistician” (http://www.automaticstatistician.com)

Getting involved in data science research

Find an applied area you are interested in, find a faculty advisor in the area, start using the techniques you've learned in this class

Anecdotally, most researchers will be interested in how data science and machine learning techniques can be applied to their domains, but you will need to spend *substantial* time learning the domain itself

Outline

Data science positions

Ethics in data science

Some final thoughts

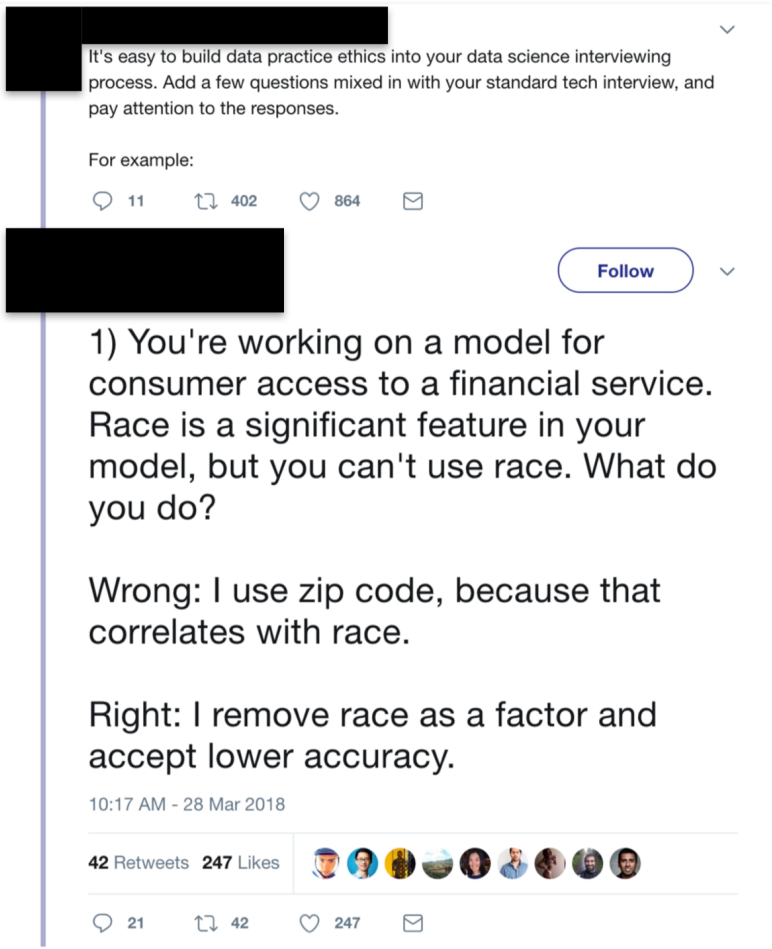
Fairness and bias in data science

https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing

Machine learning and other inference algorithms make predictions based upon past training data

If the training data itself exhibits a bias, there is a good chance the resulting models will suffer from the same bias

Machine learning “blindness”



This reflects a common thought, but it is *incorrect*

There is plenty of ability for algorithms to unintentionally introduce bias by finding proxies within the feature you *did* (accidentally) include

The “quick fix” doesn’t work

“Just remove race as a feature”

(The system analyzed in the ProPublica paper did not include race as a feature)

The problem: race is correlated with many other features we may (knowingly or unknowingly) include

We need to *include* race as an explicit feature in our models, and correct for the bias

What models are “fair”?

But how do we “correct” for the bias? Need to somehow quantify “fair” models…

One possible answer, *demographic parity*: predict the equal proportions of re-offenders for each group

- But is this really a fair model for all situations?

Another possibility (Hardt et al., 2016) *equality of opportunity*: equalize the true positive rate across different group (i.e. number of correctly predicted re-offenders divided by total number of reoffenders)

These different notions are *incompatible,* cannot have a model that satisfies both, and the “right” answer depends on the context

Privacy in data science

Opinion | THE PRIVACY PROJECT

A Brief History of How Your Privacy Was Stolen

Google and Facebook took our data — and made a ton of money from it. We must fight back.

By Roger McNamee

Mr. McNamee, a long-time tech investor, was an adviser to Mark Zuckerberg.

June 3, 2019

224

Initial excitement of “look at what cool things we can do with this data!”

Has given way to “there is no way that any company should have this much insight into our personal information”

What sort of analyses should we be doing?

Data is becoming increasingly available (especially at companies whose prime motivation, in some sense, *is* to collect this data)

Even ignoring issues of bias and fairness, what kinds of inferences / analyses do we as a society want to allow with this data?

Some thoughts from Dj Patil (former U.S. Chief Data Scientist):
https://medium.com/@dpatil/a-code-of-ethics-for-data-science-cda27d1fac1

Outline

Data science positions

Ethics in data science

Some final thoughts

The “future” of data science

Technological trends are extremely difficult to predict

Example: I honestly don’t know what’s going to happen with the recent surge in Artificial Intelligence (and I work in AI)

But I’m pretty confident in this prediction: data science (by one name or another) is here to stay

Data science for _____

Hard to find a field that isn't at least trying to develop a "data-driven" component to it

Examples I've personally worked with at least tangentially: energy systems, building management, wind power, material science, chemical engineering, aerospace, robotics, fluid dynamics, industrial manufacturing, fraud detection, weather forecasting

Whatever area you work in, chances are that area will already be influenced by these techniques (or if not, you should pioneer that advance)

What you've studied in this course

Data processing: web scraping and APIs, relational data and databases, data visualization, matrices and linear algebra, graphs and networks, free text, geospatial data (if you read the tutorial)

"Classical" learning methods: linear regression, linear classification, nonlinear methods using feature transformations, overfitting and cross validation, regularization, probability and statistics, maximum likelihood estimation, naïve Bayes, hypothesis testing

Other methods: decision trees and boosting, clustering and dimensionality reduction, mixtures of Gaussians, expectation maximization, recommender systems, deep learning, MapReduce, debugging data science

Additional courses to look into

CMU is an amazing place, and there are a huge number of courses available to those who want to pursue data science in more depth

To name a few (absolutely not exhaustive): 10-601/10-701 (Machine Learning), 36-402 (Advanced Data Analysis), 05-839 (Interactive Data Science), 10-605 (Machine Learning with Big Data Sets), 15-826 (Multimedia Databases and Data Mining), 15-780/15-781 (Artificial Intelligence), 11-641 (Machine Learning for Text Mining), 10-807 (Deep Learning)

Q&A (ambitiously)