15-388/688 - Practical Data Science: Data science positions and ethics

J. Zico Kolter
Carnegie Mellon University
Spring 2018

Outline

Data science positions

Ethics in data science

Some final thoughts

Q&A

Outline

Data science positions

Ethics in data science

Some final thoughts

Q&A

Poll: data science positions

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?

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

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)

Outline

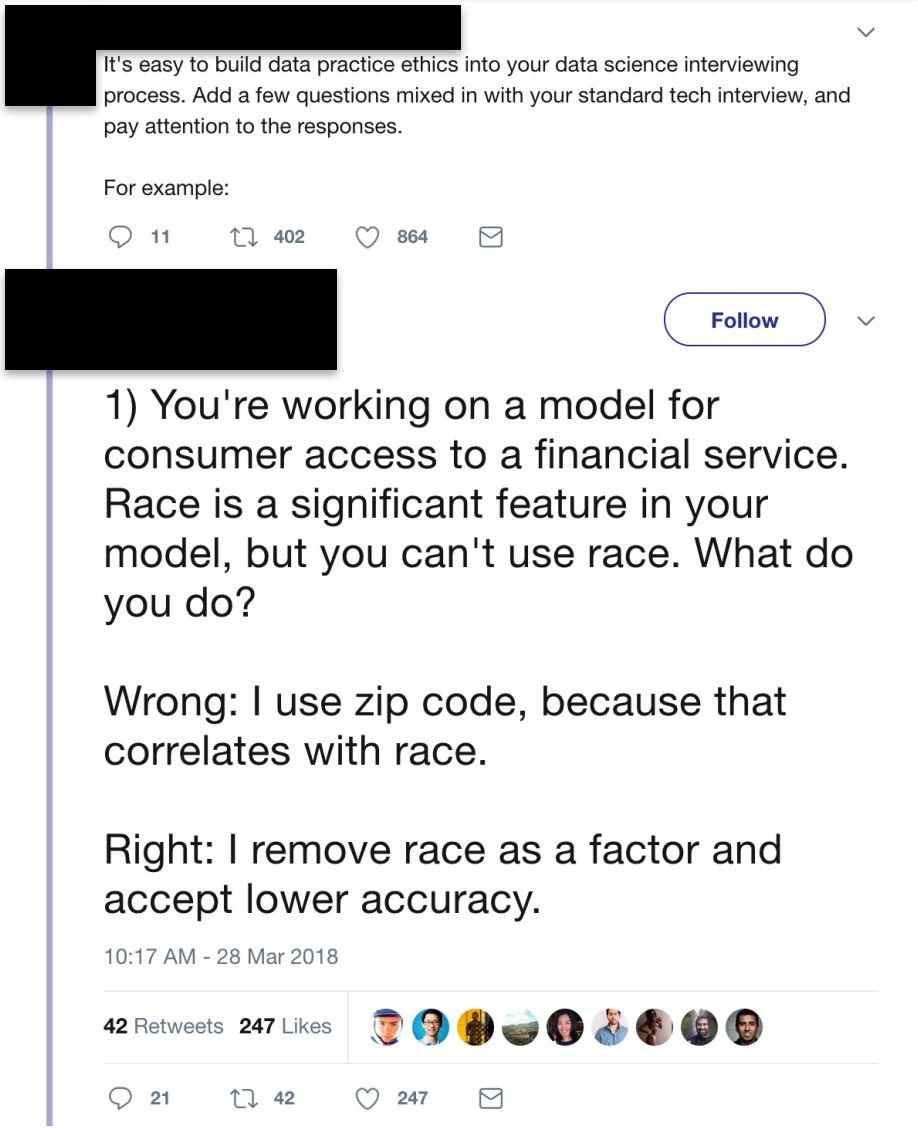
Data science positions

Ethics in data science

Some final thoughts

Q&A

Ethics in data science



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 suffers from bias, there is a good chance the resulting algorithms will suffer from the same bias

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 that we may (knowningly or unknowningly) include

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

What models are “fair”?

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

One possibility (Hardt et al., 2016), *equalized odds*: given a predicted outcome $\hat{Y}$, a true outcome Y, and a (binary) “protected attribute” A, the predicted outcome satisfies equalized odds if

$$P\left(\hat{Y}=1\middle|A=0,Y=y\right)=P\left(\hat{Y}=1\middle|A=1,Y=y\right),y\in\{0,1\}$$

E.g., if we restrict ourselves to the class of people who *really* will not reoffend, our prediction should not change based upon race

Many existing models demonstrably do not satisfy equalized odds

Privacy in data science

How Trump Consultants Exploited the Facebook Data of Millions

ed found the data firm Cambridge Analytica and worked there until 2014, has described the company
n a culture war. Andrew Testa for The New York Times

By Matthew Rosenberg, Nicholas Confessore and Carole Cadwalladr

March 17, 2018 Leer en español

(After this story was published, Facebook came under harsh criticism from lawmakers in the United States and Britain. Read the latest.)

Facebook's Role in Data Misuse Sets Off Storms on Two Continents

Maura Healey, the attorney general of Massachusetts, has announced an investigation into Facebook and the data firm Cambridge Analytica. Brian Snyder/Reuters

By Matthew Rosenberg and Sheera Frenkel

March 18, 2018

WASHINGTON — Facebook on Sunday faced a backlash about how it protects user data, as American and British lawmakers demanded that it explain how a political data firm with links to President Trump's 2016 campaign was able to harvest private information from more than 50 million Facebook profiles without the social network's alerting users.

Senator Amy Klobuchar of Minnesota, a Democratic member of the Senate Judiciary Committee, went so far as to press for Mark Zuckerberg, Facebook's chief executive, to appear before the panel to explain what the social network knew about the misuse of its data "to target political advertising and manipulate voters."

What sorts of analysis 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 about bias and fairness, what kinds of inferences / analyses do we actually *want* to do 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

Q&A

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 learning methods: decision trees and boosting, clustering and dimensionality reduction, mixtures of Gaussians, expectation maximization, recommender systems, deep learning, probabilistic models

Other: big data and 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 (if time)