15-388/688 - Practical Data Science: The data scientist position

J. Zico Kolter
Carnegie Mellon University
Fall 2016

Outline

Data science in industry

Data science in academics

Announcements

Extensions on HW6 (tomorrow midnight), final project report (12/11, midnight, no late days allowed)

Video presentations for final project scheduled for 5-8pm in Rashid auditorium, attendance is ***mandatory*** except with instructor permission

Videos to be submitted by midnight 12/13, more details to follow today on Piazza

Outline

Data science in industry

Data science in academics

Class survey

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

Outline

Data science in industry

Data science in academics

What is 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

Independent study opportunities

I've had several people ask me about the possibility of independent study projects related to a topic in data science

If there is enough interest (say, >5 students), I will consolidate in an "official" course number, with infrequent meetings throughout spring semester

Goal will be something like the class project, but on a larger scale, with the goal of producing a tangible data set and paper on the analysis