10-718: Data Analysis Course (DAC)

GHC 4303
1/14/19

Staff

Instructor: Ameet Talwalkar

TAs: Sebastian Caldas, Roy Li, Nicholay Topin

Disclaimer: DAC is **unique**!

	Traditional ML Course	**DAC**
Topics	ML methods Analysis of ML methods	EDA and Applied ML Qualitative issues (FAT, EE) Communication Skills
Interactions	Mostly Lectures	Mostly Student Presentations / Discussions
Tasks	Tightly-scoped problem sets Research projects	Open-ended, applied assignment FAT/EE presentations & writeups

Motivation

ML researchers: methods → data → problem

- Researchers often aim to develop new methods and theoretically analyze them
- Our interaction with data is often in the form of evaluating these methods on sanitized benchmarking datasets (e.g., UCI, ImageNet)
- Abstract actual domain problem/data so we can focus on methods

Motivation

Practitioners: problem ⟶ data ⟶ methods

- Practitioners focus on underlying problems and associated datasets
- Statistical analyses are a means to get answers from data
- Understanding the domain, translating a problem into a mathematical formulation, interdisciplinary collaboration, and communication are all essential (and HARD!)

Broad Goal: Appreciate practical challenges of data analysis beyond statistical modeling

Specific goals:

- Complete an open-ended, applied ML assignment
- Learn about modern societal problems related to fairness, accountability, transparency (FAT)
- Understand proper methods for empirical evaluation
- Improve communication / presentation skills

Course Activities

DAC consists of:

- Lectures
- Student presentations / discussions
- Student write-ups on presentation topics
- Semester-long data analysis assignment

Lectures (8 total)

- Introduction (today)
- FAT
- Assignment Domain Expert
- Speaking Skills
- Experimental Evaluation
- Assignment Outcomes (x3)
  - Some students will present their findings

Student Presentations

- Two individual student presentations
  - FAT and Experimental Evaluation
  - Short (15 min) and highly polished
- 50 students in the class → 100 presentations total!

Student Subgroups

- Present to subgroups of 10 students
  - Subgroup meetings in lieu of standard lectures
- Peer participation is crucial during presentations
  - Verbal: Questions / Discussion
  - Written: Feedback on presentations
- To seed discussions, students must read material in advance and submit short written summaries

Student Presentation Details

- We will provide a set of pertinent articles
  - FAT: Non-technical articles
  - EE: papers from top ML conference (e.g., N*IPS, ICML) with *substantial* experimental sections
- Everyone in subgroup must present different article

FAT Presentations

- 7-8 min: Describe topic/story/issue, presenting both sides of the argument as unbiasedly as possible
- 5 min: Pose discussion question(s) for subgroup, e.g., “which side do you support, and why?”
- 2: min: Describe your viewpoint (do this *after* audience discussion)

Experimental Evaluation (EE) Presentations

- 8 min: Describe general problem (3 min) and experimental setup (5 min)
  - Audience should already have some context here by reading ahead of time / submitting writeups
- 5 min: Critique the setup / results
  - More details to come...
- 2 min: Take questions from subgroup

Presentation Write-ups

- Prior to each subgroup meeting, each member of subgroup should read each article to be presented
- Each student must submit write-up answering 3 questions for each article
  - Details described on course website
- 4 write-ups in total

Typical Steps of Applied Data Analysis

Steps

Overview of research
Some research questions the data might answer
Description of data
Data checks / transfer
Return to questions and translating them
Present to collaborators

Simple methods to give preliminary answers
Present to collaborators

Do better / Iterate
Present to collaborators

Data Analysis Assignment, Part 1

- Everyone will work on the same open-ended problem
  - Goal: Predict travel time from NYC traffic data
- Understand domain / high-level questions
  - Domain expert will some provide context
- Acquire, explore, clean data
- Formulate quantitative statistical problem
- Propose specific analytical pipeline and experimental plan
- Discuss results during Assignment 1 Outcome Lecture
  - Converge on specific pipeline/plan

Data Analysis Assignment, Part 2

- Implement pre-defined analytical pipeline
- Evaluate it via pre-defined experimental plan
  - There will be some competition component
- Propose revised analytical pipeline / experimental plan based on outcome
- Discuss results during Assignment 2 Outcome Lecture
  - Converge on specific pipeline/plan

Data Analysis Assignment, Part 3

- Implement revised analytical pipeline
- Evaluate it via revised experimental plan
    - There will be some competition component
- Propose revised analytical pipeline / experimental plan based on outcome
- Discuss results during Assignment 3 Outcome Lecture

Schedule

- 29 putative class meetings, but only 8 lectures
- 5 subgroups, each with at most 10 students
    - Each subgroup will meet 4 times total (twice for each presentation topic)
- Overall, students will meet 12 times total
    - Attendance and participation are crucial!
- Review schedule on course website

Grading

- Assignments ½
- Presentations ¼
- Presentation Write-ups ⅛
- Attendance / Participation ⅛
    - Must attend all 4 of your subgroup meetings
    - Must attend all 3 Assignment Outcome lectures
    - Participation (verbal and written)
- No phone/laptop policy

Other Logistics

- Course website has all this info and more
  - Including link to this presentation
  - http://www.andrew.cmu.edu/course/10-718/
- Piazza for assignment discussion (see website)
- Office hours (see website)

Your Homework

- Sign up for a subgroup using sign up link on course website

Questions?

Examples of previous DAC assignments