SI 699 – Big Data Analytics

Winter 2020, Section 4

Thursdays 2:30-5:30pm, North Quad 2255

Instructor: Misha Teplitskiy (tepl@umich.edu)

Office Hours: by request

last updated: 2020-02-10

Overview

The big data analytics mastery course will require students to demonstrate mastery of data collection, processing, analysis, visualization, and prediction. To develop these skills students will work on semester-long projects that deal with large or industry-scale data sets, and solve real-world problems. Aligned with best industry practices, students will be expected to work in a fast-paced, collaborative environment, while demonstrating independence and leadership. Students must be able to create and use tools to handle very large transactional, text, network, behavioral, and/or multimedia data sets.

Learning goals

By taking this course, the students will develop and exercise the following skills:

1. Transform a real world scenario into a data analytics problem by identifying the input, the output, and different types of data needed to generate the output.
2. Formulate the problem (input -> output) as one of the typical data analysis tasks, including but not limited to pattern extraction, visualization, classification, clustering, ranking, prediction, and anomaly detection.
3. Design experiments to judge/measure the success of the data analysis task.
4. Identify the state-of-the-art algorithms and tools for particular type of data and the data analysis task.
5. Collect data needed for the analysis task from various sources.
6. Write programs to manipulate raw data into correct formats needed for the analysis.
7. Use statistical and visualization tools to describe the properties of the collected data.
8. Deal with data at scale
9. Setup, configure, customize, and execute the state-of-the-art big-data analytics tools and conduct the experiments. Implement algorithms if nothing exists off- the-shelf.
10. Validate, summarize, and present the analysis results.
11. Draw correct conclusions from the analysis results.
12. Disseminate results of big-data analytics to an audience via presentations and a final

report.

Required Prior Knowledge

The prerequisites of this course are: SI501, SI618 (data manipulation and data exploration, or equivalent), SI544 (statistics and data analysis, or equivalent), SI671 (data mining)

Furthermore, students should have taken at least two of the following (or equivalent): SI561 (natural language processing), SI608 (networks), SI649 (information visualization), SI650 (information retrieval)

Prior to enrolling in this course, students should manage the basic concepts, theory, and techniques about data structures, data mining algorithms, probability distributions, statistical tests, data visualization, statistical data analysis, and data-intensive computation. Depending on the particular projects, students should also be familiar with concepts, theory, and techniques about particular data types and application domains, such as network science, computational social science, natural language processing, information retrieval, social media, financial markets, and/or information visualization. All these knowledge and foundations can be obtained through combinations of the required and recommended courses.

Students are expected to have competency in programming, data manipulation, statistical data analysis, data mining algorithms, working with unix environments, and configuring and using state-of-the-art data mining and statistical analysis tools.

Students should also be familiar with common practices of managing individual and team projects, such as version control (e.g., git), project documentation (e.g., wiki), and progress tracking (e.g., Trello).

Outline

Projects

Students will do semester-long projects that are of real interest to industry or other stakeholders, using real- world, large-scale data, and demanding state-of-the-art techniques. Team size will be around 3 students. The projects will be of two types, (1) student-chosen and (2) external. Student-chosen projects are student lead from “A to Z”: students identify a problem, formulate it as data analytics task, and execute it. External projects are provided by real organizations that UMSI is collaborating with. A small number of such projects will be available for students to bid on. The instructor will also provide a few sample problems with large-scale data sets. For student-chosen projects, students are encouraged to propose their own problems and data sets, which needs to be approved by the instructors before Week 3.

Respondents

In addition to your role as a data analyst, you will serve as a "respondent" -- a consumer of analysis produced by other teams. As a respondent you will be responsible for commenting and providing thoughtful and constructive feedback on the presentations given by the producer team to which you are assigned. The goal is to push the producer to explore useful analysis avenues and discover limitations in their analysis.

Progress and feedback

Every team should meet with the instructor weekly to discuss project progress (which may or may not be scheduled during the class sessions). In-class sessions will be used to discuss issues that are related to all students/teams (such as tutorials of tools and algorithms and guest speakers). In most class sessions, teams will give short presentations (~10 minutes) about their progress, every other week. In particular, the first two weeks of class will be dedicated for team building and the introduction to computational environments, datasets, and sample projects. All projects and teams will be finalized in week 3 and teams will then formulate their analytics proposals. A halfway project presentation will be held in class in week 7 or 8, and a final presentation of the projects will be held in the last week of class. Members of individual teams must meet at least once every week other than the progress meeting and use different channels of coordination throughout the semester.

Every team will be assigned to a different project so that there is no direct competition among teams. Project documentation will be kept up-to-date on a running document. Teams are encouraged to check in on each others' projects throughout the semester.

Guest speakers

When appropriate, we will have guest speakers to talk about different aspects of data science in the real world.

Assignments

Assignments include 1. research proposal, 2. midterm report, 3. final report, 4. final poster, 5. 360 assessments of team. 360 assessments will be done at midterm and end of course via a short form in which you describe concrete work done by your teammates and estimate division of labor. Projects will take the form of a "consulting" project leading to a final report and presentation. The final report will consist of a 10-page (about 2500 words) document, with a 1-page executive summary at the beginning. The final presentation will consist of a poster, to be presented at a combined SI 699 poster session.

Grading

80% Projects

- 40% final report
- 25% midterm report and presentation
- 5% proposal
- 5% final presentation

- 5% “360 assessments”

20% Class participation
- 15% Weekly presentations, class attendance and participation
- 5% “Respondent” feedback

Late assignments will be penalized 1 letter-grade per day, unless you notify me in advance with a serious reason and receive an excuse.

Schedule

Week 1, Jan. 9: Introduction

- Introduction to class
- Engaged Learning Office presentation: Working with clients, NDAs/ IP
- External project descriptions
- Introductions and team building

Week 2, Jan. 16: High performance computing, project discussion, team building
- ARC-TS: Cavium cluster introduction
- Finding a good project
- Team building
- **Assignment due**: Read: Salganik 2018. *Bit by Bit*, Ch. 2 link
- ~~Assignment due:~~
    - ~~By *5pm Friday, Jan 17*: Finalize team, scope out project~~

Week 3, Jan 23: High performance computing, project design presented in class
- ARC-TS: Great Lakes cluster introduction
- ~~Presentations: Project proposals - 1st half~~

Week 4, Jan 30
- Presentations: Project proposals - 1st half
- **Assignment:**
    - By 9am, Jan 30: Prepare 5-slide presentation on project proposal

Week 5, Feb 6
- Presentations: Project proposals - 2nd half
- Project progress presentations, check-in meetings

Week 6, Feb. 13
- Presentations: 1st half

February 18: Recommended deadline to request Absentee Ballot

Week 7, Feb. 20

- Discuss Midterm report (due 10AM, March 12)
- Presentations: 2nd half

February 24: "Easy" Voter Registration deadline in Michigan. (Feb. 25-March. 10, you can still register, but only at your city/township clerk's office with proof of residency)

Week 8, Feb 27
- Guest speaker: Tarik Kurspahic - EVP of technology at *digi.me*
- Presentations: 1st half

Week 9, Mar. 5: No class (Spring break!)

March 6: Deadline for when absentee ballot request must be received
March 10: Presidential primary vote. If voting absentee, ballots must be received by 8pm

Week 10 (Mar. 12)
- **Midterm report due by March 16, 12pm**
- ~~Guest speaker: Jonathan Prantner, Chief Analytics Officer and Co-founder of *RXA*~~

Week 11 (Mar. 19) - 14 (Apr. 9): Presentations, check-in meetings

Week 15, Apr. 16
- **Assignment 1: Final presentations in class**
  - (no 1-1 meetings)
- **Assignment 2: Final paper, due by Apr. 23, 12pm**

Misc

Academic Integrity

Unless otherwise specified in an assignment all submitted work must be your own, original work. Any excerpts, statements, or phrases from the work of others must be clearly identified as a quotation, and a proper citation provided. Any violation of the School's policy on Academic and Professional Integrity (stated in the Master's and Doctoral Student Handbooks) will result in serious penalties, which might range from failing an assignment, to failing a course, to being expelled from the program. Violations of academic and professional integrity will be reported to UMSI Student Affairs. The faculty instructor determines the consequences impacting assignment or course grades; the Assistant Dean for Academic and Student Affairs may impose additional sanctions.

Accommodations for Students with Disabilities

If you think you need an accommodation for a disability, please let me know at your earliest convenience. Some aspects of this course, the assignments, the in-class activities, and the way we teach may be modified to facilitate your participation and progress. As soon as you make me aware of your needs, we can work with the Office of Services for Students with Disabilities (SSD) to help us determine appropriate accommodations. SSD (734-763-3000; http://www.umich.edu/sswd/) typically recommends accommodations through a Verified Individualized Services and Accommodations (VISA) form. I will treat any information you provide as private and

confidential.

Student Mental Health and Wellbeing

The University of Michigan is committed to advancing the mental health and wellbeing of its students. If you or someone you know is feeling overwhelmed, depressed, and/or in need of support, services are available. For help, contact Counseling and Psychological Services (CAPS) at (734) 764- 8312 and https://caps.umich.edu/ during and after hours, on weekends and holidays, or through its counselors physically located in schools on both North and Central Campus. You may also consult University Health Service (UHS) at (734) 764-8320 and https://www.uhs.umich.edu/mentalhealthsvcs, or for alcohol or drug concerns, see www.uhs.umich.edu/aodresources.

For a listing of other mental health resources available on and off campus, visit: http://umich.edu/~mhealth/