SI 699 – Big Data Analytics

Winter 2021, Section 4, Tuesdays 1-4pm

Instructor: Misha Teplitskiy (tepl@umich.edu)

GSI: Shiyan Yan (shiyansi@umich.edu)

Office Hours: Thursdays 4-5pm (Misha)

Fridays 9-10am (Shiyan)

last updated: 2021-01-19

Zoom links

- Class session: link
- 1-1 groups Misha: Tues, Thurs
- 1-1 groups Shiyan: Tues, Thurs
- Office hours Misha: link
- Office hours Shiyan: link

Overview

This course helps students develop mastery of executing a major big data project. Students will get to do the project "A to Z", from coming up with a research question and identifying and/or collecting relevant data, to processing, analyzing, and presenting it. To develop these skills students will work on semester-long projects that deal with large or industry-scale data sets that address 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. In addition to the projects, students will engage in a substantial amount of peer instruction on new computational developments: your group will deliver a tutorial on a cutting-edge big data topic to your classmates, as well as absorb tutorials from all the other groups.

Learning goals

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

Big data project design

1. Identify phenomenon of interest and formulate it as an answer-able research question
2. Identify data needed to answer the question.
3. Formulate the question as one of typical analysis tasks, including but not limited to pattern extraction, visualization, classification, clustering, ranking, prediction, and anomaly detection.

4. Identify the state-of-the-art algorithms and tools needed for analysis
5. Identify/design metrics to judge the success of your answer

Project execution

1. Write code to manipulate raw data into format needed for the analysis.
2. Use statistical and visualization tools to describe the properties of the collected data.
3. Deal with data at scale.
4. Implement algorithms or create your own if nothing exists off- the-shelf.
5. Validate, summarize, and present the analysis results.
6. Draw correct conclusions from the analysis results.
7. Disseminate results of the analysis 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 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 student-chosen and driven: you will identify a problem, formulate it as data analytics task, and execute it. The instructor will provide example projects from past years and guidance. Projects must be approved before Week 3. The final report will consist of a 10-page (about 2500 words) document, with a 1-page executive summary at the beginning. Detailed expectations will be provided later in the semester.

- Projects roster and meeting times: link

Tutorials

Over the course of the semester, each student group will deliver one 30 minute tutorial or paper review to the class on a topic related to big data analytics. Example topics include: dealing with missing data, interpretable ML, causal inference, PyTorch, image recognition, etc. This tutorial can rely on existing tutorials or research papers, no need to reinvest the wheel completely, *but it must be customized for the class,* and you must identify the source(s) you are drawing on. In particular, the sophistication (and length) of the tutorial should be appropriate for the class, i.e. not elementary and not for domain experts either. Think of how you would best learn about this topic in 30-40 minutes, and deliver it at approximately that pace. You can and, if possible, should provide in-class exercises for the audience. If needed, you can request that we install the necessary packages/software in advance and/or sign on to Great Lakes or Cavium accounts.

- Async: Respond to questions on Slack next day

1-1 meetings

Every team should meet with the instructor and/or GSRA 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.

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.

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

Grading
70% Projects
- 35% final report and presentation
- 15% midterm report and presentation
- 10% pitch presentation and proposal presentation
- 10% 1-1 meetings
- (qualitative) Peer review: https://forms.gle/pLnjWwKA7PbTLUJ17

20% Tutorial/paper presentation

10% Class participation
- 5% Class 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.

Async details
The class can be completed asynchronously, in order to accomodate students in different locations. To make this possible, we will do the following:
- All class sessions will be recorded and posted for online viewing if necessary
- You will also need to record your presentations if you cannot deliver them live and send Shiyan the recording by 11am on day of class.

For details on sync sessions and camera etiquette, see section "Zoom participation policy" below.

Schedule
Week 1, Jan. 19: Introduction
- Introduction to class (link)
    - Survey: link
    - Slack: link
    - Great Lakes: link
- Project ideas and team building: link
    - Breakout rooms + Google slides link
- Tutorial ideas: link

    - Sign up link
- [If time permits: Project management best practices: link]

Week 2, Jan. 26: High performance computing, pitches
- ARC-TS: Great Lakes introduction (slides) (video)
    - Logging into Great Lakes: https://arc-ts.umich.edu/greatlakes/user-guide/
- Tutorial ideas discussion: link
    - Sign up link
- *Assignment due*: present 1-2 slide "pitch" on your idea
    - May be done individually or in group

Week 3, Feb. 2
- Tutorial: Shiyan Yan (UMSI)
- *Assignment* due: 5-slide project proposal
- Present 5-slide project proposal (1st half)

Week 4, Feb. 9
- Tutorial 1: Team 1
- Present 5-slide project proposal (2nd half)

Week 5, Feb 16
- Speaker: Greg Mundy (NOAA, BrightHive)
- Tutorial 2: Team 8
- Tutorial 3: Team 3

Week 6, Feb. 23
- Speaker: Bryan Berend (DemocracyWorks)
- Tutorial 4: Team 7
- Tutorial 5: Team 11

Week 7, Mar. 2
- Speaker: Jake Fisher (Facebook)
- Tutorial 6: Team 12
- Tutorial 7: Team 6
- Discuss Midterm report

Week 8, Mar. 9
- Tutorial 8: Team 2
- *Assignment*: midterm report and presentation
- Present midterm (1st half)

Week 9, Mar. 16
- Tutorial 9: Team 13
- Present midterm (2nd half)

~~**Week 10, Mar. 23** - Wellness day :)~~

Week 11, Mar. 30

- Speaker Jonathan Hersh (Chapman University)
- Tutorial 10: Team 10
- Tutorial 11: Team 5

Week 12, Apr. 6

- Speaker: Arun Varghese (Coursera)
- Tutorial 12:
- Tutorial 13:

Week 13, Apr. 13

- Speaker: Aseem Bharitya and Alex Daly (Domino's)
- Tutorial 14:
- Tutorial 15 (?)

Week 14, Apr. 20

- Assignment 1: Final presentations in class
- Assignment 2: Final paper, due by Apr. 23, 12pm
- Assignment 3: peer evaluations, due by Tuesday Apr 28, midnight
    - https://forms.gle/pLnjWwKA7PbTLUJ17

Zoom participation policy

The instructors are joining the live sessions from home (just like you, most likely!) **If the network connection drops or the video freezes, please don't give up on the meeting**! Hang out for at least 5 minutes while the instructor reboots or switches to his cell phone for Internet connectivity. Thanks!

General rules for Zoom use in this class:

- **Video: we expect/encourage everyone to use the video, as but video is not required**, and may be difficult due to technical issues (see below)
- You must always use your UMich credentials/account to access this course. If you have another Zoom account, it will not work!
- Here is information on how to log into your U-M Zoom account (Links to an external site.)
- If you have already used your UM email to create a non-UM Zoom account, please see this page for information on how to migrate your account. (Links to an external site.) (Note that this can take several hours, so please do it well before the first class meeting.)
- Please only connect using a Zoom client, preferably on your laptop/desktop (better

than phone or tablet). NEVER on a web browser. It does not work as well.
- Please keep your Zoom client updated to get the newest features.
- Be sure your display name in Wolverine Access actually represents your preferred display name (see "Preferred" name in "Campus Personal Information").
- Have your camera on as much as possible during class meetings (but mute your microphone as much as possible, too) I strongly encourage you to keep your camera on for community-building purposes (but will not require it). Check out your background when you are on camera. Remove or cover items you don't want others to see
- Please set up a profile photo for your account (for when your camera is off)
- Pets are welcomed!
- Test your audio and video -- use headphones and a microphone are probably the best thing.
- Please mute when not talking. However, it is OK to make mistakes! We will forgive each other for barking dogs, crying children, sudden doorbells, etc.

What to do if you are on a low-bandwidth connection?

- Leave video off when you don't need it
- Turn off HD video (in the Zoom app video settings)
- When you screen share, only do so for as long as necessary
- If possible, use collaborative (e.g., Google) documents rather than screen sharing
- Mute your audio when not speaking (do this anyway!)
- To improve your overall Zoom performance, consider asking others at your location to limit their high-bandwidth activities while you need to be online for a course or required meeting.
- Avoid running other data-intensive programs during Zoom meetings, such as streaming video or music or other web sites with dynamic content. (Why are you watching streaming video during class?)

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/

Embedded psychologist at UMSI

https://docs.google.com/document/d/1EWAFqqK_S1VMV9OyToz5gsUUP9oqXXNSBjcKviY_XVE/edit