CSCI 6900

Mining Massive Datasets

T/R: 11:00am-12:15pm, Poultry Science 238
W: 11:15-12:05pm, Boyd GSRC 208

Dr. Shannon Quinn

Email: squinn@cs.uga.edu
Website: http://cs.uga.edu/∼squinn/mmd_s15/
Office Location: Boyd GSRC 638A
Office Hours: Mondays 9-10:30am

The course syllabus is a general plan; when (not if) deviations arise, they will be announced.

Course Description: Distributed computing and the paradigm of "big data" have garnered a significant amount of attention in recent years as costs of capturing and storing information have plummeted; analytical bottlenecks have shifted from data acquisition and curation to downstream analysis. However, this shift has created its own set of problems, the most pertinent of which is that large datasets are computationally expensive to process. Algorithms that efficiently process data that fit in memory may become prohibitively expensive to use on larger datasets. Consequently, it can be difficult to gain an intuition for the underlying data and troubleshoot issues.

This course has three primary goals. First, it is intended to provide the student with an appreciation for the issues involved in deploying classic machine learning algorithms–classification, clustering, and dimensionality reduction–to work on datasets that do not fit in main memory. Second, it is intended to provide a working knowledge of and experience with some of the current distributed frameworks and their various philosophies (e.g. Hadoop, Spark). Third, the course is intended to reinforce software engineering best-practices by providing students with hands-on opportunities to implement solutions using real-world datasets.

Prerequisites: None required; experience in machine learning (CSCI 8950) and software engineering (CSCI 4050/6050) highly recommended. You should have a good programming background. If you're unsure, make an honest assessment of yourself with this pre-requisite check: http://www.cs.cmu.edu/~wcohen/10-601/self-assessment/Intro_ML_Self_Evaluation.pdf.

Credit Hours: 4

Text(s): We will have no required text book for this class. I will provide references to recommended reading both on the website and in the lecture slides.

Topical Course Outline

1. Introduction to distributed computing
2. Review of basic machine learning, probability, and statistics
3. Classification at scale (Naïve Bayes)
4. MapReduce and associated frameworks (Hadoop)
5. Clustering at scale (K-means, spectral clustering)
6. Topic modeling at scale (LDA)
7. Distributed graph analytics (PageRank)
8. Alternative distributed frameworks (Spark)
9. Dimensionality reduction at scale (PCA, Bloom filters)
10. Final project presentations

Grade Distribution:

Presentations	10%
Assignments	40%
Midterm Exam	25%
Final Project	35%

There will be 4 assignments, each worth 10% of your final grade. Each assignment will consist of a reasonably in-depth programming assignment. The final project will consist of several parts, graded individually and aggregated to 35% of your final grade. **Your lowest assignment grade will be dropped at the end of the semester**.

If you are auditing the course, the only requirement is the presentation.

Course Policies

- **Attendance**
  - Come to lecture. If you're not in lecture to answer every question I pose to the class, and you have a bad day on the midterm, I won't have any way to know that you actually understand the material. This course has an enrollment limit of 15; I'll know when someone is missing.
  - If you cannot attend lecture, let me know ahead of time and we'll work something out.
  - Try to keep in-lecture hacking on your laptop to a minimum. I certainly understand testing out methods that we discuss in class, but I may interleave critical exam hints into the lecture if I detect a lack of attentiveness; I cannot be held responsible if these hints fall on distracted ears.
- **Assignments**
  - Assignments are due by 11:59:59pm on the noted date. Assignments turned in after this deadline will lose 25/100 points for every subsequent 24 hour-period they are late.
  - With the exception of the final project, assignments are to be completed individually, but you may collaborate with other students as long as you cite the specifics of the collaboration.
  - *The presence or absence of any form of help or collaboration, whether given or received, must be explicitly stated and disclosed in full by all involved, on the first page of their*

assignment ("I did not give or receive any help on this assignment" or "I helped [person] with [specific task]."). Collaboration without full disclosure will be handled severely; except in usual extenuating circumstances, my policy is to fail the student(s) for the entire course.

- **Final Project**

  This class will not have a final exam. In lieu of a final exam, this class will have a final project. You have the option of working with another student in a team of 2 (but are under no obligation to do so). More details on the project will be released over the course of the semester.

Academic Honesty

As a University of Georgia student, you have agreed to abide by the University's academic honesty policy, "A Culture of Honesty," and the Student Honor Code. All academic work must meet the standards described in "A Culture of Honesty" found at: `https://ovpi.uga.edu/academic-honesty/academic-honesty-policy`. Lack of knowledge of the academic honesty policy is not a reasonable explanation for a violation. Questions related to course assignments and the academic honesty policy should be directed to the instructor.

- Read "A Culture of Honesty," the UGA academic honesty policies and CS Academic Integrity policies.

- You must not allow others to copy or look at your work.

- You must not give/share your lab/project assignment work to a fellow student.

- Copying significant portions of code from a fellow student or any other source (including internet) is plagiarism and will be dealt with as such.

- If you have questions about an assignment or if you run into problems, contact your instructor/lab instructor.

- During exams, no assistance and no additional materials are allowed.

- All of your coursework must meet the aforementioned policies and rules. Students that violate any of these rules or the UGA Academic Honesty policies will be liable to a penalty. The instructor will strictly enforce Academic Honesty policies and report any violation of the aforementioned policies and rules.