COSC-254 Data Mining

Times & Location: MW 2–3.20pm, Science Center E110

Website: `http://rionda.to/courses/cosc-254-s19/`, Moodle for assignments and forum

Prerequisites: COSC-211 Data Structures

Instructor: Matteo Riondato (he/his, please call me "Matteo")
Contact: `mriondato@amherst.edu` (Only for personal messages that cannot go to the forum. Please use `[COSC254]` in front of your subject.)
Office Hours: M 3.30–5.30pm, Science Center C214. Please reserve a 15-minutes slot at `http://bit.ly/mroh19s` *by the day before (Sunday) at 4pm.*

TA: Alexander Einarsson
Office Hours: Th 3-5.00pm, Science Center E210.

Welcome! We are excited to have you aboard!

Description

This course is an *introduction to data mining*, the area of computer science that deals with the development of efficient *algorithms for extracting information from data*. We will:

- talk about the key tasks in the analysis of transactional datasets, time series, and graphs, and the most efficient algorithms to solve them;
- learn about parallel/distributed systems to perform the analysis of massive datasets;
- use *interactive notebooks* and *large-scale systems* to evaluate algorithms and analyze data.

Goals & Objectives

- Develop new ways of reasoning about the development, analysis, and evaluation of algorithms for mining large datasets;
- Acquire the ability to implement data mining algorithms for single machines and for large-scale clusters, to evaluate their performance, and to use them for the analysis of large datasets, presenting the results to a technical audience;
- Experience the use of tools and systems commonly used in industry.
- Become familiar with important data mining tasks and the algorithms to solve them.

Textbooks & Materials

We will cover *subsets* of the books:

- Leskovec, Rajamaran, Ullmann, *Mining of Massive Datasets*, Cambridge, v3.0beta (available online);
- Aggarwal, *Data Mining – The Textbook*, Springer (available from the E-reserves on Moodle).

Don't buy the books, download the PDFs from the above links. Eventually, you may want to print the sections that we cover (you can find a preliminary list on the last page of this syllabus).

We will post any *additional materials* to the course website or on Moodle E-reserves.

Assessment & Grading

- *Weekly homework assignments* will be released on Wednesday after class, and will be due on Wednesday before class. They may include quiz questions, open-ended questions, and programming exercises. Details for submission will be available on the course website.
- Three *implementation and evaluation projects* will be released throughout the course, with at least two weeks for completion.
- A *2-hours in-class final exam*.

We will evaluate you on the level of detail and rigor in your answers. For the programming assignments, we care about correctness, robustness, and efficiency.

The course grade is, in a first approximation, a weighted average of the scores in:

- the weekly homework assignments (20% in total)
- the projects (15% each)
- the final (35%)

High-quality participation in class and on the forum may increase the course grade by at most 3% of the maximum grade.

Late Submission Policy

- You may submit *one* late assignment *without penalty*, if you submit it by 11.59pm of the calendar day after it is due.
- For each subsequent late assignment or if the first late one comes later than the "extended" deadline above, the score of the assignment will be reduced by 20% of the maximum score. You may have the penalty waived by having your class dean email Matteo.

There are no exceptions to the above policy.

Expectation of Students

Active Class Participation

Active participation in class creates a inclusive environment for learning and teaching. It requires:

attending every lecture: materials build on top of each other, and *you want to be part of the ongoing discourse*, as we will make many references to previously covered materials.

asking questions: *All questions are good*: if you have one, you are likely not alone, so please ask it. The more you asks, the more everybody learns, and the more we can *adapt the course to your needs, pace, and interests*. We may not have answers to all questions: if we did, we would be teaching theology in Paris.

answering questions: Matteo will ask questions during the lecture, to *give you the opportunity of self-reflect on your immediate understanding of the materials* being covered: if you are already understanding well, you may not need to study it later; if you don't, you may want to ask some questions. Your answers, like your questions, help us *adapt the course to your needs, pace, and interests*.

limiting the use of electronic devices: There is scientific evidence that *electronic devices usage, including typing notes, impairs in-class learning*, and the more you learn in class, the less you have to study later. In-class learning may, from time to time, benefit from using a laptop, for example to test some code on an interactive notebook. You *can use a laptop in class*, provided that what you are doing is *strictly related to the course*. You *cannot use cell phones* in any way, and please have them *completely* silent (not on vibrate).

Matteo reserves the right to ask you to leave the lecture if you are not actively participating.

Collaboration Policy

Computer science is an collaborative discipline, but we must be able to *fairly assess each of you individually*. Thus, there are limitations to how and when you may collaborate with others:

- You may *not collaborate in any way* in *programming* homework assignments and the final.
- You may discuss *non-programming* homework assignments in groups of students all taking the course. You must *write up your solutions independently*. If you discuss the assignment with students taking the course, list their names on the front page of your solutions.
- You may *not directly copy or adapt solutions* from other students, from materials distributed in previous versions of this or other courses, or from any material available online. You may *not make your solutions available to anyone at any point in time*.

If you have doubts about whether you may collaborate and how, please ask on the forum. In all cases, you are bound to the Amherst College Honor Code.

Expectation of Instructor and TA

You can expect us to:

- do our best in explaining the materials and adapt them to your pace and interests;
- assess you fairly;
- create an inclusive and conducive learning environment;
- listen when you have questions, complaints, and even praise;
- help you succeed in the course;
- respect your student rights as stated in the Amherst College Honor Code.

Accessibility and Accommodations

If you have any condition that might require modification of any of the course procedures, you will need to contact Accessibility Services (`accessibility@amherst.edu` or 413-542-2337) as early as possible. *Immediately* after you have arranged your accommodations with Accessibility Services, please inform Matteo via email or in person.

Course Outline

Here is a list of topics and readings covered in the course, with a *tentative* schedule (*MMD*: Mining of Massive Datasets, *DMT*: Data Mining – The Textbook).

MapReduce and Hadoop: The system that revolutionized large-scale data analysis *Weeks 1–2*

- MMD: Sects. 2.1 to 2.4.4

Pattern Mining: Find interesting combinations of features *Weeks 2–3*

- MMD: Sects. 6.1, 6.2, 6.4.4, 6.4.5, 6.4.6
- DMT: Sects. 4.1 to 4.4

Data Streams What to do when data arrives fast, and we have little memory *Week 4*

- MMD: Sects. 4.1 to 4.4, 4.7, 6.5
- DMT: Sects. 12.1 to 12.3

Clustering: Partition the space into regions of similar elements *Weeks 5–6*

- MMD: Sects. 7.1 to 7.3
- DMT: Sects. 6.1 to 6.4, 6.6

Outlier detection: How to find the odd ones. *Weeks 6–7*

- DMT: Ch. 8

Time Series and Sequences Adding time to the data equations *Week 8*

- DMT: Sects. 14.1, 14.2, 14.4, 15.1, 15.4

Graph analysis Networks are everywhere and there is so much they can tell us *Weeks 9–12*

- MMD: Sects. 5.1 to 5.3, 10.1 to 10.4, 10.7
- DMT: Sects. 18.1 to 18.419.1 to 19.3, 19.6

Hypothesis testing: Nothing is true in science, something may be significant *Week 13*

- Readings will be posted on the course website.