INF 553: Foundations and Applications of Data Mining (Fall 2014)

About This Course

Data mining is a foundational piece of the data analytics skill set. At a high level, it allows the analyst to discover patterns in data, and transform it into a usable product. The course will teach data mining algorithms for analyzing very large data sets. It will have an applied focus, in that it is meant for preparing students to utilize topics in data mining to solve real world problems.

Lectures

- Tuesday/Thursday 11:00am-12:20pm (Location WPH 102)

Instructor

- Dr. Ann Chervenak
  Email: annc@isi.edu
- Office Hours: Thursday 12:30-1:30pm or by appointment.
- Grader/Course Supervisor: Pooja Anand <poojaana@usc.edu>

Class Communication and Collaborative Learning

Blackboard at USC will be used for most class communication (assigning and submitting homework, posting lecture slides, discussion board, etc.).

Students are *strongly* encouraged to post questions and respond to other students' postings. Active participation can help those students with borderline final grade.

Prequisites

A basic understanding engineering principles is required, including basic programming skills; familiarity with the Python language is desirable. Most assignments are designed for the Unix environment; basic Unix skills will make programming assignments much easier. Students will need sufficient mathematical background, including probability, statistics, and linear algebra. Some knowledge of machine learning is helpful, but not required.

Textbook

- Rajaraman, J. Leskovec and J. D. Ullman, *Mining of Massive Datasets*
  - Cambridge University Press, 2012.
  - Available free at: http://infolab.stanford.edu/~ullman/mmds.html

In addition to the textbook, students may be given additional reading materials such as research papers. Students are responsible for all assigned reading assignments.

Course Policy/Grading Allocations

Grading for the course will be based on student performance on programming and other homework assignments (approximately 4 to 6 assignments during the semester); two midterm examinations; a final examination (which may be a non-cumulative third mid-term exam); and class participation, including scores on weekly short quizzes and activity on class discussion forums.

- Programming Assignments, Homework: 40%
- Midterm(s), Final: 45%
- Class participation: 15% (10% bi-weekly quizzes, 5% group presentation in class, discussion board participation)

Grading Scale

- A: 95 to 100
- A-: 90 to 94
- B+: 85 to 89
- B: 80 to 84
- B-: 75 to 79
- C+: 70 to 74
- C: 65 to 69

Programming Assignments

All homework assignments are to be submitted to BlackBoard. To obtain maximum points on the homework assignment, please follow the assignment guidelines carefully.

Late Work

For each day the homework assignment is late, the student will lose 1/3 of the grade for the assignment. In extenuating circumstances, such as a serious medical ailment or a family emergency, students must communicate and make arrangement with the instructors **in advance**. Finally, in case of a serious medical ailment, an original doctor's note must accompany the late submission.

Grading Corrections

Grades are not negotiable, and any student who wastes the instructor's time with non-legitimate requests for additional points on an assignment or exams risks losing additional points as well as having their behavior affect their class participation grades.

Any legitimate request for re-grading must be submitted in writing, with a careful explanation of why it is believed that an assignment has not been properly graded.

Class Participation

Regarding the class participation grade, there are three components:

1. There will be bi-weekly quizzes on the previous lessons. The quizzes are designed to enforce class attendance, participation, and attention.
2. Students will work in groups and will present advanced concepts related to basic course material.
3. Students are expected to actively participate in the class forum. For example: asking questions and posting answers to other students' questions.

Academic Integrity

Cheating will not be tolerated. Students must do their own work, including on all programming assignments. All parties involved in cheating will receive a grade of F for the assignment and/or the course and be reported to Student Judicial Affairs and Community Standards, with no exceptions. If you have questions or concerns regarding what is permitted and not permitted in terms of collaboration or teamwork, please do not hesitate to confer with the instructor/TA for clarifications.

We will utilize the Moss software to detect software plagiarism: http://theory.stanford.edu/~aiken/moss/

ADA Statement

Reasonable accommodation will be provided to any student who is registered with the Office of Students with Disabilities and requests needed accommodation.

Course Schedule

This schedule is tentative and is subject to change.

Week	Topic	Readings	Homework	Exams
1 (1/13 to 1/15)	Introduction to Data Mining, MapReduce	Ch1: Data Mining and Ch2: Large-Scale File Systems and Map-Reduce		
2 (1/20 to 1/22)	MapReduce (cont.)	Ch2: Large-Scale File Systems and Map-Reduce	Homework 1 assigned	
3 (1/27 to 1/29)	Frequent itemsets and Association rules	Ch6: Frequent itemsets, Ch3: Finding Similar Items (section 3.5: Distance Measures)		
4 (2/3 to 2/5)	Frequent itemsets and Association rules	Ch6: Frequent itemsets	Homework 1 due, Homework 2 assigned	
5 (2/10 to 2/12)	Recommendation Systems: Content-based and Collaborative Filtering	Ch9: Recommendation systems, additional readings		
6 (2/17 to 2/19)	Recommendation Systems, Shingling, Minhashing, Locality Sensitive Hashing, Clustering	Ch9: Recommendation systems, Ch3: Finding Similar Items	Homework 2 due, Homework 3 assigned	
7 (3/3 to 3/5)	Shingling, Minhashing, Locality Sensitive Hashing	Ch3: Finding Similar Items		
8 (3/10 to 3/12)	Clustering	Ch7: Clustering		Midterm 1
9 (3/17 to 3/19)	Link Analysis: PageRank	Ch5: Link Analysis	Homework 3 due, Homework 4 assigned	
10 (3/24 to 3/26)	Link Analysis: Web spam and TrustRank, Random Walks with Restarts	Ch5: Link Analysis		
11 (3/31 to 4/2)	Analysis of Massive Graphs (Social Networks)	Ch10: Analysis of Social Networks		
12 (4/7 to 4/9)	Analysis of Massive Graphs (Social Networks)	Ch10: Analysis of Social Networks	Homework 4 due, Homework 5 assigned	
13 (4/14 to 4/16)	Web Advertising	Ch8: Advertising on the Web		
14 (4/21 to 4/23)	Mining data streams	Ch4: Mining data streams	Homework 5 due	
15 (4/28 to 4/30)	Mining data streams, Course Summary, Midterm 2			Midterm 2