CS 6220: Data Mining Techniques

Fall 2015

August 30, 2015

Location: Tuesdays 11:45 - 1:25am, Thursdays 2:50-4:30pm, Shillman Hall 210

Instructor: Olga Vitek, WVH 310F, o.vitek@neu.edu
Office hours Tuesdays 1:30-2:30 or by appointment.

Tecahing assistant: Aida Ehyaei, ehyaei.a@husky.neu.edu
Office hours TBA

Goals of the course: The course introduces basic concepts of data mining. The course covers topics such as linear regression, supervised classification, unsupervised clustering, association analysis, and analysis of data with complex dependencies. As appropriate, the course emphases both the algorithmic aspects and the aspects of statistical inference.

The course is driven by hands-on homeworks, and a project. The course will use the programming language R. In many cases the course will rely on the existing implementations of the methods, but some programming effort will also be required. At the end of the course the students will be able to (1) recognize data mining problems, (2) perform data analysis, and (3) draw valid conclusions and clearly present the results.

Pre-requisite: The course is designed for MS students in computer science. The course attempts to be as self-contained as possible. However, the mathematical and computational literacy at the beginner graduate student level is expected. Prior exposure to R is desirable but not required.

Software: The data examples, the case studies, the homeworks and the projects will use the programming language R. Access to R is required. Please install R from http://lib.stat.cmu.edu/R/CRAN/ prior to the course. Instructions for using statistical methods in R will be provided during the course.

Course web page: http://www.ccs.neu.edu/course/cp6220f15/CS6220-Fall15.html
Daily updates on the schedule, handouts and homework assignments will be posted on the course page.

Attendance: The instructor will not monitor attendance. However, you are responsible for all the material covered in class. The instructor and the TA will not repeat the material for you during office hours.

Communication: The course will be using the discussion board Piazza
piazza.com/northeastern/fall2015/cs6220 You are encouraged to ask and answer questions on the discussion board. All important announcements will be made through Piazza. Once the course begins, course-related email inquiries will be left unanswered.

Textbook: The key textbooks are:

1. J. Han, M. Kamber & J. Pei (2012). *Data mining: concepts and techniques*, 3rd Ed, The Morgan Kaufmann Series in Data Management Systems.
2. C. C. Aggarwal (2015). *Data mining*. Springer.
3. J. James, D. Witten, T. Hastie & R. Tibshirani (2013). *An introduction to statistical learning with applications in R*. Springer.

The pdf files of the three textbooks are available free of charge online.

Homework: Expect weekly homeworks during the semester. Homeworks are due in the beginning of the class on Blackboard. An authorization for all extensions must be obtained from the instructor at least 24 hours before the deadline. Any homeworks turned in afterwards without the authorization of the instructor will not receive credit.

Exams: One in-class midterm exam, and one final exam.

Project: At the end of the semester groups the students will perform a group project analyzing a real-world problem. The project will be summarized in the form of a research paper. Each report will receive written reviews by members of other groups.

Grades: All grades will be distributed via Blackboard.

Re-grades: All re-grading requests should be made in writing within 7 days after receiving the grade. The request should state the specific question that needs to be re-grades, as well as a short (1-2 sentences) explanation of why re-grading is necessary. The new grade can potentially be lower than the original grade.

Breakdown of Grade: The final grade is based on a total of 400 points broken down into homeworks (100 pts), midterm (100 pts), project (100 pts), final exam (100 pts).

The final letter grades will follow the usual scale:
90-100 = A, 80-89 = B, 70-79 = C, 60-69 = D, 0-59 = F.
Refinements to the grades (i.e., '+' and '-'), and changes to the scale can be made at any time at the discretion of the instructor.

Changes to Final Grade: All requests to change the final grade should be made within 48 hours after the grade is released. The request should be made in writing, and should state the specific reason. Following the request, the instructor will regrade **all** the documents submitted throughout the semester. The new grade can be lower than the original grade.