CS 7280: Statistics for Big Data

Spring 2015

January 13, 2015

Location: MTh 11:45 - 1:25am, Cargill Hall 094

Instructor: Olga Vitek, Meserve Hall 443, o.vitek@neu.edu
Office hours Mondays 1:30-2:30 or by appointment.

Tecahing assistant: Paul Grosu, Meserve Hall 443 pgrosu@gmail.com
Office hours Mondays 3:00-4:00 or by appointment

Goals of the course: This is a basic course in applied statistics. The course introduces concepts of statistical modeling and inference, which are particularly relevant to computer scientists, and which complement other courses such as machine learning. When appropriate, the course emphases the challenges and opportunities for statistical inference in situations that involve large and/or complex datasets.

The course is driven by practical examples, 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 the problems of inferential nature and understand the underlying principles, (2) use statistical inference in data analysis, and (3) draw valid conclusions and clearly present the results.

Pre-requisite: The course is designed for graduate students in computer science, but is also open to students from other majors. The course attempts to be as self-contained as possible and can be taken by students without an extensive prior training in statistics. 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/cp2500f14/CS7280-Spring15.html
Daily updates on the schedule, handouts and homework assignments will be posted on the course page.

Attendance: Attendance is optional, but you are responsible for all the material covered in class.

Communication: The course will be using the discussion board Piazza piazza.com/northeastern/spring2015/cs7280 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:
Kutner, Nachtsheim, Neter & Li (2005). *Applied Linear Statistical Models*, 5th Ed, McGraw-Hill.
Sharon L. Lohr (2019). *Sampling: Design and Analysis.* 2nd Ed Ed, Cengage.

Pages from additional texts will be distributed on the course website.

Homework: Homeworks are due in the beginning of the class. Any homeworks turned in afterwards will not receive credit. Expect weekly homeworks during the semester.

Exams: Two in-class midterm exams, and a final exam.

Project: At the end of the semester groups the students will perform a group project analyzing a real-world problem.

Grades: All grades will be distributed via Blackboard.

Re-grades: All re-grading requests should be made in writing one week 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 500 points broken down into homeworks (100 pts), midterm (100 pts each), 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.