PubH 7475/8475/Stat 8931 Statistical Learning and Data Mining
Spring 2015, 3 credits, A/F or S/N

Course Description: The subject of this course is closely related to machine learning, data analytics, and Big Data. This course will introduce various statistical techniques for extracting useful information (i.e. learning) from data, possibly Big Data. Topics to be covered include penalized/regularized regression, linear discriminant analysis, tree-structured classifiers, feed-forward neural networks, support vector machines, classifier ensembles (such as bagging and boosting), unsupervised learning, and Big Data. These techniques can be applied in many fields, such as in business and bioinformatics/computational biology.

Class time and location: 11:15 AM - 12:30 PM, Tu, Th, KHKH 3-115.

Number of credits: 3

Course Instructors: Dr. Wei Pan, Professor of Biostatistics
Office: A428 Mayo, Phone: 626-2705, Fax: 626-0660, Email: weip@biostat.umn.edu
Office hours: 12:30-13:30pm Tu & Th
Dr. Xiaotong Shen, Professor of Statistics
Office: 384 Ford Hall, Phone: 612-624-7098, Email: xshen@stat.umn.edu Office hours: 10:15am-11:00am, Tu & Th
TAs: Mr. Chen Gao, Email: gaoxx492@umn.edu, Office hours: 1:30-3:30 Tu, A446 Mayo.
Mr. Zhiyuan Xu, Email: xuxx0284@umn.edu, Office hours: 10-12 F, A446 Mayo.

Course homepage:
http://www.biostat.umn.edu/~weip/course/dm/s15/home.html

For whom intended: This course PubH 7475 is designed for second-year MS biostatistics/statistics graduate students, and other graduate students in public health, computer science, engineering, business and biology who have relevant statistical background. More work is expected for those who want to take it as PubH 8475/Stat 8931 for PhD credits.

Prerequisites: Statistics at the level of PubH 7405–7407 or equivalent (e.g. Stat 5303) or permission of instructor, and some programming background in using R (or another high level language such as FORTRAN, C/C++, SAS).

Objective: After taking the course, the student should have a working knowledge of using various machine learning techniques in practice.
Methods of Instruction and Work Expectations: In-class lectures are the main method of instruction. Students are expected to come to class and participate in discussions, to do assignments, and to write a report and give a presentation for a course project. For those taking the 8000 section, some extra problems at a higher and theoretical level in the assignments and project are expected. Late assignments or project report are not accepted unless with advance permission from the instructor.

Evaluation and Grading: Course evaluation will be based on class participation, homework assignments and a course project. The final grade is based on a weighted average score of a student’s performance in homework assignments, a mid-term exam and a project, with weights 40%, 20% and 40% respectively.

There are about five to six homework assignments. Each assignment involves applying and evaluating some statistical learning methods, and/or writing a reading report; those taking the 8000 section may need to do some more theoretical problems, and read and critique journal articles. We will use R programming language, which is freely available from http: //www.r-project.org; you can use whatever language/system you like, though of course it will be your sole responsibility with programming. The mid-term exam is tentatively scheduled in week 10. For the course project, possible topics include a case study (i.e. analysis of a specific data set), an empirical or theoretical comparison of a few statistical learning methods, or development/review and evaluation of a new method (not covered in class), or do a literature review/survey. Your project topic needs to be approved by an instructor in advance. The project will be done by a team of three to four students. You need to write a half-page project proposal, due in class in week 5. In the final two weeks, a short presentation on each project will be given by its team members; every team will evaluate every other’s work and presentation. A ≤ 5-page final project report, including Introduction (or Background), Methods, Results, and possibly Discussion sections, is due by 4pm on May 12 (Tu) in either instructor’s office or Biostat/Stat main office.

No late homework assignments, project proposal or project report will be accepted unless with some legitimate reasons (e.g. illness with appropriate documents) or with my approval in advance.

In a scale of 100 total points of the average score, letter grade is determined tentatively as follows:
A=90-100 points (4.0) Represents achievement that is outstanding relative to the level necessary to meet course requirements

A- = 87-89 points
B+ = 84-86 points
B = 80-83 points (3.0) Represents achievement that is significantly above the level necessary to meet course requirements

B- = 78-79 points
C+ = 75-77 points
C = 70-74 points (2.0) Represents achievements that meets the minimum course requirements

C- = 65-69 points
D = 60-64 points
F (or N) = < 60 points
S = ≥ 65 points

S = Achievement that is satisfactory will be expected to complete all assignments and receive a minimum of 65% to receive a passing score.

F (or N) Represents failure (or no credit) and signifies that the work was either (1) completed but at a level of achievement that is not worthy of credit or (2) was not completed and there was no agreement between the instructor and the student that the student would be awarded an I.

Students may change grading options without written permission as specified by the University and without penalty during the initial registration period or during the first two weeks of the semester. No W will appear on the transcript.

After the second week students are required to do the following:

- The student must contact and notify their advisor and course instructor informing them of the decision to withdraw from the course.
- The student must send an e-mail to the SPH Student Services Center (SSC). The email must provide the student name, ID#, course number, section number, semester and year with instructions to withdraw the student from the course, and acknowledgement that the instructor and advisor have been contacted.
- The advisor and instructor must email the SSC acknowledging the student is canceling the course. All parties must be notified of the student’s intent.
- The SSC will complete the process by withdrawing the student from the course after receiving all emails (student, advisor, and instructor). A W will be placed and remain on the student transcript for the course.
- After discussion with their advisor and notification to the instructor, students may withdraw up until the eighth week of the semester. There is no appeal process.
An incomplete grade is permitted only in cases of extraordinary circumstances and following consultation with the instructor. In such cases an I grade will require a specific written agreement between the instructor and student specifying the time and manner in which the student will complete the course requirements. Extension for completion of the work will not exceed one year.

Scholastic dishonesty is a violation of the student conduct code and is defined as any act that violates the rights of another student in academic work or that involves misrepresentation of your own work. Scholastic dishonesty includes (but is not limited to): cheating on assignments or examinations; plagiarizing, which means misrepresenting as your own work any part of work done by another; submitting the same paper, or substantially similar papers, to meet the requirements of more than one course without the approval and consent of all instructors involved; depriving another student of necessary course materials; or interfering with another student’s work. Scholastic dishonesty in any portion of the academic work for a course shall be grounds for awarding a grade of F or N for the entire course. Please consult the student conduct code at: http://www.umn.edu/regents/policies/academic/StudentConduct.html.

Textbook and reference:


- James G, Witten D, Hastie T, Tibshirani R (2014). *An Introduction to Statistical Learning with Applications in R*. Springer. (not required; at a lower level with R examples)

An Outline of the Course Schedule:

- Dr. Pan:
  - Introduction (1 week);
  - Linear regression, penalized regression, logistic regression and discriminant analysis (2 weeks);
  - Classification and regression trees (0.5 week);
  - Ensemble methods, including Bagging, boosting, MART and RF (1.5 weeks);
  - Neural networks and support vector machines (1 week);
  - Unsupervised learning: clustering, PCA, etc. (2 week);
• Dr. Shen:
  – More on high-dimensional data (1 week);
  – Other topics: Graphical models and network reconstruction, Recommender system (2 week);
  – Big Data (1-2 weeks);
• Student presentations (2 weeks).

Disability accommodation:
Any student with a documented disability (e.g., physical, learning, psychiatric, vision, hearing, etc.) who needs to arrange reasonable accommodations must contact the instructor and Disability Services at the beginning of the semester. All discussions will remain confidential. For further information contact the University of Minnesota Disability Services website at http://disserv3.stu.umn.edu/index2.html or call 612/626-1333 (V/TTY).