PubH 7475/8400 Statistical Learning and Data Mining

Spring 2007, 3 credits, A/F or S/N

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

Number of credits: 3

Course Instructors: Dr. Wei Pan, Associate Professor of Biostatistics
Office: A428 Mayo, Phone: 626-2705, Fax: 626-0660, Email: weip@biostat.umn.edu
Office hours: 2pm-3pm T&Th

Course homepage:
http://www.biostat.umn.edu/∼weip/course/dm/s07/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 8400 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 a higher level language, such as FORTRAN, C/C++, SAS or Splus/R.

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 for a course project. For those taking PubH 8400, some extra problems at a *higher-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 class participation, homework assignments and the project, with weights 10%, 40% and 50% respectively.

There are about five homework assignments. Each assignment involves applying and evaluating some statistical learning methods, or writing a reading report; those taking PubH 8400 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. For the course project, you can analyze a specific data set, or empirically or theoretically compare a few statistical learning methods, or develop and evaluate a new method, or do a literature survey. If you are not sure whether your chosen topic is appropriate, you are encouraged to discuss with me. The project can be done individually, or by a team of two students, in which case each participant's role must be clearly specified. You need to write a half-page project proposal, **due in class on March 27 (T)**. In the final two weeks (or so), a short presentation on each project will be given by its team member(s); everyone will evaluate every other's presentation. A $\leq$ 5-page final project report, including Introduction (or Background), Methods, Results, and possibly Discussion sections, is **due by 4pm on May 11 (F)** in my Mayo mailbox or Biostatistics frontdesk in Mayo A460.

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 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 = $\geq$ 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:

- Hastie T, Tibshirani R and Friedman J (2001). *The Elements of Statistical Learning, Data Mining, Inference and Prediction.* Springer. (required)
- Ripley BD (1996). *Pattern Recognition and Neural Networks.* Cambridge Univeristy Press.

Both books are reserved at the Biostatistics Reading Room (A460 Mayo).

An Outline of the Course Schedule:

- Introduction. (1 week)
- Linear regression, logistic regression, linear discriminant analysis, and flexible discriminant analysis (4 weeks)
- Nonparametric methods: nearest neighbor, mixture models (1 week)
- Tree-structured classifiers (1 weeks)
- Ensemble methods, including Bagging, boosting, MART and RF (2 weeks)
- Neural networks and support vector machines (2 weeks)
- Model selection (1 week)
- Unsupervised learning (1 week)
- Student presentation (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).