SYLLABUS & COURSE INFORMATION

SCHOOL OF PUBLIC HEALTH
UNIVERSITY OF MINNESOTA

PUBH 7475/8475, SECTION 001

Statistical Learning and Data Mining
Spring 2022

COVID-19 INFO

https://docs.google.com/document/d/1sc_wcOe3fmhVcAvaoyoJaKbTxKL7rdh699BlbWrGYBA/edit#

COURSE & CONTACT INFORMATION

Credits: 3
Meeting Days: M&W; for the first few weeks, zoom on Mondays and in-person on Wednesdays.
Meeting Time: 9:45 am—11:00 am
Meeting Place: MoosT 5-125 & Virtual/Zoom
Zoom: on Canvas

Instructor: Dr. Wei Pan
Email: panxx014@umn.edu
Office Hours: 11 am—12 pm M&W
Office Location: A446 Mayo; Zoom (the same Zoom link above)

TA: Mr. Andy Becker
Email: beck0822@umn.edu
Office Hours: 2-3pm M&W
Zoom link: on Canvas

TA: Mr. Jonathan Kim
E-mail: kim00225@umn.edu
Office Hours: 10-11am Tu&Th
Zoom Link: on Canvas

COURSE DESCRIPTION

The subject of this course is closely related to machine learning and data science with an emphasis on statistical aspects/views. This course will introduce various statistical/computational techniques for supervised learning and unsupervised learning. Topics to be covered include basic concepts (such as training versus test errors, cross-validation, bias-variance trade-off), penalized/regularized regression, linear discriminant analysis, tree-structured classifiers, neural networks, support vector machines, classifier ensembles (such as bagging and boosting), unsupervised learning (dimension reduction, clustering analysis, network analysis). These techniques can be applied in many fields, such as in business and bioinformatics/computational biology.

Course home pages: http://www.biostat.umn.edu/~weip/course/dm/s22/home.html
Please visit regularly.

COURSE PREREQUISITES

Statistics at the minimum level of PUBH 7405–7406 or equivalent (e.g. Stat 5303), preferably at a higher level, or permission of instructor; familiarity with programming in R (or Python if you are willing to learn by yourself).

COURSE GOALS & OBJECTIVES

After taking the course, the student should have a working knowledge of using various machine learning techniques in practice. For 8475, one is expected to be able to do methods research in the field.

METHODS OF INSTRUCTION AND WORK EXPECTATIONS

In-class lectures are the main method of instruction. Students are expected to come to class for active learning, e.g. participating in discussions, to do (reading and written) assignments, and to (co-)write a report and present for a course project towards the end of the semester. For those taking 8475, 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 legitimate reasons or advance permission from the instructor.

This is a 3 credit course. The University expects that for each credit, you will spend a minimum of three hours per week attending class or comparable online activity, reading, studying, completing assignments, etc. over the course of a 15-week term. Thus, this course requires approximately [3 * 45] hours of effort spread over the course of the term in order to earn an average grade.

Learning Community

School of Public Health courses ask students to discuss frameworks, theory, policy, and more, often in the context of past and current events and policy debates. Many of our courses also ask students to work in teams or discussion groups. We do not come to our courses with identical backgrounds and experiences and building on what we already know about collaborating, listening, and engaging is critical to successful professional, academic, and scientific engagement with topics.

In this course, students are expected to engage with each other in respectful and thoughtful ways.

In group work, this can mean:

- Setting expectations with your groups about communication and response time during the first week of the semester (or as soon as groups are assigned) and contacting the TA or instructor if scheduling problems cannot be overcome.
- Setting clear deadlines and holding yourself and each other accountable.
- Determining the roles group members need to fulfill to successfully complete the project on time.
- Developing a rapport prior to beginning the project (what prior experience are you bringing to the project, what are your strengths as they apply to the project, what do you like to work on?)

In group discussion, this can mean:

- Respecting the identities and experiences of your classmates.
- Avoid broad statements and generalizations. Group discussions are another form of academic communication and responses to instructor questions in a group discussion are evaluated. Apply the same rigor to crafting discussion posts as you would for a paper.
- Consider your tone and language, especially when communicating in text format, as the lack of other cues can lead to misinterpretation.

Like other work in the course, all student to student communication is covered by the Student Conduct Code (https://z.umn.edu/studentconduct).

COURSE TEXT & READINGS

Hastie T, Tibshirani R and Friedman J (2009). *The Elements of Statistical Learning, Data Mining, Inference and Prediction.* Springer, 2nd edition. Available on-line. (required)

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)

François Chollet with J. J. Allaire (2018). Deep Learning with R. Manning. (not required; DL examples in R interface to Keras) (François Chollet (2018). Deep Learning with Python. Manning.)

COURSE OUTLINE/WEEKLY SCHEDULE

Week	Topic	Readings	Activities/Assignments
Week 1: 1/19	• Introduction	• Class notes (&textbook, papers)	• Class website
Week 2: 1/24, 1/26	• Overview; linear models for classification	• Class notes	• HWK1 assigned; class website
Week 3: 1/31, 2/2	• Penalized regression	• Class notes	• HWK2 assigned; class website
Week 4: 2/7, 2/9	• Penalized regression; PCR/PLS	• Class notes	• Class website
Week 5: 2/14, 2/16	• CART; Bagging; BMA; Stacking	• Class notes	• HWK3 assigned; class website
Week 6: 2/21, 2/23	• RF; Boosting	• Class notes	• Class website
Week 7: 2/28, 3/2	• SVM; FNN	• Class notes	• HWK4 assigned; class website
Spring Break: 3/7, 3/9	• No class	•	•
Week 8: 3/14, 3/16	• FNN&CNN	• Class notes	• Class website
Week 9: 3/21, 3/23	• RNN; In-class exam	• Class notes	• Class website
Week 10: 3/28, 3/30	• Clustering	• Class notes	• Proposal due; Class website
Week 11: 4/4, 4/6	• Clustering; semi-supervised	• Class notes	• HWK5 assigned; Class website
Week 12: 4/11, 4/13	• Network analysis	• Class notes	• Class website
Week 13: 4/18, 4/20	• Other topics; Student presentation	• Class notes	• Class website; Slides
Week 14: 4/25, 4/27	• Student presentations	•	• Slides
Week 15: 5/2	• Student presentations	•	• Slides; report due on 5/6

SPH AND UNIVERSITY POLICIES & RESOURCES

The School of Public Health maintains up-to-date information about resources available to students, as well as formal course policies, on our website at www.sph.umn.edu/student-policies/. Students are expected to read and understand all policy information available at this link and are encouraged to make use of the resources available.

The University of Minnesota has official policies, including but not limited to the following:

- Grade definitions
- Scholastic dishonesty
- Makeup work for legitimate absences
- Student conduct code
- Sexual harassment, sexual assault, stalking and relationship violence
- Equity, diversity, equal employment opportunity, and affirmative action
- Disability services
- Academic freedom and responsibility

Resources available for students include:

- Confidential mental health services
- Disability accommodations
- Housing and financial instability resources
- Technology help
- Academic support

EVALUATION & GRADING

Course evaluation will be based on class participation, homework assignments, a midterm exam and a course project. **The final grade is based on a weighted average score of a student's performance in class participation, homework assignments, a mid-term exam and a final project, with weights 10%, 40%, 20% and 30% respectively.**
There are 5-**6 homework assignments**. Each assignment involves applying and evaluating some statistical learning methods, and/or writing a reading report; those taking 8475 may need to do some more theoretical or computational 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 as an in-class (NOT take-home) on **March 23, Wednesday**. For the final 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/implementation and evaluation of a new/existing method (e.g. not covered or emphasized in class), or do an extensive literature review/survey on a topic. Your final project topic may be discussed with and approved by the instructor in advance. A project proposal will be due by Week 9. The project will be done by a team of **two-three** students; *team work is strongly encouraged*. In the final two weeks, a short presentation on each project will be given by its team members. **A ≤ 5-page final project report for a whole team, including Introduction (or Background), Methods, Results, and possibly Discussion sections, is due by 4pm (tentatively on May 6). Each student is required to write a short critique on each presentation (not given on the same day as one's own) and submit with the same deadline as that for the final project report.**
No late homework assignments and project reports will be accepted unless with some legitimate reasons (e.g. illness with appropriate documents) or with my approval in advance.

Grading Scale
The University uses plus and minus grading on a 4.000 cumulative grade point scale in accordance with the following, and you can expect the grade lines to be drawn as follows:

% In Class	Grade	GPA
93 - 100%	A	4.000
90 - 92%	A-	3.667
87 - 89%	B+	3.333
83 - 86%	B	3.000
80 - 82%	B-	2.667
77 - 79%	C+	2.333
73 - 76%	C	2.000
70 - 72%	C-	1.667
67 - 69%	D+	1.333
63 - 66%	D	1.000
< 62%	F	

- A = achievement that is outstanding relative to the level necessary to meet course requirements.
- B = achievement that is significantly above the level necessary to meet course requirements.
- C = achievement that meets the course requirements in every respect.
- D = achievement that is worthy of credit even though it fails to meet fully the course requirements.
- F = failure because 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 (Incomplete).
- S = achievement that is satisfactory, which is equivalent to a C- or better
- N = achievement that is not satisfactory and signifies that the work was either 1) completed but at a level that is not worthy of credit, or 2) not completed and there was no agreement between the instructor and student that the student would receive an I (Incomplete).

Evaluation/Grading Policy	Evaluation/Grading Policy Description
Scholastic Dishonesty, Plagiarism, Cheating, etc.	You are expected to do your own academic work and cite sources as necessary. Failing to do so is scholastic dishonesty. Scholastic dishonesty means plagiarizing; cheating on assignments or examinations; engaging in unauthorized collaboration on academic work; taking, acquiring, or using test materials without faculty permission; submitting false or incomplete records of academic achievement; acting alone or in cooperation with another to falsify records or to obtain dishonestly grades, honors, awards, or professional endorsement; altering, forging, or misusing a University academic record; or fabricating or falsifying data, research procedures, or data analysis (As defined in the Student Conduct Code). For additional information, please see https://z.umn.edu/dishonesty The Office for Student Conduct and Academic Integrity has compiled a useful list of Frequently Asked Questions pertaining to scholastic dishonesty: https://z.umn.edu/integrity. If you have additional questions, please clarify with your instructor. Your instructor can respond to your specific questions regarding what would constitute scholastic dishonesty in the context of a particular class-e.g., whether collaboration on assignments is permitted, requirements and methods for citing sources, if electronic aids are permitted or prohibited during an exam. Indiana University offers a clear description of plagiarism and an online quiz to check your understanding (http://z.umn.edu/iuplagiarism).
Late Assignments	Needs prior approval with suitable document
Attendance Requirements	Class participation is required
Extra Credit	N/A