SYLLABUS & COURSE INFORMATION

SCHOOL OF PUBLIC HEALTH
UNIVERSITY OF MINNESOTA

PUBH 8475/STAT 8056

Advanced Topics on Machine Learning
Spring 2024

COURSE & CONTACT INFORMATION

Credits: 3 credits
Meeting Day(s): Monday and Wednesday
Meeting Time: 9:45am – 11:00am
Meeting Place: Health Sciences Edu Ctr 2-132

Instructor: Wei Pan, PhD
Email: panxx014@umn.edu
Office Hours: Monday and Wednesday, 11:00am-11:30am, Mayo D378 (during the weeks as the instructor)
Office Location: University Office Plaza, 2221 University Avenue SE, Room 235, Minneapolis

Instructor: Xiaotong Shen, PhD
Email: xshen@umn.edu
Office Phone: 612-624-7098
Office Hours: Monday and Wednesday, 11:00 am-12:00 pm (during the weeks as the instructor)
Office Location: Ford 391

TA: Michael Anderson
Email: and02709@umn.edu
Office Hours: Tuesday and Friday, 3:00pm-4:00pm
Office Location: University Office Plaza, 2221 University Avenue SE, Room TBA, or zoom https://umn-private.zoom.us/j/7013228968

COURSE DESCRIPTION

It covers a range of emerging topics in machine learning, data analytics, and big data. This course will introduce various statistical and computational techniques for prediction and inference. These techniques are directly applicable to Business Analytics, Engineering, and Biomedical Research. This course requires basic knowledge of machine learning and data mining (e.g., Pubh7475, Stat8053 or a similar course). Topics include the following:

Data exploration and data science;
Optimization for machine learning;
High-dimensional analysis: prediction and inference;
Deep neural network learning: basics (FNN, CNN, RNN/LSTM); advanced topics (Transformers, Diffusion models, etc);
Recommender systems: personalized prediction;
Undirected and directed graphical models;
Unstructured data and text mining: Numerical embedding and language models;
Causal Machine Learning.

Course home pages: Pub 8475- http://www.biostat.umn.edu/~weip/course/dm/s24/home.html
Stat 8056- http://www.stat.umn.edu/~xshen/stat8056.htm

Please visit regularly.

COURSE PREREQUISITES

Pubh7475 or Stat8053 or a similar course, or permission of instructor; familiarity with programming in R or Python.

COURSE GOALS & OBJECTIVES

After taking the course, the student should have a working knowledge of using some state-of-the-art of AI/M techniques and learning other emerging ones in practice.

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, doing (reading and written) assignments, and (co-) writing a report and presenting for a course project towards the end of the semester. Late assignments or project reports 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 a 15-week term. Thus, this course requires approximately [3 * 45] hours of effort spread over the term to earn an average grade.

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

COURSE TEXT & READINGS

No textbook. Slides and published research papers will be shared. Following is a list of suggested (optional) readings, which will be updated as the course progresses during the semester.

1. Introduction

1) McKinsey Global Institute June 2011 Big data: The next frontier for innovation, competition, and productivity.
2) Donoho D. (2015), 50 Years of Data Science. JCGS.
3) Breiman L. (2001), Statistical Modeling: The Two Cultures (with comments and a rejoinder by the author). Statist. Sci. 16, iss. 3, 199-231.
4) Hand, D.J. (2006), Classifier Technology and the Illusion of Progress (with comments and a rejoinder by the author). Statist. Sci. 21, Iss. 1, 1-34.
5) S. Guha, R. Hafen, J. Xia, J. Rounds, J. Li, B. Xi, and W. S. Cleveland (2012), Large complex data: divide and recombine (D&R) with RHIPE, Stat 1, 53-67.
6) Cleveland W.S. (2001, republished 2014), Data science: An action plan for expanding the technical areas of the field of statistics. Statistical Analysis and Data Mining 7, Iss. 6, 414-417

2. Optimization

1) Breiman, L. and Cutler, A. (1993). A Deterministic algorithm for global optimization. Mathematical Programming, 58, 179--1993.
2) Horst, R., and Thoai, N. V. (1999). DC programming: Overview. Journal of Optimization Theory and Application. 103, 1-43.
3) Chen, Y., Ye, Y., and Wang, M. (2018). Approximation hardness for a class of sparse optimization problems. Journal of Machine Learning Research, 20, 1-27.

3. High-dimensional Analysis

1) Fan J, Li R (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. Journal of the American Statistical Association 96 (456), 1348-1360.
2) Zou H (2006), The Adaptive Lasso and Its Oracle Properties. JASA, 101, 418-1429.
3) Zou H, Hastie T (2005), Regularization and variable selection via the elastic net. Journal of the Royal Statistical Society: Series B, 67, 301-320.
4) Austin E, Pan W, Shen X. (2013). Penalized Regression and Risk Prediction in Genome-Wide Association Studies. Stat Anal Data Min. 6(4). doi: 10.1002/sam.11183.
5) Zhu Y, Shen X, Pan W (2013). Simultaneous grouping pursuit and feature selection over an undirected graph. JASA, 108, 713-725.
6) Kim S, Pan W, Shen X (2013). Network-based penalized regression with application to genomic data. Biometrics. 69(3), 582-593.
7) Friedman J, Hastie T, Hoefling H, Tibshirani R (2007). Pathwise Coordinate Optimization. The Annals of Applied Statistics, 2(1), 302–332.
8) Jerome Friedman, Trevor Hastie, Robert Tibshirani (2010). Regularization Paths for Generalized Linear Models via Coordinate Descent. Journal of Statistical Software, 33(1), 1-22.
9) S. Boyd, N. Parikh, E. Chu, B. Peleato, and J. Eckstein, (2011). Distributed optimization and statistical learning via the alternating direction method of multipliers. Foundations and Trends in Machine Learning, 3(1):1-122.
10) Shi C, Song R, Chen Z, Li R. (2019). Linear hypothesis testing for high dimensional generalized linear models. Ann Stat, 47(5), 2671-2703.
11) Zhu Y, Shen X, Pan W. (2020). On High-Dimensional Constrained Maximum Likelihood Inference. JASA, 115(529), 217-230.
12) Dezeure R, Buhlmann P, Meier L and Meinshausen N (2015). High-Dimensional Inference: Confidence Intervals, p-Values, and R-Software hdi. Stat Sci, 30(4), 533-558.
13) Fan J, Lv J (2008). Sure independence screening for ultrahigh dimensional feature space. JRSS-B 70, 849-911.
14) Wu, C., Xu, G., Shen, X., and Pan, W. (2020). A regularization-based adaptive test for high-dimensional generalized linear models. Journal of Machine Learning Research}. 21(128), 1-67.

4. Deep Learning

1) LeCun et al (1998). Gradient-based learning applied to document recognition. Proc of IEEE. (Comment: Section I. p.5-7 most helpful to understand convolutional NNs.)
2) Krizhevsky A, Sutskever I, Hinton G. (2012). ImageNet Classification with Deep Convolutional Neural Networks. NeurIPS.
3) Zhou J and Troyanskaya OG (2015). Predicting effects of noncoding variants with deep-learning based sequence model. Nature Methods, 12, 931-934.
4) Silver et al. (2016). Mastering the game of Go with deep neural networks and tree search. Nature, 529, 484-489.
5) Xiao M, Shen X, Pan W. (2019). Application of deep convolutional neural networks in classification of protein subcellular localization with microscopy images. Genetic Epi, 43(3), 330-341.

6) Zhuang Z, Shen X, Pan W. (2019). A simple convolutional neural network for prediction of enhancer-promoter interactions with DNA sequence data. Bioinformatics, 35(17), 2899-2906.
7) Fan J, Ma C, Zhong Y. (2019). A Selective Overview of Deep Learning. arXiv:1904.05526.
8) Bernhard Schölkopf, Francesco Locatello, Stefan Bauer, Nan Rosemary Ke, Nal Kalchbrenner, Anirudh Goyal, Yoshua Bengio. (2021). Towards Causal Representation Learning. arXiv:2102.11107
9) Zech JR, Badgeley MA, Liu M, Costa AB, Titano JJ, Oermann EK (2018) Variable generalization performance of a deep learning model to detect pneumonia in chest radiographs: A cross-sectional study. PLoS Med 15(11): e1002683. doi:10.1371/journal.pmed.1002683.
10) Chaofan Chen, Oscar Li, Daniel Tao, Alina Barnett, Cynthia Rudin, Jonathan K. Su (2019). This Looks Like That: Deep Learning for Interpretable Image Recognition. Advances in Neural Information Processing Systems 32 (NeurIPS 2019).
11) Ramprasaath R. Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, Dhruv Batra. Grad-CAM: Visual Explanations from Deep Networks via Gradient-based Localization. International Journal of Computer Vision (IJCV) 2019.
12) Stephanie Clark, Rob J Hyndman, Dan Pagendam, Louise M Ryan. (2020). Modern strategies for time series regression. International Stat Rev, 88(S1), S179-S204.
13) Volodymyr Mnih et al. (2015). Human-level control through deep reinforcement learning. Nature, 518, 529-533.
14) Dai, B., Shen, X., and Pan, W. (2022). Significance tests of feature relevance for a black-box learner. IEEE Transactions on Neural Networks and Learning Systems.
15) Sohl-Dickstein, J., Weiss, E., Maheswaranathan, N., and Ganguli, S. (2015). Deep unsupervised learning using nonequilibrium thermodynamics. International Conference on Machine Learning. 2256--2265.
16) Kingma, D. P. and Dhariwal, P. (2018). Glow: Generative flow with invertible 1 by 1 convolutions. Advances in neural information processing systems, 31.
17) Kotelnikov, A.and Baranchuk, D. and Rubachev, I., and Babenko, A. (2022).TabDDPM: Modelling tabular data with diffusion models. arXiv preprint arXiv:2209.15421.

5. Recommender systems

1) Bi, X., Qu, A., Wang, J., and Shen, X. (2017). A group-specific recommender system. Journal of American Statistical Association, 112, 1344-1353.
2) Dai, B., Wang, J, Shen, X., and Qu, P. (2020). Smooth neighborhood recommender systems. The Journal of Machine Learning Research. 20(16),1-2.
3) Mazumder, R., Hastie, T, and Tibshirani, R. (2010). Spectral regularization algorithms for learning large incomplete matrices. Journal of Machine Learning Research, 11, 2287-2322.

6. Graphical Models

1) Mazumder, R. and Hastie, T. (2012). The graphical lasso: New insights and alternatives. Electronic Journal of Statistics, 6, 2125-2149.
2) Guo, J., Levina, E., Michailidis, G. and Zhu, J. (2010). Joint estimation of multiple graphical models. Biometrika, 98, 1-15.
3) Yuan, Y., Shen, X., Pan, W., and Wang, Z. (2019). Reconstruction of a directed acyclic Gaussian graph. Biometrika. 106, 109-125.

7. Causal ML

1) Ishwaran H, Lu M. (2019). Standard errors and confidence intervals for variable importance in random forest regression, classification, and survival. Stat Med. 38(4):558-582.
2) Lu M, Sadiq S, Feaster DJ, Ishwaran H. (2018). Estimating Individual Treatment Effect in Observational Data Using Random Forest Methods. J Comput Graph Stat. 27(1), 209-219.
3)Vincent Dorie, Jennifer Hill, Uri Shalit, Marc Scott, and Dan Cervone. (2019). Automated versus Do-It-Yourself Methods for Causal Inference: Lessons Learned from a Data Analysis Competition. Statist Sci, 34, 43-68.
4)Li, C., Shen, X., and Pan, W. (2020). Likelihood inference for a large causal network. Journal of American Statistical Association. 113, 1--16.
5) Chen, L., Li, C., Shen, X., and Pan, W. (2023). Discovery and Inference of a Causal Network with Hidden Confounding. Journal of American Statistical Association.

8. Semi-supervised learning:

1) Ting Chen, Simon Kornblith, Mohammad Norouzi, Geoffrey Hinton, (2020). A Simple Framework for Contrastive Learning of Visual Representations. ICMK 2020.
2) Peng Liu , Yusi Fang, Zhao Ren, Lu Tang, George C. Tseng (2021). Outcome-Guided Disease Subtyping for High-Dimensional Omics Data. arXiv:2007.11123
3) Wagstaff et al (2001). Constrained K-means Clustering with Background Knowledge.
4) Liu B, Shen X, Pan W (2013). Semi-supervised spectral clustering with application to detect population stratification. Frontiers in Genetics. 4:215. doi:10.3389/fgene.2013.00215.
5) Wang J, Shen X, Pan W. (2009). On efficient large margin semisupervised learning: method and theory. Journal of Machine Learning Research. 10, 719-742.
6) Wang, J., Shen, X., and Pan, W. (2006). On transductive support vector machines. Contemp. Math., 43, 7-19.
7) Wei Pan, Xiaotong Shen, Aixiang Jiang, and Robert P. Hebbel (2006). Semi-supervised learning via penalized mixture model with application to microarray sample classification. Bioinformatics, 22, 2388-2395.

9. Network analysis:

1) Neuman MEJ. Detecting community structure in networks.
2) Blondel VD, Guillaume J-L, Lambiotte R, Lefebvre E. (2008). Fast unfolding of communities in large networks. arXiv:0803.0476

3) Zhao Y, Levina E, Zhu J (2012). Consistency of community detection in networks under degree-corrected stochastic block models. Ann. Statist. Volume 40, Number 4 (2012), 2266-2292.
4) Fortunato S (2010). Community detection in graphs. Physics Reports 486, 75-174.
5) David Meunier, Renaud Lambiotte and Edward T. Bullmore (2010). Modular and hierarchically modular organization of the brain networks. Front. Neurosci., 4, 200.

COURSE OUTLINE/WEEKLY (TENTATIVE) SCHEDULE

Week	Topic	Readings	Activities/Assignments
Week 1, Jan 15-19	• Data exploration and data science (Shen)	• Selected papers in Group 1.	• Readings
Week 2, Jan 22-26	• Optimization for machine learning (Shen)	• Papers in Group 2	• Readings
Week 3, Jan 29-Feb 2	• High-dimensional analysis (Pan)	• Selected papers in Group 3.	• Assignment 1 • Readings
Week 4, Feb 5-9	• Deep learning Basics: FNN, CNN (Pan)	• Selected papers in Group 4	• Readings
Week 5, Feb 12-16	• Python tutorial: Deep Learning (Guest: Liu)	Selected papers in Group 4	• Readings
Week 6, Feb 19-23	• DL Advanced: RNN, LSTM, Models (Shen)	• Selected papers in Group 4	• Readings
Week 7, Feb 26-Mar 1	• Natural Language Processing and models (Shen)	• Selected papers in Group 4	• Assignment 2 • Readings
Week 8, Mar 4-8	• Spring Break (No Class)		
Week 9, Mar 11-15	• Undirected and directed graphical models (Pan)	• Selected papers in Group 6	• Readings
Week 11, Mar 18-22	• Diffusion Models, Normalizing Flows, Synthetic Data, and Applications (Shen)	• Papers in Group 4	• Readings
Week 12, Mar 25-29	• Recommender systems (Shen)	• Selected papers in Group 5	• Readings
Week 13, Apr 1-5	• Causal ML (Pan)	Selected papers in Group 7	• Assignment 3 • Readings
Week 13, Apr 8-12	• Semi-supervised learning (Pan)	• Selected papers in Group 8	• Readings
Week 14, Apr 15-19	• Student presentations	•	• Readings
Week 15, Apr 22-26	• Student presentations	•	• Critiques
Week 16, Apr 29	• Student Presentations	•	• Critiques

LAND ACKNOWLEDGEMENT

The School of Public Health at the University of Minnesota Twin Cities is built within the traditional homelands of the Dakota people. Minnesota comes from the Dakota name for this region, Mni Sóta Maḳoce, which loosely translates to the land where the waters reflect the skies.

It is important to acknowledge the peoples on whose land we live, learn, and work as we seek to improve and strengthen our relations with our tribal nations. We also acknowledge that words are not enough. We must ensure that our institution provides support, resources, and programs that increase access to all aspects of higher education for our American Indian students, staff, faculty, and community members.

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, and a final project, with weights of 10%, 40%, and 50% respectively.**
There are about 3 **homework assignments**. Each assignment involves applying and evaluating some statistical learning methods, and/or writing a reading report; the students may need to do some more theoretical or computational problems, and read and critique journal articles. 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 may be undertaken individually or by a team of 2-3 students, with a strong preference for collaboration. In the final 2-3 weeks, a 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 to be submitted during the last class session (tentatively on April 29). Each student is required to write a short critique on each presentation (not given on the same day as one's own) and submit it 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 (https://plagiarism.iu.edu/certificationTests/).
Late Assignments	
Attendance Requirements	
Makeup Work for Legitimate Reasons	If you experience an extraordinary event that prevents you from completing coursework on time and you would like to make arrangements to make up your work, contact your instructor within 24 hours of the missed deadline if an event could not have been anticipated and at least 48 hours prior if it is anticipated. University policy recognizes that there are a variety of legitimate circumstances in which students will miss coursework and that accommodations for makeup work will be made. This policy applies to all course requirements, including any final examination. Students are responsible for planning their schedules to avoid excessive conflicts with course requirements. 1. Instructors may not penalize students for absence during the academic term due to the following unavoidable or legitimate circumstances: illness, physical or mental, of the student or a student's dependent; medical conditions related to pregnancy; participation in intercollegiate athletic events; subpoenas; jury duty; military service; bereavement, including travel related to bereavement; religious observances; participation in formal University system governance, including the University Senate, Student Senate, and Board of Regents meetings, by students selected as representatives to those bodies; and activities sponsored by the University if identified by the senior academic officer for the campus or the officer's designee as the basis for excused absences. 2. Voting in a regional, state, or national election is not an unavoidable or legitimate absence. 3. Instructors are expected to accommodate students who wish to participate in party caucuses, pursuant to the Board of Regents resolution (see December 2005 Board of Regents Minutes, p 147.) 4. For circumstances not listed in (1), the instructor has primary responsibility to decide on a case-by-case basis if an absence is due to unavoidable or legitimate circumstances and grant a request for makeup work. Because this course is entirely online and all materials are available to students from the first day of the term, we expect students to plan accordingly if travel or access to the internet will cause them to miss a deadline. Note that our deadlines are generally set for 11:55 p.m. CST, so traveling to a different time zone will require additional planning. Further, circumstances that qualify for making up missed work will be handled by the instructor on a case-by-case basis; they will always be considered but not always granted. For complete information, view the U of M's policy on Makeup Work for Legitimate Absences (https://policy.umn.edu/education/makeupwork).
Extra Credit	

CEPH COMPETENCIES

Competency	Learning Objectives	Assessment Strategies
Evidence-based Approaches to Public Health	Apply suitable quantitative methods to analyze public health data	Homework assignments, readings, and final project
Public Health & Health Care Systems	Discuss structural bias, social inequities, and health inequities	Class discussions, readings
Planning & Management to Promote Health	NA	
Policy in Public Health	NA	
Leadership	Apply negotiation and mediation skills to create a vision, empower others, and foster collaboration	The final project as a team
Communication	Effective scientific communication both in writing and through oral presentation	Class discussions; final course project (oral presentation and written report)
Interprofessional Practice	Perform effectively on interprofessional teams	Final project with a team of 2-3 members from different majors.