Columbia University
MAILMAN SCHOOL
OF PUBLIC HEALTH

Topics in Statistical Learning & Data Mining
P9120

CLASS SESSIONS

Thursdays, 1:00 – 3:50 PM;
Classroom: Hammer 305

INSTRUCTOR

Min Qian, PhD
212-305-6448, mq2158@cumc.columbia.edu
722 W. 168th St, Room 645, New York, NY, 10032
Office Hours: by appointment

TEACHING ASSISTANT

Yuqi Miao, ym2771@cumc.columbia.edu
Bin Yang, by2303@cumc.columbia.edu
Office hours:
11:30pm-12:30pm on Tuesdays: https://columbiacuimc.zoom.us/j/2248966521
8-9am on Fridays: https://columbiacuimc.zoom.us/j/4698005235

COURSE DESCRIPTION

The aim of this course is to provide students a systematic training in key topics in modern supervised statistical learning and data mining. For the most part, the focus will remain on a theoretically sound understanding of the methods (learning algorithms) and their applications in complex data analysis, rather than proving technical theorems. Applications of the statistical learning and data mining tools in biomedical and health sciences will be highlighted.

PREREQUISITES

This course is generally intended for Biostatistics PhD students in their second year or higher. Methodologically inclined DrPH or MS students of Biostatistics, or graduate students from other departments may be allowed to take the course at Instructors' permission. Students are expected to **have a strong background in calculus and linear algebra, linear and logistic regression models, probability theory and statistical inference**. Some working knowledge of **optimization techniques** and familiarity with **R and Python programming** would be desirable.

REFERENCES

[ESL] The Elements of Statistical Learning: Data Mining, Inference, and Prediction (2nd Edition)
ISBN: 9780387848570
Publisher: Springer
Authors: Trevor Hastie, Robert Tibshirani, Jerome Friedman
Publication date: 2009

[ISL] An introduction to statistical learning with applications in R
ISBN: 978-1-461-47137-0
Publisher: Springer

Authors: Gareth James, Daniela Witten, Trevor Hastie, Robert Tibshirani
Publication date: 2013

[DL] Deep Learning
ISBN: 9780262035613
Publisher: MIT press
Authors: Ian Goodfellow, Yoshua Bengio and Aaron Courville
Publication date: 2016

Deep Learning Specialization - course slides from DeepLearning.AI
https://www.deeplearning.ai/resources/

[RL] Reinforcement learning: An introduction (2nd Edition)
ISBN: 9780262039246
Publisher: MIT press
Authors: Richard S. Sutton, Andrew G. Barto
Publication date: 2018

COURSE LEARNING OBJECTIVES

Students who successfully complete this course will be able to:

- Describe and identify Supervised learning, Unsupervised learning and Reinforcement learning problems.
- Conduct regression analysis using modern model selection, shrinkage and regularization techniques.
- Understand the rationale of modern classification methods developed in computer science, and analyze classification problems using those tools.
- Apply unsupervised learning (e.g. cluster analysis, dimension reduction, etc.) techniques to appropriate statistical problems.
- Understand the connection between reinforcement learning and medical decision-making.
- Choose an appropriate method and apply it to solve any specific biomedical or public health research problem.
- Implement the above tools using standard statistical software to perform data analysis.

ASSESSMENT AND GRADING POLICY

Student grades will be based on:
Homework……..……...40%
Class Attendance and Participation……10%
Weekly mini quizzes……..…… 15%
Group Paper Presentation…..……..…….…… 10%
Final Project……...…… 25%

Homework assignments (there will be 4 of them in total) may include analyses of real data sets, running simulation studies and/or derivations of theoretical properties. While the analytical/theoretical part can be handwritten, the data analysis/simulation part must be typed in using LaTeX or MS Word. Data analysis/simulation results should be incorporated in the main text, while the source code should be included in the Appendix. Homework assignments should be submitted **online** on due dates. **Late homework will not be accepted.**

Each student will present a research paper as a group relevant to the course once in the semester. **Students will be randomly grouped and assigned to the topic at the 4rd session (September 26th).**

The individual final project consists of an **in-depth exploration of a machine learning method**. The topic should be chosen by the individual student in consultation with the instructor. The project report should include a through literature review of the method, technical exercise of relevance or simulation study or real data analysis (comparison with other methods) and conclude with your results and discussion. **A short (up to one page) proposal is due on December 2nd. The final project report (up to five pages) is due on December 22nd.**

COURSE STRUCTURE

The class sessions will be lectures, and each lecture will have a break in the middle. Some lectures will include a component of student presentation. This syllabus is designed to provide an overview of the course structure and organization. Please note that the new Courseworks website will have the most uptodate information on lecture slides, readings and assignments. Students are responsible for checking this site before each class.

MAILMAN SCHOOL POLICIES AND EXPECTATIONS

Students and faculty have a shared commitment to the School's mission, values and oath. http://mailman.columbia.edu/about-us/school-mission/

Academic Integrity

Students are required to adhere to the Mailman School Honor Code, available online at http://mailman.columbia.edu/honorcode.

Disability Access

In order to receive disability-related academic accommodations, students must first be registered with the Office of Disability Services (ODS). Students who have or think they may have a disability are invited to contact ODS for a confidential discussion at 212.854.2388 (V) 212.854.2378 (TTY), or by email at disability@columbia.edu. If you have already registered with ODS, please speak to your instructor to ensure that they have been notified of your recommended accommodations by Meredith Ryer (mr4075@cumc.columbia.edu), Assistant Director of Student Support and Mailman's liaison to the Office of Disability Services.

TENTATIVE COURSE SCHEDULE (as of September 23rd, 2024)

Please see the lecture section of Courseworks to download the lecture slides.

Session 1 – Overview to Statistical Machine Learning	
9/5/24	Learning Objectives: Describe and identify Supervised learning, Unsupervised learning and Reinforcement learning problems. Before class: Read syllabus, Chapters 1 & 2 of [ESL] and lecture notes

	During class: Live lecture

Session 2 – Linear Regression Methods	
9/12/24	Learning Objectives: Explain the connection of Least squares, Ridge Regression, Principal Components Regression; Understand Bias-Variance tradeoff; Apply principles of subset selection, AIC and BIC to appropriate statistical problems. Before class: Read Chapters 3 and 7 of [ESL] and lecture notes. During class: Live lecture After class: mini quiz (due at 9pm EST on the following Monday)

Session 3 – Variable selection and resampling methods	
9/19/24	Learning Objectives: Investigate Lasso and other simultaneous model selection and parameter estimation methods; conduct model selection and statistical inference using resampling methods such as Cross-validation and Bootstrap. Before class: Read Chapters 3 & 7 of [ESL] and lecture notes. During class: Live lecture After class: mini quiz (due at 9pm EST on the following Monday) hw1 (due on 10/5/24)

Session 4 – Classification Methods	
9/26/24	Learning Objectives: Understand the connection of Linear Discriminant Analysis, Logistic Regression, and Support Vector Machines Before class: Read Chapters 3 and 12 of [ESL] and lecture notes During class: Live lecture, random assignment of students to group paper presentation topics After class: mini quiz (due at 9pm EST on the following Monday)

Session 5 – Tree-based classification methods	
10/3/24	Learning Objectives: Understand and implement tree-based methods to perform classification and regression tasks Before class: Read Chapters 9, 10 and 15 of [ESL] and lecture notes During class: Live lecture After class: mini quiz (due at 9pm EST on the following Monday)

Session 6 – Deep Learning I

10/10/24	Learning Objectives: Understand the principles of deep learning and neural network models Before class: Read Chapter 11 of [ESL], Chapters 6, 7 of [DL] and lecture notes During class: Live lecture, Lab session for Python After class: mini quiz (due at 9pm EST on the following Monday) HW2 (due on 10/26/24)

Session 7 – Deep Learning II	
10/17/24	Learning Objectives: Understand and apply convolutional neural networks to model image data Before class: Read Section 10.3 of [ISL], Chapter 9 of [DL] and lecture notes During class: Live lecture, Lab session for Python After class: mini quiz (due at 9pm EST on the following Monday)

Session 8 – Deep Learning III	
10/24/24	Learning Objectives: Understand and apply recurrent neural networks to model sequence data Before class: Read Sections 10.4, 10.5 of [ISL], Chapter 10 of [DL] and lecture notes During class: Live lecture, Lab session for Python After class: mini quiz (due at 9pm EST on the following Monday)

Session 9 – Reinforcement Learning I	
10/31/24	Learning Objectives: Discuss Bandits and Contextual Bandits problems Before class: Read Chapters 1-2 of [RL] and lecture notes. Submit HW3. During class: Live lecture, Lab session for Python After class: mini quiz (due at 9pm EST on the following Monday HW3 (due on 11/16/24)

Session 10 –Reinforcement Learning II	
11/7/24	Learning Objectives: Discuss reinforcement learning framework and methods in infinite horizon setting. Before class: Submit final project proposal; Read Chapters 3, 4, 6 of [RL] and lecture notes

	During class: Live lecture, group paper presentation After class: mini quiz (due at 9pm EST on the following Monday)

Session 11 – Reinforcement Learning III	
11/14/24	Learning Objectives: Discuss Reinforcement Learning framework, methods and the applications in health applications Before class: Read lecture notes During class: Live lecture, group paper presentation After class: mini quiz (due at 9pm EST on the following Monday) Final project proposal (up to 1 page, due on 12/02/24)

Session 12 – Unsupervised Learning I	
11/21/24	Learning Objectives: Discuss multiple testing Before class: Read Section 18.7 of [ESL] and lecture notes. During class: Live lecture, group paper presentation After class: mini quiz (due at 9pm EST on the following Monday) HW4 (due on 12/9/24)

Session 13 – Unsupervised Learning II	
12/5/24	Learning Objectives: Discuss collaborative filtering and content-based filtering for recommender systems Before class: Read lecture notes. During class: Live lecture, group paper presentation After class: mini quiz (due at 9pm EST on the following Monday)

Final Project	
12/22/24	Submit final project report