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

Yinjun Zhao, yz3503@cumc.columbia.edu
Regular Office Hours: TBD
Extra Office Hours in homework due weeks: TBD
Zoom link: TBD

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 programming** would be desirable.

TEXTBOOKS

The Elements of Statistical Learning: Data Mining, Inference, and Prediction (2nd Edition)
ISBN: 978-0-387-84857-0
Publisher: Springer
Authors: Trevor Hastie, Robert Tibshirani, Jerome Friedman
Publication date: 2009

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

Reinforcement learning: An introduction (2nd Edition)
ISBN: 978-0-262-03924-6
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 Participation……………………..………………	20%
(12% for mini-quiz from sessions 2 – 12, and 8% for attendance and participation)	
Group Book Chapter Presentation…..……..……..……	10%
(8% for your presentation, 2% for group performance)	
Final Project Presentation……………………………..	10%
(8% for your presentation, 2% for your feedback to your peers' presentations)	
Final Project Report…………………………………...	20%

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** at the beginning of the class on due dates. **Late homework will not be accepted.**

Each student will have to present a book chapter relevant to the course once in the semester. **Students will be randomly grouped and assigned to the topic at the 4rd session (September 28th).**

The individual final project consists of an in-depth exploration of a machine learning/statistical 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 November 16th.** Each student is required to **do a 5- to 10-minute presentation describing your final project, record it, and submit the video presentation by December 17th**. Your presentation will be evaluated by anonymous peers. **The final project report (up to five pages) is due on December 24th.**

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

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

Session 1 – Introduction to Statistical Machine Learning	
9/7/23	Learning Objectives: Describe and identify Supervised learning, Unsupervised learning and Reinforcement learning problems. Before class: Read syllabus, Chapters 1 & 2 of Textbook ESL, and lecture notes During class: Live lecture, Q&A time

Session 2 – Methods in Linear Regression I

9/14/23	<u>Learning Objectives</u>: Apply Least squares, Ridge Regression, Principal Components Regression and Partial least squares to linear regression problems <u>Before class</u>: Read Chapters 3 of Textbook ESL and lecture notes. <u>During class</u>: Live lecture, Q&A time <u>After class</u>: mini quiz (due at 9pm EST on the following Monday)

Session 3 – Methods in Linear Regression II	
9/21/23	<u>Learning Objectives</u>: Understand Bias-Variance tradeoff; Apply principles of subset selection, AIC and BIC to appropriate statistical problems; conduct model selection using resampling methods (e.g. Cross-validation; Bootstrap methods) <u>Before class</u>: Read Chapters 3 & 7 of Textbook ESL and lecture notes <u>During class</u>: Live lecture, Q&A time, random assignment of students to group book chapter presentation topics <u>After class</u>: mini quiz (due at 9pm EST on the following Monday); hw1 (due on 10/5/23)

Session 4 – Methods in Linear Regression III	
9/28/23	<u>Learning Objectives</u>: Investigate Lasso and other simultaneous model selection and parameter estimation methods <u>Before class</u>: Read Chapter 3 of Textbook ESL and lecture notes <u>During class</u>: Live lecture, Q&A time <u>After class</u>: mini quiz (due at 9pm EST on the following Monday)

Session 5 – Basis Expansions and Smoothing Methods	
10/5/23	<u>Learning Objectives</u>: Conduct regression using Smoothing and Regression Splines, Kernel methods <u>Before class</u>: Read Chapters 5 & 6 of Textbook ESL and lecture notes; submit HW1. <u>During class</u>: Live lecture, Q&A time <u>After class</u>: mini quiz (due at 9pm EST on the following Monday); HW2 (due on 10/26/23)

Session 6 – Linear Methods for Classification

10/12/23	Learning Objectives: Discuss linear methods for classification, e.g. Linear Discriminant Analysis and Logistic Regression. Before class: Read Chapter 4 of Textbook ESL and lecture notes During class: Live lecture, Q&A time After class: mini quiz (due at 9pm EST on the following Monday)

Session 7 – Support Vector Machines	
10/19/23	Learning Objectives: Discuss Maximum Margin Classifiers and Kernel Trick Before class: Read Chapter 12 of Textbook ESL and lecture notes; submit HW2. During class: Live lecture, Q&A time After class: mini quiz (due at 9pm EST on the following Monday)

Session 8 – Trees, Bagging, Boosting	
10/26/23	Learning Objectives: Discuss tree-based methods to perform classification and regression Before class: Read Chapters 9 & 10 of Textbook ESL and lecture notes During class: Live lecture, Q&A time After class: mini quiz (due at 9pm EST on the following Monday); HW3 (due on 11/16/23)

Session 9 – Neural Networks and Deep Learning	
11/2/23	Learning Objectives: Discuss Neural Networks and Deep Learning Before class: Read lecture notes During class: Live lecture, Q&A time After class: mini quiz (due at 9pm EST on the following Monday); final Project proposal (up to 1 page, due on 11/16/23)

Session 10 – Unsupervised Learning	
11/9/23	Learning Objectives: Discuss unsupervised learning problems: association rules, density estimation, Clustering analysis, HMM, etc. Before class: Read Chapters 6 and 14 of Textbook ESL and lecture notes; Submit HW3. During class: Live lecture, Q&A time

	After class: mini quiz (due at 9pm EST on the following Monday.

Session 11 – High Dimensional Problems/Reinforcement Learning	
11/16/23	Learning Objectives: Understand difficulties in p >> N problems; apply Multiple Testing and FDR; Bandits and Contextual Bandits Before class: Submit final project proposal; Read Chapter 18 of Textbook ESL and lecture notes During class: Live lecture, Q&A time After class: mini quiz (due at 9pm EST on the following Monday); HW4 (due on 12/7/23)

Session 12 – Reinforcement Learning	
11/30/23	Learning Objectives: Discuss Reinforcement Learning and its applications Before class: Read lecture notes. During class: Live lecture, Q&A time, discussion of final project presentation After class: mini quiz (due at 9pm EST on the following Monday)

Final Project	
12/17/23	Submit final project presentation video
12/24/23	Submit final project report and anonymous peer review of final project presentation