Course Information

GU4241/GR5241 Statistical Machine Learning
417 International Affairs Building
Friday 2:40pm - 5:25pm

Instructor

Name: Linxi Liu
Email: ll3098 [at] columbia [dot] edu
Office: SSW 901C
Office hours: Thursdays, 4:30pm - 6:00pm

Overview

This class provides an introduction to Machine Learning and its core algorithms, with an emphasis on the statistical properties of the learning methods. Topics include linear methods for regression and classification, model selection and regularization, kernels, decision trees, boosting, bagging, random forests, suport vector machines, principle component analysis, clustering, and Bayesian models. The class also gives a brief introduction to convex optimization and Markov chain Monte Carlo. Advanced topics may include neural networks, spectral clustering and non-linear dimension reduction methods.

All resources from class will be posted on Canvas https://courseworks.columbia.edu/welcome/. You need your UNI and password to log into the system.

Course Organization

This semester, there are two types of lectures. The regular lecture is from 2:40pm to 5:25pm on Friday, and we have separate labs/review sessions during the week. Each student must have registered for one section. You are REQUIERED to attend ONE lab/review session every week.

Every week I will upload the course materials on Canvas, which may include lecture notes, tutorials, suggested reading, and assignments. Students are expected to check emails at least once every 12 hours during the week and every 24 hours over the weekends.

Below is a tentative schedule.

- Week 1 (1/25): Course logistic and introduction, frequency vs. Bayes, linear algebra
  - Assignment 1 release (due 1/31)

- Week 2 (2/1): Principle component analysis, prototype methods and nearest-neighbors
  - Tutorial: Basic data processing using R and Python
  - Assignment 2 release (due 2/14)
- Week 3 (2/8): Linear methods for regression and classification
  - Tutorial: Applying PCA and k-means clustering to Zipcode data
- Week 4 (2/15): Support vector machines I
  - Tutorial: Linear methods for regression and classification
  - Assignment 2 due. Assignment 3 release (due 2/28)
- Week 5 (2/22): Introduction to convex optimization
  - Tutorial: SVM
- Week 6 (3/1): Support vector machines II, cross-validation and Bootstrap
  - Tutorial: Stochastic gradient descent
  - Assignment 3 due. Assignment 4 release (due 3/12)
- Week 7 (3/8): Model selection and regularization
  - Tutorial: *cross-validation and Bootstrap*
- Week 8 (3/15): **Midterm exam**, *basis expansion and kernels (tentative).*
  - Tutorial: glmnet
  - Assignment 4 due.
- Week 9 (3/29): Decision trees, Bagging and Random Forest
  - Tutorial: Data visualization or review
  - Assignment 5 release (due 4/11)
- Week 10 (4/5): Boosting, neural networks
  - Tutorial: Tree-based methods
- Week 11 (4/12): Neural networks (continued), clustering
  - Tutorial: *Boosting*
  - Assignment 5 due. Assignment 6 release (due 4/30)
- Week 12 (4/19): Clustering (continued), model order selection
  - Tutorial: neural networks
- Week 13 (4/26): Text models, information theory
  - Tutorial: *Clustering Analysis*
- Week 14 (5/3): Missing Data, review

– Tutorial: Tweets classification

– Assignment 6 due.

Textbook

T. Hastie, R. Tibshirani and J. Friedman. *The Elements of Statistical Learning.* Second Edition, Springer, 2009

References

G. James, D. Witten, T. Hastie and R. Tibshirani. *An Introduction to Statistical Learning with Applications in R.* Springer, 2013

K. P. Murphy. *Machine Learning: a Probabilistic Perspective.* MIT Press, 2012.

J. Shawe-Taylor and N. Cristianini. *An Introduction to Support Vector Machines and Other Kernel-Based Learning Methods.* Cambridge University Press, 2000.

D. Barber. *Bayesian Reasoning and Machine Learning.* Cambridge University Press, 2012.

Teaching Assistants

We have a course mailing list: gr5241_gu4241_course_staff [at]columbia [dot] edu

For any course-related inquiries, please send them to the mailing list. Please DO NOT email the instructor or the TAs in person. Any email directly sent to the instructor or the TAs WILL NOT get replied.

Name: Gabriel Loaiza

Name: Adi Bousso Dieng

Name: Shuaiwen Wang

Name: Timothy Jones

Name: Peter Lee

Office Hours

TBA

Graders

TBA

Grading

Your overall course grade will be determined as a weighted average of the following categories:

10%	Attendance	
20%	Homework assignments	*The lowest score will be dropped.*
30%	Midterm exam	**Fri, Mar. 15, 2019, 2:40pm - 4:00pm,** *in lecture.*
40%	Final exam	**Fri, May 10, 2019, 1:10pm - 4:00pm,** *Location TBA.*

Programming Language

We will mainly use R and Python for this course. R is a free software which can be downloaded at https://www.r-project.org. And Python can be downloaded at https://www.python.org

Exams

In general, **NO MAKE-UP EXAMES** are granted. Make-up exams will be given only in rare cases of emergency. If an emergency occurs on the exam day, you must contact the instructor *before* the exam (or arrange for someone else to do so). We will not approve any exam rescheduleing requests based on personal reasons such as travel, leisure, or to ease exam week schedules. We will not approve any exam rescheduleing requests for students who take another class whose lectures or final exam occur at the same time as those of our class. No make-up exams will be granted to a student who contats us after the exam is over. No special accommodations will be made for students who arrive late to exams, regardless of the reason (missing a bus; overslept; sick; etc.). If you need to miss an exam due to a sudden severe illness, injury, traumatic event, etc., after consulation with the instructor it is possible that you will be given an **Incomplete** in the course and asked to complete the course in a future semester.

Homeworks

There are six assignments in total. According to the tentative schedule, most of them will be due on Thursday. Please submit your homework electronically on Gradescope (Entry code: 97EDWR). Note that we **DO NOT** accept late homework. However, the lowest score will **NOT** be counted in your final grade.

See next page for a suggested reading list.

Suggested Reading

Week	Lecture	ELS	ISL
1	1: Course logistic and introduction 2: Frequency vs. Bayes, linear algebra	Chapter 2 Sections 2.4, 8.3	Chapter 2
2	3: Principle component analysis 4: Prototype methods and nearest-neighbors	Section 14.5 Sections 2.3, 2.5, 13.3, 14.3.6	Section 10.2 Sections 2.2.3, 10.3.1
3	5: Linear methods for regression 6: Linear methods for classification	Sections 3.2, 3.3 Chapter 4	Chapter 3 Chapter 4
4	7: Support vector machines I	Section 12.2	Sections 9.1, 9.2, 9.5
5	8: Introduction to convex optimization		
6	9: Support vector machines II 10: Cross-validation and Bootstrap	Section 12.3 Sections 7.10, 7.11	Sections 9.3, 9.4 Chapter 5
7	11: Model selection and regularization I 12: Model selection and regularization II	Sections 7.4—7.7, 3.4 Section 3.4	Chapter 6 Chapter 6
8	13: Basis expansion and kernels 14: Midterm in class (Lecture 1-13)	Sections 5.2, 5.3, 5.4, 5.8	Chapter 7
9	15: Decision trees 16: Bagging and Random Forest	Section 9.2 Sections 9.2, 8.7, 15.2, 15.3	Section 8.1 Section 8.2
10-11	17: Boosting 18-19: Neural networks	Sections 10.1—10.7 Chapter 11	Section 8.2
11-12	19-21: Clustering 22: Model order selection	Sections 14.3, 8.5	Section 10.3
13	23: Text models 24: Information Theory		
14	25: Missing Data 26: Review		