PS 919: Syllabus
Introduction to Machine Learning
Spring 2020 *

Adeline Lo

1 Course details

Class meets Wednesdays 3:30pm-5:25pm, Ogg Room 422, North Hall
Professor: Adeline Lo
Office: 322C, North Hall
Contact: aylo@wisc.edu
Office hours: Wednesdays 2:25pm-3:25pm, 5:30pm-6:30pm

This course serves as a graduate-level introduction to machine/statistical learning. It will cover some, but not all, common techniques to collect, analyze and utilize large and unstructured data for social science questions. The general goal of this course is to introduce students to modern machine learning techniques and provide the skills necessary to apply the methods widely. Note, this course does not utilize `Python` or `Julia`. All computation work is conducted in `R`.

2 Overview

The following are the broad and specific learning outcomes to the course:

- Introduction to basic topics in statistical and machine learning. By the end of the semester, students should:
  - know the mathematical theory behind each machine/statistical learning approach,
  - be able to read extensions off of original ideas introduced in the course,
  - identify when is a responsible use of each covered approach,

*Last edited: January 18, 2020

  - and explain the findings and results of an application to a colleague.

- Students should also have expanded working knowledge of R programming. Specifically students should have:
  - less reliance on canned approaches and transitioned to either building from scratch or active recoding of existing code;
  - increased computational efficiency through code vectorization and parallelizing;
  - Handled different kinds of data, both structured and unstructured including: case-control, text, image, network, and high dimensional

2.1 Prerequisites

Formally, students should have taken PS 812, 813 and 818 and be comfortable working in the `R` language. Instructor approval may override course requirements on a case-by-case basis. Auditors should note that assessments will not be provided to them.

Informally, and even more importantly, the most essential prerequisite is a willingness to work hard on possibly unfamiliar material. Learning statistics and programming can be like learning new languages, which require time and dedication. Similar to studying languages, fluency and comfort come from daily practice and consistent effort.

Students should arrive to the first day of class having completed the following:

- Successfully downloaded and opened `R` and `RStudio`
- Installed `knitr` in `RStudio`
- Downloaded and opened `LaTex`
- Successfully knitted the example .rmd file in `RStudio` (simply click and open up a new .rmd file and it will populate a template) in both pdf and html form
- Obtained a copy of the main suggested text for the course *Elements of Statistical Learning*, available online, and perused Chapter 2

2.2 Class and Lab

Instruction for this course is conducted via two avenues: class and lab. Class lectures compose the first hour and ten minutes (3:30-4:40pm) and will typically focus on statistical material. Lab follows directly after for the remainder of the course time (4:40pm-5:25pm) and will typically focus on practical problem solving and/or computational skills. Both are essential to the learning process.

2.3 Course website

The course heavily relies on communication of information, materials and questions through `Canvas`. Assignments will be distributed and submitted through it, and students are expected to submit questions (and answer their colleagues' questions!) on the forum.

3 Materials

3.1 Computational tools

The best way, and often the only way, to learn about data analysis and new statistical procedures is by doing. We will therefore make extensive use of a flexible (open-source and free) statistical software program called `R`, `RStudio`, and a number of companion packages. Problem sets and the takehome quiz will be completed in R Markdown. The class assumes a proficient level of `R` progamming.

3.2 Readings

There is no required textbook for the class, as the main text will be lecture slides, but the following title is exceptionally useful and pertinent to the materials we will cover:

ESL: Hastie, Trevor, Robert Tibshirani and Jerome Friedman. *The Elements of Statistical Learning.* A free version is available online!

4 Assignments

There are five types of assignments in this course:

1. **Problem sets**: learning statistics and programming takes consistent practice. The course has four problem sets, which are described below.

2. **Lab exercises**: we will have weekly lab exercises which we will begin together in the course on Wednesday and are due (in finished knitted and .rmd form on Canvas) on the next day (Thursday) by 9am. These constitute a part of each lecture's participation.

3. **Quiz**: there will be a take-home quiz in the first few weeks of the semester

4. **Midterm**: there will be an in-class midterm in the middle of the semester

5. **Final group project**: you will work in small groups to complete a final in-class presentation at the close of the semester. More details will be provided in the weeks to come, but you should begin searching for interesting topics/data that you would

like to analyze in groups of **3 or 4** and schedule meetings with me to discuss feasibility of your proposed topics in the first couple weeks of class.

6. **Participation**: as this is a seminar style class, you will be expected to come prepared to contribute in questions and in serving as a public goods providing citizen in the course. You can do this via asking or answering questions regarding the course material in person and online (Canvas), as well as proactively helping one another during lab sessions, submissions of lab exercises and support of one another outside of class on collaborated activities. Simply sitting in Ogg listening quietly for class/lab sessions will not suffice. And, if you are benefiting from a colleague's public goods support and goodwill, please be sure to acknowledge it as well! Actively assisting your classmates in class, lab or Canvas will constitute 10% of your final grade. I take this aspect of your performance very seriously, not only in my assessment of your final grade but also in any letters/recommendations for which you might ask of me.

4.1 Problem sets

Learning by watching and not doing is hard when learning how to use any new tool, so in order to do more and practice with regularity you will have **four** problem sets. The assignments will be a mix of analytic problems, computer simulations and data analysis.

Assignments must be completed in R Markdown, which allows you to show both your answers and the code you used to arrive at them. You will need to submit both the .rmd file and the knitted html/pdf.

Problem sets will be made available on Canvas starting Wednesday immediately after class and are due Monday the following week by 9am Central. Solutions will be made available two days after, through Canvas. Working through the problem sets include looking at the solutions key so please remember to do this portion!

Problem sets are graded out of 0-5 points. I reserve the right to add bonus points for aesthetics including presentable graphs, clear code, nice formatting and well-written answers.

There will be 4 problem sets in total, constituting 40% of your grade. Late problem sets drop a grade by 1 point (out of the total 5) each late day, with a maximum of 2 late days, after which I will not accept the problem set anymore. I do not want to hold up the class and will not wait for everyone to submit their problem sets in order to post the solutions key.

Code Conventions: Throughout the course, students will receive feedback on their code from myself and other students. Therefore, consistent code conventions are critical. Good coding style is an important way to increase the readability of your code (even by a future you!). I strongly recommend you follow the code conventions developed by Hadley Wickham and implemented in the package `lintr`, which is built into `RStudio`.

Collaboration Policy: Unless otherwise stated, I encourage students to work together on the assignments, but you should write your own solutions (this includes code). That is, no copy-and-paste from other people's code. You would not copy-and-paste from someone's paper, and you should treat code the same way. However, I strongly suggest that you make a solo effort at all the problems before consulting others. Finally, you should credit your colleagues' help when appropriate!

4.2 Quiz

There will be a takehome quiz in the first few weeks of the course. It, and the problem sets due before it, are styled to help you assess your progress in the course as early as possible; if you are having trouble with the first two weeks of material and/or do poorly on the quiz, I would recommend you come speak to me immediately in office hours and consider deferring taking the course until after you've spent more time on statistics and/or programming in R.

The quiz will be posted immediately after class and then due two days after by midnight. It is "open-book" in the sense that you can use the slides, your notes, books, and internet resources to answer the questions. However, the quiz must be completed *by yourself.* It will be the approximate length of a problem set (although I caution that it might take longer if you are used to collaborating on the problem sets). I encourage you to start early.

4.3 Midterm

There will be an in-class miderm in week 8 of the class. It will be a traditional closed-book exam focusing on the theoretical aspects of the material you have covered up to that point.

4.4 Final group project

You will work in groups of 3 or 4 to ask a social science question of a dataset that will culminate in a final presentation which will be presented at the end of the semester. I will provide more details on the project later in the semester.

4.5 Grading

- Problem sets: 4 total, 40%

- Quiz: 5%
- Midterm: 15%
- Final group project: 30%
- Participation: 10%

5 How to Learn in this Course

If you find this course challenging, you are not alone. Statistics and programming can be (differently!) challenging and we cover a lot of ground. I have confidence in your abilities as smart and engaged researchers who can handle it. Below are some details on forms of support that I offer in this class.

Your primary responsibilities in this class are to *work hard* and *communicate* with me about what you need. You cannot learn if you aren't putting in the time. I also can't help if I don't know there's a problem. Students who require alternative accommodations should schedule a meeting with me within the first two weeks of class to appropriately address their needs. (see "Accommodations for Students with Disabilities" section at end)

5.1 Resources for Getting Help

Below are a few main sources of support for this class.

1. Class and Lab

   I encourage you to be an active participant in class and section. Ask questions if you don't understand something that is happening. Ask questions to *better* understand something that is happening.

2. Readings and Slides

   If you are studying alone and hit something you don't understand, you should turn to the slides first. There will be a fair amount of material in the slides and they are intended to be used and reviewed multiple times, not just seen once during lecture.

3. Canvas

   We will be using the Canvas discussion board for communication in this class, but also as a source to post and answer questions about the material. You will not be required to post, but the system is designed to get you help quickly and efficiently from classmates and I. **Unless the question is of a personal nature or completely specific to you, you should not resort to emailing me directly**; instead, you should post your questions on Canvas. I will be monitoring the discussion board, but I encourage you to help your classmates as well. Likely a significant amount of overlap will exist for both things people want to know more about and things people have just figured out. This is particularly true given the heavy emphasis on programming in this class.

4. Instructor office hours

   My office hours are in my office 322C North Hall immediately before and after each class session.

5. Problem Set/Lab Keys

   Soon after the problem sets/labs are due, I will post the key. It may be tempting to immediately turn focus towards the next problem set/lab, but if you were uncertain about anything in the material, I recommend you check the key to lock down core concepts. Some of the material builds directly on previous concepts!

6 Class Schedule

Note, this WILL change as we roll through the semester, though no exam/final project presentation dates will change. Please check Canvas regularly for updates.

Week 1: Introduction & Classification

January 22, 2020 Suggested readings prior to class: ESL Ch 2

- Lecture: Introduction to the course and classification, intro to k-means/least squares
- Lab: k-means

Week 2: Regression

January 29, 2020 Suggested readings prior to class: ESL Ch 3.2, 4.4

- Lecture: linear regression and classification, optimization methods
- Lab: Prediction using linear methods
- Problem Set 1 given after class; due following Monday by 9am.

Week 3: Shrinkage methods

February 5, 2020 Suggested readings prior to class: ESL Ch 3.4
Fun/advanced reading: Bien, Jacob, Jonathan Taylor, and Robert Tibshirani. (2013) “A lasso for hierarchical interactions”. *The Annals of Statistics*.

- Lecture: shrinkage approaches
- Lab: application of lasso
- Take-home Quiz given after class; due end of Friday before midnight.

Week 4: Model assessment & selection

February 12, 2020 Suggested readings prior to class: ESL Ch 7.2-5, 7.10-7.11

- Lecture: bias/variance, model complexity, training and testing error, cross validation
- Lab: cross validation

Week 5: Additive models and trees

February 19, 2020 Suggested readings prior to class: ESL Ch 9.2, 15.1-15.4
Fun/advanced reading: Athey, Susan and Guido Imbens. (2016) “Recursive partitioning for heterogeneous causal effects”. *PNAS*.

- Lecture: regression trees/classification trees, random forests
- Lab: random forests

Week 6: Neural networks, Model inference/Averaging

February 26, 2020 Suggested readings prior to class: ESL Ch 8.2-8.7, ESL Ch 11.3-5
Fun/advanced reading: Dempster, Art, Nan Laird and Don Rubin. (1977) “Maximum Likelihood from Incomplete Data via the EM Algorithm”. JRSS.

- Lecture: neural networks, fitting, issues in neural networks, EM/MCMC
- Lab: neural networks
- Problem Set 2 given end of class, due following Monday by 9am.

Week 7: Ensembles and boosting, High Dimensional problems

March 4, 2020 Suggested readings prior to class: ESL Ch 16.1-16.3, 18.1

Fun/advanced reading: Lo, Adeline, Herman Chernoff, Tian Zheng and Shaw-hwa Lo. (2016). "Framework for making better predictions directly". *PNAS*.

- Lecture: Ensembles and boosting, issues in HD data
- Lab: model inference techniques; high dimensional data

Week 8: In-class Midterm

March 11, 2020

UW Madison Spring recess

Week 9: Clustering, PCA

March 25, 2020 Suggested readings prior to class: ESL Ch 14.3, 14.5

- Lecture: Clustering, PCA
- Lab: PCA example
- Problem Set 3 given end of class; due following Monday by 9am.

Week 10: Text as data

April 1, 2020 Fun/Advanced reading: King, Gary, Patrick Lam, and Margaret Roberts. (2017) "Computer-Assisted Keyword and Document Set Discovery from Unstructured Text." *AJPS*.

- Lecture: preprocessing, dictionary
- Lab: text example

Week 11: Text modeling

April 8, 2020 Fun/advanced reading: Roberts, Margaret, Brandon Stewart and Edoardo M. Airoldi. (2016) "A Model of Text for Experimentation in the Social Sciences". *JASA*.

- Lecture: clustering, topic models
- Lab: topic model
- Problem Set 4 given end of class; due following Monday by 9am.

Week 12: Image analysis[1]

April 15, 2020

- Lecture: image analysis
- Lab: image data

Week 13: Social network analysis

April 22, 2020 Fun/advanced reading: Zhang, Yan, A.J. Friend, Amanda Traud, Mason Porter, James Fowler and Peter Mucha. (2008). "Community Structure in Congressional Cosponsorship Networks". *Physica A.*

- Lecture: descriptive statistics and key concepts in social network analysis
- Lab: network analysis example

Week 14: Wrap up and group presentations

April 29, 2020

- Group presentations

Acknowledgements

This course was developed on the shoulders of giants, in some cases borrowing directly from materials developed by amazing colleagues in political science, economics, and statistics. I am extremely grateful to everyone who has contributed directly, or indirectly. Lecture slides and related circulated materials should have appropriate citations – please send me an email if you believe they are incorrectly citing or lacking in citation rigour. Individuals include but are not limited to: Ines Levin, Peter Orbanz, Rochelle Terman, Michelle Torres, Tian Zheng.

All errors that remain are my own.

ACADEMIC INTEGRITY

By enrolling in this course, each student assumes the responsibilities of an active participant in UW-Madison's community of scholars in which everyone's academic work and behavior are held to the highest academic integrity standards. Academic misconduct compromises

[1] Shortened class: class will end 5pm

the integrity of the university. Cheating, fabrication, plagiarism, unauthorized collaboration, and helping others commit these acts are examples of academic misconduct, which can result in disciplinary action. This includes but is not limited to failure on the assignment/course, disciplinary probation, or suspension. Substantial or repeated cases of misconduct will be forwarded to the Office of Student Conduct & Community Standards for additional review. For more information, refer to `https://conduct.students.wisc.edu/academic-integrity/`.

ACCOMMODATIONS FOR STUDENTS WITH DISABILITIES

The University of Wisconsin-Madison supports the right of all enrolled students to a full and equal educational opportunity. The Americans with Disabilities Act (ADA), Wisconsin State Statute (36.12), and UW-Madison policy (Faculty Document 1071) require that students with disabilities be reasonably accommodated in instruction and campus life. Reasonable accommodations for students with disabilities is a shared faculty and student responsibility. Students are expected to inform faculty [me] of their need for instructional accommodations by the end of the third week of the semester, or as soon as possible after a disability has been incurred or recognized. Faculty [I], will work either directly with the student [you] or in coordination with the McBurney Center to identify and provide reasonable instructional accommodations. Disability information, including instructional accommodations as part of a student's educational record, is confidential and protected under FERPA. `http://mcburney.wisc.edu/facstaffother/faculty/syllabus.php`.

DIVERSITY & INCLUSION

Diversity is a source of strength, creativity, and innovation for UW-Madison. We value the contributions of each person and respect the profound ways their identity, culture, background, experience, status, abilities, and opinion enrich the university community. We commit ourselves to the pursuit of excellence in teaching, research, outreach, and diversity as inextricably linked goals. The University of Wisconsin-Madison fulfills its public mission by creating a welcoming and inclusive community for people from every background – people who as students, faculty, and staff serve Wisconsin and the world. `https://diversity.wisc.edu/`