Int. to Statistical Machine Learning – CSC/DSCC 265/465

Location: Wegmans Hall 1400

Lecture Time: Mondays, Wednesdays, 2:00 – 3:15 PM

Course Description

This course is designed to introduce you to modern machine learning concepts, techniques, and algorithms. Topics include regression, classification, unsupervised and supervised learning, kernels, support vector machines, feature selection, clustering, and network analysis. The course will help you further develop your programming skills, teach you the mathematical foundations of algorithms and ask you to code them. The class will both emphasize individual work and teamwork by helping students solidify their knowledge through a final project.

Instructor

Instructor: Cantay Caliskan, *PhD*

E-mail: cantay.caliskan@rochester.edu

Office Location: Wegmans Hall, ***1205***
Office Hours: Mon (4:30 – 6:30 PM), Wed (6 – 7 PM)

Teaching Assistants

Class TAs: Sai J. Meka (smeka@ur.rochester.edu) (Head TA)
Xinyu Cai (xcai12@u.rochester.edu)
Ryan Hilton (rhilton2@u.rochester.edu)
Yiheng Mao (ymao16@u.rochester.edu)
Duy Pham (dpham4@u.rochester.edu)
Amanda Pignataro (apignata@u.rochester.edu)
Vaarya Srivastava (vsrivast@u.rochester.edu)

Qihang Tang (qtang5@u.rochester.edu)

TA Office Locations: Please check *BlackBoard*.
TA Office Hours: Please check *BlackBoard*.
If you are not able to make the office hours, you are encouraged to contact me or your TAs. For office hours

Course Goals

- Understand the mathematical background of algorithms
- Code algorithms from scratch and comparing your code's performance to pre-packaged algorithms
- Apply your (hard-earned) code to empirical problems
- Develop an understanding of different statistical philosophies
- Handle missing data with both simple and sophisticated techniques
- Describe structures in data using unsupervised algorithms
- Reduce dimensionality of data for easier storage, computation, and analysis
- Estimate algorithm run times and efficiently handle large computations
- Understand both the 'breadth' and the 'depth' of the material
- Develop an understanding of what a good ML problem/puzzle might be

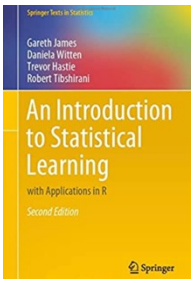
with your instructor, please reserve a 20-min slot on *Google Calendar*.

> *"The world is one big data problem."*
> ***Andrew McAfee***

Course

Course Website: https://learn.rochester.edu

Prerequisites: DSCC/CSC/STAT 262 or STAT 212 or STAT 213 or equivalent introductory statistics background. CSC161 or CSC171 or equivalent introductory programming background.



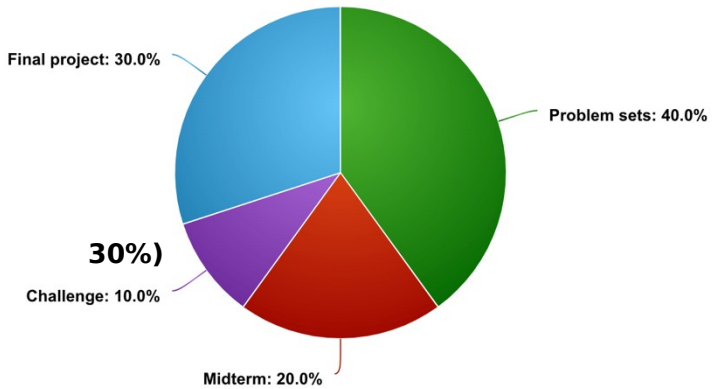
Textbook: We will be using a textbook:
An Introduction to Statistical Learning: with Applications in R 2nd edition (2021)
All class material will be shared on *BlackBoard*.

Required Software: The class will use the Anaconda distribution of *Python* for programming.

Download Anaconda for PC and Mac:
https://www.anaconda.com/distribution/#download-section

Background**:** This class will expect familiarity with statistics, scripting or coding in general. Some guidance will be provided on how to accomplish certain tasks using your programming skills.

Assignments and Grading



1. Problem sets (**40%**)
   i. Eight assignments (5% each)
2. Midterm (**20%**)
3. Challenge (**10%**)
4. Final Project (5%+10%+15% =
   i. Planning and preparation (5%)
   ii. Final presentation (10%)
   iii. Final report (15%)

Problem Sets (40%)

Problem sets will be an opportunity for you to think more deeply about what we have learned in the previous week. These exercises will both have theoretical and empirical components.

Problem sets allow up to 2 people to work together. Discussing the class material with fellow students is highly encouraged, although code sharing on problem sets is strictly not allowed. Problem set files will be evaluated for plagiarism using a code checker. Problem sets will involve theoretical and computational problems and writing paragraphs interpreting the results and putting them in context.

Midterm (20%)

Midterm will test your ability to answer relatively rigorous, and mostly theoretical questions in a limited amount of time. The exam will be cumulative – you will be responsible for the material that we have covered until midterm. If the class is in person, you will have an in-class midterm, if not we are planning to have a take-home exam.

Challenge (10%)

The challenge is a programming project that will give you a chance to apply your knowledge and develop an ML solution to a real-world problem. The goal is to digest what you have learned in the weekly problem sets to the next level and spend a few weeks implementing machine learning algorithms from scratch and applying them to a challenging dataset. More information will be shared later in the term. Challenge can be completed in a **team of two people.**

Final Project (30%)

While labs are more structured and focused more intensely on skills in the weeks prior, the final project will be more comprehensive and offer more flexibility in terms of your design, data choices, and approach while demonstrating the mastery of the breadth of skills discussed in the course. Potential topics for the project will be discussed with the instructor for suitability and suggestions. Topics can arise from any of the datasets we use in class, or from outside sources. Final reports will discuss the data, analytical approach, and visual presentation of the results.

Final project will be a **group project**.

Rules and Policies

Grading Policy: I will be using a curved grade policy for the class. When debating on which grade to assign for people between two possible grades, I will look at:

- **(i)** the general quality of the assignments provided by the person in question,
- **(ii)** the class and online participation,
- **(iii)** the general enthusiasm and the quality of interest that the student has (quality of questions in office hours, motivation to improve his/her performance).
- **(iv)**

Technology Policy: Please be respectful with your use of laptops and technology in class. I request you only use them for class related purposes, as I and others may find them distracting. Cell phones should be kept silent and away, and you can expect the same from me.

Software: All projects in this course will be analyzed using **Python**, an open source data analysis language

Academic Honesty: Academic honesty, the cornerstone of teaching and learning, lays the foundation for lifelong integrity. Academic dishonesty is intellectual theft. It includes, but is not limited to, providing or receiving assistance in a manner not authorized by the instructor in the creation of work to be submitted for evaluation. This standard applies to all work ranging from homework assignments to major exams. I will assume that you are familiar with the Code of Academic Integrity. To learn more about it, please go to: http://www.rochester.edu/college/honesty/.

Disability Accommodations: Any student who feels he or she may need an accommodation based on the impact of a disability should contact me privately as soon as possible to discuss his or her specific needs. I rely on the Office of Disability Resources in Taylor Hall to verify the need for reasonable accommodations based on the documentation on file in that office.

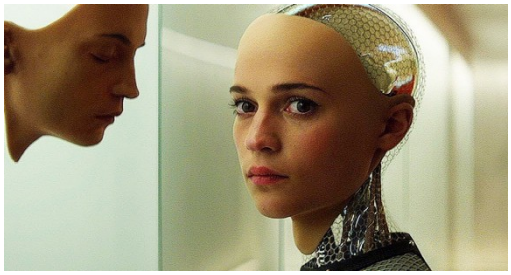
Senator Jayapal: "Does Amazon ever access and use policy seller data when making business decisions?"

Bezos: "I can't guarantee you that that has never been violated."

Course Schedule

Weeks	Topics	Readings	Assignments
Week 1 Jan. 12	**Introduction to Course and Review** Topics, policies, questions	Chapter 1	No assignment
Week 2 Jan. 19	**Math Review** Review of linear algebra and probability theory	Chapter 2	PS 1
Week 3 Jan. 24, 26	**Supervised Learning I: Regression** Regression, linear hypothesis, gradient descent	Chapter 3	PS 2
Week 4 Jan. 31, Feb. 2	**Supervised Learning II: Classification/Regularization** Finding patterns, bias/variance	Chapter 4	PS 3
Week 5 Feb. 7, 9	**Unsupervised Learning I: Clustering** Clustering, k-means, Gaussian mixtures	Chapter 12	PS 4

Week 6 Feb. 14, 16	**Unsupervised Learning II: Dimension Reduction** PCA, T-SNE	**No reading**	PS 5
Week 7 Feb. 21, 23	**Application I: Practical ML** Feature selection and feature pre-processing	Article on feature selection	PS 6
Week 8 Feb. 28, Mar. 2	**Probabilistic Models: LDA and Bayesian Methods** Gen. linear models, generative vs. discriminative	Revise Chapter 4	PS 7
Week 9 Mar. 7, 9	**SPRING BREAK**		No assignment
Week 10 Mar. 14, 16	**Support Vector Machines** Methods, kernels, loss functions	Chapter 9	**Midterm**
Week 11 Mar. 21, 23	**Application II: Missing Value Estimation** Frequentist and Bayesian approaches	Article on missing value imputation	PS 8
Week 12 Mar. 28, 30	**Application III: Ethics in Machine Learning** Bias, fairness, transparency, accountability	Article on bias and fairness in ML	Challenge
Week 13 Apr. 4	**Network Analysis I: Descriptive Concepts** Understanding the data – **Class is cancelled on Apr. 6!**	**No reading**	Challenge
Week 14 Apr. 11, 13	**Network Analysis II: Models** Clustering, tie prediction	**No reading**	Final Project Preparation
Week 15 Apr. 18, 20	**Review, Final Project (Presentations)**	**No reading**	**Final Project**
Week 16 Apr. 25, 27	**Final Project (Presentations)**	**No reading**	**Final Project**


How does the future you want to create look like?