MATH7243-Machine Learning and Statistical Learning Theory-Spring2023

Instructor: He Wang
Class time and room: see Canvas
Office Hours: see Canvas
Email: he.wang@northeastern.edu
TA: See Canvas.

Some recommended textbooks:

1. An Introduction to Statistical Learning, with applications in R, by G. James, D. Witten, T. Hastie, R. Tibshirani. https://www.statlearning.com/
2. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction* by Trevor Hastie, Robert Tibshirani, Jerome Friedman. https://web.stanford.edu/~hastie/ElemStatLearn/
3. *Hands-On Machine Learning with Scikit-Learn and TensorFlow: Concepts, Tools, and Techniques to Build Intelligent Systems* (2ed edition)- by Aurélien Géron https://github.com/ageron/handson-ml2
4. *Pattern Recognition and Machine Learning* by Christopher Bishop
5. *Linear algebra and learning from data* by Gilbert Strang
6. *Machine Learning: A Probabilistic Perspective* by Kevin P. Murphy
7. *Understanding Machine Learning: From Theory to Algorithms* by Shai Shalev-Shwartz, Shai Ben-David
8. Other online sources will be provided on Canvas.
9. My recommended book list (include linear algebra and probability) https://docs.google.com/spreadsheets/d/1j8LcyTxBTZ5Nh0-P1rb6NT2x_2iXFp1j8ptpvHYQWGA/edit?usp=sharing

Prerequisite: Basic knowledge (undergraduate level) about linear algebra, multivariable calculus, probability and statistics are required. Graduate level of applied linear algebra, probability1 is preferred.

Overview

Introduces both the mathematical theory of learning and the implementation of modern machine-learning algorithms appropriate for data science. Modeling everything from social organization to financial predictions, machine-learning algorithms allow us to discover information about complex systems, even when the underlying probability distributions are unknown. Algorithms discussed include regression, decision trees, clustering, and dimensionality reduction. Offers students an opportunity to learn the implications of the mathematical choices underpinning the use of each algorithm, how the results can be interpreted in actionable ways, and how to apply their knowledge through the analysis of a variety of data sets and models.

Practicum: **Algorithms and Implementation:**

The first portion of this course will focus on defining and applying the a variety of learning algorithms. We will start the class discussing the problem of learning a partially complete data labeling (Supervised Learning) and develop algorithms such as regression, support vector machines, and decision trees. We will then move to detecting patterns in unlabeled data using clustering and principle component analysis. By the end of the course, students should be familiar with the mathematical structure of these algorithms, and implications of any choices made in implementing them.

The machine learning is almost defined by its application. We will have frequent in class labs wherein we implement the algorithms we have discussed in the theory portion on real world data sets. In these labs, you will learn to use Python to obtain data, view it, clean it, analyze it, and finally use machine learning algorithms to try to solve a problem. The labs are open ended, and students are expected to use their own ideas and intuition to try to get an honest, best fit on the data sets given. The lab portion will cumulate in a final project, where you (or possibly your team) will attempt a novel machine learning project within an area of interest.

Theory: Frameworks and Communication:
After we have a handle on high level methods of machine learning, we will develop we will use the PAC (partially approximately correct) model to define data, sampling, training, and the metrics we will use to evaluate a machine learning algorithms success. The PAC framework will allow us to discuss each component of the machine learning process, and the practical effects of all the relevant choices. It will also allow us to state and discuss the No Free Lunch Theorem (the bias/complexity trade off) about the limits of learnability. We will cover the VC dimension and the conditions under which we can guarantee an algorithm can learn a problem. Finally, we will analyze regularization, gradient decent and ensambling as extensions of the PAC model, and show how they can often lead to better results without violating the theorems above.

The final portion of the class is about the communication of machine learning results and ideas. As an applied mathematician, part of your job will to communicate your results to coworkers, bosses or investors that don't understand the math as well as you do. To this end, you will be expected to communicate clearly and effectively in both your labs and in your final project. The goal of your final project is to produce something you can put on a resume, a git-hub, or give as a presentation.

I will be available to help you throughout this process with writing, editing and communication. This course requires Python, Jupyter Notebook Server and github, all of which are free and open source.

Grade Breakdown

- **Written Homework** (20%) - There will be three or four written assignments which will focus on theory.
- **Labs** (25%) - There will be roughly 6 labs. Labs will focus on the implementation of algorithms on real world data sets. Class time will be allotted for labs, but students may finish labs at home. In each lab, we will fit a real world data set using the algorithms of techniques introduced in that weeks' theory lecture. Labs will be graded out of 10 pts, 6 pts for completion, 4 pts for communication. There will also be a standing bonus of 2 pts for going above and beyond and exploring an interesting aspect of the parameter space, or getting a really good fit.
- **Midterm** (15%)
- **Attendance and class participation:** (10%)
- **Final Project** (30%) - The final project will consist of a proposal (1 page), middle stage progress report(2-3 pages), project report (roughly 5 pages) and presentation (roughly 20 minutes with poster or slides). Project groups should contain 3-5 people.

I am more than happy to discuss possible projects in any of these categories with you.

Masters students: This class features an XN project with an industry partner. Masters students are encouraged to participate in this project. (https://careers.northeastern.edu/experiential-network/)

PhD students: If you would like to propose your own project, it can take one of three forms:

- A computational analysis of a data set using sufficiently complicated or novel techniques from this course.
- A theoretical presentation of a topic not covered in this course with a case study.
- Thesis or Lab project.

Excellent can consider to submit the poster to RISE(https://www.northeastern.edu/rise/about/)

Rough project timeline: (Exact due day and time will be announced on Canvas.)

- Group Selection: As early as possible
- Project Proposal Deadline: Early Feb.
- Progress Report Paper Draft: After learned Resampling and validation
- Presentations: Last class
- Final paper: After Presentation

Late submission policy. Late submission within a week of any assignment without permission will receive at most 90% of the grade. Late submission within a week but after the posting of the solution will receive at most 70% of the grade. Other late submissions will depend on instructor's discretion.

Machine Learning Algorithm Roadmap

- **Supervised** - Labeled training data

Regression - Predicting continuous values
Linear Regression
Nonlinear Regression and Functional Fitting
Radial Basis Functions
Neural Networks
Classification - Predicting discrete values
Logistic Regression
Support Vector Machines (SVM)
Decision Trees
Neural Networks

- **Unsupervised** - Unlabeled data, pattern detection and descriptive modeling

Principle Component Analysis
Clustering
Neural Networks
TDA

Further topics in Machine Learning (Not covered in this class)
Semi-Supervised - Image segmentation, sparse training data
Neural Networks (Regional-CNN, U-NET)
k-means Clustering
Graph Partitioning
Implementation
Model Ensembles
Online Learning
Large Data Sets
Distributed Computations
Natural Language Processing
Reinforcement learning

Collaboration: You are welcome, even encouraged, to collaborate on the homework and lab assignments, though we urge you to first attempt working out all of the problems by yourself. However, you are expected to write answers **yourself** and understand everything that you hand in. **Copy** results from any sources will be considered as violating Academic Integrity. Collaboration is not allowed on the quizzes and exams.

Academic Integrity Policy: Cheating will not be tolerated. All incidents of cheating will be reported. From the Academic Integrity Policy: (see http://www.northeastern.edu/osccr/academic-integrity-policy/)

> "A commitment to the principles of academic integrity is essential to the mission of Northeastern University. The promotion of independent and original scholarship ensures that students derive the most from their educational experience and their pursuit of knowledge. Academic dishonesty violates the most fundamental values of an intellectual community and undermines the achievements of the entire University.
>
> As members of the academic community, students must become familiar with their rights and responsibilities. In each course, they are responsible for knowing the requirements and restrictions regarding research and writing, examinations of whatever kind, collaborative work, the use of study aids, the appropriateness of assistance, and other issues."

Title IX Policy: The University strictly prohibits sex or gender discrimination in all university programs and activities. Information on how to report an incident of such discrimination (which includes sexual harassment and sexual assault) is located at http://www.northeastern.edu/titleix .

Inclusion and Diversity: I value all students regardless of their background, country of origin, race, religion, gender, sexual orientation, ethnicity, or disability status, and am committed to providing a climate of excellence and inclusiveness within all aspects of the course. If there are aspects of your culture or identity that you would like to share with me as they relate to your success in this class, I would be happy to meet to discuss. Also, if you have any concerns in this area or are facing any special issues or challenges, I encourage you to discuss the matter with me as you feel comfortable, with assurance of full confidentiality (only exception being mandatory reporting of NU Academic Integrity Policy violations and Title IX sex and gender discrimination).

Students with disabilities: Students who have disabilities who wish to receive academic services and accommodations should follow the standard Disabilities Resource Center (DRC) procedures (see http://www.northeastern.edu/drc/getting-started-with-the-drc/).
College of Science Policies: The current College of Science Academic Course Policies is available at https://cos.northeastern.edu/wp-content/uploads/2012/10/COS-teaching-policies-April-2017.pdf .

TRACE: Every student is expected to complete the online TRACE survey at the end of the semester.

Tentative Schedule

Weeks		Topics	Assignment	Lab	Project
Week 1		Introduction to ML			
Week 2		Matrix Derivatives, Linear Regression	HW1		
Week 3		Ridge Regression, Lasso Regression, Gradient Decent	HW2	Lab 1	
Week 4		Logistic Regression		Lab 2	
Week 5		LDA, QDA	HW3	Lab 3	Proposal Due
Week 6		Resampling			
Week 7		Naïve Bayes, k-Nearest Neighbors			
Week 8		ANN, CNN +Lab	Midterm1	Lab 4	
Week 9	Spring Break				
Week 10		RNN+ Lab			Check In
Week 11		Support Vector Machines		Lab 5	
Week 12		PCA, Clustering	HW4		
Week 13		Clustering			Rough Draft
Week 14		Gaussian Mixtures		Lab 6	
Week 15		TDA, Presentation For Projects			Presentations