Fundamentals of Machine Learning (CSCI-UA.473)
(Fall 2024)

Overview of the Course

Machine Learning (ML) has tremendously impacted our society and economy over the years. In the future, its impact is only expected to grow even further. This course introduces undergraduate students to a diverse set of fundamental concepts of ML to enable them to apply these approaches to solve real-world problems. No prior knowledge of ML is required. However, a thorough understanding of the concepts of probability, linear algebra, and some form of calculus is a must to understand the contents of the course.

This semester’s class will focus on the main learning framework, namely supervised learning. We will describe the main learning algorithms, from linear models to modern NNs, and cover their main statistical and computational properties. We will end the course with an overview of modern ML methodologies beyond supervised learning.

Target Audience

This course is aimed at 3rd- or 4th-year undergraduate students in computer science. Please contact Romeo Kumar <kumar _at_ cs [dot] nyu [dot] edu>, the student advisor at the CS Department, directly for any questions on enrollment and eligibility criteria for the course.

Course Structure and Expectations from Students

The following section describes the course structure and what is expected from students for each part of the course.

Lectures and Labs

- Every week, there will be two kinds of sessions. The main lecture will take place on Tuesdays and Thursdays, where the theoretical concepts of ML will be discussed. On Fridays, there will be two recitation/lab sessions. Recitations will focus on the practical implementation of the lecture content, using real data as much as possible. **The contents of both the recitations will be the same.** Hence, students are only required to attend one, informed by their other commitments.
- Students are expected to attend and actively participate in all the lectures and during the lab sessions. Active participation in lab sessions is necessary to learn machine learning concepts successfully. It is vital to get your hands dirty with the code and play around with it if you want to learn the material and build an intuition of the concepts. While the slides and lab notebooks will be made available after the lecture in order to avoid note-taking during the class, there is no substitute for actual class presence. Only the lectures and labs before the add/drop course period will be recorded. After that, students are expected to attend in person; no recording is available.

- **Laptop policy**: Laptops are strongly discouraged during the main lectures. Students are required to bring their laptops to the lab sessions, which will involve hands-on coding. **Without the laptop, you will not be allowed to attend the recitation/lab sessions.**
- Machine learning is a vast and growing field and it is unreasonable to expect that a semester-long course can cover all the topics in depth. While there is no official book for the course, the students are strongly encouraged to read the supplementary material provided (or pointed) to within the class.

Homeworks

- Homework assignments are designed to build students' skills and conceptual proficiency. As a result, most assignments involve a mix of theory and practice.
- For the practical part of the homework, we will work with real-world datasets as much as possible to effectively demonstrate how we can solve real-world problems using ML approaches.
- H**omework 0** will be handed out on the first day of the class (September 3rd, 2024) and due on September 12th, 2024, before the class. This homework will help the students calibrate themselves and assess whether they have sufficient background to succeed in this course. **It will contribute 10% towards the final grade.**
- There will be six additional homework assignments throughout the semester. Every assignment will mix theory and practice. The due date for each assignment will be two weeks before the start of the next class. For instance, if the assignment is handed out on October 30th after the class, it'll be due on November 13th **before the class**.
- Students should **consult the course syllabus if they have** questions about assignment parameters/due dates. If the answer is not on the syllabus, students should contact the TA or the instructor.
- While students are encouraged to discuss/brainstorm assignments with their peers, **each student should write their own assignment**. In their assignments, students should acknowledge the names of other students they worked with. Copying assignments is forbidden, and NYU has a strict policy on academic integrity . **If you are not sure whether a certain form of collaboration is admissible**, **you are required to ask for clarification before submitting**. Furthermore, the goal of this course is for you to learn machine learning concepts and not to merely get good grades. Cheating will certainly not help you accomplish the goal of this course, which is to learn the fundamentals of machine learning.
- Use of AI assistants (chatGPT, etc): we have a strict policy against the use of chatGPT for the completion of the homework, unless specifically stated. **If you have used AI assistants at any parts of the homework (e.g. to rephrase, or fix orthography), you must disclose it**. Students are warned that the grading team will use statistical tools to detect cliques as well as patterns of similar answers. Unauthorized use of AI assistance will result in academic misconduct, and will carry an automatic 0 in the corresponding homework assignment.
- If you are incorporating any text/figures from an external source, you have the obligation to make it clear. Without doing so is considered plagiarism. The standard way of calling out the external source is by adding an explicit statement within your work that calls out (cites) which

external source you used for your material. This not only includes the text in your assignment but also includes the descriptions of your baseline model and your data, for example. Failure to do so will result in a zero grade for the submitted work and, except in cases of obvious mistakes, a University investigation.

- The theoretical part of all the homework assignments **should be handwritten, scanned, and the PDF output should be submitted**. No other format will be accepted.
- For the practical part of the homework assignments, students should submit Python Notebooks.
- We have devised a reasonable yet firm policy around late assignment submissions. Please see the policy details below.

Final Exam

- At the end of the semester, there will be one final exam.
- The final exam will be a **closed-book exam**.
- You will be provided with a printed copy of the question paper, which you will be expected to write your answers on using a pen. You will return the printed copy at the end of the exam.
- No electronics will be allowed inside the exam room. This includes cell phones, tablets, computers, or any other device connecting you to the internet.
- The exam will test the students' theoretical understanding and no coding will be involved.

Late Assignment Submission Policy

- We understand that unforeseen circumstances can arise which can lead to delays in assignment submission. Keeping that in mind, we'll implement the following policy to handle late assignments.
- Each student will be given three grace days to submit their assignment without penalty. They will be free to use these grace days at their convenience. Some examples of how a student could use these grace days are:
  - Student_1 submits three assignments, each one day late. They will not be penalized for any assignment.
  - Student_2 submitted one assignment that was 1 day late and another that was 2 days late. They will not be penalized for these two assignments.
  - Student_3 submits one assignment three days late. They will not be penalized for this single assignment.
- Once the student exhausts their three graces days, they will receive:
  - 75% grade if their assignment is late by one day
  - 50% grade if their assignment is late by two days
  - 0% grade if their assignment is late by more than two days
- In each scenario, the clock will be rounded up to the nearest 24 hours. For example, if the student is late in submitting an assignment by 3 hours from the deadline, it'll be considered as late by one day. If the student submits an assignment 25 hours, it'll be considered late by two days.

Other Logistics

- Students will use their NYU credentials to complete all tasks and communications for this course (sending emails etc.)

General Information

Lectures

- Day and times: Tuesdays and Thursdays 2:00 PM - 3:15 PM
- Location:
- All lectures will be in person
- The lectures will be delivered on the whiteboard, with occasional slides.
- The lecture notes will be made available after each lecture.

Recitation/Labs

- Day and times: Fridays 9:30 AM - 10:45 AM and Fridays 2:00 PM - 3:15 PM
- Location: 60 5th Avenue, Room 150
- All labs will be in person
- Both labs will have the same content. Students should attend one of the sessions depending on what is suitable for them.
- Labs will involve implementing the theory discussed during the lectures
- Students attending the labs must bring their laptops
- Lab sessions will use Python 3 and PyTorch
  - ○ Installation instructions for classic Jupyter or Jupyter Lab (Necessary for Labs)
  - ○ Installation instructions for PyTorch (Necessary for Labs)
  - ○ Installation of python packages (Necessary for Labs)
  - ○ Anaconda to install necessary packages (Highly recommended)
  - ○ Consider using a virtual environment (Advanced and not necessary)
- The lab materials will be distributed via Brightspace before the lab session

Instructors

- Lectures
  - ○ Name: Joan Bruna
  - ○ Email: bruna@cims.nyu.edu
  - ○ Office: 60 5th Avenue, Room 612
  - ○ Preferred Communication Method: Email
- Recitation/Labs
  - ○ Name: Zhe Zeng
  - ○ Email: z.zhe@nyu.edu
  - ○ Office: 60 5th Avenue, Room 340

    - Preferred Communication Method: Email

Assistants

- Teaching Assistant/Lab Assistant (Name and Email): Lei Chen lc3909@nyu.edu.
- Tutor 1 (Name and Email):
- Tutor 2 (Name and Email):
- Graders: anonymous and no contact (to ensure unbiased grading)

Office Hours and Location

- **Instructor (Joan Bruna):** Mondays 2:00 - 3:00 PM in room 612, 60 5th Ave
- **Recitation Leader (Zhe Zeng):** Fridays 11:30 AM - 12:30 PM in room 340, 60 5th Ave
- **Teaching Assistant (Lei Chen)**: Wednesdays 1:00 - 2:00 PM in room 350, 60 5th Ave
- **Tutor 1 (Veronica Zhao)**: Fridays 10:45 - 11:15 AM and 3:15 - 3:45 PM in room 350, 60 5th Ave
- **Tutor 2 (?)**: Fridays 10:45 - 11:15 AM and 3:15 - 3:45 PM in room 350, 60 5th Ave

Grading and Assignments

- Assessment will be based on two primary factors: homework and a final exam.
- There will be six (6) homework assignments throughout the semester (including Homework 0)
- Each homework assignment will contribute 10% towards the final grade. Thus, homework assignments will account for 60% of the final grade.
- The final exam will contribute the remaining 40% of the final grade
- An extra 10% credit will be allocated to students who forward solutions of recitation exercises. The exercises must be solved in class and submitted by the end of the recitation day (each Friday). The grading will be completion-based.
- However, note that you **must pass** the final exam to pass the course
- All homework assignments should be handwritten, scanned and submitted as pdfs.
- The due date for the homework will be two weeks from the date assigned AND before the lecture. For example, assignment given on 10/30 after the class will be due on 11/13 **before** the class
- There is a reasonable but firm late acceptance policy. Please see the details above
- **Date of the final exam**: TBD

Course Website

- Brightspace
    - Lecture notes, homework release and submissions, broadcast announcements
    - The link to the course page is here.
- CampusWire
    - The primary platform of communication between the students and instructor/recitation leader/lab assistant/tutors/graders, etc

  - Will be (and should be used) to clarify individual doubts
  - The link to the course chat is here. The code is 3846.

Books

The following resources will be useful but do not need to be purchased

- [ML Concepts] The Elements of Statistical Learning: Trevor Hastie, Robert Tibshirani, and Jerome Friedman
- [ML Foundations] Learning Theory from First Principles, Francis Bach.
- [ML Concepts] Pattern Recognition and Machine Learning: Christopher Bishop
- [Coding] Introduction to Machine Learning with Python: Andreas C. Müller & Sarah Guido

Other Material for Review

- Lecture notes by Kyunghyun Cho
- Linear Algebra and Vector Calculus
- Probability Theory

Prerequisites

A strong foundation in linear algebra, vector calculus, and introductory probability (standard probability distributions, continuous and discrete variables, expectations, and conditional expectation). Mathematical maturity and comfort with coding algorithms is required.

Please assess your fit for the course by attempting Homework 0. Homework 0 will be provided to you at the end of the first lecture. If most of the questions are not approachable, you may not have the right math background for this course and will likely struggle. **Note that Homework 0 will be considered a part of your final grade. It will have a 10% contribution towards it.**

- Required
  - Data Structures (CSCI-UA.102)
  - Linear Algebra (MATH-UA.140)
  - Probability and Statistics (MATH-UA.235)
- Recommended
  - CSCI-UA 310 Basic Algorithms
  - DS-GA 1001 Introduction to Data Science
    - Exercise materials are highly recommended.
  - DS-GA 1002 Statistical and Mathematical Methods

Schedule

Below is the tentative syllabus. The contents of each lecture might change as the course progresses.

Week	**Lectures**	**Recitation/Lab**	**Homework**

	Tuesdays	**Thursdays**	**Fridays**	
1 (09/03)	Introduction ● Course logistics ● Overview of machine learning. ● No-free-lunch theorem	Introduction ● Representative Examples (discrete math, weather forecast, CV) Learning paradigms ● **Supervised** ● Unsupervised ● Semi-supervised ● Self-supervised ● Reinforcement Learning ● Generative vs Discriminative model examples	Lab Logistics ● Installing Anaconda ● Installing PyTorch ● PyTorch Tutorial (non-autograd) Math Primer 1 ● Linear Algebra - basics	[HW0 (math basics) released 9/3]
2 (09/10)	Warmup: Linear Least Squares ● Setup ● Examples	Linear Least Squares ● The Normal Equations	Math Primer 2 ● Linear Algebra - advanced ● Vector Calculus - creating tensors/dataset ● Probability and Statistics *Required reading: Linear Algebra and Vector Calculus and Probability Theory*	[HW 0 due on 09/16/2024 EOD]
09/17	Course Add/Drop Deadline			
3 (09/17)	Linear Least Squares: statistical analysis ● Fixed Design Setting ● Overfitting	Linear Least Squares: Ridge regularization PCA	[Lab 1] Linear Models for Regression ● Weather forecasting in NYC ● Dataset handling: Train/Validation/Test Splitting ○ Random shuffling ○ Proportion of splits ● Coding the model ● Loss function (mean squared error loss) ● Optimization (only closed form solution) ● Regularization ○ Closed form solution with regularization ● Model analysis and interpretation: what do the weights mean?	HW1 released 09/19
4 (09/24)	Main Ingredients of Supervised Learning: Statistical Modeling ● Data Distribution ● Loss Function ● Model Class	Main Ingredients of SL (contd) ● Empirical Risk ● ERM ● Regularization	[Lab 2] Linear models for classification Credit card approval prediction ● Train/Validation/Test split ● Building a data loader for iterative optimization ● Coding the basic model ● Loss functions ○ Binary output (logistic regression) ○ Multi-class output (softmax)	

			● Optimization ● Regularization ● Model analysis and interpretation	
5 (10/01)	Decomposition of Error in Supervised Learning ● Approximation ● Generalization ● Optimization Bias-Variance Tradeoff	Different Learning Regimes ● Underparametrised Regime ● Overparametrised regime ● Regularization Parametric vs Non-parametric learning Examples	[Lab 3] Underfitting / Overfitting	HW1 due 10/03
6 (10//08)	The Curse of Dimensionality ● In Approximation ● In Estimation	CoD (contd) ● In Optimization	[Lab 4] CoD in Practice	HW2 released 10/10
7 (10/15)	Approximation Aspects ● Universal Approximation Theorems ● Rates of approximation*	Approximation Aspects (contd) ● Linear vs Non-Linear Approximation	[Lab 5] Polynomial Regression Sparse Feature selection	
8 (10/22)	Guest Lecture: Statistical Notions ● Monte-Carlo Estimation ● Uniform Concentration	Statistical Notions (contd) ● Examples ● Practical considerations: Cross-Validation	[Lab 6] Monte-Carlo Methods Cross-Validation Model Selection	HW2 due 10/24
9 (10/29)	Optimization Notions ● Quadratic Optimization ● Convex Optimization ● Non-convex optimization	Optimization Notions ● Gradient Descent	[Lab 7] Artificial Neural Networks TODO move later ● PyTorch Autograd, GPU computations ● The MNIST dataset ● Data normalization and preparation ● Adapting the data loader from Lab 2 to work with images ● Implementing a multi-layer perceptron ● Training using stochastic gradient descent ● Stochastic vs mini-batch vs full-batch gradient descent ● Regularizing neural networks: weight decay ● Visualizing the results	HW3 released 10/31
10 (11/05)	Gradient Descent in the context of ERM	Stochastic Gradient Descent ● Practical Considerations ● Basic analysis	[Lab 8] Advanced Deep Learning Architectures TODO move later ● Convolutional neural network for the MNIST dataset ● Language modeling dataset: Sentiment analysis (classification) ● Preparing and representing the text dataset ● Linear BOWs models	

			● Transformer Networks for the text dataset	
11 (11/12)	Non-parametric SL. Local Averaging methods: Nearest Neighbors	Non-parametric SL: kernel methods Random Feature Expansions	[Lab 9] Non-Parametric models - 1 ● Nearest neighbor classifiers ● Decision trees	HW3 due 11/12 HW4 released 11/12
12 (11/19)	Linear Models: Classification ● Logistic Regression ● Multi-Class Classification: softmax ● Probabilistic Interpretation	Linear Models: classification The support vector machine	[Lab 10] Support Vector Machines and Radial Basis Function Networks ● SVMs with different kernels ● SVM visualizations	
13 (11/26)	Neural Networks: motivation and history	Thanksgiving	thanksgiving	HW4 due 11/27
14 (12/03)	Neural Networks: single-hidden layer.	Modern NNs in practice (beyond MLPs, CNNs, transformers)	[Lab 11] Backpropagation	HW5 released 12/03
15 (12/10)	Beyond SL: unsupervised learning PCA K-Means	Beyond SL: generative modeling Examples Successes and Limitations	[Lab 12] Unsupervised Learning ● Principal component analysis on a real dataset: Eigen faces ● General auto-encoders ● PCA implemented as auto-encoder ● K-means	
16 (12/18)	Exam prep Q&A	***** Final Exam *****		HW5 due 12/18

University Policies

Academic Integrity

Work you submit should be your own. Please consult the CAS academic integrity policy for more information: https://cas.nyu.edu/content/nyu-as/cas/academic-integrity.html. Penalties for violations of academic integrity may include failure of the course, suspension from the University, or even expulsion.

Religious Observance

As a nonsectarian, inclusive institution, NYU policy permits members of any religious group to absent themselves from classes without penalty when required for compliance with their religious obligations. The policy and principles to be followed by students and faculty may be found here: The University Calendar Policy on Religious Holidays (http://www.nyu.edu/about/policies-guidelines-compliance/policies-and-guidelines/university-calendar-policy-on-religious-holidays.html)

Disability Disclosure Statement

Academic accommodations are available to any student with a chronic, psychological, visual, mobility, learning disability, or who is deaf or hard of hearing. Students should please register with the Moses Center for Students with Disabilities at 212-998-4980.

NYU's Henry and Lucy Moses Center for Students with Disabilities
726 Broadway, 2nd Floor
New York, NY 10003-6675
Telephone: 212-998-4980
Voice/TTY Fax: 212-995-4114
Web site: http://www.nyu.edu/csd