Carnegie Mellon University
HeinzCollege
INFORMATION SYSTEMS • PUBLIC POLICY • MANAGEMENT

Carnegie Mellon University
95-828 Machine Learning for Problem Solving
Spring 2021

HOME
SYLLABUS
ASSIGNMENTS
COURSE POLICY
RESOURCES

CLASS MEETS:

There are two sections of the course offered in Spring 2021.
Time:

- **Section A & B:** Tue & Thu 10:10AM - 11:30AM

Place: Both sections ONLINE @Zoom (*Calendar invitations with link sent individually*)

WEEKLY RECITATION:

Time: Fri 10:10AM - 11:30AM
Place: ONLINE @Zoom (*See link on Canvas*)

PEOPLE:

Instructor: Leman Akoglu

- Online office hour: **FRI 9-10PM EDT**; also, by appointment
- Email: *invert* (cs.cmu.edu @ lakoglu)

Teaching Assistants:

Ziyi Chen

- Office hour: **TUE 9-10AM EDT**
- Email: *invert* (andrew.cmu.edu @ ziyichen)

Lingxiao Zhao

- Office hour: **THU 7-8PM EDT**
- Email: *invert* (andrew.cmu.edu @ lingxia1)

Please find all the Zoom links to office hours on Canvas.

Graders:

Jinchen Xie

- Email: *invert* (andrew.cmu.edu @ jinchenx)
- Office hours: by appointment

Mansi Goyal

- Email: *invert* (andrew.cmu.edu @ mansigoy)
- Office hours: by appointment

Gautham Kumar Vedam

- Email: *invert* (andrew.cmu.edu @ gvedam)
- Office hours: by appointment

COURSE DESCRIPTION:

Machine Learning (ML) is centered around automated methods that improve their own performance through learning patterns in data, and then using the uncovered patterns to predict the future and make decisions. ML is heavily used in a wide variety of domains such as business, finance, healthcare, security, etc. for problems including display advertising, fraud detection, disease diagnosis and treatment, face/speech recognition, automated navigation, to name a few.

"If I had an hour to solve a problem I'd spend 55 minutes thinking about the problem
and 5 minutes thinking about solutions." -- Albert Einstein
"A problem well put is half solved." -- John Dewey

This course is designed to give a graduate-level student a thorough grounding in the methodologies, technologies, and best practices used in machine learning. The main premise of the course is to equip students with the intuitive understanding of machine learning concepts grounded in real-world applications. The course is to help students gain the practical knowledge and experience necessary for recognizing and formulating machine learning problems in the wild, as well as of applying machine learning techniques effectively in practice. The emphasis will be on learning and practicing the machine learning process, more than learning theory.

"All models are wrong, but some models are useful." -- George Box

As there exists no universally best model, we will cover a wide range of different models and learning algorithms, which have varying speed-accuracy-interpretability tradeoffs. In particular, the topics include supervised learning:

linear models, decision trees, ensemble methods, kernel methods, nonparametric learning, and unsupervised learning: density estimation, clustering, and dimensionality reduction. The class will include biweekly homework each containing a mini-project (i.e., a problem solving assignment that involves programming) in addition to other conceptual and technical questions, a midterm, a final exam, and a case study at the end of the course. The case study will give students a chance to dig into a substantial problem using a large dataset and apply machine learning tools they have learned throughout the course.

Prerequisites

This course does not assume any prior exposure to machine learning theory or practice. Students are expected to have the following background: • Basic knowledge of probability • Basic knowledge of linear algebra • Basic programming skills • Familiarity with Python programming and basic use of NumPy, pandas and matplotlib.

Learning Objectives

By the end of this class, students will

- learn the main concepts, methodologies, and tools for machine learning
- be able to recognize machine learning tasks in real-world problems
- develop the critical thinking for comparing and contrasting models for a given task
- learn the best practices for reliably performing model selection and evaluation
- gain experience with implementing ML solutions in Python and applying them to various real world datasets

BULLETIN BOARD and other info

- For course materials, assignments, announcements, and grades please see the Canvas.
- For submitting homework electronically, you will use Gradescope.
- For questions and discussions please use Piazza. Here is the link to signup.
- Carnegie Mellon 2020-2021 Official academic calendar

TEXTBOOK:

There is no required textbook for the course. I will post course notes and slides for each lecture as well as some code examples (Jupyter notebooks) on Canvas. See Resources for a list of recommended books that could help supplement your understanding of the course material.

MISC - FUN:

Fake (ML) protest @G20 Summit (2009)   ML demonstration @PittMarathon (2019)

Carnegie Mellon University
HeinzCollege
INFORMATION SYSTEMS • PUBLIC POLICY • MANAGEMENT

Carnegie Mellon University
95-828 Machine Learning for Problem Solving
Spring 2021

HOME
SYLLABUS
ASSIGNMENTS
COURSE POLICY
RESOURCES

Course Syllabus (download as pdf)

LECTURES:

I will provide course notes as well as slides for each lecture. Those will be uploaded to Canvas **before** the lecture. Feel free to print them and bring them to class with you for annotating.

You may also benefit from the recommended books (listed under Resources, see left tab) to further your understanding. To stay on track, make sure to read the course notes in a timely fashion, and follow up with questions in lectures, office hours, recitations, and/or Piazza.

RECITATIONS:

There will be a recitation session held by one of the TAs on Fridays. The recitation will review the week's material and answer any questions you might have about the course material, including homework.

Week	Lectures	Notes
Week 1	**INTRO TO MACHINE LEARNING** [+] **DATA PREPARATION** [+]	**HW 0 out** • Python and Jupyter setup **Recitation 1** • Python setup • Data prep
	PART I: SUPERVISED LEARNING	
Week 2	**LINEAR REGRESSION (LR)** [+]	**Recitation 2** Data prep demos • Linear Algebra review
Week 3	**MODEL SELECTION** [+]	**Recitation 3** • Linear Reg. demos • Convex optimization basics **HW 1 out** • EDA • LR • Model selection • LogR
Week 4	**LOGISTIC REGRESSION (LogR)** [+]	**Recitation 4** Bias-Variance trade-off • Cross-validation
Week 5 • Week 6	**NON-PARAMETRIC LEARNING** [+] **MODEL EVALUATION** [+]	**Recitation 5** LogR • Gradient descent review and demos **HW 2 out** • Non-parametric learning • Model evaluation • DT **Recitation 6** • kNN • Kernel regression • Model evaluation
Week 7	**DECISION TREES (DT)** [+]	**Recitation 7** • DT review and demos
Week 8	**Midterm Review** **Midterm Exam**	Exam will be during class on Thur. Duration: 80 minutes. You can only bring your own notes up to 2 A4-size sheets. No electronics. Friday NO RECITATION

Week 9	**NO CLASS: Spring Break**	
Week 10	**ENSEMBLE METHODS** [+] **NAIVE BAYES (NB)** [+]	HW 3 out • Ensembles • NB • SVM **Recitation 8** Random Forest • Boosting • NB Case Study out • Dataset provided, Tasks recommended
Week 11	**SUPPORT VECTOR MACHINES (SVM)** [+]	**Recitation 9** • SVM and Kernels
Week 12 •	**NEURAL NETWORKS (NN)** [+]	HW 4 out • Kernels • Neural Nets • Density estimation **Recitation 10** • NNs • Back-propagation
Week 13	PART II: UNSUPERVISED LEARNING **DENSITY ESTIMATION** [+]	
	Thur NO CLASS: Spring carnival	Friday NO RECITATION
Week 14 • Week 15	**CLUSTERING** [+] **DIMENSIONALITY REDUCTION** [+]	HW 5 out • Clustering • EM • Dimensionality reduction **Recitation 11** • Density estimation • hierarchical clustering • k-means **Recitation 12** • EM • Dim. reduction
Week 16	**Case Study & Final Review**	**Recitation 13** • Case Study review • Final Q&A

Carnegie Mellon University
95-828 Machine Learning for Problem Solving
Spring 2021

HOME
SYLLABUS
ASSIGNMENTS
COURSE POLICY
RESOURCES

Coursework

Coursework consist of (grading in parentheses):

- 5 Homework (9% each)
- 1 Midterm exam (15%)
- 1 Final exam (25%)
- 1 Case Study (15%)

HOMEWORK:

Homework will be posted on Canvas. Each homework will consist of two parts: (1) a set of conceptual questions, and (2) programming. For the programming part, we will provide a code template (and sometimes partial code as well) in a Jupyter notebook. You will have two weeks to complete each homework assignment.

Getting help: You can visit the instructor and the TAs during office hours as well as post questions on Piazza to get help on the assignments. Regarding help from fellow students, see the note on collaboration below.

Collaboration: Collaboration and study groups are allowed and encouraged. All assignments are to be done by study groups, 2-3 students each. The Case Study can be done in groups of up to 4 members. Each group uploads a single submission on Gradescope. Please see the collaboration policy for details.

Submitting: We ask that you submit two files per homework: (1) a pdf file with your answers to the conceptual questions, and (2) the Jupyter notebook we provide as a template with all your code that you filled in. Both files (.pdf and .ipynb) are to be uploaded **electronically only** on Gradescope (no hard copy print outs).

Homework assignments are **due at the beginning of the class** on the day it is due. You can upload your files multiple times, but note that we will use the latest upload date as the submission date, which may factor into your slip days accordingly. Please see the late submission policy for details.

IMPORTANT DATES:

Assignment	Note	Out	Due	Weight
Homework 0	Setting up Python and Jupyter	Feb 2	n/a	0%
Homework 1	EDA, LR, Model selection	Feb 16	Mar 2	9%
Homework 2	LogR, Model eval., Non-parametric, DT	Mar 2	Mar 16	9%
Midterm Exam	(in class)	Mar 18	--	15%
Homework 3	Ensemble models, NB, SVM	Mar 23	Apr 6	9%
Homework 4	Kernels, Neural nets, Density estimation	Apr 6	Apr 20	9%
Homework 5	Clustering, EM, Dimensionality reduction	Apr 20	May 4	9%
Case Study	Mini 4	Mar 22	May 7	15%
Final Exam		TBD TBD	--	25%

EXAMS:

There will be a midterm exam (in class) and a final exam (to be scheduled by the University).

Note: For the midterm, you are allowed to bring with you 2 A4-size sheets, containing your own notes (hand-written or typed). You can use both sides of each sheet. For the final, you are allowed to bring up to 5 A4-size sheets (double sided), again containing your own notes. Use of any computers or other electronic devices during the exams is not allowed. The tentative dates are posted above, the finalized dates will be announced during the semester.

CASE STUDY:

Starting the second half of the course (after Spring break), we will provide you with information describing a large dataset and a list of potential questions to address based on this dataset. We will also release the dataset after Spring break. You will be given the second half of the semester to complete your analysis and modeling on the data. Particularly, you will be expected to carefully choose to apply the techniques and tools you have learned throughout the course to address the problems of interest using machine learning.

The Case Study will consist of 3 Phases. In Phase I, you will think about the data and the problem at hand and brainstorm. In Phase II, you will do hands-on data cleaning, preparation and exploratory analysis and data understanding. In Final Phase III, you will build predictive models using various machine learning tools you have learned throughout the course.

Evaluation: We will assess your case study outcomes in terms of your analytical approach to the problems, and not only based on the quality of your results. That is, the emphasis will be on evaluating how methodical you were in your analysis in terms of the tools you chose to apply, in the way you draw conclusions from your own results, and the sequence of steps you took based on your analyses and intermediate results. We will also assess if you used the best practices in building your solutions, including proper model selection, model comparisons to appropriate baselines, choice of evaluation metrics, and so on.

Teams: The Case Study can be done in groups of up to 4 students. We recommend forming groups of 4, but groups of 2-3 students should also be fine. We do not recommend single-member teams given the amount of workload. You can use Piazza for communication toward finding team members. **Submitting:** You are asked to submit a single Jupyter notebook, composed of all your code and results, along with a pdf file with answers to specific questions. All submissions will be made on Canvas.

Carnegie Mellon University
HeinzCollege
INFORMATION SYSTEMS • PUBLIC POLICY • MANAGEMENT

Carnegie Mellon University
95-828 Machine Learning for Problem Solving
Spring 2021

HOME
SYLLABUS
ASSIGNMENTS
COURSE POLICY
RESOURCES

Course Policies

LECTURES

- All devices such as laptops, cell phones, noisy PDAs, etc. should be turned off for the duration of the lectures and the recitations, because they may distract other fellow students.
- Students who would like to use their laptops during the course are strongly encouraged to sit at the back-most row of the classroom.
- Please come to all lectures on time and leave on time, again so that there are no distractions to the classmates.

PRE-REQUISITES

This course does not assume any prior exposure to machine learning theory or practice. Students are expected to have the following background:

- Basic knowledge of probability
- Basic knowledge of linear algebra
- Working knowledge of basic computing principles
- Familiarity with Python programming and basic use of NumPy, pandas and matplotlib

ASSIGNMENTS

- Assignments are due at the * **beginning of lecture** * on the due date.
- The due date of assignments are posted at the assignments page.
- Assignments will be posted on Canvas.
- Students should submit their homework solutions (a pdf file with answers to conceptual questions and a Jupyter notebook with answers to programming questions) **only electronically** via Gradescope (no print outs).

Important Note: As we reuse problem set questions, covered by papers and webpages, we expect the students not to copy, refer to, or look at the solutions in preparing their answers. Since this is a graduate-level class, we expect students to want to learn and not google for answers. The purpose of problem sets in this class is to help you think about the material, not just give us the right answers.

Therefore, please restrict attention to the class notes, slides, and the supplementary books mentioned on the resources page when solving problems on the problem set. If you do happen to use other material, it must be acknowledged clearly with a citation on the submitted solution.

Questions and Re-grade requests

- You should use Piazza for all your questions about the assignments and the course material. Instructor and TA(s) will do their best to answer your questions timely.
- Regrade requests should be done in **writing/email**,
  - within **2 days** after graded assignments are distributed
  - to the **grader students** specified on the front page (see Graders under People), and specifying
    - the question under dispute (e.g., 'HW1-Q.2.b')
    - the extra points requested (e.g., '2 points out of 5')
    - and the justification (e.g., 'I forgot to divide by variance, but the rest of my answer was correct')
  - In the remote case there is no satisfactory resolution, please contact the instructor.

Homework Grading and Solutions

- All homework will be graded online through Gradescope. Graders will provide comments and feedback on the deductions they have made accordingly.
- We will post solutions to the assignments on Canvas, 4 days after the due date (to account for students using slip days, see below).

Late submission policy

- No delay penalties, for medical/family/etc. **emergencies** (bring written documentation, like doctor's note).
- Each student is granted '**4 slip days**' total for the whole course duration, to accommodate for coinciding deadlines/interviews/etc. That is, no questions asked, if the total delay is 4 days or less.
  - You can use the extension on any assignment during the course (unless otherwise stated). For

instance, you can hand in one assignment 4 days late, or 4 different assignments 1 day late each.
    - Late days are rounded up to the nearest integer. For example, a submission that is 4 hours late will count as 1 day late.
    - After you have used up your slip days, any assignment handed in late will be marked off **25% per day of delay**.
- To use slip days:
    - upload your homework solutions on **Gradescope to mark the time of submission**
    - You can upload your modified files multiple times at different points in time. However, please note that we will use **your latest upload date** as the date of submission, even if you have modified only a small part of your files.

Collaboration policy

You are encouraged to discuss homework problems with your fellow students. However, the work you submit must be your own. You must acknowledge in your submission any help received on your assignments. That is, you must include a comment in your homework submission that clearly states the name of the student, book, or online reference from which you received assistance.

Submissions that fail to properly acknowledge any help from other students or non-class sources will receive NO credit. Copied work will receive NO credit. Any and all violations will be reported to the Heinz College administration and may appear in the student's transcript.

Academic integrity

All students are expected to comply with CMU's policy on academic integrity. Please read the policy and make sure you have a complete understanding of it.

EMAIL

Piazza should be used for general course and assignment related questions. For other types of questions (e.g., to report illness, request various permissions) please contact the instructor directly via email.

Please make sure to **include '95828' in the subject line** of your email.

AUDITING

Auditing is not allowed. Only those students who are officially enrolled to take the course for credit are allowed to sit in class.