Carnegie Mellon University
HeinzCollege
INFORMATION SYSTEMS • PUBLIC POLICY • MANAGEMENT

Carnegie Mellon University
95-828 Machine Learning for Problem Solving
Spring 2024

HOME
SYLLABUS
ASSIGNMENTS
COURSE POLICY
RESOURCES

CLASS MEETS:

There are two sections of the course offered in Spring 2023.
Time:

- **Section A:** Tue & Thu 11:00AM - 12:20PM
- **Section B:** Tue & Thu 2:00PM - 3:20PM

Place: Both sections in HBH A301. *Link to Zoom (optional) on Canvas*

WEEKLY RECITATION:

Time: Fri 2:00PM - 3:20PM
Place: HBH A301 (*Also see Zoom link on Canvas*)

PEOPLE:

Instructor: Leman Akoglu

- Office hour (Mini A3): **THU 12:50PM - 1:50PM** (starts Jan 23)
- Office: **HBH 2118C**, office ph. 412-268-30 four three
- Email: *invert* (andrew.cmu.edu @ lakoglu)

Teaching Assistants:

Xueying Ding
- Office hour: **MON 11:30AM to 12:30PM EDT**
- Email: *invert* (andrew.cmu.edu @ xding2)

Xiaobin Shen
- Office hour: **THU 10-11AM EDT**
- Email: *invert* (andrew.cmu.edu @ xiaobins)

Zijun Ding
- Office hour: **WED 5:15-6:15PM EDT**
- Email: *invert* (andrew.cmu.edu @ zijund)

***Zijun and Xiaobin** have reserved **HBH 2108** to hold their in-person OHs.*
***Xueying**'s OHs will be on **Zoom/online** (please find link on Canvas).*

Graders:

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

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

Longyang Xu
- Email: *invert* (andrew.cmu.edu @ longyanx)
- 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 2023-2024 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)

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 5:30-7pm. 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** [+]
	• The Learning Problem, Terminology
	• Canonical Learning Problems
	  ◦ Supervised Learning
	  ▪ Regression
	  ▪ Classification (binary vs. multi-class)
	  ◦ Unsupervised Learning
	  ▪ Density estimation
	  ▪ Clustering
	  ▪ Dimensionality reduction
	• ML applications in the real world
	• What does it mean to learn?
	  ◦ A key ML concept: Generalization
	  ◦ vs. Overfitting
	• Course Logistics
		HW 0 out • Python and Jupyter setup

	DATA PREPARATION [+]
	• Python for ML Intro
	• Feature Engineering
	• Preliminary Data Analysis
	  ◦ EDA: exploratory data analysis
	  ▪ 1D: bar chart, histogram, box plot;
	  ▪ 2D: scatter plot, heat map and contourmap;
	  ▪ >3D: parallel coordinates, radar plot
	• Data Cleaning and Transformation
	  ◦ Handling missing values
	  ▪ mean/median, kNN, model-driven imputation
	  ◦ Transforming feature types and feature values
		Recitation 1 • Python setup • Data prep

- OHE: one-hot-encoding
- normalization
- log-transform

PART I: SUPERVISED LEARNING

Week 2 **LINEAR REGRESSION (LR)** [+]

Recitation 2 Data prep demos • Linear Algebra review

- Formalizing the Learning Problem
  - loss functions
  - data generating distribution
  - models, parameters, hyperparameters
  - optimization algorithms
- Supervised Learning Cycle
- Linear models and Parameters
- Closed-form opt. for squared loss
- Interpreting coefficients
- Regularization
- Shrinkage methods: Ridge & Lasso regression
- Beyond linearity
  - Non-linear basis expansions
  - Local regression (*)
  - GAMs: Generalized Additive Models
- Practical issues:
  - feature scaling
  - categorical features, OHE
  - outliers & high-leverage points
  - collinearity
  - high dimensions

Week 3 **MODEL SELECTION** [+]

Recitation 3 • Linear Reg. demos • Convex optimization basics

HW 1 out • EDA • LR • Model selection • LogR

- What is a good model?
- Overfitting and Generalization
- Decomposition of error
  - estimation vs. approximation error
- Bias-Variance tradeoff
- Regularization
- Separation of training and test data
- CV: Cross Validation

Week 4 **LOGISTIC REGRESSION (LogR)** [+]

Recitation 4 Bias-Variance trade-off • Cross-validation

- Classification vs. Regression
- 0-1 loss
- Convex surrogate loss functions & logistic loss
- Decision rule and boundary
- Intro to convex optimization basics
- Gradient descent optimization
- LR with >2 classes

- Kernel Logistic Regression (*)

NON-PARAMETRIC LEARNING [+]

Week 5

- k Nearest Neighbors (kNN) classifier
  - decision boundaries

Week 6

- kNN regression
- Local regression
- Locally-weighted linear regression

- Comparison of LR/LogR with kNN

- Practical issues:
  - curse of dimensionality
  - intelligibility
  - computational efficiency
  - distance functions

Recitation 5 LogR • Gradient descent review and demos

HW 2 out • Non-parametric learning • Model evaluation • DT

Recitation 6 • kNN • Kernel regression • Model evaluation

MODEL EVALUATION [+]

- Evaluation metrics
  - Cost of false positives and false negatives
  - Confusion matrix
  - Visualizing model performance
    - ROC, precision-recall, lift, profit curves

- Debugging your model
  - train/test mismatch
  - analyzing error, ablative analysis
  - class imbalance and resampling strategies

- Creating baseline methods for comparison
- Statistical comparison of models

Week 7

DECISION TREES (DT) [+]

- Classification trees
- Regression trees
- Regularization and pruning

- Trees vs. Linear models

- Practical issues:
  - handling missing values

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 [+]

- Combining multiple models
- Bagging
- Random Forests
- Boosting

HW 3 out • Ensembles • NB • SVM

Recitation 8 Random Forest • Boosting • NB

Case Study out • Dataset provided, Tasks recommended

NAIVE BAYES (NB) [+]

- Classification by density estimation
- Conditional independence
- MLE, Regularization via priors and MAP
- Generative vs. Discriminative models
- Gaussian NB (*)

Week 11

SUPPORT VECTOR MACHINES (SVM) [+]

- SVM formulation
  - construction of the max-margin classifier
- The non-separable case
  - hard vs. soft-margin SVM
  - slack variables
- Hinge loss
- SVMs with >2 classes
- Relation to LR
- Intro to dual optimization
- SVM dual
- The Kernel trick
  - From feature combinations to kernels
  - Kernel SVM
  - Interpreting SVM dual and its solution
  - (*) Kernel Logistic Regression

Recitation 9 • SVM and Kernels

Week 12

NEURAL NETWORKS (NN) [+]

- Representation
  - Perceptron
  - single- & multi-layer networks
  - multiclass classification
- Learning
  - Backpropagation algorithm
  - Regularization

HW 4 out • Kernels • Neural Nets • Density estimation

Recitation 10 • NNs • Back-propagation

Week 13

PART II: UNSUPERVISED LEARNING

DENSITY ESTIMATION [+]

- Parametric
  - Gaussian/Poisson/etc

- Gaussian/Poisson/etc.
  - MLE: Maximum Likelihood Estimation
  - MAP: Maximum A Posteriori estimation

- Non-parametric
  - Histograms
  - KDE: Kernel Density Estimation

Thur NO CLASS: Spring carnival

Friday NO RECITATION

Week 14

CLUSTERING [+]

- Similarity/distance functions
- Hierarchical clustering

HW 5 out • Clustering • EM • Dimensionality reduction

Recitation 11 • Density estimation • hierarchical clustering • k-means

Week 15

- k-means clustering
- Mixture models
- EM: Expectation Maximization

Recitation 12 • EM • Dim. reduction

DIMENSIONALITY REDUCTION [+]

- Unsupervised embedding techniques
  - PCA: Principal Component Analysis
  - Kernel PCA
  - t-SNE
  - MDS: Multi-Dimensional Scaling
- Supervised reduction techniques
  - Feature selection
    - forward selection
    - backward selection

Week 16

Case Study & Final Review

Recitation 13 • Case Study review • Final Q&A

Last modified by Leman Akoglu, 2019

Carnegie Mellon University
HeinzCollege
INFORMATION SYSTEMS • PUBLIC POLICY • MANAGEMENT

Carnegie Mellon University
95-828 Machine Learning for Problem Solving
Spring 2024

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.
