Machine Learning – CS74020
Spring 2025 – GC CUNY

Week 01
Introduction to Machine Learning and Class Mechanics

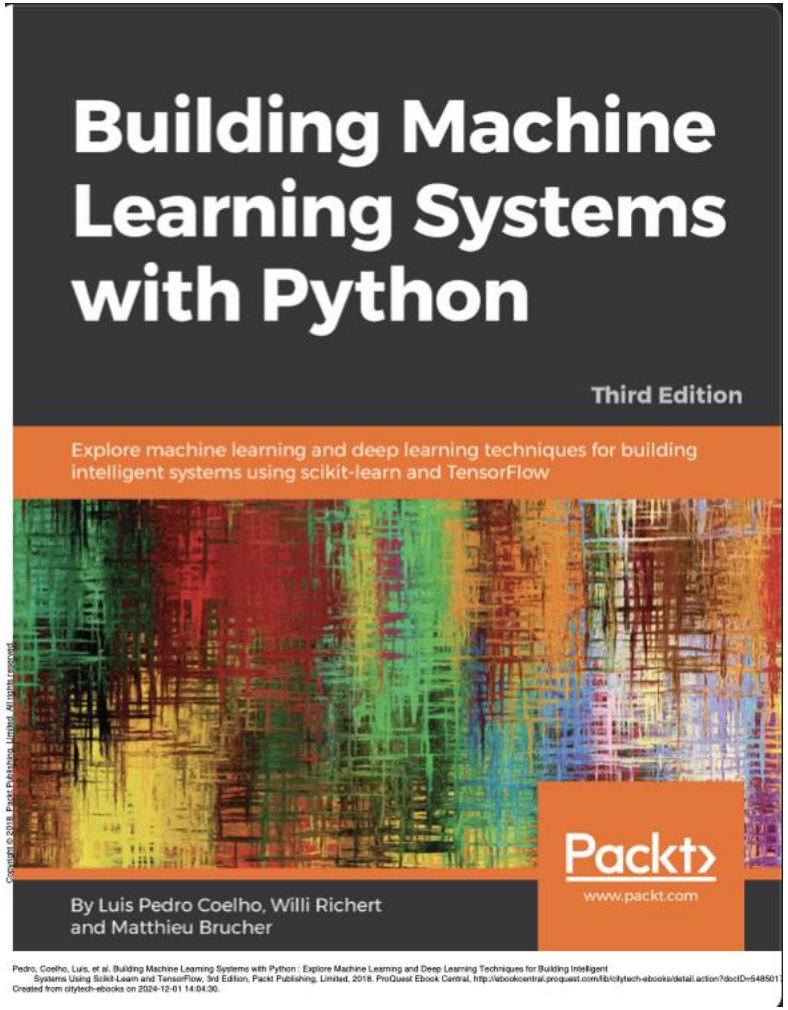
Pegah Khosravi, PhD
Assistant Professor of Biomedical AI
New York City College of Technology (City Tech)
Faculty Member, Biology and Computer Science, CUNY Graduate Center

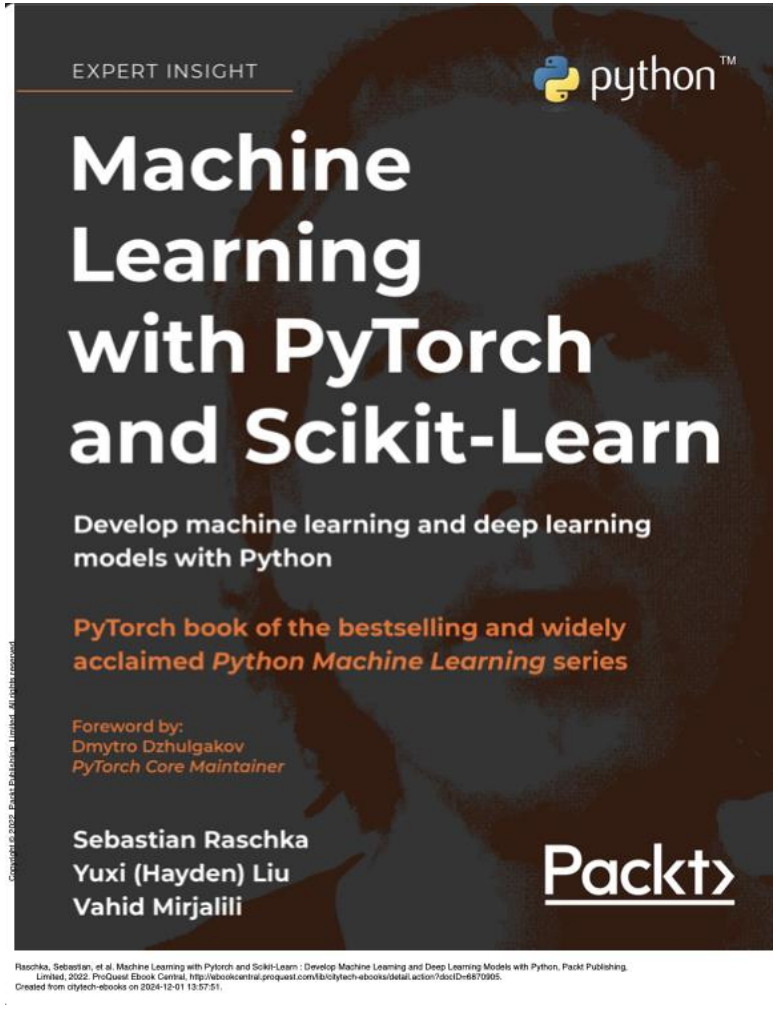
CUNY GRADUATE CENTER

In this session, we will cover:

- Class Mechanics
- The three main types of Machine Learning (ML):
  - Supervised Learning
  - Unsupervised Learning
  - Semi-supervised Learning
- Generative vs. Discriminative models
- The basic workflow of ML projects
- Hands-on: Loading and exploring a dataset using Python

Textbooks


Building Machine Learning Systems with Python
Third Edition
Explore machine learning and deep learning techniques for building intelligent systems using scikit-learn and TensorFlow
By Luis Pedro Coelho, Willi Richert and Matthieu Brucher
Packt
www.packt.com


EXPERT INSIGHT
python
Machine Learning with PyTorch and Scikit-Learn
Develop machine learning and deep learning models with Python
PyTorch book of the bestselling and widely acclaimed Python Machine Learning series
Foreword by:
Dmytro Dzhulgakov
PyTorch Core Maintainer
Sebastian Raschka
Yuxi (Hayden) Liu
Vahid Mirjalili
Packt

Grading Policies

ASSIGNMENT		DESCRIPTION	POINTS
Lab - Lecture	In class participation	Active participation in coding and analysis	15%
	Final Project	Comprehensive project due at the end of the course	15%
	Quizzes	Two quizzes assessing course topics	20%
	Exam 1	Midterm: Covers material from the first half of the course	25%
	Exam 2	Final: Covers material from the entire course	25%
Total			100%

Important Dates

- Class: In Person – GC6421 – Wednesdays 2 to 4 PM EST
- Quiz 1: Week 4 – 03/05
- Midterm: Week 7 – 03/19
- Quiz 2: Week 11 – 04/23
- Final Project: Week 14 – 05/14
- Final Exam: Week 15 – 05/21

What is Supervised Learning?

The model learns from labeled data (input-output pairs): Supervised learning is a core paradigm in ML, where the model is trained on a labeled dataset. Each training example consists of an input-output pair where:

- Input (X): A set of features (vectors) representing the data
- Output (Y): Corresponding target values (labels), which can be categorical or continuous

- Examples:
  - Predicting house prices (Regression)
  - Classifying emails as spam or not (Classification)

- Key Characteristics:
  - Training Data
  - Loss Function
    - Mean Squared Error (MSE) for regression
    - Cross-Entropy Loss for classification
  - Optimization: Techniques like Stochastic Gradient Descent (SGD)
  - Evaluation Metrics
    - Classification: Accuracy, Precision, Recall, F1-score, ROC-AUC
    - Regression: Mean Absolute Error (MAE), Root Mean Squared Error (RMSE)

What is Unsupervised Learning?

The model identifies patterns in unlabeled data: The primary goal of unsupervised learning is to understand the underlying structure of the data, group similar data points, and reduce complexity where needed.

- Examples:
  - Clustering: K-Means, Hierarchical Clustering
  - Dimensionality Reduction: PCA, t-SNE

- Key Characteristics:
  - No Labels:
    - The dataset lacks predefined labels or target values.
    - The model explores the data's intrinsic patterns without guidance
  - Applications:
    - Clustering: Grouping data into similar clusters based on their features
    - Dimensionality Reduction: Simplifying datasets while retaining the most important information
  - Evaluation:
    - Metrics like silhouette score (for clustering) and explained variance (for dimensionality reduction) are commonly used.

What is Semi-supervised Learning?

Combines a small amount of labeled data with a large amount of unlabeled data: The central idea is that labeled data can guide the learning process, while the unlabeled data helps capture the broader structure of the dataset.

- Examples:
  - Label Propagation
  - Self-training
  - Co-training
  - Graph-based Method

- Key Characteristics:
  - Small Labeled Dataset
  - Large Unlabeled Dataset
  - Applications:
    - Medical imaging
    - Text classification
    - Speech recognition
  - Evaluation

Generative vs. Discriminative Models

- Generative Models and Discriminative Models are two types of machine learning models, but they focus on different tasks:

- Think of a detective solving a case:
  - Generative Model (storyteller): The detective tries to understand the full story of how the crime happened, recreating every detail (generating all possibilities).
  - Discriminative Model (decision-maker): The detective focuses only on the evidence at hand to decide whether the suspect is guilty or not (making a classification).

Aspect	Generative Models	Discriminative Models
Goal	Model P(X,Y)	Model P(Y|X)
Uses	Data generation, unsupervised learning	Classification, supervised learning
Examples	GANs, Naive Bayes, VAEs	Logistic Regression, SVMs, Neural Networks
Focus	How data is generated and structured	How to separate or classify data effectively

Generative Models

What they do?
Generative models aim to model the joint probability distribution P(X, Y) or the marginal probability P(X). This means they learn the structure of the data itself and how input features X and labels Y are related. By capturing this distribution, they can generate new plausible data samples.

Why they're useful?
They can generate new samples of data that resemble the original dataset. For example:

- A generative model trained on images of cats can create entirely new, realistic-looking images of cats.
- Generative models can also model uncertainty, learn variations in data, and fill in missing information.

Examples:

- **Naive Bayes**: A simple generative model for classification.
- **GANs (Generative Adversarial Networks)**: Used to create realistic images.
- **Variational Autoencoders (VAEs)**: Used for data generation and reconstruction.

Discriminative Models

What they do?
Discriminative models focus on modeling the conditional probability distribution P(Y|X). Instead of learning the full data distribution, they focus on drawing decision boundaries between classes, making them better suited for classification and prediction tasks.

Why they're useful?
They are designed to efficiently separate or classify data and are often more accurate for tasks like supervised learning. Unlike generative models, they do not generate new data but focus on making precise classifications.

- For example, a discriminative model trained on cat and dog images will classify a new image as either a cat or a dog rather than generating a new cat image.

Examples:

- **Logistic Regression**: Classifies data into categories based on features.
- **Support Vector Machines (SVMs)**: Finds the best boundary to separate data.
- **Neural Networks (e.g., CNNs, RNNs)**: Excellent for tasks like image or text classification.

Machine Learning Workflow

A typical ML project follows these steps:

1. **Data Collection:** Gather data relevant to the problem.
2. **Preprocessing:** Clean and prepare the data.
3. **Model Training:** Use algorithms to find patterns in the data.
4. **Model Evaluation:** Assess performance on unseen data.
5. **Deployment:** Use the model in a real-world application.