STATGR5244/GU4244: Unsupervised Learning

YoonHaeng Hur
Columbia University

Lecture: Tuesdays and Thursdays, 8:40 - 9:55 AM
Location: TBA
Instructor: YoonHaeng Hur (yoonhaeng.hur@columbia.edu)
Office hours: TBA
Teaching Assistant: TBA
Course Description: Unsupervised Learning is a masters level course on foundations, methods, practice, and applications in machine learning from data without associated labels or outcomes. This course will focus on dimension reduction and clustering techniques, *including this year's selected topic on generative models. Unlike previous years, this year's focus will be on rigorous mathematical foundations of unsupervised learning and data science.* The course will also emphasize conceptual understanding and practical applications of unsupervised learning in data visualization, exploratory data analysis, data pre-processing, and data-driven discovery using Python programming.
Prerequisites:

STAT GR 5206 Statistical Computing and Intro to Data Science
STAT GR 5241 Statistical Machine Learning (strongly recommended)
STAT GR 5205 Linear Regression (recommended)
STAT GR 5203 Probability (recommended)
Students should also be familiar with linear algebra

Learning Outcomes: Students completing this course will learn mjaor unsupervised learning techniques with a strong emphasis on mathematical foundations and practical applications using Python programming. Students will learn (i) linear dimension reduction techniques such as principal component analysis (PCA) and extensions, (ii) manifold learning techniques such as spectral embedding, t-SNE, UMAP, and autoencoders, (iii) clustering techniques such as k-means, spectral clustering, hierarchical clustering, (iv) mixture models and the expectation-maximization (EM) algorithm, and (v) generative models such as generative adversarial networks (GANs), normalizing flows, diffusion models, and flow matching.
Requirements and Grading: This course requires active participation in class, completion of in-class quizzes, and two exams (midterm and final). Grades will be decomposed as follows:

- In-class quizzes: 30%
- Midterm exam: 40%
- Final exam: 30%

In-class quizzes will be given at the end of the class and will consist of 1-2 short questions based on the material covered in that day's lecture, which will take no more than 10 minutes to complete; see below for the tentative course and quiz schedule, which is subject to change. The midterm exam will be held during class time on October 22, 2024, and the final exam will be held during the final exam period (TBA). All quizzes and exams will be open-book and open-notes but no digital devices will be allowed.
Recommened Textbooks:

Deep Learning: Foundations and Concepts (2024), Christopher M. Bishop and Hugh Bishop
- Chapters 2-3 for basic probability
- Chapter 15 for clustering, mixture models, and EM algorithm
- Chapter 16 for PCA and dimension reduction

- Chapters 17-20 for generative models and autoencoders

Pattern Recognition and Machine Learning (2006), Christopher M. Bishop
- Chapter 2 for basic probability
- Chapter 9 for clustering, mixture models, and EM algorithm
- Chapter 12 for PCA

Probabilistic Machine Learning: An Introduction (2022), Kevin P. Murphy
- Chapters 2-3 for basic probability
- Chapter 20 for dimension reduction
- Chapter 21 for clustering and mixture models

An Introduction to Statistical Learning: With Applications in Python (2023), Gareth James, Daniela Witten, Trevor Hastie, Robert Tibshirani, and Jonathan Taylor
- Chapter 12 for PCA and clustering

Mathematical Foundations for Data Analysis (2021), Jeff M. Phillips
- Chapter 1 for basic probability
- Chapter 3 for basic linear algebra
- Chapter 7 for dimension reduction
- Chapter 8 for clustering

Topics and Tentative Course Schedule: The schedule before and after the Thanksgiving break is to be determined based on the progress of the course. The topics and tentative schedule are as follows (subject to change with reasonable advance notice by the instructor):

Date	Topic	Notes
Sep 8	Intro + Basic Probability	
Sep 10	Basic Probability and Linear Algebra	Quiz 1
Sep 15	Linear Algebra and PCA I	Quiz 2
Sep 17	Linear Algebra and PCA II	Quiz 3
Sep 22	Dimension Reduction and Manifold Learning I	
Sep 24	Dimension Reduction and Manifold Learning II	Quiz 4
Sep 29	Dimension Reduction and Manifold Learning III	
Oct 1	Dimension Reduction and Manifold Learning IV	Quiz 5
Oct 6	Clustering and Mixture Models I	
Oct 8	Clustering and Mixture Models II	Quiz 6
Oct 13	Clustering and Mixture Models III	
Oct 15	Clustering and Mixture Models IV	Quiz 7
Oct 20	Clustering and Mixture Models V	Quiz 8
Oct 22	Midterm Exam	
Oct 27	Generative Models I	
Oct 29	Generative Models II	Quiz 9
Nov 3	No class (Election Day)	
Nov 4	Generative Models III	
Nov 10	Generative Models IV	Quiz 10
Nov 12	Generative Models V	
Nov 17-19		
Nov 24-26	No class (Thanksgiving)	
Dec 1-3		