DATA SCI 415: Syllabus and Course Outline

Snigdha Panigrahi

451 West Hall
psnigdha@umich.edu

Team: Instructor

Snigdha Panigrahi

- Lecture: Mon, Wed 2:30-4pm, UMMA Aud
- Office hour: Mon 9:30-10:30am, 451 West Hall

Team: GSIs

Alexander Kagan

- Lab, Tues 10:00-11:30
- Office hour: TBA

Samuel Rosenberg

- Lab, Tues 11:30-1.00
- Office hour: TBA

Material

- Textbook: James, Witten, Hastie and Tibshirani (2015) An Introduction to Statistical Learning. Springer.
- Lecture notes; Use textbook as a supplement to the Lectures
- Lab Assignments, Homework Problem Sets
- Software: Python

Assessment

- 5 (long) homework problem sets, 2 exams, and regular lab assignments
- First exam: Oct 22, 2:30-4pm, UMMA Aud
- Second exam: Dec 10, 4.00-6.00pm, UMMA Aud
- Homework 10%, labs 20%, first exam 30%, second exam 40%

Academic Integrity

- A random subset of the assigned homework problems will be graded.
- Similar questions might also appear on your graded lab assignments!!
- Think of the homework as extra practice for your labs—completing it on your own will help you perform better on the graded lab assignments.
- If you use external sources, you must cite and credit them. Otherwise, you get no credit.
- No late homework.

Homework

- Homework will be submitted electronically through Canvas as a pdf
- Jupyter notebooks for code are a part of the submission
- Posted and due on Fridays

Exams

- Exams are closed book and do not involve a computer
- You are allowed to bring one standard size sheet of paper, writing whatever you want on both sides, and a calculator
- Exams will not involve any coding, though they may require understanding code snippets
- 1 Practice Exam will be provided before each Exam

Labs

- Labs will be instructed by your GSIs
- First Lab: Sept 2
- Bear in mind! Labs will be graded and assignments must be done in Labs.

Discussion beyond lectures and OHs

- Piazza policy: GSIs will rotate
- Stay tuned for more on this policy: The GSIs will announce this during the Labs

Syllabus

- Nature of data and Exploratory data analysis
- Regression: Linear Methods
- Classification
  - Logistic regression, LDA and QDA
- Regression and Classification: Non linear methods
  - GAMs, Splines
  - Tree-based methods
  - Ensemble methods
- Dimension reduction
- Unsupervised learning: Clustering, PCA
- Deep Learning

Prerequisites

- Multivariate calculus (MATH 215)
- Linear algebra (MATH 214 or MATH 217)
- At least one upper-level statistics course (e.g., STATS 401, STATS 412, STATS 425, STATS 426)
- Some programming knowledge

Relevance: in the age of Gen AI

- Gen AI models can generate vast amounts of data quickly
- But, they are limited in their abilities to extract knowledge from data
- Data scientists can understand patterns and trends and make insightful predictions and inferences based on this data or outputs of Gen AI models

Relevance: in the age of Gen AI

- Gen AI models are only as good as the data they are trained on
- If the data contains biases, these models will reflect these biases in downstream results
- Data scientists can understand and correct for biases in the data: help analyze data with their critical thinking skills!!!

Relevance: in the age of Gen AI

- While Gen AI models are powerful, they are still limited in their abilities to solve complex problems: you need to combine them with statistical reasoning and other modeling tools to make sense of them
- Data scientists can understand and improve Gen AI models
- Task at hand is especially even more important if these models get used in high-stakes applications like healthcare, where there is very little room for errors

Relevance: in the age of Gen AI

Understanding of Statistical Learning/ Machine Learning concepts and algorithms are important

For example, statistical concepts such as bias-variance tradeoff carry over to all of machine learning, both old and new.