Statistics, Machine Learning, and Data Mining

Syllabus MA322 Section OL1 Fall 2024

Info

3 Credits
Room: None, Since OL course

Instructor Information

calvin_williamson@fitnyc.edu
office: B831 Science and Math
office hours: T 5-6, W 11-12, R 10-12

Description

This is an introduction to statistical techniques for machine learning and data mining. It emphasizes mathematical methods and computer applications related to automated learning for prediction, classification, knowledge discovery and forecasting in modern data science. Special emphasis will be given to the collection, mining, and analysis of massive data sets. (G2: Mathematics) Prerequisite(s): MA 222 and mathematic proficiency (see beginning of Mathematics section)

Outcomes

Upon completion of this course students will be able to:

1. Describe the concepts of machine learning and identify examples of its use in data science.
2. Employ statistical software to collect data, create training and test sets, and perform predictions.
3. Create regression models for predicting outcome variables in terms of predictors.
4. Explain the contributions of Google in understanding web scale data and the structure of the internet.
5. Identify the characteristics of massive data sets and describe the tools needed to analyze them.
6. Analyze decision tree models and display them with appropriate graphics.
7. Use recommendation systems software and understand how it makes suggestions based on similarity measures.
8. Perform classifications for data sets using nearest neighbor and probabilistic algorithms.

9. Collect text data and use text mining software to perform sentiment analysis.

Course Materials

Textbook

Some readings are from an OER textbook that is free. No other textbook is required.

Software

We will be using Google Spreadsheets, Google Colab Notebooks and Datacamp for all work in this course. Since these are web-based applications that are free and so there is NO OTHER SOFTWARE required for the course besides a web browser.

Topics

Regression

- Simple Regression
- Multiple Regression
- Applications
- Conjoint Analysis

Introduction to Python

- Google Colab Notebook
- Using LLM as Coding Assistant
- Calculations
- Variables
- DataTypes
- Lists
- Dictionaries
- Functions
- Dataframes
- f-Strings

Introduction to Large Language Models (LLMs)

- LLM Examples
- ChatGPT, GPT-4o, Gemini, Claude, Mistral
- Completions, APIs

Prompt Engineering

- Prompting
- Prompt Chaining
- Roles and Personas
- Chain of thought
- Few-shot and zero-shot Learning

Machine Learning

- Classification, Accuracy
- Training, Testing
- Decision Trees

Evaluation

Your grade will come from a set of assignments (one per week) due on Fridays and some Datacamp courses you complete.

- 15 Spreadsheet Assignments, roughly one per week (94%)
- DataCamp Courses (6%)

Spreadsheet Assignments (94%) (using Google Spreadsheets)

- For each module you will have to complete some problems using a Google Spreadsheet which you share with me.
- Due dates are every Friday at 11:59 EST PM
- There are no late submissions allowed for any reason. Any work time stamped as occuring after the due dates on each problem cannot be counted for any credit.
- For each assignment I will choose only some parts to grade, but you will not know which parts ahead of time are graded so you need to do all the assignments. Share the document or assignment with me as soon as you start the module. This process will be explained in the first assignment.

DataCamp Courses (6%)

This course will use some material at datacamp.com, a website for learning data science. You will have free login to the site for 6 months beginning the first day of our course. You will be completeing 2 or 3 required courses in datacamp for topics related to our course. But you can also take any of the courses in datacamp for free, related to our course or not. There are many courses at all levels from beginner to more advanced about data science, programming, AI, etc.

AI Policy

You may use any AI tool (ChatGPT, Gemini, Claude, and others) to work on material for this course. However keep in mind the accuracy of these tools for mathematics and statistics is still in question, with some AI better than others. It is beneficial to understand the limitations and be comfortable with working with AI, so you are encouraged to use these tools and evaluate critically how much they assist you. If they keep you from understanding what is really going on, you will have problems on the work in the course. So use with caution.

Course Policies

Modules

This course is organized into 7 modules:

Each module lasts two week.s A module becomes available on the first day it is assigned. (See the course schedule for the dates.) When a module is finished it will remain open so you can refer to it but you will not be able to do further work in that module.

There is no way to make-up any module work once a module is finished, so stay up-to-date with the modules, otherwise you will lose the credit for work in that module.

Module Activities - Overview

For each week you will follow essentially the same activities listed below:

1. Watch a demo video from the instructor talking about the topics and techniques
2. Look at any accompanying documents or references
3. Do the assignments