Syllabus for Mathematics of Data Science

AS.110.205, Spring 2024

Instructor Information

Professor Ratigan

Email: cratiga1@jhu.edu this is by far the best way to contact me.

Zoom: https://jhu-cty.zoom.us/my/cratigan Passcode: 918057

Office Hours: TBA, otherwise by appointment

Course Materials

The course will require internet access for Zoom office hours and accessing the course website. Additionally, the course will require the student to work with python notebooks through Jupyter Notebooks using a miniconda distribution and requires the textbook: *An Introduction to Statistical Learning with applications in Python* (available for free online).

Course Schedule

The course is thirteen weeks broken up into the following content, note the readings will skip around a bit in the textbook.

Week	Topic	Book Chapters
Week 1	Introduction to Python programming: basic datatypes, list comprehension, slicing, control flow, recursion, etc.	N/A, various sites will be used for practice
Week 2	Introduction to Pandas and Matplotlib for data cleaning, analysis and visualization; Exploration of some Classic Datasets.	Ch 2
Week 3	Feature engineering, performance metrics, Intro to Regression	Parts of Ch 3
Week 4	Advanced Regression Models; train_test_split/cross_validation; Project 1: Modeling Housing Prices	Parts of Ch5, 7 & 8
Week 5	Introduction to Probability using statsmodels; Law of Large Numbers, Central Limit Theorem.	Ch 4
Week 6	Introduction to Classification Problems and confusion matrices	Ch 4
Week 7	Bayesian Statistics and baseline models.	Ch 5
Week 8	Ensemble Methods; Project 2: Classification	Ch 8
Week 9	Intro to unsupervised learning/clustering	Ch 12
Week 10	PCA and Dimensionality Reduction	Ch 12
Week 11	Intro to Neural Networks	Ch 10
Week 12	Working with image, text and time series data	Other Sources
Week 13	Bias-Variance Tradeoff, ROC Analysis, start final project.	TBA
Week 14	Final Project Due	Depends on project

Online Lectures

Prerecorded lectures will be posted for the week in each module togeher with iteractive python notebook (.ipynb) files for you to follow along in. Installation instructions for using .ipynb files can be found in the course information module and will be handled during the orientation session.

There will be live synchronous sessions throughout the semester through zoom, with dates posted in the course. During these live synchronous components, homework problems and project review will take place. Links to the recordings of each live online sessions will be posted in Canvas and emailed to the class.

For more information regarding Zoom, please see the Zoom Student Quick Start Guide.

Course Description

This course is designed for students of all backgrounds to provide a solid foundation in the underlying mathematical, programming, and statistical theory of data science. In today's data driven world, it is increasingly important to be data literate. To this end, the course will motivate the fundamental concepts used in this growing field. While discussing the general theory behind common methods of data analysis, there will be numerous applications to real world data sets. In particular, the course will use Python libraries to create, import, and analyze data sets. There are no mathematical prerequisites for this course, though prior knowledge of statistics and/or programming can be helpful.

Course Goals

This course will prepare you to understand the opportunities, challenges and limitations of data science and machine learning. By teaching you mastery in the techniques, theory, and applications of the data science process this course will serve as a stepping stone to navigating real world issues in our AI driven world.

Course Objectives

By the conclusion of this course, you are expected to have gained the ability to

- Understand the theory underlying modern data science
- Fit and train a model to solve regression and classification problems
- Use various metrics to evaluate the effectiveness of a model
- Understand the distinction between regression and classification, supervised and unsupervised learning.
- Communicate effectively regarding real world data.

Course Textbooks

Required: *An Introduction to Statistical Learning* by Gareth James et al. (available for free online)

Optional: *The Grammar of Graphics* by Leland Wilkinson ISBN 0387245448 (recommended)

Student Coursework Requirements

Homework

Will be assigned each week as Python Notebook problem sets, these are to be completed and then uploaded to Canvas for review.

Projects

There will be three projects applying the skills students developed to actual datasets. The first project will have students analyze a well-known dataset, the final project will have the students work on a modeling problem of their choice.

Quizzes

There will be a short (20 minute) online quiz each week. There are two attempts at the quiz and the questions are pulled randomly during each attempt. The higher of the two attempts counts for the grade. Your lowest quiz grade will be dropped.

Discussion Forums

There will be weekly discussion forums to give students the opportunity to read and write about the material in the course.

Grading

The final grade for the class will be given as a weighted average as follows.

Homeworks: 25% (lowest dropped)

Project 1: 15%

Project 2: 15%

Project 3: 15%

Quizzes: 20% (lowest dropped)

Discussion Forums: 10%

The letter grades are assigned as follows based on your final weighted average:

A: 90-100

B: 80 - 89

C: 70 - 79

D: 63 - 69

F: < 63

"+" and "-" will be determined at the end of the semester.

Students with disabilities

Students with documented disabilities or other special needs who require accommodation must register with Student Disability Services. After that, remind the instructor of the specific needs at least one week prior to each exam; the instructor must be provided with the official letter stating all the needs from Student Disability Services.

There may be a student in this class who requires the services of a note taker. This is an opportunity to share notes through the Student Disability Services Office. If you are interested in performing this service, please register as a notetaker with Student Disability Services.

Academic Integrity

Academic Misconduct Policy
All students are required to read, know, and comply with the Johns Hopkins University Krieger School of Arts and Sciences (KSAS) / Whiting School of Engineering (WSE) Procedures for Handling Allegations of Misconduct by Full-Time and Part-Time Graduate Students.
This policy prohibits academic misconduct, including but not limited to the following: cheating or facilitating cheating; plagiarism; reuse of assignments; unauthorized collaboration; alteration of graded assignments; and unfair competition. You may request a paper copy of this policy at this by contacting jhep@jhu.edu.

JHU ethics statement

"Undergraduate students enrolled in the Krieger School of Arts and Sciences or the Whiting School of Engineering at the Johns Hopkins University assume a duty to conduct themselves in a manner appropriate to the University's mission as an institution of higher learning. Students are obliged to refrain from acts which they know, or under circumstances have reason to know, violate the academic integrity of the University. [The JHU Code of Ethics]"

Ethical violations include cheating on exams, plagiarism, reuse of assignments, improper use of the Internet and electronic devices, unauthorized collaboration, alteration of graded assignments, forgery and falsification, lying, facilitating academic dishonesty, and unfair competition.