Elements of Data Science: A First Course Fall 2023

Time: Monday 7:00pm - 9:30pm
Instructor: Andi Cupallari, PhD
Email: ac5562@columbia.edu

Textbook: Python Data Science Handbook by JakeVanderPas
Machine Learning with PyTorch and Scikit-Learn by Raschka, Liu and Mirjalili

Prerequisite(s):

- Introductory programming class as well as basic familiarity with Python 3.
- Basic familiarity with the command line.

Course Description

This course is designed as an introduction to elements that constitute the skill set of a data scientist. The course will focus on the utility of these elements in common tasks of a data scientist, rather than their theoretical formulation and properties. The course provides a foundation of methodology with applied examples to analyze large engineering, business, and social data for data science problems. Hands-on experiments with Python will be emphasized.

The amount of material covered in class is meant to introduce you to the topic and make you familiar with the problems you will face when working with data. Industry experts (guest speaker) will join us to share their experience in at least one of the classes.

The midterm and final will be in class.

Homework assignments are take-home, due two weeks after they are announced. No extension to be provided.

Topics include:

- Python Data Science Tools
- Data Cleaning, Exploration and Visualization
- Hypothesis Testing and Statistical Modeling
- Classification, Regression and Clustering
- Dimensionality Reduction and Topic Modeling
- Model Evaluation and Model Selection
- Feature Engineering and Feature Selection
- Natural Language Processing
- Data processing and delivery using ETL and APIs
- Dealing with Time Series Data

Assignments and Grading

Weekly Quiz	10%
Homework Assignments (Four, equally weighted at 10% each)	40%
Midterm Exam	25%
Final Exam	25%
TOTAL	100%

Quality of Performance	**Letter Grade**	**Range %**	**GPA/Quality Pts.**
Excellent - work is of exceptional quality	A+	99 - 100	4.33
	A	93 - 98.99	4.0
	A-	90 - 92.99	3.67

Good - work is above average	B+	87 - 89.99	3.33
Satisfactory	B	83 - 86.99	3.0
Below Average	B-	80 - 82.99	2.67
Poor	C+	77 - 79.99	2.33
	C	73 - 76.99	2.0
	C-	70 - 72.99	1.67
	D	65 - 69.99	1.0
	D-	60 - 64.99	0.67
Failure	F	< 60	0.0

Weekly Outline

Week	**Topic**
Sep 11	Introduction to Data Science Problems and Tools
Sep 18	Python and Numpy
Sep 26	Pandas, Visualization and Data Exploration
Oct 2	Hypothesis Testing
Oct 9	Intro to Machine Learning
Oct 16	Midterm
Oct 23	Machine Learning Models
Oct 30	Model Evaluation and Selection
Nov 6	*Academic Holiday*
Nov 13	Data Cleaning and Feature Engineering
Nov 20	Joining Data, Dimensionality Reduction and Imbalanced Classes
Nov 27	NLP, Sentiment Analysis and Topic Modeling
Dec 4	Clustering and Recommendation Systems
Dec 11	Timeseries, Data Processing and Delivery/ Final exam Q&A
Dec 18	Final