ECE 4710J: Introduction to Data Science

Ailin Zhang (ailin.zhang@sjtu.edu.cn)

Spring, 2022

Course Description

Data science is a combination of data, computation and analytical thinking, and it is redefining processes in problem solving and decision making. In this class, we will explore key areas of data science including question formulation, data collection and cleaning, visualization, statistical inference, predictive modeling, and decision making.

The course puts a strong emphasis on solving real-world data driven problems. To be more specific, the course will cover languages for transforming, querying and analyzing data; algorithms for machine learning methods including regression, classification and clustering; principles behind creating informative data visualizations; and statistical concepts of measurement error and prediction.

Prerequisites

While we are not enforcing prerequisites during enrollment, it is strongly recommended that you have basic understanding/ knowledge of the following aspects. Furthermore, all of the prerequisites will be used starting very early on in the class/ homework.

- Foundations of Math and Statistics

  Linear algebra, probability and statistics are essential. We will need some basic concepts like linear operators, eigenvectors, derivatives, and integrals to enable statistical inference and derive algorithms.

- Computing

  We will use python as the computing language for teaching and homework. You need to be familiar with python programming (e.g., for loops, lambdas, debugging, and complexity)

  You can use the following tutorial to pick up your python skill.

  General Python: `https://docs.python.org/3.9/tutorial/index.html`
  Numpy and Pandas: `https://cs231n.github.io/python-numpy-tutorial/`

Grading Policy

The typical JI grading scale will be used. I reserve the right to curve the scale if there are less than 30% of students with grades ≥ A. The grade will count the assessments using the following proportions:

- **30%** Homework (5-7 submissions)
- **20%** Project
- **20%** Midterm
- **30%** Final
- **3%*** Extra Credit

Course Agenda and Timeline

The agenda is tentative and subject to change. The bullet points are key concepts you should grasp after each week, and also as a study guide before exams.

Week 01 Recap and Fundamentals

- Introduction
- Sampling and Probablity

Week 02 Estimation and Bias

- Estimators and Bias
- Jupyter notebook

Week 03 Data Acquisition and Manipulation

- Sampling
- Randomness

Week 04 Data Manipulation

- Pandas
- Regex

Week 05 Data Preprocessing

- Data cleaning

Week 06 Data Preprocessing

- Data visualization (matplotlib, seaborn)

Week 07 Modeling and **Midterm**

- General overview of modeling

Week 08 Feature Engineering

- Feature generation
- KDE

Week 09 Regression

- Linear regression
- Ordinary Least Squares

Week 10 Bias and Variance

- Regularization
- Gradient descent

Week 11 Classification

- Logistic regression
- Model Evaluation

Week 12 Classification

- Decision Tree/ Random Forest
- Boosting

Week 13 Unsupervised learning

- PCA
- Clustering

Week 14 Clustering and Review

- **Project due**
- Review for final

We want you to succeed!

If you are feeling overwhelmed, visit our office hours and talk with us, and we want to help you succeed.