ENGR 5963 Data Centric AI and Visualization

5000 level 2 Semester Hour Course

Course Description

Garbage-In Garbage Out (GIGO) may be the most widely used maxim in machine learning, but how does one assess the quality of data in an analysis pipeline? In this course students will learn the theory and practice of assessing the quality of a dataset, detecting and fixing issues with the data and visualizing important relationships. Class sessions are broken into two parts, the first half covering theory and the second half a hands-on lab in python.

The first week of the course covers the statistical analysis of data. The cleaning, sampling, imputation, and normalization of data is covered. Data reduction techniques and the creation of synthetic data will also be covered.

The second week covers graphical and modeling techniques for exploring data, with an emphasis on visualization, interpretation, and clear communication of findings. The choice of the right chart for a particular question is covered. The principles of visual design, including typography, contrast, balance, emphasis, movement, white space, proportion, hierarchy, repetition, rhythm, pattern, unity, and variety are covered.

Learning Objectives

Learning objectives for the course are:

- Descriptive statistics
- Probability distributions
- Imputing data
- Normalizing and scaling data
- Data reduction
- Sampling, bootstrapping and confidence intervals
- Pseudo-labeling
- Synthetic data
- Error analysis
- Data drift and concept drift
- Charts for comparing values
- Data visualization
- Exploratory data analysis (EDA)
- Compositional charts
- Distribution charts
- Charts for trends
- Charts for relationships
- Principles of visual design

Course Prerequisites

Programming experience with python. An intro to machine learning course would be useful.

Qualifications

PhD in Computer Science from UCLA with a minor field in statistics
MS in Information Design and Visualization from Northeastern University

Preferences

I prefer the classes meet in 3 hour sessions in the afternoons or late mornings.
Giving students the option of on-ground or remote through NuFlex makes the most sense to me.