Elements of Data Science - F22

Midterm Review

This is intended as a guide and is not guaranteed to be comprehensive.

Material considered fair for the exam is anything from class and slides.

Data Science Tools

- Data Science workflow
- Jupyter+Ipython Notebooks
- conda Virtual Environments
- using Git to pull code and materials

Python Intro/Review, Numpy, and Pandas

- Importing modules
- Defining functions
- String formatting
- What are Exceptions and how do we catch them?
- Using `assert`
- Basic Python data types
- Collections module: Counter, defaultdict
- Python flow control: `if: elif: else:` , `for x in xs:`
- Sorting with `lambda` functions as the key
- List Comprehensions
- Numpy
  - arrays
  - indexing/slicing
  - Boolean masks and bitwise operations
- Pandas
  - Series
  - DataFrames
  - indexing/slicing
  - `.loc[]`
  - `.iloc[]`
  - `.describe()`
  - `.info()`
  - `.shape`

Visualization and Data Exploration

- Matplotlib
  - plotting using `matplotlib`
  - using `plt.subplots()`
  - modifiying plots using `ax`

- Variable Types (numeric,categorical,ordinal)
- Central tendencies
  - mean
  - median
- Spread
  - variance
  - std deviation
  - IQR
- Correlation
  - Pearson Correlation Coefficient
- Univarite Plotting
  - histogram
  - boxplots
- Bivariate Plotting
  - scatterplot
  - barplot
  - jointplot
  - pairplot

Confidence Intervals and Hypothesis Testing

- Random Sampling vs. Population Distribution
- Sample Statistic
- Confidence Intervals
  - Bootstrap Sampling
- Normal (Gaussian) Distribution
  - Standard Normal Distribution
  - Z-Score
- Central Limit Theorem
- Hypothesis Testing
  - Type I and II error
  - Significance and Power
  - Permutation Tests
  - One-tailed vs Two-tailed
  - p-values
  - A/B Test
- Multi-Armed Bandit
  - benefits of using
  - greedy
  - epsilon-greedy

Intro to ML

- "Dimensions" of ML
  - Interpretation vs. Prediction
  - Learning Paradigms (SL,UL,etc.)

    - Regression vs Classification
    - Binary, Multiclass, Multilabel Classification
- sklearn common functions
    - `.fit()`
    - `.predict()`
    - `.predict_proba()`

Machine Learning Models

- Simple Linear Regression
    - Interpreting Coefficients of OLS
    - Colinearity
- Multiple Linear Regression
- Logistic Regression
- Concept of Gradient Descent
- Perceptron/Multilayer Perceptron
- k-Nearest Neighbor
- Decision Trees
- Ensembles
    - Random Forest
    - Gradient Boost
    - Stacking
- Multiclass, Multilabel and One vs. Rest Classification