Elements of Data Science - F21

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?
- 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
- Central tendencies
  - mean
  - median

- Spread
  - variance
  - std deviation
  - IQR
- Correlation
  - Pearson Correlation Coefficient
- Univarite Plotting
  - histogram
  - boxplots
- Bivariate Plotting
  - scatterplot
  - barplot
  - jointplot
  - pairplot

Hypothesis Testing

- Random Sampling vs Population Distribution
- Sample Statistic
- Confidence Intervals
- Normal (Gaussian) Distribution
  - Standard Normal Distribution
  - Z-Score
- Central Limit Theorem
- Bootstrap Sampling
- A/B Test
- Hypothesis Testing
  - Type I and II error
  - Significance and Power
  - Permutation Tests
  - One-tailed vs Two-tailed
  - p-values
- Calculating “How many observations?”
  - what 4 values are related?
- 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
  - Residuals in linear models
  - Interpreting Coefficients of OLS
  - Colinearity
- Multiple Linear Regression
- Logistic Regression
- Concept of Gradient Descent
- Perceptron/Multilayer Perceptron
- One vs. Rest for Multiclass/Multilabel
- k-Nearest Neighbor
- Decision Trees
- Ensembles
  - Random Forest
  - Gradient Boost
  - Stacking