Elements of Data Science - S20

Midterm Review

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

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

Note sheet allowed: 1 8x11in sheet, 2-sided

Data Science Tools

- Data Science workflow
- Jupyter+Ipython Notebooks
- conda Virtual Environments
- uses for Git

Python Intro/Review Numpy and Pandas

- Importing modules
- Defining functions
- String Formatting
- What are Exceptions?
- Using assert
- Basic Python data types
- Collections modlue: Counter, defaultdict
- Python flow control: if: elif: else: , for x in xs:
- Sorting with lambda functions
- List Comprehensions
- Numpy
  - arrays
  - indexing/slicing
  - Boolean masks and bitwise operations
- Pandas
  - Series
  - DataFrames
  - indexing/slicing
  - .describe

Visualization and Data Exploration

- Matplotlib
  - plotting using matplotlib
  - plt vs ax level plotting
  - using plt.subplots()
- Variable Types
- Central tendencies
  - mean
  - median
- Spread
  - variance

- std deviation
- skew
- IQR
- Correlation
  - Pearson Correlation Coefficient
- Univarite Plotting
  - hist
  - seaborn.distplot()
  - boxplots
- Bivariate Plotting
  - scatter
  - jointplot
  - pairplot
- Categorical Plotting
  - bar
  - catplot

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

Modeling, Prediction Model Evaluation and Selection

- Dimensions of ML
  - Interpretation vs Prediction
  - Learning Paradigms (SL,UL,etc.)
  - Regression vs Classification
  - Binary, Multiclass, Multilabel
- sklearn common functions
  - .fit
  - .predict
  - .predict_proba

- .score
- Generalization
  - Train/Test split
  - stratification
- Overfitting/Underfitting
  - Bias/Variance Tradeoff
- Baseline Models
- Tuning Hyperparameters and Model Selection
  - k-Fold Cross Validation
  - Grid Search
- Plotting Model Fit
  - Validation Curve
  - Learning Curve
- Metrics: Classification
  - Confusion Matrix
  - Accuracy/Error
  - Precision
  - Recall
  - Precision-Recall Curve
  - F1 Score
  - ROC Curve (FPR vs TPR)
  - ROC AUC
- Metrics: Regression
  - R^2
  - Adjusted R^2
  - Mean Squared Error
  - RMSE

Machine Learning Models

- Concept of Gradient Descent
- k-NearestNeighbor
- Naive Bayes (generally)
- Simple Linear Regression
  - Residuals in linear models
  - Interpreting Coefficients of OLS
  - Colinearity
- Multiple Linear Regression
- Logistic Regression
  - Interpreting Coefficients of LogReg
- Decision Trees
- Ensembles
  - Random Forest
  - Gradient Boost
  - Stacking