Elements of Data Science - F19

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, with handwritten notes

Sections below are grouped by which slides the concepts predominantly appear in.

Introduction to Data Science problems and tools

- Data Science workflow
- Jupyter+Ipython Notebooks
- conda Virtual Environments
- Basic Python data types and structures
- List Comprehensions
- Numpy
  - arrays
  - indexing/slicing
- Pandas
  - Series
  - DataFrames
  - indexing/slicing
  - .describe,.info

Data Exploration and Visualization

- Variable Types
- Central tendencies
  - mean
  - median
- Spread
  - variance
  - std deviation
  - skew
  - IQR
- Correlation
  - Pearson Correlation Coefficient
- Plotting
  - plt vs axis level plotting
  - Plotting real valued variables
    - histogram
    - scatter
    - sns.distplot
- Plotting categorical variables
  - bar
  - catplot
- Plotting interactions

    - jointplot
    - pairplot

Statistical Modeling and Hypothesis Testing

- Random Sampling vs Population Distribution
- Sample Statistic
- Central Limit Theorem
- Normal (Gaussian) Distribution
    - Standard Normal Distribution
    - Z-Score
- Confidence Intervals
- Bootstrap Sampling
- A/B Test
- Hypothesis Testing
    - Type I and II error
    - Significance and Power
    - Permutation Tests
    - p-values
- Calculating “How many observations?”
    - what 4 values are related?
- Multi-Armed Bandit
    - greedy
    - epsilon-greedy

Regression and Classification

- Concept of Gradient Descent
- Interpreting Coefficients of OLS
- Residuals
- Colinearity
- Regression Models
    - Simple Linear Regression
    - Multiple Linear Regression

Prediction and Sklearn

- Interpretation vs Prediction
- Classification vs Regression
- Regularization
    - Ridge (l2)
    - Lasso (l1)
    - ElasticNet
- sklearn common functions
    - .fit
    - .predict
    - .predict_proba
    - .score
- Classification models
    - (know models conceptually)
    - (know differences between models)
    - logistic regression

    - svm (very generally)
    - k Nearest Neighbor
    - decision trees
    - naive bayes (very generally)
    - neural networks (very generally)
    - ensembles
        - random forest
        - gradient boost
        - stacking

Model Evaluation and Selection

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