Elements of Data Science - F21

Final Review

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

Material considered fair-game for the exam is anything from class, readings 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
  - .info()
  - .describe()
  - .agg()
  - .groupby()

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
    - skew
    - IQR
- Correlation
    - Pearson Correlation Coefficient
- Univariate Plotting
    - histogram
    - boxplot
- Bivariate Plotting
    - scatterplot
    - 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
- 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
- One vs. Rest for Multiclass/Multilabel Classification
- k-Nearest Neighbor
- Decision Trees
- Ensembles
  - Random Forest
  - Gradient Boost
  - Stacking

After the Midterm

Model Evaluation

- Generalization
  - Train/Test split
  - stratification
- Overfitting/Underfitting
  - Bias/Variance Tradeoff
- Baseline/Dummy Models
- Tuning Hyperparameters and Model Selection
  - k-Fold Cross Validation
  - Grid Search
- Metrics: Classification
  - Confusion Matrix
  - Accuracy/Error
  - Precision
  - Recall
  - F1 Score
  - ROC Curve
  - ROC AUC
- Metrics: Regression
  - $R^2$
  - Mean Squared Error
  - RMSE
  - Adj-$R^2$
- Regularization
  - Ridge
  - LASSO
  - ElasticNet

Data Cleaning

- Duplicates
- Missing Data
- Dummy Variables
- Rescaling
- Dealing With Skew

- Removing Outliers

Feature Engineering

- Binning
- One-Hot encoding
- Derived Features

Dimensionality Reduction

- Feature Selection
  - LASSO
  - Tree Based Models Feature Importance
  - Univariate Tests
  - Recursive Feature Selection
- Feature Extraction
  - PCA

NLP and Topic Modeling

- What is a corpus?
- Tokens and Tokenization
- Vocabulary
- Bag Of Words representation
- n-grams
- Term Frequency
- Document Frequency
- Stopwords
- TfIdf
- Latent Dirichlet Allocation (general concept)
  - per document topic distribution
  - per topic term distribution

Clustering

- k-Means
- Hierarchical Agglomerative Clustering
  - linkage

Recommendation Engines

- Content-Based Filtering
- User-Based Collaborative Filtering
- Issues
- Evaluating

Timeseries

- unique characteristics of timeseries data
- timeseries in pandas
- indexing with a DatetimeIndex
- converting data to datetime with pd.to_datetime()
- Shifting
- Resampling/Frequencies

- Upsampling vs Downsampling
- Moving/Rolling Window functions

Dealing with Imbalanced Data

- Random Oversampling minority class
- Random Undersampling majority class
- SMOTE and ADASYN (general concept)

Data Processing (ETL and API)

- Different filetypes handled by pandas
- sklearn Pipelines
    - ColumnTransformer
- What can we use the flask python library for?

SQL

- Relational Databases (Normalization/Denormalization)
- SQL
    - SELECT
    - LIMIT
    - WHERE
    - ORDER BY