Elements of Data Science - F22

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 with a focus on material covered after the midterm.

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[]`
  - `.info()`
  - `.describe()`
  - `.shape`
  - `.agg()`
  - `.groupby()`

Visualization and Data Exploration

- Matplotlib

    - plotting using `matplotlib`
    - using `plt.subplots()`
    - modifiying matplotlib axes using `ax`
- Variable Types (numeric,categorical,ordinal)
- Central tendencies
    - mean
    - median
- Spread
    - variance
    - standard deviation
    - IQR
- Correlation
    - Pearson Correlation Coefficient
- Univariate Plotting
    - histogram
    - boxplot
- 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 Tests
- Multi-Armed Bandit
    - benefits of using
    - greedy vs. 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
- k-Nearest Neighbor
- Decision Trees
- Ensembles
  - Random Forest
  - Gradient Boosting
  - Stacking
- Perceptron/Multilayer Perceptron
- Multiclass, Multilabel and One vs. Rest Classification

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 for Classification
  - Accuracy/Error
  - Confusion Matrix
  - Precision
  - Recall
  - F1 Score

    - ROC Curve
    - ROC AUC
- Metrics for Regression
    - R^2
    - Adjusted-R^2
    - Mean Squared Error
    - Root Mean Squared Error
- Regularization
    - Ridge
    - LASSO
    - ElasticNet

Data Cleaning

- Dealing with Duplicates
- Dealing with Missing Data
- Dummy Variables
- Rescaling
- Dealing with Skew
- Detecting/Removing Outliers

Feature Engineering

- Binning
- One-Hot Encoding
- Derived Features

Joining Datasets

- pandas `df.join()` and `pd.merge()`
- Join Types
    - LEFT
    - RIGHT
    - INNER
    - OUTER

Dimensionality Reduction

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

Sklearn Pipelines

- .fit_transform() on train and .transform() on test
- GridSearch on Pipelines
- ColumnTransformer

NLP and Topic Modeling

- What is a corpus?
- Tokens and Tokenization
- Vocabulary
- Bag Of Words Representation
- n-grams
- Term Frequency
- Document Frequency
- Stopwords
- TfIdf
- Sentiment Analysis as Classification
- Topic Modeling with Latent Dirichlet Allocation (general concept)
  - per document topic distribution
  - per topic term distribution

Clustering

- k-Means
  - Within Cluster Sum of Squared Distances
- Hierarchical Agglomerative Clustering
  - linkage types
  - dendrogram representation

Recommendation Engines

- Content-Based Filtering
- User-Based Collaborative Filtering
- Issues
- Evaluating
  - Precision and Recall at K

Dealing with Imbalanced Data

- Random Undersampling majority class
- Random Oversampling minority class
- Oversample Synthetic Minority Items
  - SMOTE and ADASYN (general concepts)

Timeseries

- unique characteristics of timeseries data

- datetimes in pandas
- indexing with a DatetimeIndex
- converting data to datetime with pd.to_datetime()
- Shifting/Lagging
- Resampling and Frequencies
- Upsampling vs. Downsampling
- Moving/Rolling Window functions

Model Delivery via Flask API

- What can we use the flask python library for?

Extracting Data (ETL, APIs and Databases)

- Different filetypes handled by pandas

SQL

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