Elements of Data Science - S20

Final 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.

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

After The Midterm

Feature Selection

- Model Based
- Low Variance

- Univariate
- Recursive

Data Cleaning

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

Feature Engineering

- Binning
- One-Hot Encoding
- PolynomialFeatures

Dimensionality Reduction

- Unsupervised vs Supervised Learning
- Uses of dimensionality reduction
- Principle Components Analysis (PCA)
- Eigenfaces

Clustering

- k-Means
- Heirarchical Agglomerative Clustering
  - linkage

NLP and Topic Modeling

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

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 column to datetime
- Shifting
- Resampling
- Upsampling vs Downsampling
- Moving Window functions

Data Processing (ETL and API)

- What does ETL stand for?
- Difference between csv and json
- What can we use the python library flask for?

Datamanagement (SQL and NoSQL)

- benefits of RDBMS over flat-files
- Normalization/De-Normalization
- SQL
  - SELECT
  - FROM
  - AS
  - WHERE
  - LIMIT
  - COUNT
  - GROUP BY
  - ORDER BY
  - Subqueries
  - JOINs
- General difference between RDBMS and NoSQL dbs

Imbalanced Data

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