INFO 2950 Midterm-to-Final Lecture Topics, Fall 2023

Below is a list of broad topics covered in the second half of the course in 2950 (details of these topics, even if not listed, may still be on the final exam even if not explicitly listed on this handout). The final exam will cover both first and second halves of the course. Topic length does not necessarily correspond to weight in the final exam.

Correlation & Interaction Effects

- Correlations:
    - Spearman and Pearson Correlation calculation and interpretation
    - Ranks & Monotonicity
    - Correlation Matrices
- Interaction Effects
    - Variable interactions in a model
    - Interpreting coefficients in the presence of interactions
    - Visualizing Interaction Effects

Overfitting, Train/Test Splits & Cross-Validation

- Cause and solution of overfitting
    - Feature selection
- Train/Test split strategies
    - Implement in Python
    - Validation sets
- Importance of data partitioning
- Implementing cross validation in Python
- Bias vs Variance balance
- Best model selection methods

Evaluation Metrics and Linear Regression

- Techniques and considerations for binary outcome data
- Metrics: Precision, Recall, F1 score, ROC-AUC
    - Significance in model evaluation
    - Trade-offs between different metrics
- Non-binary outcome data metrics
    - RMSE, MAE, MSE, MAPE

Probability Distributions

- Binomial distribution
    - Sequences, probabilities, counts, polling
    - Margin of error

- Geometric Distribution
- Negative Binomial Distributions
- Poisson Distribution
    - Count data
- Normal distribution
    - 68 / 95 / 99.7 rule

Hypothesis Testing, Bootstrapping & Permutation

- Hypothesis testing in the context of regression analysis
    - Read and interpret regression tables
    - Determine the significance of findings
- Permutation tests implementation
    - Difference between bootstrap and permutation models

Joint/Marginal/Conditional Probabilities & Bayes Rule

- How to apply and calculate:
    - Joint probability
    - Marginal probability
    - Conditional probability
- Bayes' theorem
    - Conditional probability assessments

Clustering and Collaborative Filtering

- K-means clustering
    - Measuring distance
        - Euclidean, Manhattan, Cosine
    - Global mean vs cluster mean
- Similarity
- Collaborative filtering
    - Item-item vs. user-user

Matrix Multiplication & Singular Value Decomposition

- Matrix Multiplication
- SVD
    - Always possible to decompose a mxn matrix A into the multiplication of three unique matrices

Working with Text Data

- Naive Bayes
- Bag-of-words

    - Bag-of-words failure
- TF-IDF
- Word embeddings
    - Values of word embeddings over Bag-of-words and TF-IDF

Experimentations

- A/B test
- Hypothesis testing
- Multi-armed bandits
- The differences between Confidence versus Power
- Confidence intervals
- Reasons to preregister hypothesis
- Experimentation pitfalls

Classifications

- Linear regressions versus neural networks (nonlinearities)
- Perceptrons
- Stochastic Gradient Descent
- Hyperparameters
    - The difference between Hyperparameters versus Parameters
    - Hyperparameters tuning strategies

Discussion Sections

- Github
    - Reasons for using github
    - Github repo, clone, add, commit, push, pull
- Webscraping
    - Ethics of scraping
    - Python Libraries for webscraping
- Hypothesis testing
    - Visual intuition
    - Multiple hypothesis testing
    - Simulation
    - T-test
- Probability distributions
- Regular Expressions
    - Definition
    - When to use them