CS-498 Data Science – Final Exam Review Study Guide

Final objectives since midterm:

- Regularization and overfitting
  - Bias-variance tradeoff
  - Methods for regularization
- Unsupervised learning, clustering
  - K-means
    - Methods for determining the optimal number of clusters
  - Hierarchical Agglomerative Clustering
  - Linkage: single link, etc.
    - Data preprocessing and normalization
  - Measures of clustering quality
- Decision trees, Random Forests, Bagging
  - Decision tree decision boundary
  - Recursive binary splitting, Gini index, cross-entropy/info gain
  - Overfitting
  - Bagging/bootstrap aggregation
  - Random Forests
- Gradient Boosting, boosted decision trees
  - Gradient boosting concepts
  - XGBoost/boosted decision trees
  - Overfitting
- Introduction to time-series
  - Stationarity of time series - Trend, seasonal, and cyclical
  - Visual and statistical tests (Dickey-Fuller) tests for stationarity
  - Making a time-series stationary – log, MA, exp MA, differencing, decomposition
  - ARIMA forecasting
- Dimensionality reduction
  - PCA
- Data Science process, class project

Midterm objectives:

- Data science concepts, process and objectives
  - Technological convergence
  - Scientific method
  - Data Science process
- Exploratory data analysis and visualization
  - Data types – numeric, categorical, ordinal
  - Data manipulation (with NumPy stack)
  - Principals of visualization
  - Applying and interpreting visualization to augment numerical analysis
- Probability and statistical inference

- - Sampling, distributions
  - Measures of central tendency
  - Variance, standard deviation
- Linear regression, multivariate analysis
  - Coefficient/factor analysis for linear regression
  - Polynomial regression
  - RSS, R^2, MSE, RMSE
- Machine learning, supervised learning
  - Machine learning concepts
  - KNN
  - Logistic regression
  - Cross-entropy loss
  - Odds, log odds, coefficient/factor analysis for logistic regression
- Classification model evaluation and metrics
  - Confusion matrix
  - Accuracy, precision, recall/sensitivity, specificity
  - TPR, FPR, ROC curve