Data Collection

- What is proper data?
  - Bias (Selection bias)
  - Randomness
  - Size vs. Quality
- How do you collect data?
  - Sampling Strategies
    - Random Sampling
    - Stratified Sampling
    - Systematic Sampling
    - Quota Sampling
    - Snowball sampling
- Domain Knowledge

EDA

- Structured vs. unstructured data
- Data Cleaning and Validation
- Type of Data
  - Categorical Handling
  - Scaling
  - Normalization
  - Transformation
- Descriptive Analytics
- Correlations and Associations
- Visualization
- Imputations
- Outliers and Influential Values
- Feature Creation, Selection and Dimension Reduction
  - PCA, FA, Correspondence Analysis
  - Variable Importance and Stepwise Selection

~~**Simulations and Sampling Distributions**~~

- ~~Monte Carlo Simulation~~
- ~~Probability Basics~~
- ~~Normal Distribution~~
  - ~~Confidence Intervals~~
  - ~~Hypothesis Testing~~
  - ~~CLT~~
- ~~Long-tailed Distributions~~
- ~~Student t's Distribution~~
- ~~Binomial Distribution~~
- ~~Chi-square~~
- ~~F-distribution~~
- ~~Poisson, Exponential, Weibull~~

Statistical Tests

- A/B Experiments

- Hypothesis Tests
- Permutation Testing/ Bootstrapping
- significance levels and p-values
- normal tests and t-tests
- ANOVA tests/F-tests
- Chi-Square test
- Power and Sample Size

Regression and Prediction

- Simple Linear Regression
- Multiple Linear Regression
- Model Selection and Stepwise Regression
- Weighted Regression
- Factor Variables, Interaction Variables and Main effects
- Interpreting the Regression Equation
    - Correlated Predictors
    - Multicollinearity
    - Confounding Variables
    - Variable Importance
    - Prediction Interval
- Regression Assumptions/Diagnostics
    - Heteroskedasticity
    - Non-normality
    - Residual plots
    - Nonlinearity
- Other regression types

Classification

- Imbalanced Data Issues
    - Over and Undersampling
- Naive Bayes
- Logistic Regression and GLM
    - Logit
    - Interpreting Coefficients and Odds Ratios
- Model Assessments
    - Confusion Matrix
    - Precision, Recall, Specificity
    - ROC and AUC
    - Lift

Statistical Machine Learning

- kNN
    - Choosing K
    - Distance Metrics
    - kNN as a feature engine
- Tree Models and random forests
    - recursive partitioning algorithm
    - homogeneity and impurity

    - stopping conditions
    - bagging
- Boosting and Regularization
- Hyperparameters and CrossValidation

Unsupervised Learning

- K-means clustering
    - selecting K
    - interpret clusters
    - measures of dissimilarity
- Hierarchical clustering
    - dendrogram
    - agglomerative algorithm
    - Gower’s distance
- Model-based clustering
    - multivariate normal
- Problems with clustering mixed data

Model Selection and Model Explanation/Interpretability

Machine Learning Concepts

(Database Basics)