Draft Studylist Final

k-Nearest Neighbors

- You must manually calculate the Euclidian Distance between a testing record and 3 observations from the training dataset for two predictor variables and then find the nearest neighbor for k=1.
- Euclidean Distance between a testing record and 9 observations is given together with the outcome for the 9 observations. You have to manually find the 3 nearest neighbors and make a prediction.
- In a diagram with two predictors for a binary classification that shows both classes color-coded for the training data and also shows one record to be classified. Use k=5 and k=1 to classify the record.

Linear Regression

- Calculating regression coefficients: not relevant
- Understand (!) the logic of the MSE (e.g., identify the error for an individual prediction in the formula, identify an individual prediction in the formulas, ee why MSE is the mean of squared errors)
- Understand the role of the betas (parameters and why their choice determines the MSE)
- interpret the coeficients and the p-values from a fitted regression

Logistic Regression

- When a graph of a linear or logistic regression line is given in a diagram, find the probabilities for YES/NO
- When a graph of a linear or logistic regression line is given in a diagram, find the decision boundary (e.g. every income greater 100k is predicted as YES, smaller 100k results in NO).

  - which data points are predicted Yes which NO
  - which data points are predicted correctly which are predicted incorrectly

Neural Networks

- Make a prediction when a Neural Network with parameters (βs) is given
  - in algebraic form
  - in a diagram
- Calculate the MSE for a prediction (testing observation),when a Neural Network with parameters (βs) is given
  - in algebraic form
  - in a diagram

Decision Tree

- Predict with a decision tree when spliting rules are given.
  - Classification model
  - Regression model
- Understand the meaning of the meaning of the three numbers in each node.
- Interpret a decision tree (terminal nodes and other nodes).
- Understand a confusion matrix and be able to calculate accuracy, sensitivity, and specificity)
- Know what a splitting variable and a splitting value are
- Not on final: Gini impurity and variance decrease

Random Forest

- How does a Random Forest generates a prediction from multiple decision trees (you might have to calculate the Random Forest Prediction from 5 or 10 trees.
  - Classification
  - Regression
- How do we ensure the decision trees of a random forest different?
- Random Subspace method

- Bagging (i.e., Bootstrapping)
- How can you create a bootstrap dataset from an original training dataset?

Interpretation

- Know the difference between
  - local/global interpretation
  - model specific/ model agnostic
- Interpret a Variable importance Plot
- SHAP values:
  - Understand why SHAP values are local/model agnostic interpreters
  - Understand the relation between (when numerical values are give):
    * average prediction
    * SHAZP values for all predictor variables for the observaion
    * predicted value for that observation