Lecture Notes - 11/12

Bootstrapping:

- Resampling data (with replacement) and using the resampled data to estimate distribution.
- We can use resampled data to calculate test statistics, confidence intervals, etc.
- Allows us to more effectively create estimates for a smaller dataset.

Bias and Variance:

- Bias is a measure of how much the model's prediction differs from the target.
- Variance measures the amount of fluctuation between the predictions from different training data.
- There is a tradeoff between bias and variance; we generally cannot have both a low bias and low variance.
- Ensemble idea: average results from several models with high variance and low bias.
    - Important that models be diverse so as to not be wrong in the same way.
    - Variance is reduced by averaging variance of many models.

Bagging (Bootstrap Aggregation):

- Bootstrap original data to create many training datasets, then run learning algorithm on each new dataset independently.
- For each test example, we will classify based on the majority classification from the bootstrapped training datasets.
- Random Forests: subset of bagging where classification models are in the form of decision trees.

Reviewing for Midterm 2:

- Confusion matrices with multiples classes; create a matrix with axes true class vs. predicted class.
    - Works for any number of classes, not just 0 and 1.
    - ROC curve does not work if classes are not positive and negative.
- Entropy vs. Classification error; entropy is a much more effective way to create decision trees / select classification.

- Statistics review:
    - Central Limit Theorem: for a large enough sample, sampling distributions of the mean are approximately normal
    - p-value: probability of observing a result more extreme than ours assuming the null hypothesis to be true