

# Foundations of Machine Learning

Brett Bernstein

August 22, 2018

## Trees, Bootstrap, Bagging, and RFs

### Trees

#### Trees Learning Objectives

- Be able to describe the structure of a binary tree (ex: put bounds on number of leaves given height; describe the geometry of the resulting prediction function; etc.).
- Give pseudocode for finding the optimal split for (a) a continuous feature, and (b) a categorical feature for a binary classification problem.
- Describe some reasonable strategies for controlling the complexity of a tree.
- In particular, describe the regularization approach used in CART (pruning and use of number of leaves as complexity measure), recognize the cost complexity criterion as our standard regularized ERM.
- Recall the entropy, Gini, and misclassification error splitting criteria. Give some intuition around preference for Gini/entropy (i.e. purity measures) over misclassification.

#### Trees Concept Check Questions

1. (a) How many regions (leaves) will a tree with  $k$  node splits have?  
(b) What is the maximum number of regions a tree of height  $k$  can have? Recall that the height of a tree is the number of edges in the longest path from the root to any leaf.  
(c) Give an upper bound on the depth needed to exactly classify  $n$  distinct points in  $\mathbb{R}^d$ . [Hint: In the worst case each leaf will have a single training point.]
2. This question involves fitting a regression tree using the square loss. Assume the  $n$  data points for the current node are sorted by the first feature. Give pseudocode with  $O(n)$  runtime for optimally splitting the current node with respect to the first feature.

3. Suppose we are looking at a fixed node of a classification tree, and the class labels are, sorted by the first feature values,

4, 1, 0, 0, 1, 0, 2, 3, 3.

We are currently testing splitting the node into a left node containing 4, 1, 0, 0, 1, 0 and a right node containing 2, 3, 3. For each of the following impurity measures, give the value for the left and right parts, along with the total score for the split.

- (a) Misclassification error.
- (b) Gini index.
- (c) Entropy.

## Bootstrap and Bagging

### Bootstrap and Bagging Learning Objectives

- Recall from basic statistics concepts related to an estimator (e.g. bias) and its variance.
- Describe (outside the context of bagging/RFs) how the bootstrap is a useful method for estimating the variance of an estimator, and have some intuition on how it can be applied across many problems.
- Again recalling basic statistics, understand why bagging (averaging predictions) reduces variance.
- Recalling that the bootstrap ignores an expected 37% of data in each bootstrap sample, explain how we can use out-of-bag observations to approximate test performance.
- Describe how RF reduces correlation between trees using column sampling while training on bootstrap samples.

### Bootstrap and Bagging Concept Check Questions

1. Let  $X_1, \dots, X_n$  be an i.i.d. sample from a distribution with mean  $\mu$  and variance  $\sigma^2$ . How large must  $n$  be so that the sample mean has standard error smaller than .01?
2. Let  $X_1, \dots, X_{2n+1}$  be an i.i.d. sample from a distribution. To estimate the median of the distribution, you can compute the sample median of the data.
  - (a) Give pseudocode that computes an estimate of the variance of the sample median.
  - (b) Give pseudocode that computes an estimate of a 95% confidence interval for the median.