

Week 9 Lecture: Concept Check Exercises

Trees

- How many regions (leaves) will a tree with k node splits have?
 - 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.
 - 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.]
- 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.
- 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.

- Misclassification error.
- Gini index.
- Entropy.

Bagging

- Let X_1, \dots, X_n be an i.i.d. sample from a distribution with mean μ and variance σ^2 . How large must n be so that the sample mean has standard error smaller than .01?
- 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.
 - Give pseudocode that computes an estimate of the variance of the sample median.
 - Give pseudocode that computes an estimate of a 95% confidence interval for the sample median.

Boosting

1. (★) Show the exponential margin loss is a convex upper bound for the $0 - 1$ loss.
2. Show how to perform gradient boosting with the hinge loss.
3. Suppose we are using gradient boosting. On each step we can do a better job of fitting the pseudoresiduals if we allow for deeper trees. Why might deep trees be discouraged while gradient boosting?