

1 Decision Trees

Consider constructing a decision tree on data with d features and n training points where each feature is real-valued and each label takes one of m possible values. The splits are two-way, and are chosen to maximize the information gain. We only consider splits that form a linear boundary parallel to one of the axes. We will only consider a standalone decision tree and not a random forest (hence no randomization). Recall the definition of information gain:

$$IG(\mathbf{node}) = H(S) - \frac{|S_l|H(S_l) + |S_r|H(S_r)}{|S_l| + |S_r|},$$

where S is set of samples considered at **node**, S_l is the set of samples remaining in the left subtree after **node**, and S_r is the set of samples remaining in the right subtree after **node**.

- (a) Prove or give a counter-example: In any path from the root to a leaf, the same feature will never be split on twice. If false, can you modify the conditions of the problem so that this statement is true?

- (b) Prove or give a counter-example: The information gain at the root is at least as much as the information gain at any other node.

Hint: Think about the XOR function.

- (c) Suppose that a learning algorithm is trying to find a consistent hypothesis when the labels are actually being generated randomly. There are d Boolean features and 1 Boolean label, and examples are drawn uniformly from the set of 2^{d+1} possible examples. Calculate the number of samples required before the probability of finding a contradiction in the data reaches $\frac{1}{2}$. (A contradiction is reached if two samples with identical features but different labels are drawn.)

- (d) Intuitively, how does the bias-variance trade-off relate to the depth of a decision tree?

2 Concerns about Randomness

One may be concerned that the randomness introduced in random forests may cause trouble. For example, some features or sample points may never be considered at all. In this problem we will be exploring this phenomenon.

- (a) Consider n training points in a feature space of d dimensions. Consider building a random forest with T binary trees, each having exactly h internal nodes. Let m be the number of features randomly selected (from among d input features) at each treenode. For this setting, compute the probability that a certain feature (say, the first feature) is never considered for splitting in any treenode in the forest.
- (b) Now let us investigate the possibility that some sample point might never be selected. Suppose each tree employs $n' = n$ bootstrapped (sampled with replacement) training sample points. Compute the probability that a particular sample point (say, the first sample point) is never considered in any of the trees.

- (c) Compute the values of the two probabilities you obtained in parts (b) and (c) for the case where there are $n = 2$ training points with $d = 2$ features each, $T = 10$ trees with $h = 4$ internal nodes each, and we randomly select $m = 1$ potential splitting features in each treenode. You may leave your answer in a fraction and exponentiated form, e.g., $\left(\frac{51}{100}\right)^2$. What conclusions can you draw about the concern mentioned in part (b)?