

---

## DSC 140A - Discussion 10

---

### Quick Review

#### Decision Trees

Internal nodes : ask yes/no questions about a single feature

Leaf nodes : designate class labels

Goal : find a sequence of questions that leads to correct classifications

Heuristic : at each step, choose a question that minimizes uncertainty  $p_L u(S_L) + p_R u(S_R)$

Reduce over-fitting : pruning vs. early stopping

#### Boosting (AdaBoost)

Goal : combine several 'weak learners' (decision stumps)  $H_t(\vec{x})$  to create a single 'strong learner'  $H(\vec{x})$

Margin :  $r_t = \sum_{i=1}^n \omega_i^{(t)} y_i H_t(\vec{x}^{(i)}) \in [-1, 1]$  (positive for correct predictions, negative for incorrect predictions)

Alpha :  $\alpha_t = \frac{1}{2} \ln \frac{1+r_t}{1-r_t}$  (metric of performance)

Update weights :  $\omega_i^{(t+1)} \propto \omega_i^{(t)} \cdot e^{(-\alpha_t y_i H_t(\vec{x}^{(i)}))}$  ( $\propto$  because weights are normalized)

Combine classifiers :  $H(\vec{x}) = \sum_{t=1}^T \alpha_t H_t(\vec{x})$

#### Random Forests

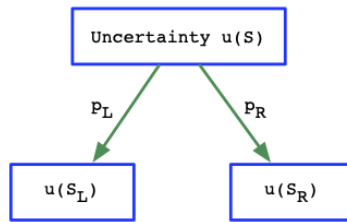
Goal : combine several decision trees\* to create a single classifier based on majority vote

\*Key properties :

1. each decision tree is fit using a bootstrapped sample of  $n$  points
2. at each node, restrict to one of  $k = \sqrt{d}$  features, chosen randomly

**Problem 1.**

Let  $u(S)$  be the uncertainty score for a set of labeled points,  $S$ . Consider a split of the following form:



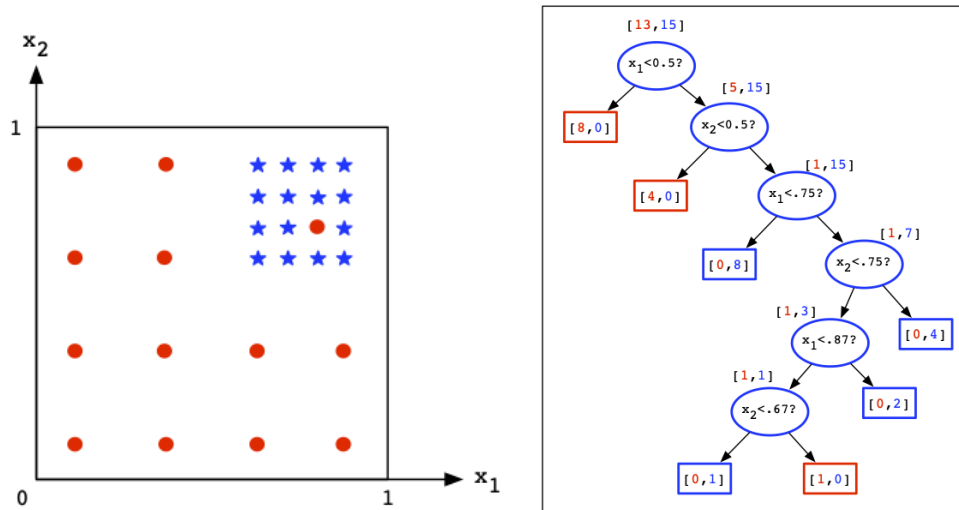
Now, let us assume that  $S$  contains 27 data points belonging to class  $+$  and 17 data points belonging to class  $-$ . We'll represent this as  $[27, 17]$ .

After the split, the left branch is given by  $[18, 4]$  and the right branch is given by  $[9, 13]$

Compute the initial uncertainty before the split,  $u(S)$ , as well as the resulting uncertainty after the split,  $p_L u(S_L) + p_R u(S_R)$ , using the formulas associated with mis-classification rate, Gini index, and also entropy.

**Problem 2.**

Recall the following decision tree over the given data points we learned during lecture.



This tree seems to overfit on our training data and we would like to address this problem via early stopping. Suppose we wanted to utilize early stopping with an uncertainty threshold of 0.2.

Calculate the uncertainty at each step in the tree using the Gini index and determine at what point we would have "early stopped". Sketch this new tree as well as the decision boundaries on the graph.

**Problem 3.**

Below, at figures A, B, and C, we show three weak learners (decision stumps) over our data. For each, determine which data points will have the largest weight increase based on the boundary drawn at that step. In figure D, approximately sketch the final classifier by combining the weak learners shown in A, B, and C.

