# DSC 140A - Discussion 07

## 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

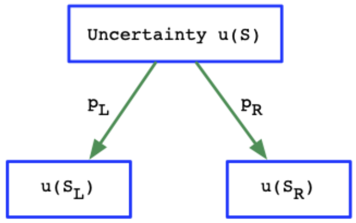
**Boosting (AdaBoost)**
Goal : combine several 'weak learners' (decision stumps) $H_t(\vec{x})$ to create a single 'strong learner' $H(\vec{x})$
Margin : $r_t = \sum_{i=1}^{n} \omega_i^{(t)} y_i H_t(\vec{x}^{(i)}) \in [-1, 1]$ (positive for correct predictions, negative for incorrect predictions)
Alpha : $\alpha_t = \frac{1}{2} \ln \frac{1+r_t}{1-r_t}$ (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.

**Solution:** We will denote $p$ as the probability of data points from class $+$, and $1-p$ as the probability of data points from class $-$.
The uncertainty scores $u(S)$ are associated with the following formulas:

1. Mis-classification rate : $u(S) = \min\{p, 1-p\}$
2. Gini index : $u(S) = 2p(1-p)$
3. Entropy : $u(S) = p \log \frac{1}{p} + (1-p) \log \frac{1}{1-p}$

Computing the uncertainty scores in these three settings yields:

1. Mis-classification rate
   $u(S) = \min\{\frac{27}{44}, \frac{17}{44}\} = \frac{17}{44} = \boxed{0.3864}$

   $$p_L u(S_L) + p_R u(S_R) = \frac{22}{44}\left(\min\{\frac{18}{22}, \frac{4}{22}\}\right) + \frac{22}{44}\left(\min\{\frac{9}{22}, \frac{13}{22}\}\right) = \frac{22}{44}\left(\frac{4}{22}\right) + \frac{22}{44}\left(\frac{9}{22}\right) = \frac{13}{44} = \boxed{0.2955}$$
2. Gini index
   $u(S) = 2\left(\frac{27}{44}\right)\left(\frac{17}{44}\right) = \boxed{0.4742}$

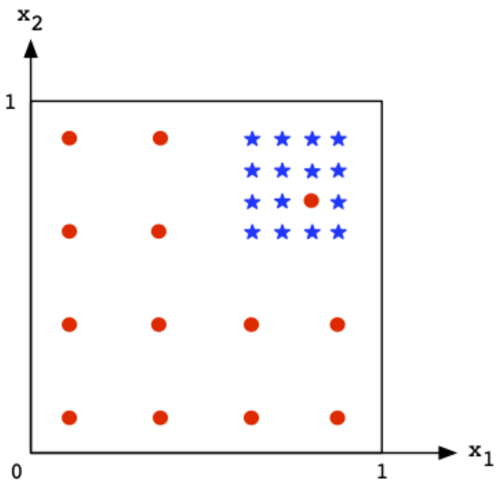
   $$p_L u(S_L) + p_R u(S_R) = \frac{22}{44}\left(2\left(\frac{18}{22}\right)\left(\frac{4}{22}\right)\right) + \frac{22}{44}\left(2\left(\frac{9}{22}\right)\left(\frac{13}{22}\right)\right) = 0.1488 + 0.2418 = \boxed{0.391}$$
3. Entropy

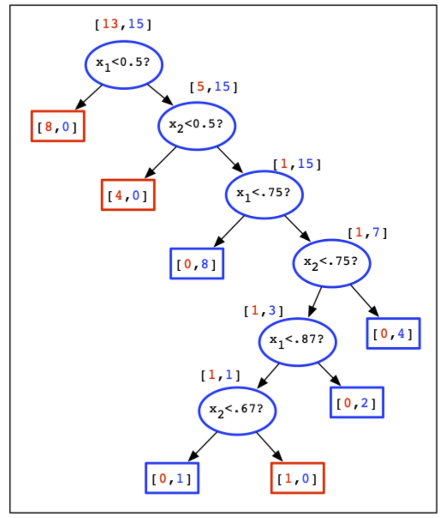
   $$u(S) = \left(\frac{27}{44}\right) \log\left(\frac{44}{27}\right) + \left(\frac{17}{44}\right) \log\left(\frac{44}{17}\right) = \boxed{0.9624}$$

   $$p_L u(S_L) + p_R u(S_R) = \frac{22}{44}\left(\left(\frac{18}{22}\right) \log\left(\frac{22}{18}\right) + \left(\frac{4}{22}\right) \log\left(\frac{22}{4}\right)\right) + \frac{22}{44}\left(\left(\frac{9}{22}\right) \log\left(\frac{22}{9}\right) + \left(\frac{13}{22}\right) \log\left(\frac{22}{13}\right)\right)$$
   $$= 0.3420 + 0.4880 = \boxed{0.83}$$

**Problem 2.**

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



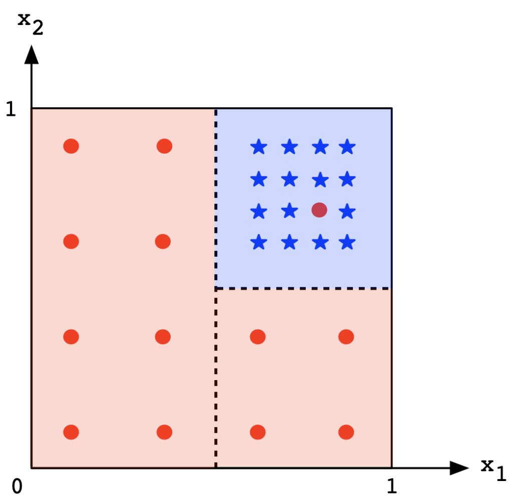



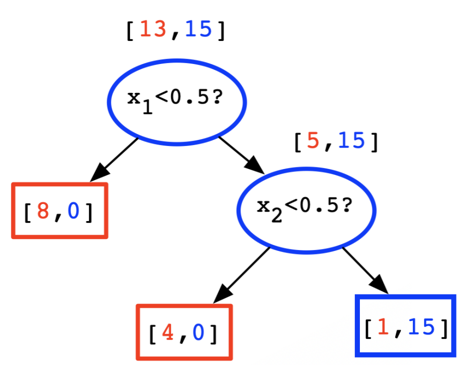

This tree seems to overfit on our training data and we would like to address this problem via early stopping. Suppose we wanted to utilize early stopping with an uncertainty threshold of 0.2.
Calculate the uncertainty at each step in the tree using the Gini index and determine at what point we would have "early stopped". Sketch this new tree as well as the decision boundaries on the graph.

**Solution:**
At step [13, 15], $u(S) = 2\left(\frac{13}{28}\right)\left(\frac{15}{28}\right) = 0.4975$. This value is not $< 0.2$, so we cannot early stop yet.

At step [5, 15], $u(S) = 2\left(\frac{5}{20}\right)\left(\frac{15}{20}\right) = 0.375$. This value is not $< 0.2$, so we cannot early stop yet.

At step [1, 15], $u(S) = 2\left(\frac{1}{16}\right)\left(\frac{15}{16}\right) = 0.1172$. This value is $< 0.2$, so we can early stop.
Even though we are classifying one point in our training set incorrectly, we still make this node a leaf node for class $-$.
We do not need to calculate the uncertainty for leaf nodes because there will be no branching past those points and the uncertainty $u(S) = 0$.
The new tree and decision boundaries are shown below.

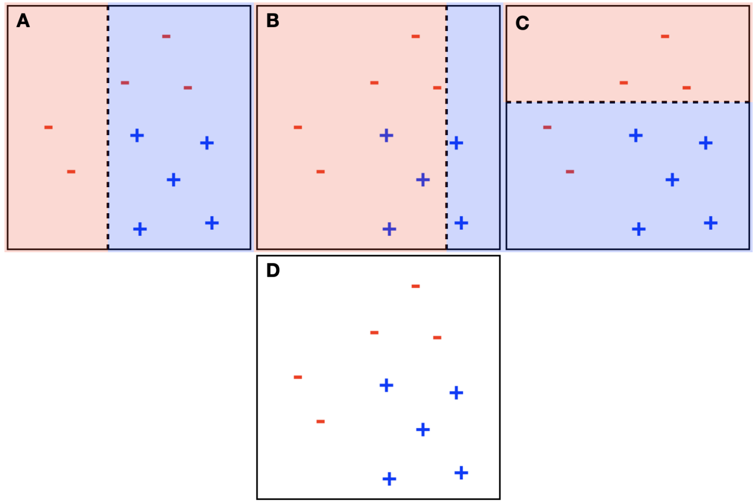



**Problem 3.**

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.




**Solution:** The data points that will receive the largest weight increase are those that are mis-classified by the given decision stump. This is due to the sign of the exponent of the update rule.
The final classifier will be a linear combination of all decision stumps. While the exact values of $alpha_t$ are not given, we can simply approximate the final classifier to achieve zero error as shown below.