Homework 3: Decision trees

1. [1pt] Given the dataset (where each sample has 3 feature values: a, b, and c) below, compute the gini impurity for the condition $S1 : a > 10$. Please show the estimations of $Pr(class = +1|a > 10)$ and $Pr(class = -1|a > 10)$. If you do not show these two but a final result, you will get no point.

Sample Number	feature a	feature b	feature c	class
1	12	3	5	+1
2	4	7	6	+1
3	5	4	8	+1
4	6	6	7	+1
5	7	5	1	-1
6	8	2	2	-1
7	9	6	3	-1
8	11	8	1	-1

- 2 samples in total satisfy threshold $S1 : a > 10$
- $Pr(class = +1|a > 10) = 0.5$. Since 1 sample (#1) belongs to class +1 given $a > 10$
- $Pr(class = -1|a > 10) = 0.5$. Since 1 sample (#8) belongs to class -1 given $a > 10$
- $g(a > 10) = 1 - 0.5^2 - 0.5^2 = 0.5$

2. [1pt] Do the same for a condition $S2 : a \leq 5$. Again, intermediate steps need to be shown.

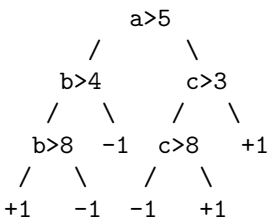
- 2 samples in total satisfy threshold $S2 : a \leq 5$
- $Pr(class = +1|a \leq 5) = 1$. Since 2 samples (#2,#3) belongs to class +1 given $a \leq 5$
- $Pr(class = -1|a \leq 5) = 0$. Since 0 sample belongs to class -1 given $a \leq 5$
- $g(a \leq 5) = 1 - 1^2 - 0^2 = 0$

3. [2pt] Based on the results from the two problems above, compute the expectation for gini impurity for the feature a and the threshold 5. Please show the estimatons of the probabilities of both conditions, i.e., $P(a > 5)$ and $P(a \leq 5)$. If you just show a final result, no point.

- $g(a > 5) = 1 - Pr^2(class = +1|a > 5) - Pr^2(class = -1|a > 5) = 1 - 0.33^2 - 0.67^2 = 0.44$
- $g(a \leq 5) = 1 - 1^2 - 0^2 = 0$
- $P(a > 5) = 6/8 = 0.75$
- $P(a \leq 5) = 2/8 = 0.25$

- $E(a,5) = P(a>5)g(a>5) + P(a\leq 5)g(a\leq 5) = 0.75 * 0.44 + 0.25 * 0 = 0.33$

4. [1pt] Using the decision tree below, decide the classification outcomes for all samples in Problem 1. Left branch is True and right branch is False. Present your result as a two-column table.



By two-column, this is it:

sample	prediction
1	-1
2	+1
3	+1
4	-1
5	-1
6	-1
7	-1
8	-1