**Submission Date: 3rd December 2021**

# 1 Classification and Regression Trees (CART)

1. What are the general advantages and disadvantages of decision trees ? Can you elaborate on the disadvantages and provide a short description ?

2. Can you mention a few of the most important hyperparameters of the decision trees ?

3. Given the following dataset:

| Ball Color | Ball Width | Sport |
|---|---|---|
| Yellow | 6 | Tennis |
| Yellow | 6 | Tennis |
| White | 22 | Football |
| Brown | 22 | Football |
| Brown | 22 | Basketball |

   - What is the initial split that gives the highest information gain ?
   - Build the entire decision tree based on the given dataset where, $K = 3$, $N = 5$, $V = \{\text{Tennis}, \text{Football}, \text{Basketball}\}$, $p(V) = \{\frac{2}{5}, \frac{2}{5}, \frac{1}{5}\}$

4. Assume that you are given a decision tree that splits the space as in figure 1 left, where light blue means that it assigns samples in that region to class 1 and white to class 0. Similarly, the right figure shows the ground truth, following the same color convention.

   - Determine the decision tree splits and the depth of the tree. Assume that $x_1$ and $x_2$ are the names for the horizontal and vertical axes.
   - Device a metric for assessing the quality of the prediction from the decision tree, and compute it. Consider 1 as the perfect prediction.

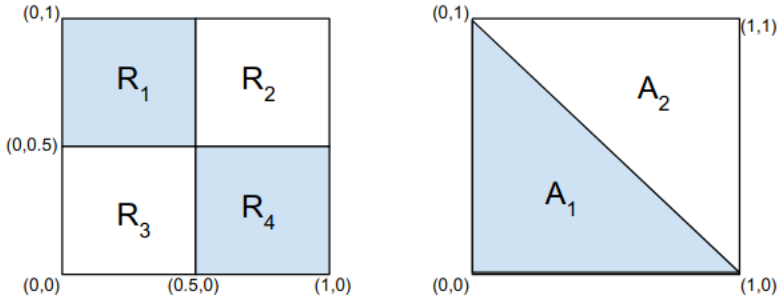


Figure 1. Left: Decision tree boundaries, Right: Ground truth.

5. If we consider the ground truth region from the previous point, which would be the best split for a decision tree with depth equals to 1 (decision stump)? In other words, which is the best $x$ that splits the region if we consider a partition like in figure 2? Try to find a formulation that is consistent with the metric specified in the last point.

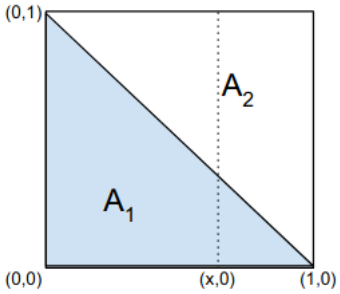


Figure 2. Decision Stump.

6. What would be the next split? (Hint: The next split should be on axis $x_2$)