1 Classification and Regression Trees (CART): Theory

1. The advantages:
   - Can easily handle categorical variables.
   - It is an interpretable model.
   - Can handle unimportant features.

   The disadvantages:
   - Deterministic models.
   - Can overfit to the training data.
   - Unstable (data, splits).
2. Some of the most important hyperparameters are:
   - Minimum number of samples per split.
   - Minimum number of samples in a leaf.
   - Total number of nodes
   - Split criterion.
   - Leaf model.
   - Maximum depth of the tree.
3. One could make a decision tree overfit in the following ways:
   - Having too many decision nodes.
   - Having an unbalanced tree.
   - Having only a few instances per node.

2 Classification and Regression Trees (CART): Hands-On

Task 2.1

Order of classes: (Tennis, Football, Basketball)

- Initial State:

$$\begin{aligned} H(V) &= -\left(\frac{2}{5}\cdot\log_2\frac{2}{5}+\frac{2}{5}\cdot\log_2\frac{2}{5}+\frac{1}{5}\cdot\log_2\frac{1}{5}\right) \\ &\approx 1.5204 \end{aligned}$$

- Considering to split on Ball Color equal to Yellow, we get:

  Left split:

$$\begin{aligned} H\left(V^{(L)}\right) &= -\left(1 \cdot \log_2 1 + 0 + 0\right) \\ &= 0 \end{aligned}$$

  Right split:

$$\begin{aligned} H\left(V^{(R)}\right) &= -\left(0 + \frac{2}{3} \cdot \log_2 \frac{2}{3} + \frac{1}{3} \cdot \log_2 \frac{1}{3}\right) \\ &\approx 0.92 \end{aligned}$$

  Information Gain:

$$\begin{aligned} I &= N \cdot H(V) - N^L \cdot H\left(V^{(L)}\right) - N^R \cdot H\left(V^{(R)}\right) \\ &\approx 5 \cdot 1.5204 - 2 \cdot 0 - 3 \cdot 0.92 \\ &= 4.842 \end{aligned}$$

- Considering to split on Ball Color equal to White, we get:

  Left split:

$$\begin{aligned} H\left(V^{(L)}\right) &= -\left(0 + 1 \cdot \log_2 1 + 0\right) \\ &= 0 \end{aligned}$$

  Right split:

$$\begin{aligned} H\left(V^{(R)}\right) &= -\left(\frac{2}{4} \cdot \log_2 \frac{2}{4} + \frac{1}{4} \cdot \log_2 \frac{1}{4} + \frac{1}{4} \cdot \log_2 \frac{1}{4}\right) \\ &= 1.5 \end{aligned}$$

  Information Gain:

$$\begin{aligned} I &= N \cdot H(V) - N^L \cdot H\left(V^{(L)}\right) - N^R \cdot H\left(V^{(R)}\right) \\ &\approx 5 \cdot 1.5204 - 1 \cdot 0 - 4 \cdot 1.5 \\ &= 1.602 \end{aligned}$$

- Considering to split on Ball Color equal to Brown, we get:

  Left split:

$$\begin{aligned} H\left(V^{(L)}\right) &= -\left(0 + \frac{1}{2} \cdot \log_2 \frac{1}{2} + \frac{1}{2} \cdot \log_2 \frac{1}{2}\right) \\ &= 1 \end{aligned}$$

Right split:

$$H\left(V^{(R)}\right) = -\left(\frac{2}{3}\cdot\log_2\frac{2}{3} + \frac{1}{3}\cdot\log_2\frac{1}{3} + 0\right)$$
$$\approx 0.92$$

Information Gain:

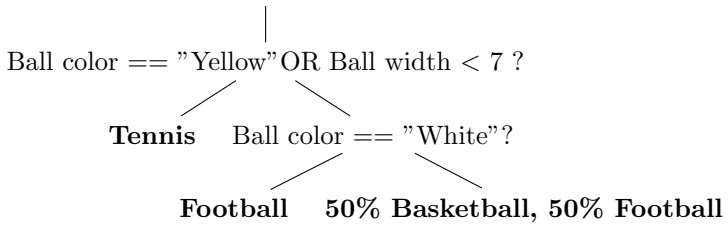
$$I = N\cdot H(V) - N^L\cdot H\left(V^{(L)}\right) - N^R\cdot H\left(V^{(R)}\right)$$
$$\approx 5\cdot 1.5204 - 2\cdot 1 - 3\cdot 0.92$$
$$= 2.842$$

- Considering to split on Ball Size (anywhere between 6 and 22), we get:

  The same as splitting on Ball Color equal to Yellow.

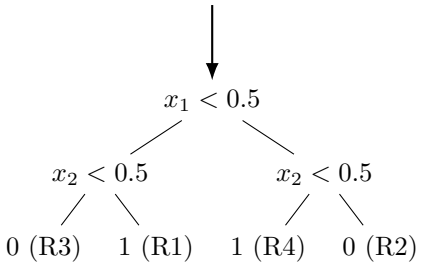
The initial split can happen at Ball Color equal to Yellow or with the Ball Size (anywhere between 6 and 22), since they have the biggest information gain.

The final decision tree is as follows (left branch is for "Yes"):



Task 2.2

- The decision has depth 2 and looks like this (left branch is for "Yes"):



- The simplest metric to assess the quality of the prediction from the decision tree, based on the ground truth, is to compute the overlapping area between the light blue regions (overlap between R1 $\bigcup$ R4 and A1) and

white regions (overlap between R2 $\bigcup$ R3 and A2). Where these regions intersect corresponds to a correct prediction region. The intersection is shown in light green in Figure 2. It is easy to notice that the area of this region is 0.5. Therefore, the quality of the prediction is 0.5 (where 1 stands for a perfect prediction).

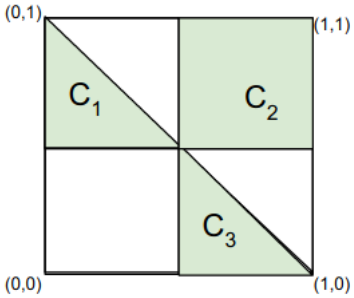
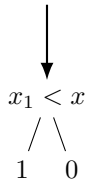


Figure 2. Overlap between regions. C1 $\bigcup$ C3 represent the overlap between R1 $\bigcup$ R4 and A1, C2 represents the overlap between R2 $\bigcup$ R3 and A2.

Task 2.3

We consider a decision stump with decision variable x_1 and threshold x:



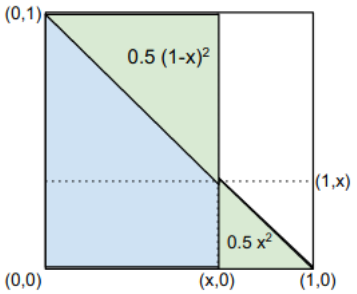


Figure 3. Green region is the incorrectly classified region.

Following the logic from the previous question, we can notice that the area of region corresponding to incorrect predictions (green region in Figure 3) is described as:

$$A = \frac{1}{2}x^2 + \frac{1}{2}(1-x)^2$$

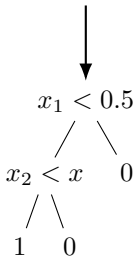
This area corresponds to the quality of the prediction. As we want to minimize it, we solve the following optimization problem:

$$x_* = \text{argmin}_{x\in[0,1]} \frac{1}{2}x^2 + \frac{1}{2}(1-x)^2$$

After computing the derivative and solving $\frac{dA}{dx} = 0$, we obtain $x_* = \frac{1}{2}$.

Task 2.4

Similarly, we can compute the next split on the vertical axis (x_2). We assume a decision tree of the form



After following a similar logic, we obtain that the split is at $x_2 = 0.75$. Note that we could also choose $x_2 = 0.25$ if we would perform the split in the right part of the tree instead of the left.

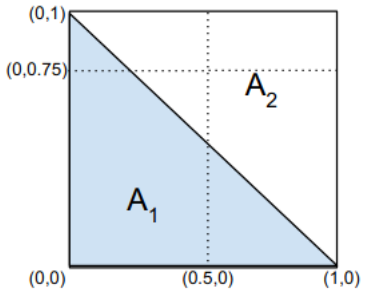


Figure 4.