Machine Learning
Exercise Sheet 5 Solution

Alper Hamza Ari 5771973 | Chenqi Hao 5781375 | Danil Skokov 5779466 | Said Orfan Haidari 5781295

November 19, 2023

1 Classification and Regression Trees (CART): Theory

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

   - `Advantages:` It is interpretable. It can be used to solve both regression and classification problems.
   - `Disadvantages:` The model can overfit to the given data easily. This is because it is a high-variance model. It is very unstable against noise or small changes in data, meaning that when you make small changes in the data, you might end up with very different tree.

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

   - `max_depth:` It states how much we want the tree to grow. The deeper a tree grows, the more it gets complex.
   - `min_leaf:` It means minimum number of points that should be in a leaf node. In other words, a decision node can be split into two if both children nodes have at least `min_leaf` many points.
   - `split_criterion:` It describes how to measure quality of the splits, e.g. *entropy* method will measure split quality by analyzing *information gain*

3. How can you make a decision tree overfit? Give three possibilities and explain.

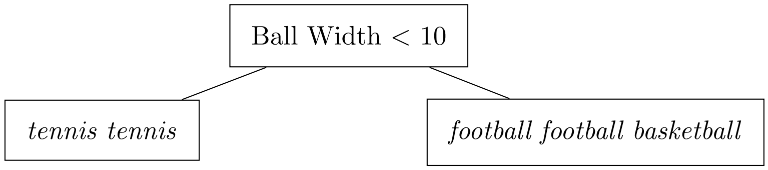
   i The model overfits if `max_depth` is very high. In other words, tree nodes will keep splitting until very high depths, or until each leaf becomes complete pure.

   ii If `min_leaf` is low, then nodes will keep splitting too much.

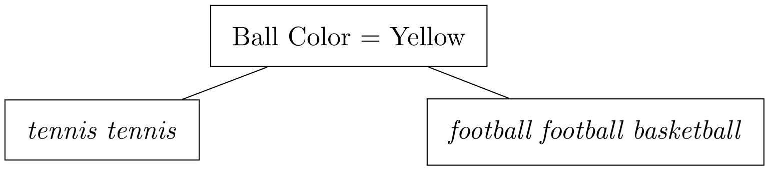
   iii If the training data is small, then it is highly that predictions will fail.

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

1. The initial split that gives maximum *information gain*: (left arrow = True)



Or another way:



Both splits above will result it highest information gain which is calculated as follows:

$$H(V) = -(\frac{2}{5}log_2(\frac{2}{5}) + \frac{2}{5}log_2(\frac{2}{5}) + \frac{1}{5}log_2(\frac{1}{5})) = 1.52 \quad (1)$$

$$P(V_{\text{left}}) = \{1, 0, 0\} \quad (2)$$

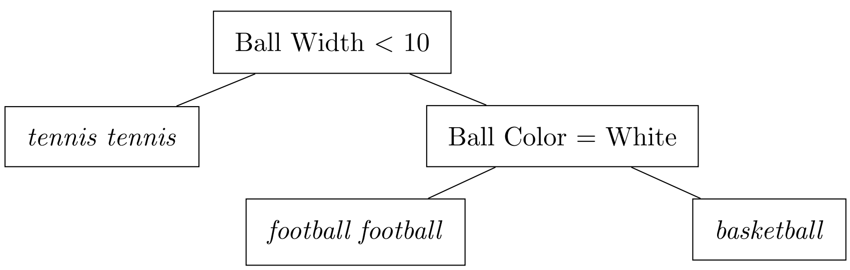
$$H(V_{\text{left}}) = -(1 \cdot log_2(1) + 0 \cdot log_2(0) + 0 \cdot log_2(0)) = 0 \quad (3)$$

$$P(V_{\text{right}}) = \{0, \frac{2}{3}\}, \frac{1}{3} \quad (4)$$

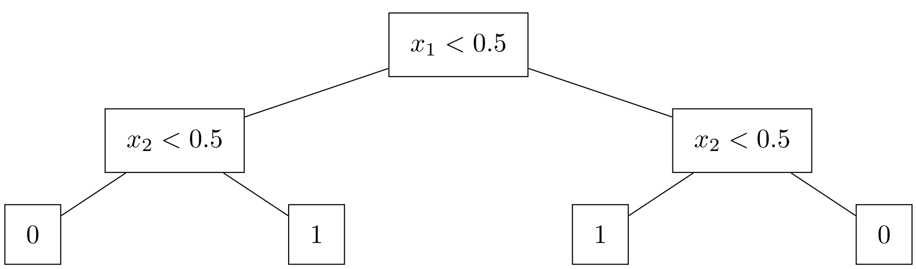
$$H(V_{\text{right}}) = -(0 \cdot log_2(0) + \frac{2}{3} \cdot log_2(\frac{2}{3}) + \frac{1}{3} \cdot log_2(\frac{1}{3})) = 0.92 \quad (5)$$

$$\text{gain} = 5(1.52) - 2(0) - 3(0.92) = 4.84 \quad (6)$$

And an optimal and complete decision tree is: (left arrow = True)



2. The decision tree corresponding to the sketch could be the following: (left arrow = True)



And to measure the quality of predictions from the decision tree we can calculate intersected areas of prediction model and ground truth. The ratio between intersection and total area will give us the accuracy.

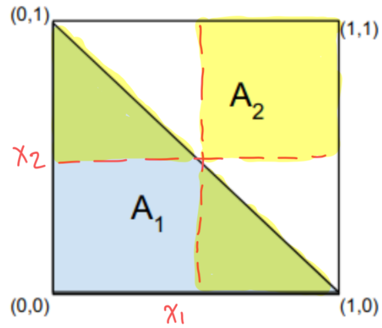


Figure 1: Intersection of ground truth and the model, denoted in yellow

The ratio gives us accuracy of 0.5.

3. Considering ground truth from the previous question, we want to find the best split candidate over x_1 axis.

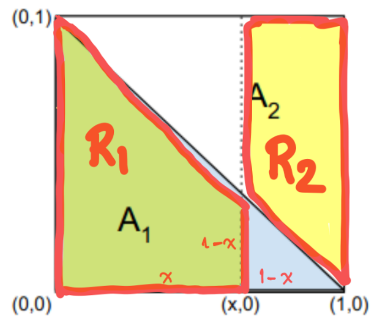
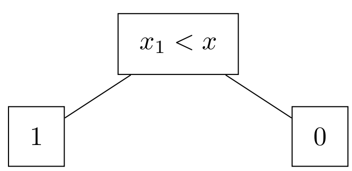


Figure 2: Area to consider and maximize in terms of x_1

Which has the corresponding 1-depth decision tree: (left arrow = True)



We need to compute following areas:

$$\begin{aligned} R_1 &= \frac{1}{2} - \frac{(1-x)^2}{2} \\ R_2 &= (1-x) - \frac{(1-x)^2}{2} \end{aligned} \tag{7}$$

And our objective is:

$$\begin{aligned} \underset{x}{\operatorname{argmin}}(R_1 + R_2) &= \underset{x}{\operatorname{argmin}}\left((1-x) + \frac{1}{2} - 2 \cdot \frac{(1-x)^2}{2}\right) \\ &= \underset{x}{\operatorname{argmin}}(-x^2 + x + \frac{1}{2}) \end{aligned} \tag{8}$$

Taking derivative with respect to x and setting to 0:

$$
\begin{aligned}
\frac{d(R_1+R_2)}{dx} &= -2x+1=0 \\
x_1 &= \frac{1}{2} \text{ will maximize the yellow area}
\end{aligned} \tag{9}
$$

4. A second split on x_2 axis and taking intersections with ground truth:

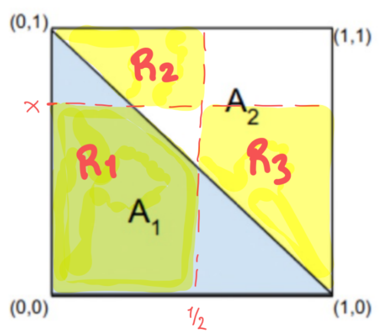
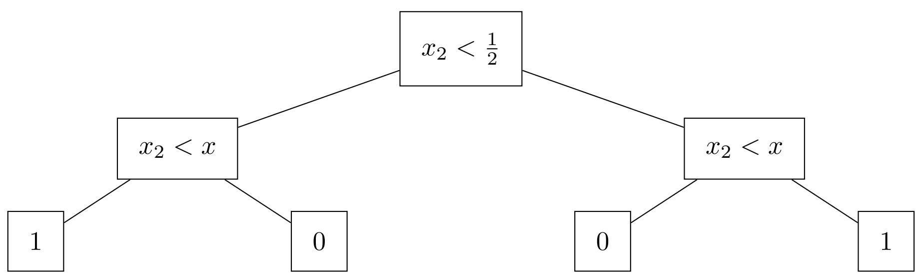


Figure 3: Area to consider and maximize in terms of x_2

Which has the corresponding 2-depth decision tree: (left arrow = True)



And our objective is:

$$
\underset{x}{\operatorname{argmin}}(R_1+R_2+R_3) = \underset{x}{\operatorname{argmin}}\left(\frac{-3}{2}(1-x)^2+x^2+\frac{1}{2}\right) \tag{10}
$$

Taking derivative with respect to x and setting to 0:

$$
\begin{aligned}
\frac{d(R_1+R_2+R_3)}{dx} &= -3(1-x)+x=0 \\
&= 4x-3=0 \\
x_2 &= \frac{3}{4} \text{ will maximize the yellow area}
\end{aligned} \tag{11}
$$

Finally, the complete decision tree should look as follows: