universität freiburg

Assignment 05

Classification and Regression Trees

Machine Learning WS23/24

Assignment 05
Solution

Theory

Theory

Task 1.1: Advantages and Disadvantages of Decision Trees

Advantages

- Can easily handle categorical variables.
- Interpretable models.
- Can handle unimportant features.

Disadvantages

- Deterministic models.
- Prone to overfitting to the training data (need to restrict growth).
- Unstable: Minor changes in data can lead to dastically different trees („high variance estimator“).

Theory

Task 1.2: Hyperparameters

Examples:

- Minimum number of samples per split (decision node).
- Minimum number of samples per leaf.
- Total number of nodes.
- Split criterion.
- Leaf model.
- Maximum depth of tree.

Theory

Task 1.3: Overfitting

Decision trees can overfit in various ways:

- **Too many decision nodes:** At some point we just fit to noise/outliers using uninformative criterions.
- **Unbalanced tree:** Focus on very specific features with (generally) few samples in each decision.
- **Only few instances/datapoints per node:** Decision trees work well if we can bypass noise/outliers by averaging inside leaves.

Hands-On

Hands-On

Task 2.1: Optimal Split

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

N = 5, V = {Tennis, Football, Basketball}, K = |V| = 3

What is the initial split that gives the highest information gain?

- **Information Gain:**

$$I(C) = N \cdot \mathrm{H}(V) - N^{(L)} \cdot \mathrm{H}(V^{(L)}) - N^{(R)} \cdot \mathrm{H}(V^{(R)})$$

- **Entropy:** Compute H(V) and $H(V^L)$, $H(V^R)$ for each potential split

$$H(V) = -\sum_{k=1}^{K} p(v_k) \log_2 p(v_k)$$

- **Possible Splits:**

$$\text{color} \in \{\text{yellow}\} \quad \text{vs.} \quad \text{color} \in \{\text{white}, \text{brown}\}$$
$$\text{color} \in \{\text{white}\} \quad \text{vs.} \quad \text{color} \in \{\text{yellow}, \text{brown}\}$$
$$\text{color} \in \{\text{brown}\} \quad \text{vs.} \quad \text{color} \in \{\text{yellow}, \text{white}\}$$

$$\text{ball-width} < 7 \quad \text{vs.} \quad \text{ball-width} \geq 7 \quad \text{(or any w. between 7 and 22)}$$

Hands-On

Task 2.1: Optimal Split

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

N = 5, V = {Tennis, Football, Basketball}, K = |V| = 3

What is the initial split that gives the highest information gain?

- **Step 1: Compute Initial Entropy**

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

Hands-On

Task 2.1: Optimal Split

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

N = 5, V = {Tennis, Football, Basketball}, K = |V| = 3

What is the initial split that gives the highest information gain?

- **Step 2: Split at Color = Yellow**

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

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

$$I(\text{Color}=\text{Yellow}) = N\cdot\mathrm{H}(V)-N^{(L)}\cdot\mathrm{H}(V^{(L)})-N^{(R)}\cdot\mathrm{H}(V^{(R)}) \approx 5\cdot 1.5204-2\cdot 0-3\cdot 0.92 = 4.842$$

Hands-On

Task 2.1: Optimal Split

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

N = 5, V = {Tennis, Football, Basketball}, K = |V| = 3

What is the initial split that gives the highest information gain?

- **Step 2: Split at Color = White**

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

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

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

Hands-On
Task 2.1: Optimal Split

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

N = 5, V = {Tennis, Football, Basketball}, K = |V| = 3

What is the initial split that gives the highest information gain?

- **Step 2: Split at Color = Brown**

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

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

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

Hands-On

Task 2.1: Optimal Split

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

N = 5, V = {Tennis, Football, Basketball}, K = |V| = 3

What is the initial split that gives the highest information gain?

- **Step 2: Split at Width < (anywhere between 6 and 22)**

  Same as Color = Yellow!

Hands-On

Task 2.1: Optimal Split

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

N = 5, V = {Tennis, Football, Basketball}, K = |V| = 3

What is the initial split that gives the highest information gain?

- **Step 3: Decide**

$$
\begin{aligned}
I\left(\text{Color}=\text{Yellow}\right) &= 4.842 \\
I\left(\text{Color}=\text{White}\right) &= 1.602 \\
I\left(\text{Color}=\text{Brown}\right) &= 2.842 \\
I\left(\text{Width}<7\right) &= 4.842
\end{aligned}
$$

➜ Two options: *Color = Yellow* or *Width < 7*

(or any width between 7 and 22)

Hands-On

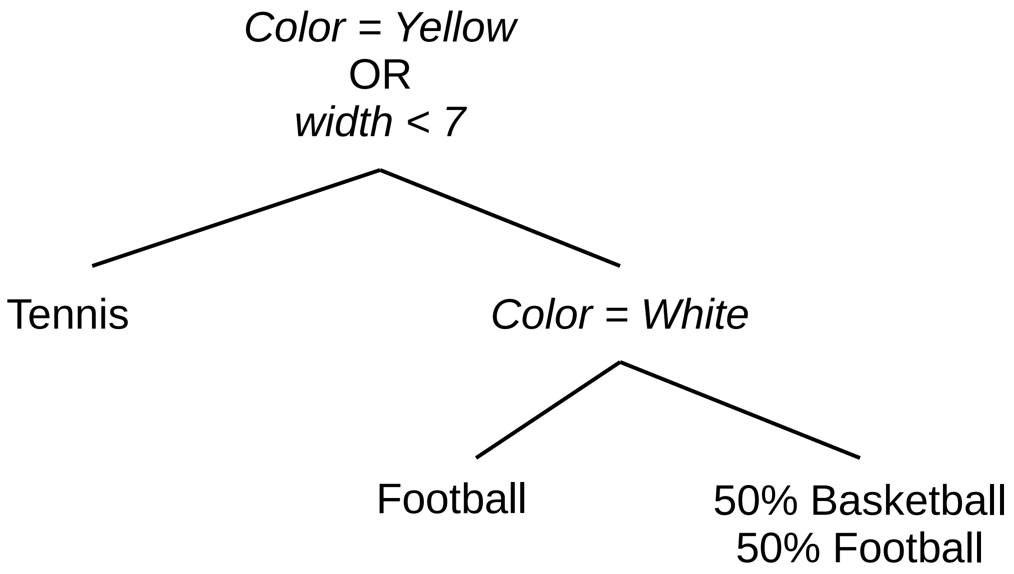
Task 2.1: Final Tree

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

$N = 5$, $V = \{\text{Tennis, Football, Basketball}\}$, $K = |V| = 3$

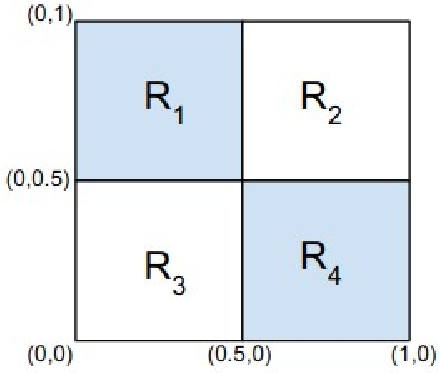
How does the final tree look like?

(left path means „yes“)

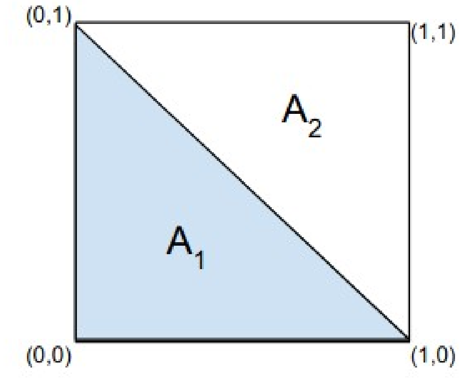


Hands-On

Task 2.2: Splits and Depth



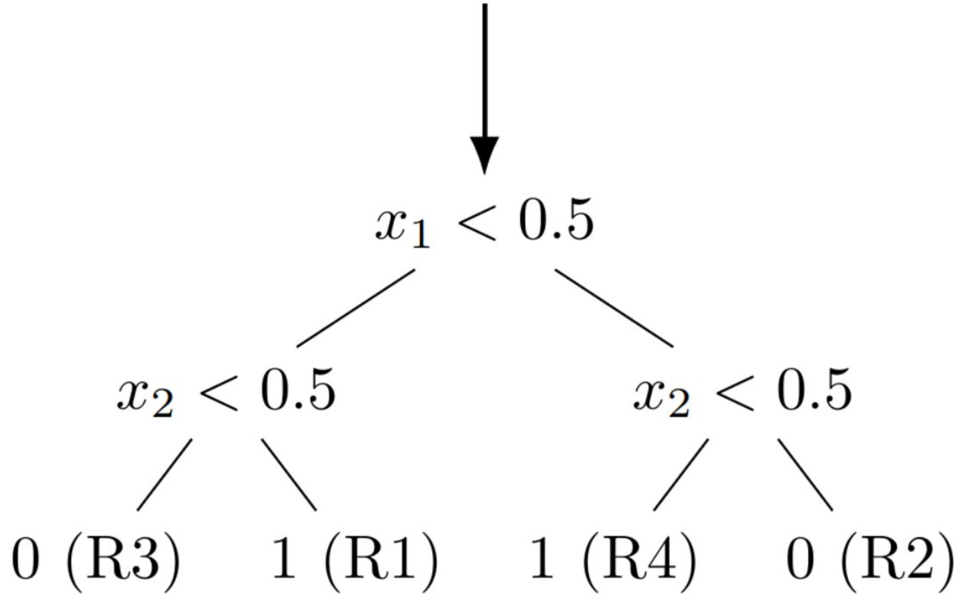
Tree Boundaries



Ground Truth

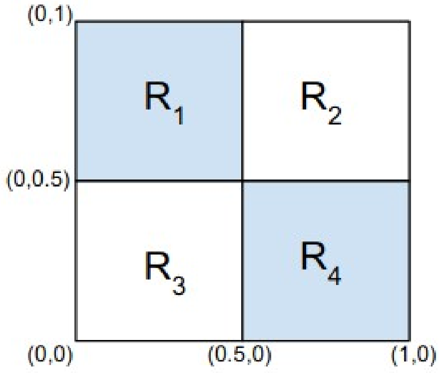
What are the splits and depth of the given decision tree?

- The depth of the tree is 2.
- It is structured as follows (left path means „yes“):

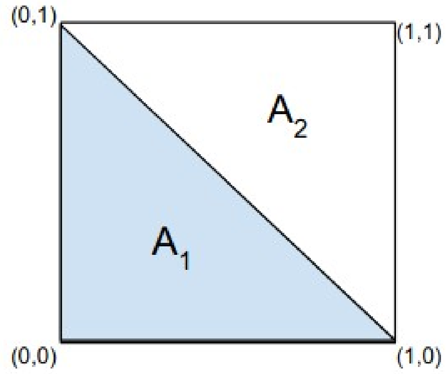


Hands-On

Task 2.2: Metric



Tree Boundaries



Ground Truth

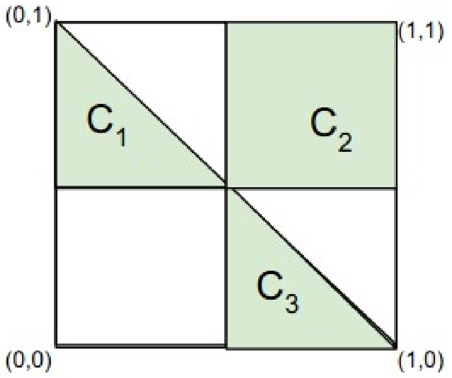
Design a metric for assessing the quality of the prediction:

- **Idea:** *Measure overlap between correctly classified regions.*

  I.e.: Compute overlap between

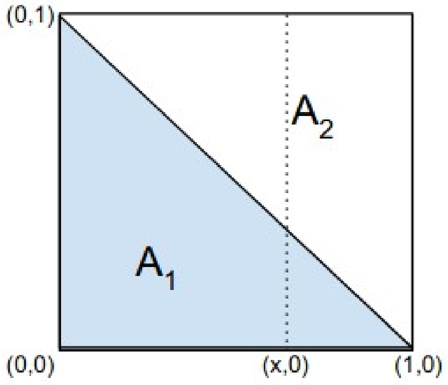
  $$R_1 \cup R_4 \text{ and } A_1 \qquad (C_1, C_3)$$
  $$R_2 \cup R_3 \text{ and } A_2 \qquad (C_2)$$

  (then sum up and normalize by whole area)

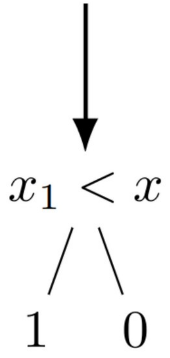


- **In this case:** Overlap is 0.5. Best would be 1, worst 0.

Hands-On
Task 2.3: Best initial Split



Decision Stump



Which x_1 value leads to the best (initial) split?

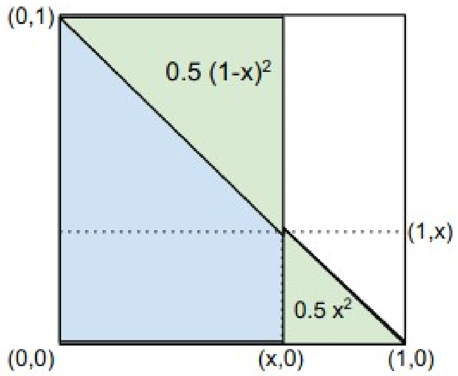
- **Idea:** Similar to before, measure overlap
  (Now between incorrectly classified regions because that is easier)

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

- **Minimize incorrectly classified area (green):**

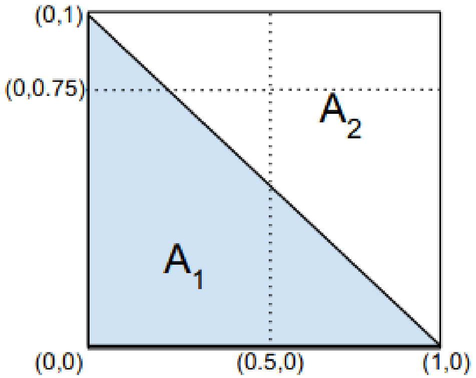
$$x^* = \operatorname{argmin}_x(A)$$

- **Result:** Set $\frac{dA}{dx} = 0$ and obtain $x^* = \frac{1}{2}$

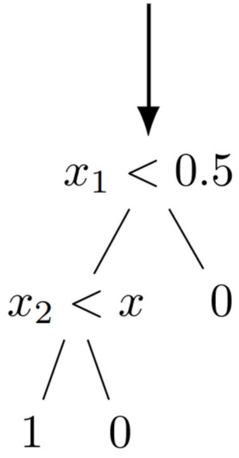


Hands-On

Task 2.4: Best secondary Split



Decision Stump



Which x_2 value leads to the best second split?

- **Idea:** Same as in Task 2.3.
- **Even easier:** Just consider top-left/bottom-right squares independently
  - [0.0, 0.5] x [0.5, 1.0] (top left)
  - [0.5, 1.0] x [0.0, 0.5] (bottom right)
  - Perform exactly the same optimization as in Task 2.3
- **Result:** $x_{2,L}* = \frac{3}{4}$ (left) or $x_{2,R}* = \frac{1}{4}$ (right)