Exercise Sheet

Learning Goals

- Decision Trees
- Impurity functions
- The ID3 algorithm

1. (a) What is overfitting?

> **Solution:** Reasons for overfitting are:
>
> - a non representative training set
> - the use of a learning algorithm with a lower inductive bias.
>
> Then the model will derive "patterns" from the noise in the example set that are not present in the feature space. The model generalizes worse and the performance on the training set fails to generalize wellgeneralization
> A concept h overfits the training examples D if there is a concept h' with
>
> 1. a higher error rate on the training examples
> 2. a lower error rate on unseen examples.

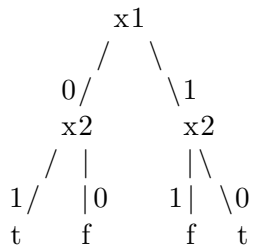
(b) Check all correct statements.

- ☐ A short training time leads to overfitting.
- √ **A smaller decision tree generalizes more than a bigger decision tree.**
- √ **The generalization capability of a decision tree depends on the training set.**
- ☐ Information theory compensates the negative impacts of small or biased training sets.

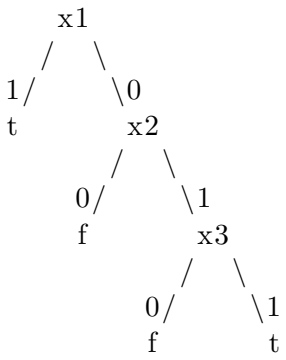
2. Draw a decision tree for each of the following three boolean functions. (All variables draw from the set $\{\texttt{FALSE}, \texttt{TRUE}\}$ and you may use $\{0, 1\}$ instead.)

A. $x_1 \wedge \neg x_2$ B. x_1 XOR x_2 C. $x_1 \vee (x_2 \wedge x_3)$

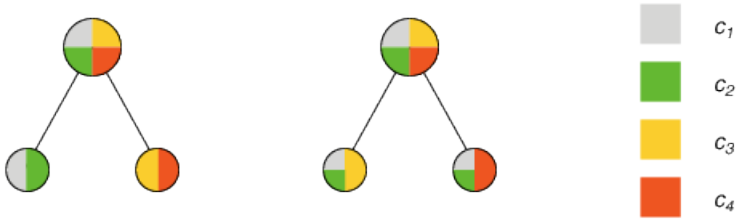
> **Solution:**
>
> - A
>
> ```
>        x1
>    1/    \0
>    x2     f
> 1/   \0
> f     t
> ```
>
> - B



- C

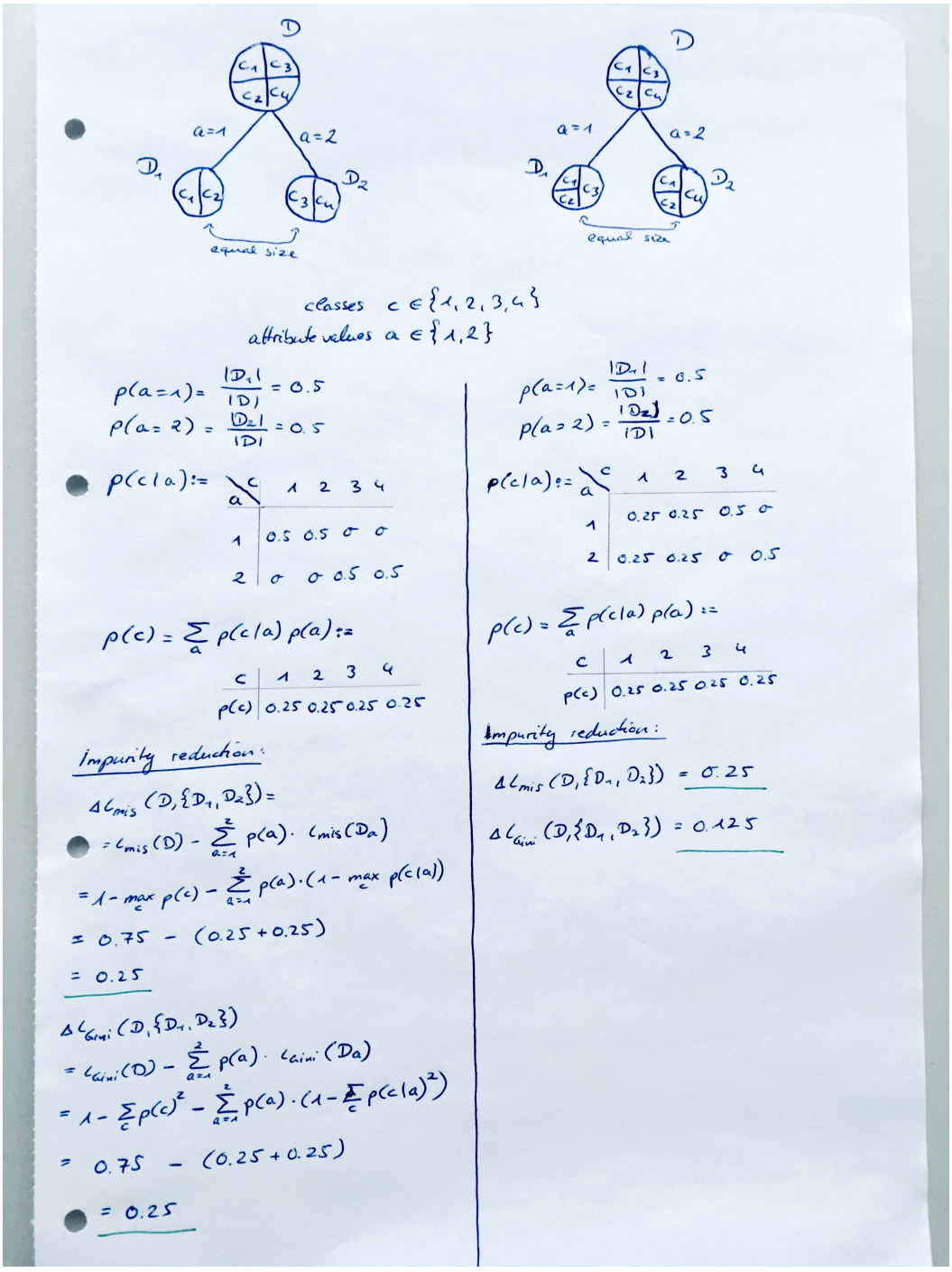


3. Consider the following illustration of two possible splittings of an example set D with four classes $C = \{c_1, c_2, c_3, c_4\}$.



(a) Compute the drop in impurity $\Delta\iota$ for left and right split, respectively, using the misclassification rate $\iota_{misclass}$ as well as the Gini impurity ι_{Gini}. (Note: This will yield 4 values: one for every combination of "split" and "impurity measure".)

Solution: .



(b) Explain the result.

Solution: The left split is preferable as it has a higher chance of leading to pure nodes in fewer splits. The misclassification rate only pays attention to the most frequent class ($\max_c$) - regardless of whether the remaining probability mass is scattered across only one or several other classes. The Gini criterion takes into account the probability mass from all classes and prefers situations in which the probability mass is scattered across only a few dominating classes. According to the Gini criterion we would choose the left split over the right split. According to the misclassification rate none of the splits is preferable over the other.

4. Given the following training set with dogs data:

Color	Fur	Size	Class
brown	ragged	small	well-behaved
black	ragged	big	dangerous
black	smooth	big	dangerous
black	curly	small	well-behaved
white	curly	small	well-behaved
white	smooth	small	dangerous
red	ragged	big	well-behaved

Use the ID3 algorithm to determine a decision tree, whereas the attributes are to be chosen with the maximum average information gain *iGain*:

$$iGain(D,A) \equiv H(D) \; - \sum_{a \in A} \frac{|D_a|}{|D|} \cdot H(D_a) \quad \text{with} \quad H(D) = -p_{\oplus}\log_2(p_{\oplus}) - p_{\ominus}\log_2(p_{\ominus})$$

Solution: Determine the values of $H(D|$ Attribute $)$:

- Attribute *Color*: $m = 7$ (Number of objects)

Color	well-behaved	dangerous	Probability
brown	1	0	$P(\text{brown}) = 1/7$
black	1	2	$P(\text{black}) = 3/7$
white	1	1	$P(\text{white}) = 2/7$
red	1	0	$P(\text{red}) = 1/7$

$$\begin{aligned} H(D|\text{ Color}) &= -\Big[\frac{1}{7}(1\log_2 1 + 0\log_2 0) + \frac{3}{7}(\frac{1}{3}\log_2(\frac{1}{3}) + \frac{2}{3}\log_2(\frac{2}{3})) \\ &\quad + \frac{2}{7}(\frac{1}{2}\log_2(\frac{1}{2}) + \frac{1}{2}\log_2(\frac{1}{2})) + \frac{1}{7}(1\log_2 1 + 0\log_2 0)\Big] \\ &= -\Big[0 + \frac{1}{7}(\log_2(\frac{1}{3}) + 2\log_2(\frac{2}{3})) + \frac{1}{7}(\log_2(\frac{1}{2}) + \log_2(\frac{1}{2})) + 0\Big] \\ &= -\Big[\frac{1}{7}(\log_2(\frac{1}{3}) + 2\log_2(\frac{2}{3})) + \frac{2}{7}\log_2(\frac{1}{2})\Big] \approx 0.679 \end{aligned}$$

- Attribute *Fur*:

Fur	well-behaved	dangerous	Probability
ragged	2	1	$P(\text{ragged}) = 3/7$
smooth	0	2	$P(\text{smooth}) = 2/7$
curly	2	0	$P(\text{curly}) = 2/7$

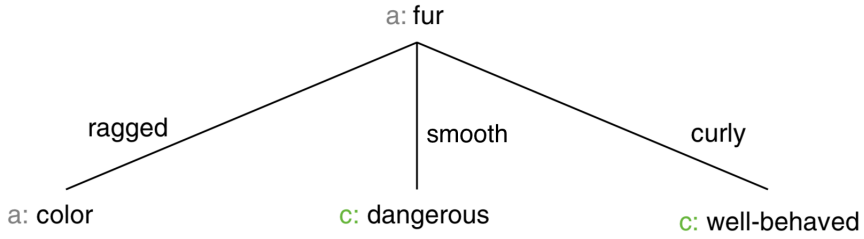
$$\begin{aligned} H(D|\text{Fur}) &= -\Big[\frac{3}{7}(\frac{2}{3}\log_2(\frac{2}{3}) + \frac{1}{3}\log_2(\frac{1}{3})) + \frac{2}{7}(0\log_2 0 + 1\log_2 1) \\ &\qquad\qquad + \frac{2}{7}(1\log_2 1 + 0\log_2 0)\Big] \\ &= -\frac{3}{7}(\frac{2}{3}\log_2(\frac{2}{3}) + \frac{1}{3}\log_2(\frac{1}{3})) = -\frac{1}{7}(2\log_2(\frac{2}{3}) + \log_2(\frac{1}{3})) \\ &\approx 0.394 \end{aligned}$$

- Attribut *Size*:

Size	well-behaved	dangerous	Probability
small	3	1	$P(\text{small}) = 4/7$
big	1	2	$P(\text{big}) = 3/7$

$$\begin{aligned} H(D|\text{Size}) &= -\left[\frac{4}{7}(\frac{3}{4}\log_2(\frac{3}{4}) + \frac{1}{4}\log_2(\frac{1}{4})) + \frac{3}{7}(\frac{1}{3}\log_2(\frac{1}{3}) + \frac{2}{3}\log_2(\frac{2}{3}))\right] \\ &= -\frac{1}{7}\left[(3\log_2(\frac{3}{4}) + \log_2(\frac{1}{4})) + (\log_2(\frac{1}{3}) + 2\log_2(\frac{2}{3}))\right] \\ &\approx 0.857 \end{aligned}$$

$H(D|A)$ minimal for $A =$ Fur $\Rightarrow$ Choose Attribute *Fur*.



Nodes fur=smooth and fur=curly are already pure. Node fur=ragged is not pure yet, so we have to evaluate the impurity reduction for all remaining attributes (except "fur") for node fur=ragged. Since there are two classes compute the related subtree recursively.

- Attribute *Color*: $m = 3$ (Number of objects)

Color	well-behaved	dangerous	Probability
brown	1	0	$P(\text{brown}) = 1/3$
black	0	1	$P(\text{black}) = 1/3$
red	1	0	$P(\text{red}) = 1/3$

$$\begin{aligned} H(D|\text{ Color}) &= -\left[\frac{1}{3}(1\log_2 1 + 0\log_2 0) + \frac{1}{3}(0\log_2 0 + 1\log_2 1) + \frac{1}{3}(1\log_2 1 + 0\log_2 0)\right] \\ &= 0 \end{aligned}$$

- Attribute *Size*:

Size	well-behaved	dangerous	Probability
small	1	0	$P(\text{small}) = 1/3$
big	1	1	$P(\text{big}) = 1/3$

$$\begin{aligned} H(D|\text{ Size}) &= -\left[\frac{1}{3}(1\log_2 1 + 0\log_2 0) + \frac{2}{3}(\frac{1}{2}\log_2(\frac{1}{2}) + \frac{1}{2}\log_2(\frac{1}{2}))\right] \\ &= -\frac{2}{3}\log_2(\frac{1}{2}) = \frac{2}{3} \end{aligned}$$

$H(D|A)$ minimal for $A = \text{Color} \Rightarrow$ choose attribute *Color*. Resulting tree: