Lab Class ML:III

By 06.12.2012 solutions for the following exercises have to be submitted: 1c, 2, 3b+c, 5a, 7a, 9

Exercise 1 : Decision Trees

Construct for each of the following boolean functions a decision tree. Note: The target concept is the set of all models, i.e., set of interpretations (0/1 assignments to the boolean variables) that fulfill a formula.

(a) $A \wedge \neg B$

(b) $A \vee (B \wedge C)$

(c) A *XOR* B

(d) $(A \wedge B) \vee (C \wedge D)$

Exercise 2 : Decision Trees

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*:

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

Exercise 3 : Decision Trees (Background)

(a) For the construction of a decision tree almost always an top-down greedy search in the hypothesis space is employed. Explain the term Greedy Search (synonymously: search with a greedy strategy). What are its advantages and what are its disadvantages? When is a greedy strategy useful? Which alternative strategies exist?

(b) The inductive bias of the Candidate-Elimination algorithm is based on a different mechanism than the inductive bias of the ID3 algorithm. Explain this statement by analyzing the rationale of the inductive bias of each algorithm.

(c) Which time complexity has the ID3 algorithm? Explain your answer.

(d) Explain the statement from Thereom 1 (see lecture notes). “The problem to decide for a set of examples D whether or not a decision tree exists whose external path length is bound by b, is NP complete.”

Exercise 4 : Decision Trees (C4.5)

In 1993 Quinlan introduced with the C4.5 algorithm a successor of the ID3 algorithm. The C4.5 algorithm eliminates various deficits of the ID3 algorithm. Inform yourself about these deficits and how they are addressed by the C4.5 algorithm.

Exercise 5 : Decision Trees (Overfitting)

(a) What is overfitting?

(b) Why is the example set D partitioned in a test set and a training set? Is such a partitioning necessary to avoid overfitting?

(c) an approach to avoid overfitting is the use of so called post-pruning algorithms: initially, an oversized decision tree is constructed, which then is generalized by means of pruning. explain different pruning strategies (e.g. reduced-error pruning, rule post pruning).

Exercise 6 : Decision Trees (Overfitting)

Which statement is true?

- ☐ 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.

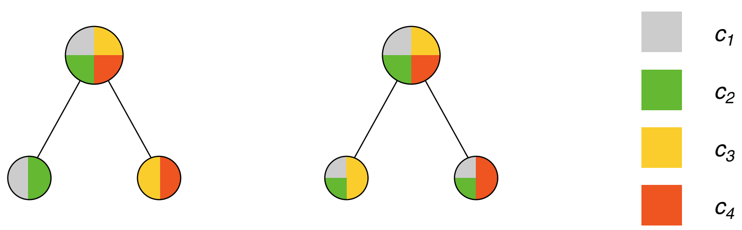
Exercise 7 : Pruning

a) Discuss disadvantages of stopping in comparison to pruning for the construction of decision trees.

b) Explain the use of different example sets for (1) the construction of a decision tree and (2) the selection of a decision tree from a set of pruning candidates.

Exercise 8

Let D be a set of examples over a feature space X and a set of classes $C = \{c_1, c_2, c_3, c_4\}$. Consider the following illustration of two possible splittings.

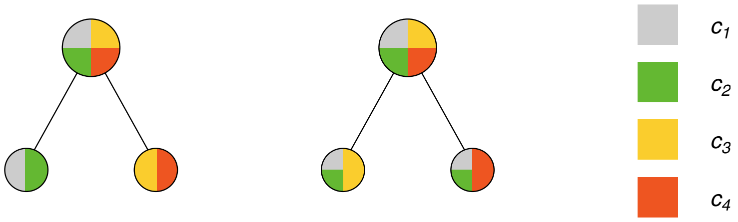


(a) Compute $\Delta\iota(\{D_1, D_2\}, t)$ with the misclassification rate $\iota_{misclass}$ and the Gini criterion ι_{Gini} as splitting criterion.

(b) Interpret the results.

Exercise 9

Let D be a set of examples over a feature space X and a set of classes $C = \{c_1, c_2, c_3, c_4\}$. Consider the following illustration of two possible splittings.



Consider a $4x4$ class confusion cost matrix that quantifies the misclassification cost $\textit{cost}(c' \mid c)$, where

$$\textit{cost}(c' \mid c) \begin{cases} \geq 0 & \text{if } c' \neq c, c \in C \\ = 0 & \text{otherwise} \end{cases}$$

Develop two class confusion cost matrices such the left tree (splitting) is preferred under one matrix, while the right tree (splitting) is preferred under the other.