Decision Trees

Exercise 1 : Impurity

(a) Verify that the arguments $\frac{|\{(\mathbf{x},c_1)\in D\}|}{|D|}, \ldots, \frac{|\{(\mathbf{x},c_k)\in D\}|}{|D|}$ of the impurity function indeed form a $k-1$-simplex.

Answer

Non-negativity holds. We also have

$$\sum_{i=1}^{k} \frac{|\{(\mathbf{x},c_i)\in D\}|}{|D|} = \frac{|D|}{|D|} = 1$$

because the datasets implied by the classes form a partition of D (every sample has exactly one class).

(b) Consider the $\arg\min$ of a impurity function ι from the definition of impurity functions. For which D is $\iota(\frac{|\{(\mathbf{x},c_1)\in D\}|}{|D|}, \ldots, \frac{|\{(\mathbf{x},c_k)\in D\}|}{|D|})$ minimal?

Answer

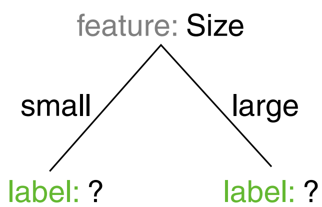
The impurity function is minimal if there is a single class $\hat{c}$ with $\forall(\mathbf{x},c)\in D : c = \hat{c}$ because then we get:

$$\iota(\frac{|\{(\mathbf{x},c_1)\in D\}|}{|D|}, \ldots, \frac{|\{(\mathbf{x},\hat{c})\in D\}|}{|D|}, \ldots, \frac{|\{(\mathbf{x},c_k)\in D\}|}{|D|}) = \iota(0,\ldots,1,\ldots,0)$$

Exercise 2 : Cost functions

Consider the set of training examples describing mushrooms, and the simple one-level decision tree given below:

	Color	Size	Points	Edibility
1	red	small	yes	toxic
2	brown	small	no	edible
3	brown	large	yes	edible
4	green	small	no	edible
5	red	large	no	edible



(a) Determine the labels of all nodes using the cost function $\textit{cost}(c', c)$ (cf. ML:VI-36):

$$\textit{cost}(c', c) = \begin{cases} 1 & \text{if } c' \neq c, c \in C \\ 0 & \text{otherwise} \end{cases}$$

Answer

$$\textit{label}(t_{\texttt{root}}) = \underset{c'\in\{\texttt{edible},\texttt{toxic}\}}{\text{argmin}} \big(1\cdot \textit{cost}(c',\texttt{toxic}) + 4\cdot\textit{cost}(c',\texttt{edible})\big) = \texttt{edible}$$

$$\textit{label}(t_{\texttt{Size=small}}) = \underset{c'\in\{\texttt{edible},\texttt{toxic}\}}{\text{argmin}} \big(1\cdot \textit{cost}(c',\texttt{toxic}) + 2\cdot\textit{cost}(c',\texttt{edible})\big) = \texttt{edible}$$

$$\textit{label}(t_{\texttt{Size=large}}) = \underset{c'\in\{\texttt{edible},\texttt{toxic}\}}{\text{argmin}} \big(0\cdot \textit{cost}(c',\texttt{toxic}) + 2\cdot\textit{cost}(c',\texttt{edible})\big) = \texttt{edible}$$

(b) Devise a new cost function that ensures that, for the same tree structure, none of the poisonous mushrooms in the training set are classified as edible.

Answer

$$\mathit{cost}_2(c', c) = \begin{cases} 3 & \text{if } c' = \texttt{edible}, c = \texttt{toxic} \\ 1 & \text{if } c' = \texttt{toxic}, c = \texttt{edible} \\ 0 & \text{otherwise} \end{cases}$$

(c) Compute the misclassification costs of the tree for both cost functions.

Answer

$\mathsf{Err}_{\mathsf{cost}}(T, D) = \frac{3}{5} \cdot (\frac{2}{3} \cdot 0 + \frac{1}{3} \cdot 1) + \frac{2}{5} \cdot (\frac{2}{2} \cdot 0) = \frac{1}{5}$

$\mathsf{Err}_{\mathsf{cost}_2}(T, D) = \frac{3}{5} \cdot (\frac{2}{3} \cdot 1 + \frac{1}{3} \cdot 0) + \frac{2}{5} \cdot (\frac{2}{2} \cdot 0) = \frac{2}{5}$

Exercise 3 : Decision Trees (Background)

(a) For the construction of a decision tree almost always a 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?

Answer

Greedy strategies optimize locally. The search space (the space of considered decision trees) is irreversibly made smaller in each step based on a local criterion. Advantages: fast. Disadvantages: might be suboptimal (does not consider the complete search space). Alternatives: complete search, backtracking.

(b) The inductive bias of the Candidate Elimination algorithm is based on a different mechanism than the inductive bias of the ID3 algorithm. Compare both mechanisms.

Answer

The Candidate Elimination algorithm is based on the version space. The model function is always consistent with the example set. The version space is completely searched. ID3 considers a complete hypothesis space (all possible decision trees), but only searches it partially (because it uses a greedy strategy).