Foundations of Artificial Intelligence

Prof. Dr. J. Boedecker, Prof. Dr. W. Burgard, Prof. Dr. F. Hutter, Prof. Dr. B. Nebel, Dr. rer. nat. M. Tangermann
T. Schulte, M. Krawez, R. Rajan, S. Adriaensen, K. Sirohi
Summer Term 2020

University of Freiburg
Department of Computer Science

Exercise Sheet 10

Due: Monday, July 20, 2020

Exercise 10.1 (Decision Trees)

No	Age	Engine power [kW]	**Risk**
1	< 25	< 100	low
2	< 25	> 200	high
3	≥ 25	> 200	high
4	≥ 25	$100 - 200$	low
5	< 25	$100 - 200$	high
6	≥ 25	< 100	low

Consider the data on car insurance risk in the table above. Produce a decision tree, which correctly classifies the insurance risk for the examples given, using the attributes *Age* and *Engine Power* in order of decreasing *information gain*. Give detailed calculations that justify the order in which the attributes are tested.
You can make use of the following values:
$\log_2(\frac{1}{3}) \approx -\frac{3}{2}$, $\log_2(\frac{2}{3}) \approx -\frac{1}{2}$, $\log_2(\frac{1}{2}) = -1$, $\log_2(1) = 0$.

Exercise 10.2 (Best practices in ML)

When doing machine learning, it is good practice to split the dataset into a training/validation/test set.

- Which subset(s) should you use for the following tasks:
  (a) training models (R & D)[1]
  (b) guard against overfitting (R & D)
  (c) model selection (R & D)
  (d) progress reports (R & D)
  (e) train the model (product)[2]
  (f) evaluating the model (product)
- Which of these subsets should always be fixed a priori (before even looking at the data)?

Note: The exercise sheets may be worked on in groups of up to three students.

[1] R & D: During research and development
[2] product: For the final product/publication