Heinrich Heine Universität Düsseldorf

HEIDERLAB

Exercise 8 on Machine Learning WS 2023/24
Prof. Dr. Dominik Heider
M. Sc. Arsam (Mohammad) Tajabadi
Submission: Until 20.12.2023, 23:59 on Ilias

Task 1: Decision trees (6 points)

a) Use the procedure from the lecture to construct a decision tree for the following data set. Only use the attributes that are suitable for constructing a decision tree. The Termination attribute represents the target variable. Use the information gain as the selection measure for the attributes and calculate it. How do you decide the order of the attributes? Draw the resulting tree. What do the leaf nodes look like? (3 P.)

Customer Number	Professional Status	Contract Duration	Termination
2	Civil servant	Low	Yes
3	Employee	Low	Yes
4	Self-employed	Medium	No
5	Employee	Medium	No
6	Non-employed	High	No
7	Employee	Medium	No
8	Civil servant	High	No
9	Self-employed	High	Yes
10	Non-employed	Medium	No

b) Classify the following examples using the decision tree you have created. Which of the customers is at the risk of termination? (1 P.)

Customer Number	Professional Status	Contract Duration
11	Self-employed	Low
12	Civil servant	Medium

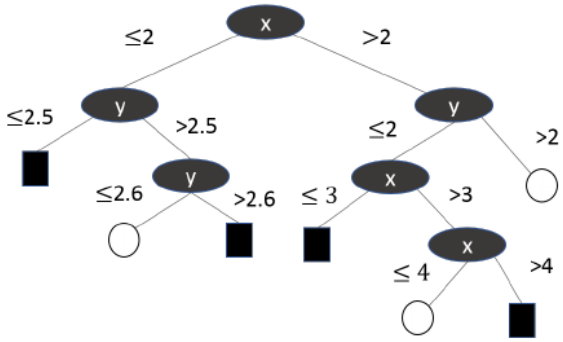
c) Derive decision rules from your tree. Is there any unexpected rule? (0,5 P.)

d) Describe the difference between information gain, gain ratio, and Gini index. (1 P.)

e) Calculate the Gini index for the customer data set given above for the attribute Contract Duration. (0,5 P.)

Task 2: Decision trees (3 Points)

a) Specify the partitioning for the following decision tree. Draw a coordinate system for this. Make sure that the partition identifiers match the tree leaves. (2 P.)



b) Which partition can be described as overfitted? In the decision tree, mark the nodes or subtrees that must be removed to avoid overfitting. (0,5 P.)

c) What does the generalization capability of a classifier mean? (0,5 P.)