**Data Mining I** Summer term 2020

Otto von Guericke University Magdeburg

Faculty of Computer Science

Knowledge Management & Discovery Lab

Exercise 3

1. Describe a potential problem of using Information Gain as splitting criterion in decision

trees. Explain why the Gain Ratio measure is a way to overcome this problem.

If a column has multiple values like ID from 1…30 then on choosing it as a node and splitting values for further decision tree shall result in a big busy tree . this shall overfit and not abel to generalize .Information gain shall be highest for the attribute since gini or entropy shall be less because of large number of distinct values.

Using gain ratio we can penalize the gain by dividing gain/split info. “The split information measures the entropy of splitting a

node into its child nodes and evaluates if the split results in a larger number

of equally-sized child nodes or not. For example, if every partition has the

same number of instances, then ∀i : N(vi)/N = 1/k and the split information

would be equal to log2 k. Thus, if an attribute produces a large number of

splits, its split information is also large, which in turn, reduces the gain ratio.”

2. Consider the training examples shown in Table 1 for a binary classification problem.

Calculate Information Gain and Gain Ratio for each of the attributes ID, Gender,

Graduation and Eye Color. Use multiway splits when possible.

Done on notebook

3. There exist several approaches to measure the generalization error rate of a decision

tree *T*. The simplest measure is termed as *optimistic* error estimate erro(*T*), which

is simply the error rate of *T* on the training set. In contrast, the *pessimistic* error

estimate errp(*T*) also incorporates the “complexity” of *T*, quantified by the number of

leaf nodes.

Let *k* be the number of leaf nodes and *Ntrain* be the number of training instances.

The pessimistic error estimate is computed as errp(*T*) = erro(*T*) + · *k*

*Ntrain*

, where

is a hyperparameter that makes a trade-off between minimizing training error and

minimizing model complexity. A third generalization error estimate is *reduced error*

*pruning* errrep, which is the error rate of *T* on a validation set.

Calculate erro(*T*), errp(*T*) and errrep(*T*) for each of the two decision trees in Figure

1. Assume that each leaf node is labeled according to the majority class of training

instances that reach the node. When calculating the *pessimistic* error estimate, assume

that = 2.

4. Explain the following classifier evaluation methods: (a) hold-out method, (b) random

subsampling, (c) cross-validation, (d) stratified cross-validation, (e) leave-one-out crossvalidation

and (f) bootstrap.

1

Table 1: Dataset for task 2.

ID Gender Graduation Eye Color Class

1 F College brown 1

2 F College brown 1

3 F College gray 0

4 F High School blue 0

5 F High School blue 0

6 F High School brown 0

7 F High School gray 1

8 F Middle School blue 1

9 F Middle School blue 1

10 F Middle School brown 1

11 M High School blue 0

12 M High School brown 0

13 M High School brown 0

14 M High School gray 0

15 M High School green 0

16 M High School green 0

17 M Middle School brown 1

18 M Middle School gray 1

19 M Middle School green 1

20 M Middle School green 1

Figure 1: Two decision trees generated from the same training set and validation data.

2