**Data Mining I** Summer term 2020

Otto von Guericke University Magdeburg

Faculty of Computer Science

Knowledge Management & Discovery Lab

Exercise 2

1. What is a decision tree? How is a decision tree induced?

A decision tree is a tree like tree like structure helped in decision making made of root node, internal nodes and leaf nodes. Each internal node is an attribute and each branch represents the value of the attribute. There are many algorithms to induce a decision tree which includes Hunt’s algorithm. For each internal node an attribute is decided to be tested and at the leaf/terminal node the outcome of the branch is decided which is the classification.

2. What is the difference between binary and “bushy” decisions trees?

Binary tree has 2 brances yes and no, whole busy tree shall have multiple brances . each attribute can have multiple values to be tested resulting in bushes.

3. Explain Gini index, misclassification rate and entropy.

All are impurity measures of an attribute.

Entropy = *−*

\_*c−*1

*i*=0

*pi*(*t*) log2 *pi*(*t*)*,* (3.4)

Gini index = 1 *−*

\_*c−*1

*i*=0

*pi*(*t*)2*,* (3.5)

Classification error = 1 *−* max

*i*

[*pi*(*t*)]*,*

4. How is Information Gain calculated?

Information gain of an attribute is calculated by subtracting the impurity of a node by its ratio from total impurity of the set.

5. Build a decision tree using the Information Gain for the “Defaulted Borrower” classification

problem given in Table 1. Discretize the attribute Annual Income into the

intervals [0, 100], (100, 200], (200, ). Defaulted Borrower is the target attribute.

6. The discretization of Annual Income chosen in he previous question is not ideal. What

is the optimal binary split for Annual Income according to Gini index?

7. Use Orange to investigate the effect of the size of a decision tree on its accuracy. First,

create a new workflow and add a widget to load the “Employee attrition” dataset. Use

the “Data Sampler” widget to manually create a training set and a test set. Set the

sampling type to “Fixed proportion of data” and 70%. Additionally, check the options

“Replicable (deterministic) sampling” and “Stratify sample (when possible)”. Add a

a decision tree widget. In the settings, uncheck the parameter “Do not split subsets

smaller than:”. Vary the parameter “Min. number of instances in leaves” using the

values 2, 5, 10, 25, 50, 100, 200, 500. Add a widget to measure accuracy on the training

set and on the test set, respectively. Additionally, record the number of tree nodes for

each parameter value. Visualize the results in a chart with “Min. number of instances

in leaves” on the x-axis and accuracy on the y-axis. The chart should depict 2 curves.

Interpret the results.

8. Create a confusion matrix from Table 2. Assume that “Y” serves as positive class.

Calculate accuracy, precision, recall, *F*1-measure score, sensitivity and specificity.