CS6029 – Data Mining / CS5041 – Advanced Data Warehousing and Data Mining (FALL 2022)

Dr. Rabia Maqsood

Assignment 2

Topics Covered: Classification algorithms: Decision tree, Rule-based classifiers, k-Nearest Neighbor, Model evaluation Individual Assignment
Submission Deadline: **Thursday, November 10, 2022 (in the last lecture)** *Only hand-written solutions will be accepted.*

Problem # 1: Consider the data given below in a table with *Salary* as target variable. Since we have worked on classification methods, we need to first change the numeric target variable to categorical. Let's discretize the *Salary* variable as follows:

- Level1: Less than $35,000
- Level2: $35,000 to less than $45,000
- Level3: $45,000 to less than $55,000
- Level4: above $55,000

Occupation	Gender	Age	Salary
Service	Female	45	$48,000
	Male	25	$25,000
	Male	33	$35,000
Management	Male	25	$45,000
	Female	35	$65,000
	Male	26	$45,000
	Female	45	$70,000
Sales	Female	40	$50,000
	Male	30	$40,000
Staff	Female	50	$40,000
	Male	25	$25,000

a. Construct the CART decision tree to classify *Salary* based on the other variables. Use Gini Index for splitting and perform binary attribute splits.
b. Construct the C4.5 decision tree to classify *Salary* based on the other variables. Use Gain Ratio for splitting and perform binary attribute splits.
c. Compare the two decision trees and discuss the benefits and drawbacks of each.
d. Generate the full set of decision rules for both decision trees.
e. Compare the two sets of decision rules and discuss the benefits and drawbacks of each.

Problem # 2: Consider a binary classification problem with the following set of attributes and attribute values:

- Air Conditioner = {Working, Broken}
- Engine = {Good, Bad}
- Mileage = {High, Medium, Low}
- Rust = {Yes, No}

Suppose a rule-based classifier produces the following rule set:

$$\begin{array}{|l|}\hline \text{Mileage} = \text{High} \longrightarrow \text{Value} = \text{Low} \\ \text{Mileage} = \text{Low} \longrightarrow \text{Value} = \text{High} \\ \text{Air Conditioner} = \text{Working, Engine} = \text{Good} \longrightarrow \text{Value} = \text{High} \\ \text{Air Conditioner} = \text{Working, Engine} = \text{Bad} \longrightarrow \text{Value} = \text{Low} \\ \text{Air Conditioner} = \text{Broken} \longrightarrow \text{Value} = \text{Low} \\ \hline \end{array}$$

a. Are the rules mutually exclusive?
b. Is the rule set exhaustive?
c. Is ordering needed for this set of rules?
d. Do you need a default class for the rule set?

Problem # 3: Consider a training set that contains 150 positive examples and 320 negative examples. For each of the following candidate rules,

- R1: A $\longrightarrow$ + (covers 12 positive and 4 negative examples)
- R2: B $\longrightarrow$ + (covers 45 positive and 15 negative examples)
- R3: C $\longrightarrow$ + (covers 115 positive and 100 negative examples)

Determine which is the best and worst candidate rule according to:

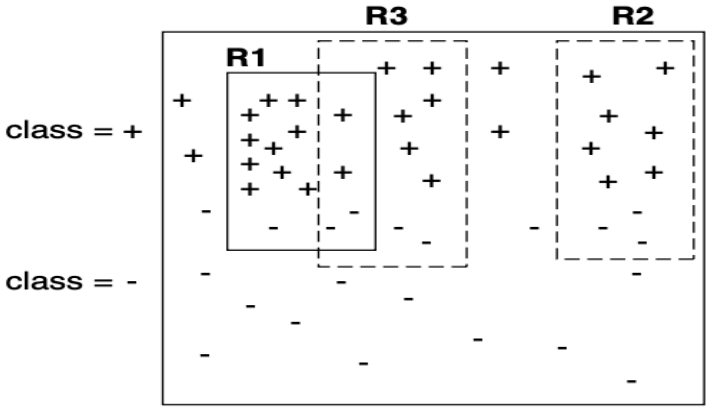
a. Rule accuracy
b. FOIL's information gain

Problem # 4: Consider the one-dimensional data set given below.

x	0.5	3.0	4.5	4.6	4.9	5.2	5.3	5.5	7.0	9.5
y	−	−	+	+	+	−	−	+	−	−

Classify the two data points $x_1 = 4.0$ and $x_2 = 7.5$ as + or – according to its 1-, 3-, 5-, and 9-nearest neighbors using: (a) unweighted majority vote, (b) distance-weighted majority vote.

Problem # 5: Consider the following diagram which shows coverage of three rules: R1, R2 and R3.



Determine which is the best and worst rule according to:

a) The rule accuracy after R1 has been discovered, where none of the examples covered by R1 are discarded).
b) The rule accuracy after R1 has been discovered, where only the positive examples covered by R1 are discarded).
c) The rule accuracy after R1 has been discovered, where both positive and negative examples covered by R1 are discarded.

Problem # 6: Consider the following data points with (x, y) dimensions, and construct the KD-Tree. Use the median value as splitting point and alternate dimensions at each point. Construct two different trees where one follows the order of x-y dimensions for splitting and the other follows y-x dimensions.

A = (65, 50), B = (60, 70), C = (70, 60), D = (75, 25), E = (50, 90), F = (90, 65), G = (10, 30), H = (80, 85), I = (95, 75)

Finally, show how the new point Z = (55, 60) can be classified using both trees? Did you find any difference in the performance of both trees? State your conclusions precisely (5-6 lines at maximum).