CS4038 – Data Mining (SPRING 2024)

Ayesha Liaqat

Assignment 2

Topics Covered: Classification algorithms: Decision tree, Rule-based classifiers, k-Nearest Neighbor, Model evaluation Individual Assignment
Submission Deadline: **Wednesday, April 3, 2024 (01:00 PM)** *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.

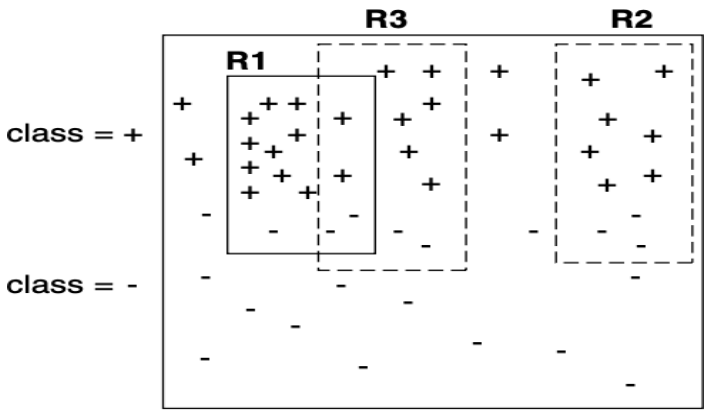
Problem # 2: Consider the following data set of Pizza outlet prediction and apply Kdtree by using the following splitting order Age and then Cheese Content. Splitting must be performed on the mean of max and min values. Once tree is constructed, draw graph to represent each split.

NAME	AGE	CHEESE CONTENT	PIZZA OUTLET
Riya	30	6.2	Pizza Hut
Manish	15	8	Dominos
Rachel	42	4	Pizza Hut
Rahul	20	8.4	Pizza Hut
Varun	54	3.3	Dominos
Mark	47	5	Pizza Hut
Sakshi	27	9	Dominos
David	17	7	Dominos
Arpita	8	9.2	Pizza Hut
Ananya	35	7.6	Dominos

Predict the Pizza outlet for the following given sample.

Harry	46	7	???

Problem # 3: 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 # 4: Consider the training examples shown in the following table for a binary classification. The table shows a training set for a problem of predicting whether a loan applicant will repay his/her loan obligation or defaulting on his/her loan.

Home Owner	Marital Status	Annual Income	Defaulted Borrower
Yes	Single	70,000	No
No	Married	60,000	No
Yes	Married	85,000	No
Yes	Single	80,000	Yes
No	Single	75,000	No
Yes	Married	90,000	No
No	Single	70,000	No
Yes	Married	100,000	No
No	Single	65,000	Yes
Yes	Married	110,000	No

Using the KNN approach, predict the class label for this test example:
X = (Home Owner = yes, Marital Status = Single, Annual Income = 120,000). Assume that K=3 and distance is L2 norm.
Classify test data using a) unweighted majority voting b) distance weighted majority voting.

Problem # 5: From the following Confusion matrix, answer the following questions:

Actual	**Predicted**		
	Roses	**Daisies**	**Tulips**
Roses	100	10	5
Daisies	15	85	20
Tulips	8	18	90

1) What is the total number of instances in the dataset, based on the confusion matrix provided?
2) What is the overall accuracy of the classifier?

3) Compute the sensitivity (recall) for each class. What insights does this provide about the classifier's performance for each class?
4) Determine the specificity for each class. How does specificity help evaluate the classifier's performance for individual classes?
5) Calculate the precision for each class. What does precision reveal about the classifier's ability to correctly identify each class?

Do the best you can Until you know better. Then when you know better, do better 😊