

Trent University
Department of Computing and Information Systems
Data Mining (COIS 4400H)
Fall 2016

Assignment 2
Due Wednesday, October 26th 2016 (noon)

Question 1 (20 points)

Given the following training and test instances classify each test instance using the k nearest neighbor classifier for k values of 1, 2, 4 and 8. Use Euclidean distance as the distance measure. Given your results, calculate the precision, recall, and f_1 measure for each value of k. Which value of k performed better? Justify your answer in terms of the metrics you calculated making certain to indicate what each metric means from a performance perspective.

Training Data

Attr. 1	Attr. 2	Attr. 3	Class
5.2	2.7	3.9	A
5.6	2.5	3.9	A
5.7	2.6	3.5	A
5.5	2.5	4	A
5.7	2.8	4.1	A
7.2	3.6	6.1	B
6	2.2	5	B
7.2	3	5.8	B
6.9	3.1	5.4	B
5.9	3	5.1	B

Test Data

Attr. 1	Attr. 2	Attr. 3	Class
6.6	2.9	4.6	A
6.7	3	5	A
5.8	2.8	5.1	B
6.7	3.3	5.7	B

Question 2 (20 points)

It is difficult to assess accuracy based on class membership when data may belong to more than one class at a time. Propose and discuss three criteria that you would use to compare the performance of different classifiers on such data.

Question 3 (20 points)

Given a decision tree, you have the option of a) converting the decision tree to rules and then pruning the resulting rules, or b) pruning the decision tree and then converting the pruned tree to rules. Which approach do you think should be preferred and why?

Question 4 (40 points)

Using Weka, analyze the dataset posted on WebCT and discuss the results (include screen shots). Use the following classifiers (using the default configurations with the exception of the classifier of your own choice) and 10-fold Cross Validation:

- a) MultilayerPerceptron
- b) IBk
- c) J48
- d) Your Choice

Based on the results of the above, answer the following questions:

- a) Which of the classifiers performed better in terms of the underrepresented class? Justify your answer.
- b) Consider your results from the IBk classifier, given the default configuration why might this classifier be a poor fit for such an unbalanced classification problem?
- c) In general terms suggest two approaches you might take to improve upon the classification of the underrepresented class. Discuss the advantages and disadvantages of each approach.