CPSC445/545/MBB334/545/CBB545 Introduction to Data Mining Spring 2007

Homework 2

(Due: Feb 20th 2007)

Non-programming Alternative

Binary Decision Trees

Using paper and pencil, briefly show the construction process of the **binary** decision trees based on the following datasets. Use information gain (the reduction in entropy) as the measure for node splitting. You may also design a reasonable criterion to stop splitting during this process.

Dataset 1: Consider the problem of learning the concept of whether or not to purchase a CD album.
Artist: possible values: BS, CA, MC.
Price: possible values: Cheap, Expensive.

Artist	Price	Class
CA	Cheap	Yes
BS	Expensive	No
MC	Cheap	No
BS	Cheap	No
CA	Expensive	Yes
MC	Expensive	Yes
CA	Cheap	Yes
MC	Expensive	Yes
BS	Cheap	No

Use the whole dataset above to construct the decision tree, and then decide whether or not to buy an expensive CD of MC.

Dataset 2: Consider the following dataset, use binary variables X1 and X2 to predict Y.

X1	X2	Class
0	0	0
0	1	1
1	0	1
1	1	0

Use **the first three cases as the training set** to construct the decision tree, and apply the resulting tree to classify the fourth case. Discuss possible ways to construct a more accurate decision tree for this dataset.

Programming Alternative

Binary Decision Trees

Write program(s) **in R** to solve the problems in the non-programming alternative section. **Instead of using packages such as rpart, you will have to implement the decision tree algorithm by yourself.** Please email the source code with compiling/running instructions to the TAs (Jiang Du: jiang.du@yale.edu; Edo Liberty: edo.liberty@yale.edu). You are not required to answer those discussion questions in the non-programming section.