Decision Tree

Instructor: Kunpeng Zhang (kzhang6@uic.edu)
TA: Minghong Xu (mxu29@uic.edu)

The Deadline: May 2, 2015 (Saturday)

Write a program in Java to implement the decision tree algorithm. You should read in a space delimited dataset in a file called dataset.txt and output to the screen your decision tree and the training set accuracy (the percentage your decision tree correctly classifies your data) in some readable format.

For example, here is the tennis dataset. The first line will contain the names of the fields. The last column is the classification attribute, and will always contain the values yes or no.

outlook temperature humidity wind playtennis
sunny hot high FALSE no
sunny hot high TRUE no
overcast hot high FALSE yes
rainy mild high FALSE yes
rainy cool normal FALSE yes
rainy cool normal TRUE no
overcast cool normal TRUE yes
sunny mild high FALSE no
sunny cool normal FALSE yes
rainy mild normal FALSE yes
sunny mild normal TRUE yes
overcast mild high TRUE yes
overcast hot normal FALSE yes
rainy mild high TRUE no

For output, you can choose how to draw the tree so long as it is clear what the tree is. You might find it easier if you turn the decision tree on its side, and use indentation to show levels of the tree as it grows from the left. For example:

---

outlook = sunny
| humidity = high: no
| humidity = normal: yes
outlook = overcast: yes
outlook = rainy
| windy = TRUE: no
| windy = FALSE: yes

You don't need to make your tree output look exactly like above: feel free to print out something similarly readable if you think it is easier to code.

Note:

1). To simplify our setting, all attributes here are categorical.

2). To obtain the best split attribute, you should the one with maximum information gain.

3). If a splitting attribute has *n* different possible values, you should split it into *n* sub-trees instead of two sub-trees.

4). For stopping criteria at a node, please use purity.

Requirements:

- Your codes must be readable and clean.
- When you submit your codes through blackboard, you need to put all source codes (.java files, NOT jar files) and some other optional files (e.g., a readme file) into one folder and name that folder as <YOUR UID>_ASSIGN8.
- Assignments not following this rule will not be graded. In addition, no resubmission after TA grades it. Late submission rule: 10% deduction for one day late. Late submission over a week is NOT acceptable.

DO NOT copy any codes from others. Otherwise, both will be penalized.