CSCI 4144, Winter 2017

Assignment 4: ID3 Algorithm Implementation

(Issue: Feb 28, Due: Mar 15)

- **TA:** Purvesh Patel (purvesh.patel@dal.ca)
- **Ass4 Tutorial:** Mar 9, 6:00-7:30PM, CS 127
- **Ass4 Help Hours:** Mar 7 & 14, 1:00-2:30PM, CS 233; Mar 9, 6:00-7:30PM, CS 127

---

1. Objectives:

1) To gain an in-depth understanding on decision tree (DT) classification.
2) To learn how to handle the main technical issues of implementing a DT algorithm.
3) You may also gain a team work experience (allowed in a group of 2 students).

2. Programming language and computer system:

The implementation is required to use a serious production implementation language, such as Java family/C family, etc. (but not a script language, such as R), which should have a compiler on bluenose.

3. Data sets and interface design requirements:

1) Your program should be able to handle data files with the following format: the first line contains column headings (i.e., attributes), and every following line contains the values that represent a tuple.
2) Five data files are provided for developing and testing your program. The datasets data1 and data3 are the same data used in Ass3. The file data2 is an extension of data1 with more tuples and few records containing some noise. The files of adult1 and adult2 are from data2 in Ass3, but the attribute "age" of adult2 has been converted into binary domain.
3) Your program should have an interface allowing the user to choose 1) a data file, 2) a target attribute from the available binary attributes of the data set.
4) The mined DT rules should be placed into an external file named as "Rules". It is encouraged to provide a usefulness measure to each mined decision rule.
5) An implementation example (executable code) is available from Ass/Ass4-demo/ID3.

4. Submit your Ass4 electronically:

1) Create a directory assign4 in your bluenose account. This directory should include the developed source code, Makefile, and README file. The README file should provide the instructions how to compile the code and run the program. It should also provide a brief description of the overall code architecture (the functions and the call relationships, etc).
2) Submit the assignment directory from your home directory by the command line: submit assign4.
3) Do not submit any data.

5. Evaluation:

Your assignment will be evaluated based upon the overall quality of the work including user interface, functionality, modularity and readability of the program, and the clarity of the README file.

** You are allowed to discuss about the algorithm and various implementation options, but you must design and generate your own code individually.

Plagiarism and Intellectual Honesty: (http://plagiarism.dal.ca) Dalhousie University defines "plagiarism as the presentation of the work of another author in such a way as to give one's reader reason to think it to be one's own." Plagiarism is considered a serious academic offence which may lead to loss of credit, suspension or expulsion from the University, or even the revocation of a degree.

Good Luck & Have Fun!