DT228-4 AI2 Assignment: Forest Type Prediction

Submission Deadline: Wednesday 15th April 2015 at 11:59 p.m.

NOTE: This document is 3 pages long. Make sure you read all 3 pages as you may lose marks if you do not follow the instructions correctly.

Task Description: In this assignment you will develop a classifier that uses geographical data to predict the type of forest in an area.

Team work: Find a partner to work on this project. (Two person teams are encouraged, though you may work alone if you prefer. No >=three person teams please).

What you are given:

I have made three files available on webcourses:

1. *datadescription.txt*: this file contains a description of the data types of the different columns in the data
2. *trainingset.txt*: 435148 training instances. This file lists a training id, the descriptive features and the target feature level for each instance.
3. *queries.txt*: 145864 query instances. This file lists a test id, the descriptive features for each instance. However, the target feature level has been overwritten with '?'

Submission Deadline and Late Submissions:

- **Deadline: Wednesday 15th April 2015 at 11:59 p.m.**
- Marks will be deducted for late submission - 10% per day late.

How do you submit your solution?

- You submit your assignment work through the Assignment Submission Form I have set up on the Webcourse module.

What you need to submit:

You need to submit **separate** 2 files (don't bundle them in a zip file submit the two files separately): (1) a solutions file, and (2) the Python code for you classifier. Details on the naming convention and format of these files are given below:

1. The **solutions file**:
   a. Naming convention: This file should be named using the following convention 'studentnumber1+studentnumber2.txt', where studentnumber1 is the student number of the first member of the team and studentnumber2 is the student number of the other member of the team. For example, the file C1234567+D9876543.txt is the correct name for the solution file for a team comprising of the students C1234567 and D9876543. If you are working by yourself name your solution file 'studentnumber.txt', e.g. C1234567.txt
   b. Contents: The file should list your classifier's target variable predictions for each of the query instances in the *queries.txt* file. Each line in the file should list one query id followed by a comma followed by your classifier's prediction for that query, i.e.:

      <tstid>,<prediction>

      The box below illustrates what someone looking at a portion of your solutions files should see

      ```
      tst100,<=50K
      tst101,>50K
      tst102,<=50K
      tst103,<=50K
      tst104, >50K
      ```

2. The **Python code** for your classifier.
   a. Naming convention: This file should be named using the following convention '**studentnumber1+studentnumber2.py**', where studentnumber1 is the student number of the first member of the team and studentnumber2 is the student number of the other member of the team. For example, the file C1234567+D9876543.py is the correct name for the Python file for a team comprising of the students C1234567 and D9876543. If you are working by yourself name your solution file 'studentnumber.txt', e.g. C1234567.py
   b. Contents: This code should expect the training and query data to be in a subdirectory of the directory its in called 'data'. It should have a main function which when run creates your solution file and stores it in a subdirectory called 'solutions'. Make sure to include your names and student numbers as comments at the top of you python code file. Your code should be appropriately commented.

Marking Scheme

Marks are awarded based on the accuracy of the classifier. The accuracy metric used will be the **average class accuracy** (harmonic mean) of the classifier.

Marks may be deducted for the following reasons:

(a) Late submission (including submissions that are incomplete by the time the deadline has passed): 10 marks per day late.

(b) Incorrect submission: 10 marks will be deducted if your submission is does not follow the stated formats. The reason that I do this is that if you do not correctly format you submission this slows down the correction process for everyone. I will be strict in deducting these marks. If your submission does not follow the guidelines these marks will be deducted. Examples of the types of errors that will result in these marks not be awarded include:

- Solution file named incorrectly
- Leaving blank lines between solutions in the solutions file
- Using incorrect labels or using the wrong case for your labels e.g., using lowercase ks
- Having trailing blank spaces after key values before commas in the solutions file
- Forgetting to put commas between the fields in the solutions file
- The solutions file not being a .txt file, for example submitting your solution as an .rtf or other file format (or bundling your two files as a zip file submission)
- Not commenting your code clearly
- Not putting your name and student number at the top of your code