AI2 Assignment

Submission Deadline: Monday 10th April 2017 at 9am

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 data to predict the outcome of a Bank marketing campaign.

Team work: For this assignment you can work on your own, or with a partner but no three person teams or larger, please. Be aware, however, that if you work as a team everyone on the team gets the same mark, irrespective of who does what. Also, if you have a falling out with your teammate I will not be involved in resolving the problem. Also, the marking is the same irrespective of whether you work by yourself or with another person.

What you are given:
There are 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*: this file contains the training instances. This file lists a training id, the descriptive features and the target feature level for each instance.
3. *queries.txt*: this file contains the 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: Monday 10th April 2017 9am**

Marks will be deducted for late submission.

- **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 3 **separate** files (don't bundle them in a zip file submit the three files separately):
(1) a solutions file,
(2) the Python code for you classifier,
(3) short 1 page description of how you solved the problem and any decisions you had to make.

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 the format of your solution file.

```
TEST1,"TypeA"
TEST2,"TypeA"
TEST3,"TypeB"
TEST4,"TypeA"
…
```

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 directory it is 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 both **your documentation** and 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 does not follow the stated formats. The reason for this is because the outputs will be automatically processed and evaluated, so if you do not correctly format you submission I will have to manually modify and tweak it to make it conform to the required format, and 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 Ts
- 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