Artificial Intelligence Project II

Due Date: 11:59PM 06/05/2017

Description:

For AI Project II, you are required to implement:

(a) a Naïve Bayesian (NB) classifier introduced in lectures

(b) a k-fold stratified CV tool, where k can vary

Test your NB classifier and C4.5 (posted on e3) on **5** different datasets from UC Irvine's Machine Learning Repository:

adult, car, isolet, page-blocks, winequality. (In dataset folder)

Conduct a 10-fold stratified CV, using your CV tool, on each dataset, and then perform a paired *t*-test (e.g. using MS Excel) between your NB classifier and C4.5.

Show the results of each 10-fold CV in a table which presents the accuracy for each fold, the average accuracy of 10 folds, and the *p*-value from your *t*-test.

Example:

Assume we have a dataset testData.data:

Attribute1	Attribute2	Class
0	0	A
1	1	B
1	2	B
0	1	A
2	0	B
1	0	A

With your CV tool, if we set **k = 3 (i.e. 3-fold cv),** It should output **6** files:

testData_cv1.data:

0	0	A
2	0	B
1	0	A
1	2	B

testData_cv1.test:

1	1	B
0	1	A

testData_cv2.data:

1	1	B
0	1	A
1	0	A
1	2	B

testData_cv2.test:

0	0	A
2	0	B

testData_cv3.data:

1	1	B
0	1	A
0	0	A
2	0	B

testData_cv3.test:

1	0	A
1	2	B

The CV your program outputted may be different from example, that is fine but it should be stratified (i.e. the proportion of class in each cv is the same as input data).

Because the design of c4.5, you also need to copy testData.names to **3** files: testData_cv1.names, testData_cv2.names and testData_cv3.names.
See example folder for another example of 10-fold CV.

Then you can run the following commands to get the predict result of each cv:
c4.5 -f testData_cv1 -u
c4.5 -f testData_cv2 -u
c4.5 -f testData_cv3 -u

Requirement:
The code of NB classifiers and CV tool **must be** implemented in C/C++ only. Using any other programming language will incur penalty.

What to turn in:
(a) Your source code of NB and CV
(b) Your report that includes the following.
   I. The tables of your experimental results
   II. The description of your way to deal with numeric attributes, e.g. temperature.

Output format:
Please including 3 section: (You will get penalty if you don't fellow this format)
1. Briefly describe the architecture of your program:
   Design of architecture, how to deal with attributes, e.g.
2. Analysis the result of NB, CV and C4.5
   A. Result: Please use **result_template.xlsx** or **result_template.ods** as template, DO NOT just paste all the output of the program!
   B. Discuss the result
3. Discuss the problem you encounter and how to solve it.

Testing environment:
Cpu: i7-6700, Memory: 8GB
Operating system: Windows 10 (64x) or Linux Mint 18.1 Serena (64x)
Compiler version: gcc: 5.4.0, g++: 5.4.0
(Please note which operating system you want us to use in your report, or we will run it on Windows 10)