CS 412: Introduction to Data Mining **Summer 2017**

Homework 4

Handed Out: July 18, 2017 *Due: August* 3, 2017 *11:59 pm*

1 General Instructions

- The programming assignment will be hosted on hackerrank (https://www.hackerrank.com/) as a programming contest. To participate in this contest, please open a hackerrank account with your illinois.edu email id. If your username in hackerrank is different from your net id, let us know by filling out your net id and username in the spreadheet (link provided in Piazza). The contest framework will allow you to verify the correctness of your submission based on a set of sample test cases. We may use additional test cases to grade your submission.
- It is OK to discuss the problems with the TAs and your classmates, however, it is NOT OK to work together or share code. Plagiarism is an academic violation to copy, to include text from other sources, including online sources, without proper citation. To get a better idea of what constitutes plagiarism, consult the CS Honor code (http://cs.illinois.edu/academics/honor-code) on academic integrity violations, including examples, and recommended penalties. There is a zero tolerance policy on academic integrity violations. Any student found to be violating this code will be subject to disciplinary action.
- Please use Piazza if you have questions about the homework. Also feel free to send TAs emails and come to office hours.

2 Programming Assignment Instructions

This question aims to provide you a better understanding of the basic concepts of classification. Participate in the programming contest hosted at hackerrank: www.hackerrank.com/homework4.

- Please read the problem description carefully.
- The input will always be valid. We are mainly testing your understanding of frequent pattern mining, not your coding skills.
- Please pay special attention to the output format. We will be using the hackerrank based autograder and it is extremely important that your generated output satisfies the requirement.
- We don't have specific constrains for this programming question. The only constrains are the standard environment constraints in hackerrank: https://www.hackerrank.com/environment.

- The grading will be based on how many total test cases you passed. You are provided with two sample test cases to test your code. For the final grading, we will use additional test cases to test your code.
- If you have any questions, post on piazza.

3 Programming Assignment (50 points)

Given a training dataset with several categorical attributes and a target label, you have to identify the best attribute for splitting the dataset.

You have to choose the best attribute based on their **information gain** and **gain ratio** values.

Input Format

First line of the input contains the number of lines (1+ number of training instances).

Second line contains the name of the attributes and the target label (last) which are comma separated.

The following lines are training instances which are comma-separated: the first categorical attribute value, second categorical attribute, and so on and finally the target label value.

Please refer to the sample input below for an example.

Constraints

NA

Output Format

The output will be the best attribute based on information gain followed by the best attribute based on gain ratio on separate lines.

Please refer to the sample output below for an example.

Sample Input 0

```
15
age,income,student,creditrating,buyscomputer?
l30,high,no,fair,no
l30,high,no,excellent,no
```

```
31to40 ,high ,no ,fair ,yes
g40,medium,no,fair ,yes
g40,low,yes,fair ,yes
g40,low,yes,excellent ,no
31to40 ,low,yes,excellent ,yes
l30 ,medium,no,fair ,no
l30 ,low,yes,fair ,yes
g40,medium,yes,fair ,yes
l30 ,medium,yes,excellent ,yes
31to40 ,medium,no,excellent ,yes
31to40 ,high ,yes ,fair ,yes
g40,medium,no,excellent ,no
```

Sample Output 0

```
age
age
```