| **CS 412: Introduction to Data Mining** | **Spring 2018** |
|---|---|
| Homework 4 | |
| *Handed Out: March* 29, 2018 | *Due: April* 11, 2018 *11:59 pm* |

# 1 General Instructions

- This assignment is due at 11:59 PM on the due date. We will be using Compass (http://compass2g.illinois.edu) for collecting the non-programming part of this assignment. Contact TAs if you face technical difficulties in submitting the assignment. We shall NOT accept any late submission!
- The programming part of this 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. Please check the assignment page on course website in a couple of days for accessing the contest.
- It is OK to discuss with your classmates and your TAs regarding the methods, but 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.
- For each non-programming question, you will **NOT** get full credit if you only give out a final result. Necessary calculation steps and reasoning are required to be shown.
- Please use Piazza if you have questions about the homework. Also feel free to send TAs emails and come to office hours.

# 2 Question 1 (10 points)

This question aims to provide you a better understanding of classification, especially for decision tree. Suppose we want to predict if a candidate will be accepted to the PhD program of some University X, given the student's information about GPA, university, publications, and recommendation. Please answer the following questions based on Figure 1.

(a) What's the information gain for the 'univ' attribute? Please show your calculation.

(b) Now suppose we want to use Gini Index as attribute selection measure. What's the Gini index for the attribute 'published'? What's the reduction in impurity in terms of Gini Index? Please show your calculation.

| id | GPA | univ | published | recommendation | accepted |
|---|---|---|---|---|---|
| 1 | 4.0 | top-10 | yes | good | yes |
| 2 | 4.0 | top-10 | no | good | yes |
| 3 | 4.0 | top-20 | no | normal | yes |
| 4 | 3.7 | top-10 | yes | good | yes |
| 5 | 3.7 | top-20 | no | good | yes |
| 6 | 3.7 | top-30 | yes | good | yes |
| 7 | 3.7 | top-30 | no | good | no |
| 8 | 3.7 | top-10 | no | good | no |
| 9 | 3.5 | top-20 | yes | normal | no |
| 10 | 3.5 | top-10 | no | normal | no |
| 11 | 3.5 | top-30 | yes | normal | no |
| 12 | 3.5 | top-30 | no | good | no |

Figure 1: Data for university admissions

**Solution.**

---

# 3 Question 2 (30 points)

This question aims to provide you a better understanding of classification, especially for Naive Bayes. Refer to Figure 1 for answering the following questions.

(a) What is the prior probability of *accepted* being *yes/no* estimated from the data?

(b) Given *accepted=yes*, what is the probability of *GPA* attribute taking each of the values $\in$ *(4.0; 3.7; 3.5)*? (You have to report probabilities for each value)
Also, calculate the probabilities for rest of the attributes (*university, published, recommendation*) taking each of its possible values given that *accepted=yes.*

(c) Calculate similar probabilities asked in (b) given that *accepted=no.*

(d) Based the results you got from (a)-(c), given a student with attributes (*GPA=3.7, university=top-20, published=yes, recommendation=good*), calculate the probability of the student being accepted. What will the probability become if the student has (

*GPA=3.7, university=top-30, publication=no, recommendation=normal*)?

# 4 Question 3 (10 points)

Suppose a sequence database D (shown in table below) contains three sequences as follows. Note (bc) means that items b and c are purchased at the same time (i.e., in the same transaction). The following questions require you to perform GSP algorithm. Let the **minimum support** be 3.

| Customer Id | Shopping Sequence |
|---|---|
| 1 | a(bc)(de)f |
| 2 | bc(ad)ef |
| 3 | a(bc)d(ab)ef |

(a) Scan database once, list length-1 sequential pattern candidates $C_1$ and the result $L_1$ after pruning.

(b) Following, generate $C_2$, and $L_2$. *Hint: do not miss any candidates in* $C_2$

(c) Now, generate $C_3$, $L_3$, $C_4$, $L_4$ and longer candidates/results until the algorithm terminates.

# 5 Question 4 (50 points)

The programming part is hosted on hacker rank.