CptS 315: Introduction to Data Mining
Homework 1

(Due date: Feb 9th midnight)

Instructions

- Please use a word processing software (e.g., Microsoft word) to write your answers and submit one zip file (PDF of answers and code) on Canvas. The rationale is that it is sometimes hard to read and understand the hand-written answers.
- All homeworks should be done individually.

Analytical Part (40 points)

Q1. Consider the following market-basket data, where each row is a basket and shows the list of items that are part of that basket.

1. $\{A, B, C\}$
2. $\{A, C, D, E\}$
3. $\{A, B, F, G, H\}$
4. $\{A, B, X, Y, Z\}$
5. $\{A, C, D, P, Q, R, S\}$
6. $\{A, B, L, M, N\}$

a) What is the *absolute* support of item set $\{A, B\}$? (3 points)

b) What is the *relative* support of item set $\{A, B\}$? (3 points)

c) What is the confidence of association rule $A \Rightarrow B$? (3 points)

Q2. Answer the below questions about storing frequent pairs using triangular matrix and tabular method.

a) Suppose we use a triangular matrix to count pairs and the number of items $n = 20$. If we store this triangular matrix as a *ragged* one-dimensional array $Count$, what is the index where count of pair $(7, 8)$ is stored? (3 points)

b) Suppose you are provided with the prior knowledge that only ten percent of the total pairs will have a non-zero count. In this case, which method among triangular matrix and tabular method should be preferred and why? (3 points)

Q3. This question is about the PCY algorithm for counting frequent pairs of items. Suppose we have six items numbered 1, 2, 3, 4, 5, 6. Consider the following twelve baskets.

1. $\{1, 2, 3\}$
2. $\{2, 3, 4\}$
3. $\{3, 4, 5\}$
4. $\{4, 5, 6\}$
5. $\{1, 3, 5\}$
6. $\{2, 4, 6\}$
7. $\{1, 3, 4\}$
8. $\{2, 4, 5\}$
9. $\{3, 5, 6\}$
10. $\{1, 2, 4\}$
11. $\{2, 3, 5\}$
12. $\{3, 4, 6\}$

Suppose the support threshold is 4. On the first pass of the PCY algorithm, we use a hash table with 11 buckets, and the set $\{i, j\}$ is hashed to $i \times j$ mod 11.

a) By any method, compute the support for each item and each pair of items. (5 points)

b) Which pairs hash to which buckets? (5 points)

c) Which buckets are frequent? (3 points)

d) Which pairs are counted on the second pass of the PCY algorithm? (2 points)

Q4. Please read the following paper and write a brief summary of the main points in at most ONE page. You can skip the theoretical parts. (10 points)

Saul Schleimer, Daniel Shawcross Wilkerson, Alexander Aiken: Winnowing: Local Algorithms for Document Fingerprinting. SIGMOD Conference 2003: 76-85
`https://theory.stanford.edu/~aiken/publications/papers/sigmod03.pdf`

Programming and Experimental Part (60 points)

Product Recommendations: The action or practice of selling additional products or services to existing customers is called cross-selling. Giving product recommendation is one of the examples of cross-selling that are frequently used by online retailers. One simple method to give product recommendations is to recommend products that are frequently browsed together by the customers.

Suppose we want to recommend new products to the customer based on the products they have already browsed on the online website. Write a program using the A-priori algorithm to find products which are frequently browsed together. Fix the support to s =100 (i.e., product pairs need to occur together at least 100 times to be considered frequent) and find itemsets of size 2 and 3.

Use the online browsing behavior dataset provided with this homework. Each line represents a browsing session of a customer. On each line, each string of 8 characters represents the id of an item browsed during that session. The items are separated by spaces.

a) Identify pairs of items (X, Y) such that the support of $\{X, Y\}$ is at least 100. For all such pairs, compute the confidence scores of the corresponding association rules: $X \Rightarrow Y$, $Y \Rightarrow X$. Sort the rules in decreasing order of confidence scores and list the top 5 rules in the writeup. Break ties, if any, by lexicographically[1] increasing order on the left hand side of the rule.

b) Identify item triples (X, Y, Z) such that the support of $\{X, Y, Z\}$ is at least 100. For all such triples, compute the confidence scores of the corresponding association rules: $(X, Y) \Rightarrow Z$, $(X, Z) \Rightarrow Y$, $(Y, Z) \Rightarrow X$. Sort the rules in decreasing order of confidence scores and list the top 5 rules in the writeup. Order the left-hand-side pair lexicographically and break ties, if any, by lexicographical order of the first then the second item in the pair.

Instructions for Code Submission and Output Format.
Please follow the below instructions. It will help us in grading your programming part of the homework.

- Supported programming languages: Python, Java, C++
- Store all the relevant files in a folder and submit the corresponding zipfile named after your student-id, e.g., 114513209.zip
- This folder should have a script file named

  `run_code.sh`

  . Executing this script should do all the necessary steps required for executing the code including compiling, linking, and execution.

[1] `https://en.wikipedia.org/wiki/Lexicographic_order`

- Assume relative file paths in your code. Some examples:

  ```
  ``./filename.txt'' or ``../hw1/filename.txt''
  ```

- The output of your program should be dumped in a file named “output.txt” in the following format (random example, not real output):

  ```
  OUTPUT A
  FRO11987 FRO12685 0.4325
  FRO11987 ELE11375 0.4225
  FRO11987 GRO94758 0.4125
  FRO11987 SNA80192 0.4025
  FRO11987 FRO18919 0.4015
  OUTPUT B
  FRO11987 FRO12685 DAI95741 0.4325
  FRO11987 ELE11375 GRO73461 0.4225
  FRO11987 GRO94758 ELE26917 0.4125
  FRO11987 SNA80192 ELE28189 0.4025
  FRO11987 FRO18919 GRO68850 0.4015
  ```

  Explanation.

  - Line 1 should have “Output A”
  - Next five lines should have the top five rules with decreasing confidence scores for part (a) of the programming question. Format: $< item1 > < item2 > < confidence >$ meaning $\{item1\} \Rightarrow item2$
  - Line 7 should have “Output B”
  - Next five lines should have the top five rules with decreasing confidence scores for part (b) of the programming question. Format: $< item1 > < item2 > < item3 > < confidence >$ meaning $\{item1, item2\} \Rightarrow item3$

- Make sure the output.txt file is dumped when you execute the script

  ```
  run_code.sh
  ```

- Zip the entire folder and submit it as

  ```
  <student_id>.zip
  ```

Grading Rubric

Each question in the students work will be assigned a letter grade of either A,B,C,D, or F by the Instructor and TAs. This five-point (discrete) scale is described as follows:

- **A) Exemplary (=100%)**.
  Solution presented solves the problem stated correctly and meets all requirements of the problem.
  Solution is clearly presented.
  Assumptions made are reasonable and are explicitly stated in the solution.
  Solution represents an elegant and effective way to solve the problem and is not overly complicated than is necessary.

- **B) Capable (=75%)**.
  Solution is mostly correct, satisfying most of the above criteria under the exemplary category, but contains some minor pitfalls, errors/flaws or limitations.

- **C) Needs Improvement (=50%)**.
  Solution demonstrates a viable approach toward solving the problem but contains some major pitfalls, errors/flaws or limitations.

- **D) Unsatisfactory (=25%)**
  Critical elements of the solution are missing or significantly flawed.
  Solution does not demonstrate sufficient understanding of the problem and/or any reasonable directions to solve the problem.

- **F) Not attempted (=0%)**
  No solution provided.

The points on a given homework question will be equal to the percentage assigned (given by the letter grades shown above) multiplied by the maximum number of possible points worth for that question. For example, if a question is worth 6 points and the answer is awarded a B grade, then that implies 4.5 points out of 6.