CSC4008 Assignment 2

Start: Oct 19 at 00:00

End: Nov 7 at 23:59

Correspondence TA/USTF: Xinyang Gao (120030046@link.cuhk.edu.cn)

Important Notes

1. The assignment is an individual project, to be finished on one's own effort.
2. The work must be submitted before the deadline. Late submissions within 2 days will apply 20% penalty. Late submissions after 2 days will not be accepted.
3. Plagiarism is strictly forbidden, regardless of the role in the process. Notably, ten consecutive lines of identical codes are treated as plagiarism. Depending on the seriousness of the plagiarism, 30%-100% marks will be deducted.
4. Please read the document carefully. No argument will be accepted on issues that have been specified clearly in the documents.

1 Products You Might Buy [80pts]

Association Rules are frequently used for Market Basket Analysis (MBA) by retailers to understand the purchase behavior of their customers. This information can be then used for many different purposes such as cross-selling and up-selling of products, sales promotions, loyalty programs, store design, discount plans and many others.

Evaluation of item sets: Once you have found the frequent itemsets of a dataset, you need to choose a subset of them as your recommendations. Commonly used metrics for measuring significance and interest for selecting rules for recommendations are:

1. **Confidence** (denoted as $\text{conf}(A \to B)$): Confidence is defined as the probability of occurrence of B in the basket if the basket already contains A:

$$\text{conf}(A \to B) = \Pr(B|A),$$

   where $\Pr(B|A)$ is the conditional probability of finding item set B given that item set A is present.

2. **Lift** (denoted as $\text{lift}(A \to B)$): Lift measures how much more "A and B occur together" than "what would be expected if A and B were statistically independent":

$$\text{lift}(A \to B) = \frac{\text{conf}(A \to B)}{\Pr(B)},$$

   where $\Pr(B) = \frac{\text{Support}(B)}{N}$ and $N =$ total number of transactions (baskets).

3. **Interest** (denoted as interest$(A \to B)$): Interest measures the difference between its confidence and the fraction of baskets that contain B:

$$\text{interest}(A \to B) = \text{conf}(A \to B) - \Pr(B).$$

(a) [10pts]

A drawback of using confidence is that it ignores $\Pr(B)$. Why is this a drawback? Explain why lift and interest do not suffer from this drawback.

(b) [10pts]

A measure is symmetrical if measure$(A \to B)$ = measure$(B \to A)$. Which of the measures presented here are symmetrical? For each measure, please provide either a proof that the measure is symmetrical, or a counterexample that shows the measure is not symmetrical.

Application in 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 online. Write a program using the Apriori 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.

Data:

- Associated data file is `browsing.txt` in `A2/`, which is the online browsing behavior dataset.
- 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. Some lines contain duplicate items. Removing or ignoring duplicates should not impact your results. Below is the first 2 lines in `browsing.txt`:

```
1    FRO11987 ELE17451 ELE89019 SNA90258 GRO99222
2    GRO99222 GRO12298 FRO12685 ELE91550 SNA11465 ELE26917
       ELE52966 FRO90334 SNA30755 ELE17451 FRO84225 SNA80192
```

(c) [30pts = 25 for rules + 5 for code]

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[1], if any, by lexicographically increasing order on the left hand side of the rule.

[1]https://softwareengineering.stackexchange.com/questions/403743/what-means-breaking-ties-in-context-of-sorting

(d) [30pts = 25 for rules + 5 for code]

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.

Tips:

- You do not need to use Spark for this program. If you want to use Spark, use Google Colab to use Spark seamlessly, e.g., copy and adapt the setup cells from Colab 0.
- For sanity check, there are 647 frequent items after 1st pass ($|L_1| = 647$).
- The top 5 pairs you should produce in part (c) all have confidence scores greater than 0.985.

2 Min-hashing [10pts]

Table 1 is a matrix with six rows.

Elements	S_1	S_2	S_3	S_4
0	0	1	0	1
1	0	1	0	0
2	1	0	0	1
3	0	0	1	0
4	0	0	1	1
5	1	0	0	0

Table 1: Input Matrix

(a) [6pts]

Compute the minhash signature for each column if we use the following three hash functions to simulate the random row permutation. For example, if we use h_1, $h_1(0) = 1$, which means row 0 is mapped to row 1 after the permutation.

1. $h_1(x) = 2x + 1 \bmod 6$;
2. $h_2(x) = 3x + 2 \bmod 6$;
3. $h_3(x) = 5x + 2 \bmod 6$;

(b) [4pts]

Which of these hash functions produce true permutations? Here, a true permutation means that the numbers generated by the hash function for row numbers form a permutation of 0 to 5.

3 Estimation via Sampling [10pts]

Assume universities are unique, but a courseID is unique only within a university (i.e., different universities may have different courses with the same ID, e.g., "CS101") and likewise, studentID's are unique only within a university (different universities may assign the same ID to different students). In the data, each tuple contains universityID, courseID, studentID, grade. Suppose we want to answer certain queries approximately from a 1/20th sample of the data. For each of the queries below, indicate how you would construct the sample to avoid biased estimation. That is, tell what the key attributes should be, and give a short explanation. You only need to explain how to construct the sample, do not need to explain how to compute the fractions based on your sample. (Hint: refer to page 18 of ch05a)

(a) [5pts] Estimate the fraction of students who have taken at least 5 courses.

(b) [5pts] Estimate the fraction of courses where at least half the students got "A."

What to submit

You need to submit the following two files to BlackBoard. Please format your files as "student_id.pdf" and "student_id.py" (or "student_id.ipynb"). For example, if your student id is 123456789, then you should submit "123456789.pdf" and "123456789.py" (or "123456789.ipynb"). You can submit several files in one submission. Don't submit them in different submission.

1. A writeup contains
    - Explanation for 1(a). [10pts]
    - Proofs and/or counterexamples for 1(b). [10pts]
    - Top 5 rules with confidence scores for 1(c). [25pts]
    - Top 5 rules with confidence scores for 1(d). [25pts]
    - Answer to 2(a). [6pts]
    - Answer to 2(b). [4pts]
    - Answer to 3(a). [5pts]
    - Answer to 3(b). [5pts]
2. Your code for 1. [10pts]