Seoul National University

M1522.001400 Introduction to Data Mining

Fall 2022, U Kang

Homework 4: Frequent Itemsets, Clustering (Chapter 6,7)

Due: Nov. 16, 23:59

**Reminders**

- Lead T.A.: Jaemin Park (ytreqw97@snu.ac.kr)
- The points of this homework add up to 100.
- Please answer *in English*. Homework written in Korean may get no points.
- **All the homework should be uploaded to the eTL in PDF format. No handwritten homework** (including scanned PDF files from handwritten papers) will be accepted.
- Whenever you are making an assumption, please state it clearly.
- If you have a question about assignments, please upload your question in eTL.

**Submissions**

- **Submission after the due day will be regarded as the late submission even if it is only one minute late.**
- If you want to use slip-days or consider late submission with penalties, please note that you are allowed *one week* to submit your homework after the due date.
- You don’t need to specify whether to use the slip-days; they are automatically used.

**Question 1 [20 points]**

Table 1 shows a supermarket transaction log with 6 distinct baskets. Each basket contains 3 to 4 items.

Answer the following questions.

| Basket | Items |
|---|---|
| 1 | a, b, c |
| 2 | a, b, d |
| 3 | b, c, d, e |
| 4 | c, e, f |
| 5 | b, c, f |
| 6 | a, c, e |

Table 1 Transaction Log

(a) [5 pt] What are the support for each of the itemsets {a}, {b}, {c}, {b, c} and {a, c, e}?

(b) [5 pt] If the minimum support is 3, give all frequent itemsets.

(c) [5 pt] What is the confidence of {a,b} $\rightarrow$ c ?

(d) [5 pt] What is the interest of {a,c} $\rightarrow$ e ?

**Question 2 [20 points]**

Trace the results of using the A-Priori algorithm on the following department store example with support threshold s=2 and confidence threshold c=0.6. Show the candidates and frequent itemsets for each database scan(pass). Enumerate all the final frequent itemsets. Also indicate the association rules that are generated. You must write down the results from each scan(pass).

| Basket | Items |
|---|---|
| B1 | Hat, Belt, Knitwear |
| B2 | Hat, Belt |
| B3 | Hat, Coat, Jacket |
| B4 | Jacket, Coat |
| B5 | Jacket, Knitwear |
| B6 | Hat, Coat, Jacket |

**Question 3 [30 points]**

Here is a collection of sixteen baskets. Each contains three of the seven items 1 through 7.

| | | | |
|---|---|---|---|
| {1, 2, 3} | {2, 3, 4} | {3, 4, 5} | {4, 5, 6} |
| {1, 3, 5} | {2, 4, 6} | {1, 3, 4} | {2, 4, 5} |
| {3, 5, 6} | {1, 2, 4} | {2, 3, 5} | {3, 4, 6} |
| {1, 2, 7} | {1, 3, 7} | {2, 4, 7} | {4, 6, 7} |

Suppose the support threshold is 4. On the first pass of the PCY Algorithm, we use a hash table with 13 buckets, and the set {i, j} is hashed to bucket $i \times j$ mod 13. Answer the following questions.

(a) [10 pt] Which pairs hash to which buckets?

(b) [10 pt] Which buckets are frequent?

(c) [10pt] Which pairs are counted on the second pass of the PCY Algorithm?

**Question 4 [30 points]**

Cluster the following 8 examples into 3 clusters using the k-means clustering and Euclidean distance based on the following units:

A = (2,10), B = (2,5), C = (8,4), D = (5,8), E = (7,5), F = (6,4), G = (1,2), H = (4,9)

Below table demonstrates the distance matrix based on the Euclidean distance.

| | **A** | **B** | **C** | **D** | **E** | **F** | **G** | **H** |
|---|---|---|---|---|---|---|---|---|
| **A** | 0 | $\sqrt{25}$ | $\sqrt{72}$ | $\sqrt{13}$ | $\sqrt{50}$ | $\sqrt{52}$ | $\sqrt{65}$ | $\sqrt{5}$ |
| **B** | | 0 | $\sqrt{37}$ | $\sqrt{18}$ | $\sqrt{25}$ | $\sqrt{17}$ | $\sqrt{10}$ | $\sqrt{20}$ |
| **C** | | | 0 | $\sqrt{25}$ | $\sqrt{2}$ | $\sqrt{2}$ | $\sqrt{53}$ | $\sqrt{41}$ |
| **D** | | | | 0 | $\sqrt{13}$ | $\sqrt{17}$ | $\sqrt{52}$ | $\sqrt{2}$ |
| **E** | | | | | 0 | $\sqrt{2}$ | $\sqrt{45}$ | $\sqrt{25}$ |
| **F** | | | | | | 0 | $\sqrt{29}$ | $\sqrt{29}$ |
| **G** | | | | | | | 0 | $\sqrt{58}$ |
| **H** | | | | | | | | 0 |

Assume that the initial centroids are A, D, and G.

Run the K-means clustering in only 1 iteration. At the end of the iteration, give following values.

(a) [9 pt] The new clusters (the units in each cluster)

Cluster 1: Cluster 2: Cluster 3:

(b) [9 pt] The centers of new clusters

Center 1: Center 2: Center 3:

(c) [12 pt] How many more iterations do we need to converge?