2018.11.20

Data Mining (2018 Fall)

Answers to Midterm Exam

1. (18%) It is important to define or select similarity measures in data analysis. However, there is no commonly accepted subjective similarity measure. Results can vary depending on the similarity measures used. Suppose we have the following 3-dimensional data set:

	A_1	A_2	A_3
x_1	0	2	0
x_2	1	2	1
x_3	6	6	3

Given a new data point, x = (2,2,1) as a query, rank the database points (x_1, x_2, x_3) based on similarity with the query using Supremum distance, Manhattan distance, and cosine similarity.

評分： 總共要寫出 9 個答案，每個答案 2 分，共 18 分。

答案： (1) Supremum distance:

d (x, x_1) = **2** d (x, x_2) = **1** d (x, x_3) = **4**

(2) Manhattan distance:

d (x, x_1) = **3** d (x, x_2) = **1** d (x, x_3) = **10**

(3) Cosine similarity:

$\cos(x, x_1) = \mathbf{\frac{2}{3}}$ $\cos(x, x_2) = \mathbf{\frac{7\sqrt{6}}{18}}$ $\cos(x, x_3) = \mathbf{1}$

2. (16%) Use the methods to normalize the following group of data

100, 120, 150, 330, 600

(a) min-max normalization by setting min=0 and max=1.

(b) z-score normalization using the mean absolute deviation

評分： 每小題 8 分，共 16 分；錯 1 個答案扣 2 分、扣完為止。

答案： (a) **0, 0.04, 0.1, 0.46, 1**

(b) **– 0.9756, – 0.8537, – 0.6707, 0.4268, 2.0732**

3. A database has five transactions. Let min_sup=0.4.

(a) (20%) Write the steps of finding frequent itemsets by using pattern-growth approach.

(b) (12%) Show the maximum itemsets and closed itemsets of the mining result.

Tid	Items-bought
T100	{A, B, C, E, H}
T200	{B, C, I}
T300	{A, B, C, D, F, G}
T400	{A, B, C, D, F}
T500	{ B, C, J}

評分： (a) 寫出 Step 1：5%；寫到 Step 2：10%；寫到 Step 3：20%。
僅 Minimum_support 寫錯，其它流程皆正確：15%。
(b) Maximum itemset 項目集正確：3%、回報 support 正確：3%。
Closed itemset 項目集及 support，每組 2%，共 6%。

答案： (a) FP-Growth Mining Method

Step 1: Find the support count of each item.

Item	Support Count
A	3
B	5
C	5
D	2
E	1
F	2
G	1
H	1
I	1
J	1

Consider items with min_sup = 0.4
(i.e. min. support count = 0.4*5 = 2)

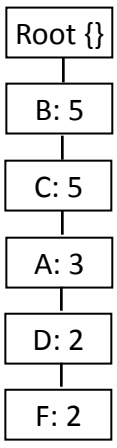
"Header table" (descending order)

Item	Support Count
B	5
C	5
A	3
D	2
F	2

Step 2: Order all items in itemset in frequency descending order.
(min_sup = 0.4) (**Note:** Consider only items with min_sup = 0.4)

Tid	Items-bought	(Ordered frequent items)
T100	{A, B, C, E, H}	B, C, A
T200	{B, C, I}	B, C
T300	{A, B, C, D, F, G}	B, C, A, D, F
T400	{A, B, C, D, F}	B, C, A, D, F
T500	{ B, C, J}	B, C

Step 3: FP-Tree construction ➔



(b) Maximum itemset: **(BCADF:2)**
Closed itemset: {**(BCADF:2), (BCA:3), (BC:5)**}

4. (20%) Suppose that a large store has a transactional database that is distributed among five locations. Propose an efficient algorithm to mine global frequent itemsets. You can present your algorithm in the form of an outline. Your algorithm should not require shipping all the data to one site.

評分：
1. 五處資料庫找出各自的頻繁項目集：5%
2. 產生全域候選項目集：5%
3. 五處資料庫計算包含全域候選項目集的 support：5%
4. 整合結果，刪除不符合 min_sup 條件者：5%。

老師作法：
1. 給定 min_sup 後，找出本地資料庫各自的 local frequent itemsets
2. global candidates 即為這些 local frequent itemsets 的聯集
3. 本地資料庫各自計算 global candidates 的 support
4. 將結果整合，將 global candidates 中 support 不符合 min_sup 條件的剃除，即得到答案。

5. (20%) In the context of data mining, the process of hiding sensitive itemsets is called data sanitization. In data sanitization, we modify the original database in some way so that sensitive itemsets are excluded from the mining result. Moreover, the sanitization process needs to ensure that the quality of the database is preserved by minimizing the impact on nonsensitive frequent itemsets. There are a lot of approaches to solve the data sanitization problem and Constraint Satisfaction Problem Model (CSP) is an important approach among them. Consider the following database. Let minsup count is 2, then the set of frequent itemsets is {A, B, C, D, E, AB, AC, AD, BC, BD, BE, CD, CE, ACD, BCE}. Assume AB and AC are identified as sensitive itemsets. Use the concept of CSP, list the inequalities to achieve the goal of data sanitization.

	A	B	C	D	E
T1	1	0	1	0	0
T2	1	0	1	1	0
T3	0	1	0	1	0
T4	0	1	1	0	1
T5	1	1	1	1	1
T6	0	0	0	1	0
T7	0	0	1	0	0
T8	1	1	0	0	0

2018.11.20

評分： SI 正確：4%、Positive Border 正確：4%；每條不等式 2%。

答案： Frequent itemset: F

Sensitive itemset: SI = **{AB, AC}** **(To be infrequent)**

Superset of "SI": SS = {AB, AC, ACD}

Sanitized Frequent itemset: F' = F – SS

Positive Border: $B^+(F')$ = **{BCE, AD, BD, CD}** **(To be frequent)**

Conditional	Inequality
AB: Infrequent	U51*U52 + U81*U82 < 2
AC: Infrequent	U11*U13 + U21*U23 + U51*U53 < 2
BCE: Frequent	1 + U52*U53 ≥ 2
AD: Frequent	U21 + U51 ≥ 2
BD: Frequent	1 + U52 ≥ 2
CD: Frequent	U23 + U53 ≥ 2