Assignment 2 Solution

(COMP3605 - Introduction to Data Analytics, 2022-2023)

Date Available: Sunday, October 23, 2022
Due Date: 11:50 PM, Sunday, November 06, 2022
Total Mark: 100 marks

Solution to Question 1 [46 marks]
a. [34 marks] *minsup* = 0.3, $N = |D| = 10$, *minsup* count = $\lceil 0.3 \times 10 \rceil = \lceil 3 \rceil = 3$
[5 marks]
1. Iteration 1: each item is a member of set of candidate 1-itemsets C_1. Algorithm scans all of transactions to count the number of occurrences of each item A (3), B (5), C (9), E (5), G (5), M (2), O (4). C_1 = {{A}:3, {B}:5, {C}:9, {E}:5, {G}:5, {M}:2, {O}:4}. Remove 1 infrequent 1-itemset {M}:2 because its support count = 2 < *min_sup* count = 3.
[5 marks]
2. Set of frequent 1-itemsets L_1 = {{A}:3, {B}:5, {C}:9, {E}:5, {G}:5, {O}:4}, $n = |L_1| = 6$.
[5 marks]
3. To discover set of frequent 2-itemsets L_2, algorithm uses the join $L_1 \bowtie L_1$ to generate a candidate set of 2-itemsets C_2. C_2 consists of 10 2-itemsets ($C(n, k) = C(6, 2) = 15$).
C_2 = {{A, B}, {A, C}, {A, E}, {A, G}, {A, O}, {B, C}, {B, E}, {B, G}, {B, O}, {C, E}, {C, G}, {C, O}, {E, G}, {E, O}, {G, O}}
4. Next, transactions in D are scanned and support count of each candidate itemset in C_2 is accumulated. C_2 = {{A, B}:1, {A, C}:3, {A, E}:1, {A, G}:1, {A, O}:2, {B, C}:4, {B, E}:2, {B, G}:3, {B, O}:3, {C, E}:4, {C, G}:5, {C, O}:3, {E, G}:2, {E, O}:2, {G, O}:1}
[5 marks]
5. Set of frequent 2-itemsets L_2 is then determined, consisting of those candidate 2-itemsets in C_2 having support count ≥ *min_sup* count = 3.
Set of frequent 2-itemsets L_2 = {{A, C}:3, {B, C}:4, {B, G}:3, {B, O}:3, {C, E}:4, {C, G}:5, {C, O}:3}.
/* removed 8 infrequent 2-itemsets: {A, B}:1, {A, E}:1, {A, G}:1, {A, O}:2, {B, E}:2, {E, G}:2, {E, O}:2, {G, O}:1 */
[5 marks]
6. $C_3 = L_2 \bowtie L_2$ = {{B, C, G}, {B, C, O}, {B, G, O}, {C, E, G}, {C, E, O}, {C, G, O}}.
- If any $(k - 1)$-subset of a candidate k-itemset is not in L_{k-1} (i.e., infrequent), then the candidate cannot be frequent and can be removed from C_k.
- The last 4 candidate 3-itemsets {B, G, O}, {C, E, G}, {C, E, O}, {C, G, O} are removed from C_3 because their 2-subsets {E, G}, {E, O}, and {G, O} are not in L_2 (i.e., the subsets {E, G}, {E, O}, and {G, O} are infrequent).
→ C_3 = {{B, C, G}, {B, C, O}}.
/*
{B, G, O} contains {G, O} $\notin L_2$, so remove {B, G, O} from C_3.
{C, E, G} contains {E, G} $\notin L_2$, so remove {C, E, G} from C_3.
{C, E, O} contains {E, O} $\notin L_2$, so remove {C, E, O} from C_3.
{C, G, O} contains {G, O} $\notin L_2$, so remove {C, G, O} from C_3.
*/
[5 marks]

7. Transactions are scanned to determine L_3, consisting of those candidate 3-itemsets in C_3 having support count ≥ *min_sup* = 3.
C_3 = {{B, C, G}:3, {B, C, O}:2}
→ set of frequent 3-itemsets L_3 = {{B, C, G}:3}.
[2 marks]
8. Algorithm uses $L_3 \bowtie L_3$ to generate a candidate set of 4-itemsets C_4.
$C_4 = \varnothing$ because L_3 contains only one 3-itemset {B, C, G}.
Thus, $L_4 = \varnothing$ and the apriori algorithm terminates.
[2 marks]
The set of all frequent itemsets found is
L = {{A}:3, {B}:5, {C}:9, {E}:5, {G}:5, {O}:4,
{A, C}:3, {B, C}:4, {B, G}:3, {B, O}:3, {C, E}:4, {C, G}:5, {C, O}:3,
{B, C, G}:3}

b. [12 marks] *minconf* = 0.75
• For frequent 3-itemset ℓ = {B, C, G}, all nonempty subsets of ℓ = {B, C, G} are *s* = {B, C}, {B, G}, {C, G}, {B}, {C}, {G}, we have rules
[2 marks]
{B, C} → {G}, $c(\{B, C\} \rightarrow \{G\}) = \sigma(\{B, C, G\}) / \sigma(\{B, C\}) = 3 / 4 = 0.75$ ✓
[2 marks]
{B, G} → {C}, $c(\{B, G\} \rightarrow \{C\}) = \sigma(\{B, G, C\}) / \sigma(\{B, G\}) = 3 / 3 = 1.0$ ✓
[2 marks]
{C, G} → {B}, $c(\{C, G\} \rightarrow \{B\}) = \sigma(\{C, G, B\}) / \sigma(\{C, G\}) = 3 / 5 = 0.6$ ×
[2 marks]
G → {B, C}, $c(G \rightarrow \{B, C\}) = \sigma(\{G, B, C\}) / \sigma(\{G\}) = 3/5 = 0.6$ ×
[2 marks]
C → {B, G}, $c(C \rightarrow \{B, G\}) = \sigma(\{C, B, G\}) / \sigma(\{C\}) = 3/9 = 0.333$ ×
[2 marks]
B → {C, G}, $c(B \rightarrow \{C, G\}) = \sigma(\{B, C, G\}) / \sigma(\{B\}) = 3/5 = 0.6$ ×

Solution to Question 2 [20 marks]
a. [10 marks] Use the majority voting technique to classify the test example *z* = 5.0 using 9-NN (i.e., *k* = 9).
• For *k* = 9, nearest neighbors are

i	1	2	3	4	5	6	7	8	9
$\mathbf{x}_i$	0.5	3.0	4.5	4.6	4.9	5.2	5.3	5.5	7.0
y_i	−	−	+	+	+	−	−	+	−
d_i	4.5	2	0.5	0.4	0.1	0.2	0.3	0.5	2

$h(C_1) = I(y_3 = C_1) + I(y_4 = C_1) + I(y_5 = C_1) + I(y_8 = C_1) = 4$,
$h(C_2) = I(y_1 = C_2) + I(y_2 = C_2) + I(y_6 = C_2) + I(y_7 = C_2) + I(y_9 = C_2) = 5$
We have $h(C_1) = 4 < h(C_2) = 5$, the class label of the test instance *z* is $y' = C_2$ (i.e., class −).

/* or

i	2	3	4	5	6	7	8	9	10
$\mathbf{x}_i$	3.0	4.5	4.6	4.9	5.2	5.3	5.5	7.0	9.5
y_i	−	+	+	+	−	−	+	−	−
d_i	2	0.5	0.4	0.1	0.2	0.3	0.5	2	4.5

$h(C_1) = I(y_3 = C_1) + I(y_4 = C_1) + I(y_5 = C_1) + I(y_8 = C_1) = 4$,
$h(C_2) = I(y_2 = C_2) + I(y_6 = C_2) + I(y_7 = C_2) + I(y_9 = C_2) + I(y_{10} = C_2) = 5$
We have $h(C_1) = 4 < h(C_2) = 5$, the class label of the test instance z is $y' = C_2$ (i.e., class −).
*/

b. [10 marks] Use the distance-weighted voting technique to classify the test example $z = 5.0$ using 9-NN (i.e., $k = 9$).
• For $k = 9$, nearest neighbors are

i	1	2	3	4	5	6	7	8	9
$\mathbf{x}_i$	0.5	3.0	4.5	4.6	4.9	5.2	5.3	5.5	7.0
y_i	−	−	+	+	+	−	−	+	−
d_i	4.5	2	0.5	0.4	0.1	0.2	0.3	0.5	2
d_i^2	20.25	4	025	0.16	0.01	0.04	0.09	0.25	4
w_i	0.049	0.25	4	6.25	100	25	11.11	4	0.25

$f(C_1) = w_3 \times I(y_3 = C_1) + w_4 \times I(y_4 = C_1) + w_5 \times I(y_5 = C_1) + w_8 \times I(y_8 = C_1)$
$f(C_1) = 4 \times 1 + 6.25 \times 1 + 100 \times 1 + 4 \times 1 = 114.25$,
$f(C_2) = w_1 \times I(y_1 = C_2) + w_2 \times I(y_2 = C_2) + w_6 \times I(y_6 = C_2) + w_7 \times I(y_7 = C_2) + w_9 \times I(y_9 = C_2)$
$f(C_2) = 0.049 \times 1 + 0.25 \times 1 + 25 \times 1 + 11.11 \times 1 + 0.25 \times 1 = 36.66$
We have $f(C_1) = 114.25 > f(C_2) = 36.66$, the class label of the test instance z is $y' = C_2$ (i.e., class +).

/* or

i	2	3	4	5	6	7	8	9	10
$\mathbf{x}_i$	3.0	4.5	4.6	4.9	5.2	5.3	5.5	7.0	9.5
y_i	−	+	+	+	−	−	+	−	−
d_i	2	0.5	0.4	0.1	0.2	0.3	0.5	2	4.5
d_i^2	4	025	0.16	0.01	0.04	0.09	0.25	4	20.25
w_i	0.25	4	6.25	100	25	11.11	4	0.25	0.049

$f(C_1) = w_3 \times I(y_3 = C_1) + w_4 \times I(y_4 = C_1) + w_5 \times I(y_5 = C_1) + w_8 \times I(y_8 = C_1)$
$f(C_1) = 4 \times 1 + 6.25 \times 1 + 100 \times 1 + 4 \times 1 = 114.25$,
$f(C_2) = w_2 \times I(y_2 = C_2) + w_6 \times I(y_6 = C_2) + w_7 \times I(y_7 = C_2) + w_9 \times I(y_9 = C_2) + w_{10} \times I(y_{10} = C_2)$
$f(C_2) = 0.25 \times 1 + 25 \times 1 + 11.11 \times 1 + 0.25 \times 1 + 0.049 \times 1 = 36.66$
We have $f(C_1) = 114.25 > f(C_2) = 36.66$, the class label of the test instance z is $y' = C_2$ (i.e., class +).
*/

Solution to Question 3 [34 marks]

• For UPGMA (Unweighted Pair Group Method with Arithmetic mean), we use the formula

group-average linkage: $d(C_i, C_j) = \dfrac{1}{n_i n_j} \sum_{\boldsymbol{p} \in C_i, \boldsymbol{p}' \in C_j} \| \boldsymbol{p} - \boldsymbol{p}' \|_2$

where $\|\cdot\|_2$ is Euclidean distance (a.k.a. L_2-norm), $n_i = |C_i|$, $n_j = |C_j|$.

$|\{p_1\}| = |\{p_2\}| = |\{p_3\}| = |\{p_4\}| = |\{p_5\}| = |\{p_6\}| = 1$

a. [4 marks]

The distance matrix M

	p_1	p_2	p_3	p_4	p_5	p_6
p_1	0.000	0.784	0.858	0.504	0.728	0.842
p_2	0.784	0.000	1.175	0.818	0.787	1.054
p_3	0.858	1.175	0.000	0.398	0.412	**0.156**
p_4	0.504	0.818	0.398	0.000	0.262	0.343
p_5	0.728	0.787	0.412	0.262	0.000	0.272
p_6	0.842	1.054	**0.156**	0.343	0.272	0.000
min	0.504	0.787	0.156	0.262	0.272	**0.156**

b. [30 marks] the **group-average linkage**

[6 marks]

Iteration 1. $n(n-1)/2 = (6\times5)/2 = 15$ distances

• The detailed computations are

$d(\{p_1\}, \{p_2\}) = [(0.1831 - 0.9624)^2 + (0.1085 - 0.1916)^2]^{0.5} = 0.784$
$d(\{p_1\}, \{p_3\}) = [(0.1831 - 0.0732)^2 + (0.1085 - 0.9594)^2]^{0.5} = 0.858$
$d(\{p_1\}, \{p_4\}) = [(0.1831 - 0.2572)^2 + (0.1085 - 0.6066)^2]^{0.5} = 0.504$
$d(\{p_1\}, \{p_5\}) = [(0.1831 - 0.4476)^2 + (0.1085 - 0.7871)^2]^{0.5} = 0.728$
$d(\{p_1\}, \{p_6\}) = [(0.1831 - 0.2292)^2 + (0.1085 - 0.9489)^2]^{0.5} = 0.842$

$d(\{p_2\}, \{p_3\}) = [(0.9624 - 0.0732)^2 + (0.1916 - 0.9594)^2]^{0.5} = 1.175$
$d(\{p_2\}, \{p_4\}) = [(0.9624 - 0.2572)^2 + (0.1916 - 0.6066)^2]^{0.5} = 0.818$
$d(\{p_2\}, \{p_5\}) = [(0.9624 - 0.4476)^2 + (0.1916 - 0.7871)^2]^{0.5} = 0.787$
$d(\{p_6\}, \{p_6\}) = [(0.9624 - 0.2292)^2 + (0.1916 - 0.9489)^2]^{0.5} = 1.054$

$d(\{p_3\}, \{p_4\}) = [(0.0732 - 0.2572)^2 + (0.9594 - 0.6066)^2]^{0.5} = 0.398$
$d(\{p_3\}, \{p_5\}) = [(0.0732 - 0.4476)^2 + (0.9594 - 0.7871)^2]^{0.5} = 0.412$
$d(\{p_3\}, \{p_6\}) = [(0.0732 - 0.2292)^2 + (0.9594 - 0.9489)^2]^{0.5} = 0.156$

$d(\{p_4\}, \{p_5\}) = [(0.2572 - 0.4476)^2 + (0.6066 - 0.7871)^2]^{0.5} = 0.262$
$d(\{p_4\}, \{p_6\}) = [(0.2572 - 0.2292)^2 + (0.6066 - 0.9489)^2]^{0.5} = 0.343$

$d(\{p_5\}, \{p_6\}) = [(0.4476 - 0.2292)^2 + (0.7871 - 0.9489)^2]^{0.5} = 0.272$

• Find two closest clusters
min{$d(\{p_1\}, \{p_2\})$, $d(\{p_1\}, \{p_3\})$, $d(\{p_1\}, \{p_4\})$, $d(\{p_1\}, \{p_5\})$, $d(\{p_1\}, \{p_6\})$,
$d(\{p_2\}, \{p_3\})$, $d(\{p_2\}, \{p_4\})$, $d(\{p_2\}, \{p_5\})$, $d(\{p_2\}, \{p_6\})$,
$d(\{p_3\}, \{p_4\})$, $d(\{p_3\}, \{p_5\})$, $d(\{p_3\}, \{p_6\})$,
$d(\{p_4\}, \{p_5\})$, $d(\{p_4\}, \{p_6\})$,
$d(\{p_5\}, \{p_6\})$}
= min{0.784, 0.858, 0.504, 0.728, 0.842,
1.175, 0.818, 0.787, 1.054,
0.398, 0.412, **0.156**,
0.262, 0.343,
0.272} = **0.156** = $d(\{p_3\}, \{p_6\})$
• The two closest clusters are $\{p_3\}$ and $\{p_6\}$. Merge $\{p_3\}$ and $\{p_6\}$ to obtain $C_1 = \{p_3, p_6\}$.
• The updated distance matrix M after merging the two closest clusters $\{p_3\}$ and $\{p_6\}$ (i.e., $C_1 = \{p_3, p_6\}$) is

	p_1	p_2	C_1	p_4	p_5
p_1	0.000	0.784	0.850	0.504	0.728
p_2	0.784	0.000	1.114	0.818	0.787
C_1	0.850	1.114	0.000	0.371	0.342
p_4	0.504	0.818	0.371	0.000	**0.262**
p_5	0.728	0.787	0.342	**0.262**	0.000
min	0.504	0.787	0.342	0.262	**0.262**

$|\{p_1\}| = |\{p_2\}| = |\{p_4\}| = |\{p_5\}| = 1$, $|C_1| = |\{p_3, p_6\}| = 2$

[6 marks]
Iteration 2. $n(n-1)/2 = (5\times4)/2 = 10$ distances

• Find two closest clusters
min{$d(\{p_1\}, \{p_2\})$, $d(\{p_1\}, \{C_1\})$, $d(\{p_1\}, \{p_4\})$, $d(\{p_1\}, \{p_5\})$,
$d(\{p_2\}, \{C_1\})$, $d(\{p_2\}, \{p_4\})$, $d(\{p_2\}, \{p_5\})$,
$d(\{C_1\}, \{p_4\})$, $d(\{C_1\}, \{p_5\})$,
$d(\{p_4\}, \{p_5\})$}
= min{0.784, $[d(\{p_1\}, \{p_3\}) + d(\{p_1\}, \{p_6\})] / (1\times2)$, 0.504, 0.728,
$[d(\{p_2\}, \{p_3\}) + d(\{p_2\}, \{p_6\})] / (1\times2)$, 0.818, 0.787,
$[d(\{p_3\}, \{p_4\}) + d(\{p_6\}, \{p_4\})] / (2\times1)$, $[d(\{p_3\}, \{p_5\}) + d(\{p_6\}, \{p_5\})] / (2\times1)$,
0.262}
= min{0.784, (0.858 + 0.842) / 2, 0.504, 0.728,
(1.175 + 1.054) / 2, 0.818, 0.787,
(0.398 + 0.343) / 2, (0.412 + 0.272) / 2,
0.262}
= min{0.784, 0.850, 0.504, 0.728,
1.114, 0.818, 0.787,
0.371, 0.342,
0.262} = **0.262** = $d(\{p_4\}, \{p_5\})$
• The two closest clusters are $\{p_4\}$ and $\{p_5\}$. Merge $\{p_4\}$ and $\{p_5\}$ to obtain $C_2 = \{p_4, p_5\}$.

• The updated distance matrix M after merging the two closest clusters $\{p_4\}$ and $\{p_5\}$ (i.e., $C_2 = \{p_4, p_5\}$) is

	p_1	p_2	C_1	C_2
p_1	0.000	0.784	0.850	0.616
p_2	0.784	0.000	1.114	0.803
C_1	0.850	1.114	0.000	**0.356**
C_2	0.616	0.803	**0.356**	0.000
min	0.616	0.803	0.356	**0.356**

$|\{p_1\}| = |\{p_2\}| = 1$, $|C_1| = |\{p_3, p_6\}| = 2$, $|C_2| = |\{p_4, p_5\}| = 2$

[6 marks]

Iteration 3. $n(n-1)/2 = (4\times3)/2 = 6$ distance

• Find two closest clusters

$\min\{d(\{p_1\}, \{p_2\}), d(\{p_1\}, \{C_1\}), d(\{p_1\}, \{C_2\}),$
$d(\{p_2\}, \{C_1\}), d(\{p_2\}, \{C_2\}),$
$d(\{C_1\}, \{C_2\})\}$
$= \min\{0.784, 0.850, [d(\{p_1\}, \{p_4\}) + d(\{p_1\}, \{p_5\})] / (1\times2),$
$1.114, [d(\{p_2\}, \{p_4\}) + d(\{p_2\}, \{p_5\})] / (1\times2),$
$[d(\{p_3\}, \{p_4\}) + d(\{p_3\}, \{p_5\}) + d(\{p_6\}, \{p_4\}) + d(\{p_6\}, \{p_5\})] / (2\times2)\}$
$= \min\{0.784, 0.850, (0.504 + 0.728) / 2,$
$1.114, (0.818 + 0.787) / 2,$
$(0.398 + 0.412 + 0.343 + 0.272) / 4\}$
$= \min\{0.784, 0.850, 0.616,$
$1.114, 0.803,$
$\mathbf{0.356}\} = \mathbf{0.356} = d(\{C_1\}, \{C_2\})$

• The two closest clusters are $\{C_1\}$ and $\{C_2\}$. Merge $\{C_1\}$ and $\{C_2\}$ to obtain $C_3 = \{C_1, C_2\} = \{p_3, p_6, p_4, p_5\}$.

• The updated distance matrix M after merging the two closest clusters $\{C_1\}$ and $\{C_2\}$ (i.e., $C_3 = \{C_1, C_2\} = \{p_3, p_6, p_4, p_5\}$) is

	p_1	p_2	C_3
p_1	0.000	0.784	**0.733**
p_2	0.784	0.000	0.959
C_3	**0.733**	0.959	0.000
min	0.733	0.959	0.733

$|\{p_1\}| = |\{p_2\}| = 1$, $|C_3| = |\{p_3, p_6, p_4, p_5\}| = 4$

[6 marks]

Iteration 4. $n(n-1)/2 = (3\times2)/2 = 3$ distances

• Find two closest clusters

$\min\{d(\{p_1\}, \{p_2\}), d(\{p_1\}, \{C_3\}),$
$d(\{p_2\}, \{C_3\})\}$
$= \min\{0.784, [d(\{p_1\}, \{p_3\}) + d(\{p_1\}, \{p_4\}) + d(\{p_1\}, \{p_5\}) + d(\{p_1\}, \{p_6\})] / (1\times4),$

$[d(\{p_2\}, \{p_3\}) + d(\{p_2\}, \{p_4\}) + d(\{p_2\}, \{p_5\}) + d(\{p_2\}, \{p_6\})] / (1\times4)\}$
= min{0.784, (0.858 + 0.504 + 0.728 + 0.842) / 4,
(1.175 + 0.818 + 0.787 + 1.054) / 4}
= min{0.784, **0.733**,
0.959} = **0.733** = $d(\{p_1\}, \{C_3\})$

• The two closest clusters are $\{p_1\}$ and $\{C_3\}$. Merge $\{p_1\}$ and $\{C_3\}$ to obtain $C_4 = \{p_1, C_3\} = \{p_1, p_3, p_6, p_4, p_5\}$.

• The updated distance matrix M after merging the two closest clusters $\{p_1\}$ and $\{C_3\}$ (i.e., $C_4 = \{p_1, C_3\} = \{p_1, p_3, p_6, p_4, p_5\}$) is

	C_4	p_2
C_4	0.000	0.924
p_2	0.924	0.000

$|\{p_2\}| = 1$, $|C_4| = |\{p_1, p_3, p_6, p_4, p_5\}| = 5$

[6 marks]

Iteration 5. $n(n - 1)/2 = (2\times1)/2 = 1$ distance

• Find two closest clusters
$\min\{d(\{p_2\}, \{C_4\}) = \min\{[d(\{p_2\}, \{p_1\}) + d(\{p_2\}, \{p_3\}) + d(\{p_2\}, \{p_4\}) + d(\{p_2\}, \{p_5\}) + d(\{p_2\}, \{p_6\})] / (1\times5)\}$
= min{(0.784 + 1.175 + 0.818 + 0.787 + 1.054) / 5} = min{4.618 / 5} = min{0.924} = 0.924.

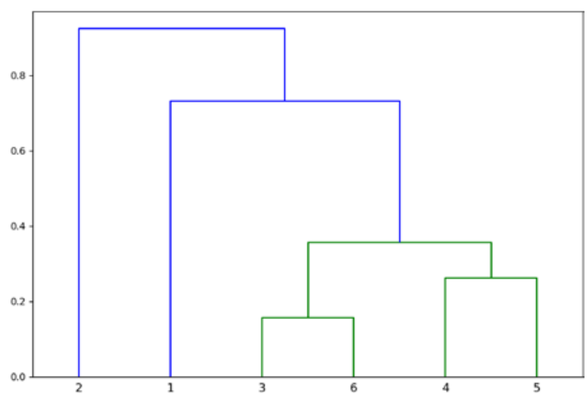
• The two closest clusters are $\{p_2\}$ and $\{C_4\}$. Merge $\{p_2\}$ and $\{C_4\}$ to obtain $C_5 = \{p_2, C_4\} = \{p_2, p_1, p_3, p_6, p_4, p_5\}$.

• The updated distance matrix M after merging the two closest clusters $\{p_2\}$ and $\{C_4\}$ (i.e., $C_5 = \{p_2, C_4\} = \{p_2, p_1, p_3, p_6, p_4, p_5\}$) is

	C_5
C_5	0.000

$|C_5| = |\{p_2, p_1, p_3, p_6, p_4, p_5\}| = 6$

We have 1 cluster, namely $C_5 = \{p_2, p_1, p_3, p_6, p_4, p_5\}$. Stop.



End of Assignment 2 Solution