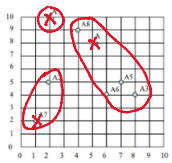
Assignment 5

Wednesday, November 24, 2021 2:50 PM

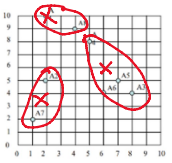
1. [20 points] By considering the following 8 2D data points below do:
   a. [15 points] Group the points into 3 clusters by using k-means algorithm with Euclidean distance. Show the intermediate clusters (by drawing ellipses on this 10 by 10 space) and centroids (by drawing marks like X on this 10 by 10 space) in each iteration until convergence. Consider the initial centroids as: C1 = A1, C2 = A4, and C3 = A7.



1st iteration

Centroid: (C1, C2, C3)								
Instance	A1	A2	A3	A4	A5	A6	A7	A8
C1 dist.	0	5	9.49	—	7.07	7.21	—	2.23
C2 dist.	—	4.24	5	0	3.61	4.12	—	1.41
C3 dist.	—	3.16	7.28	—	6.7	5.4	0	7.61
Cluster Assigned	U_1	U_3	U_2	U_2	U_2	U_2	U_3	U_2

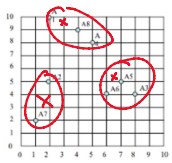
$C_1 = (2, 10)$ $C_2 = \left[(4+5+6+7+8)/5 , (9+8+5+4+4)/5\right]$ $= (6, 6)$ $C_3 = \left[(1+2)/2 , (2+5)/2\right]$ $= (1.5, 3.5)$



2nd iteration

Centroid: (C1, C2, C3)								
Instance	A1	A2	A3	A4	A5	A6	A7	A8
C1 dist.	0	5	9.49	3.60	7.07	9.21	8.06	2.23
C2 dist.	—	4.12	7.82	2.23	1.41	2	6.40	3.6
C3 dist.	—	1.58	6.51	5.7	5.7	4.52	1.58	6.0
Cluster Assigned	U_1	U_3	U_2	U_2	U_2	U_2	U_3	U_1

$C_1 = (2+4)/2 , (10+9)/2$ $= (3, 9.5)$ $C_2 = \left[(5+6+7+8)/4 , (9+5+4+4)/4\right]$ $= (6.5 , 5.25)$ $C_3 = (1.5, 3.5)$



3rd iteration

Centroid: (C1, C2, C3)								
Instance	A1	A2	A3	A4	A5	A6	A7	A8
C1 dist.	1.11	4.6	7.43	2.5	6.02	6.36	7.76	1.11
C2 dist.	—	—	—	2.13	—	—	—	—
C3 dist.	—	1.58	—	—	—	—	1.58	—
Cluster Assigned	U_1	U_3	U_2	U_1	U_2	U_2	U_3	U_1

$C_1 = \frac{2+4+5}{3} , \frac{10+9+8}{3}$ $= (3.7 , 9)$ $C_2 = \frac{6+7+8}{3} , \frac{5+4+4}{3}$ $= (7 , 4.33)$ $C_3 = (1.5, 3.5)$

b. [5 points] Calculate the SSE (Sum of Square Errors) of the final clustering.

$$SSE = \sum_{i=1}^{3} \sum_{x \in C_i} dist^2(m_i, x)$$

Sum of C_1 $= 1.97^2 + .3^2 + 1.64^2 = 6.66$

Sum of C_2 $= 1.05^2 + .67^2 + 1.05^2 = 2.65$

Sum of C_3 $= 1.58^2 + 1.58^2 = 4.99$

$= 14.3$

2. [20 points] Use the distance matrix below to perform the following operations:
   a. [14 points] Group the points by using single link (MIN) hierarchical clustering. Show your results by informing the updated distance matrix after each merging step and by drawing the corresponding dendrogram that should clearly present the order in which the points are merged.

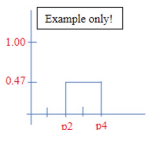
	p1	p2	p3	p4	p5
p1	0.00	0.10	0.41	0.55	0.35
p2	0.10	0.00	0.64	0.47	0.98
p3	0.41	0.64	0.00	0.44	0.85
p4	0.55	0.47	0.44	0.00	0.76
p5	0.35	0.98	0.85	0.76	0.00

Solution format:

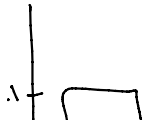
(1st iteration) Suppose the first two points to be merged are p2 and p4, then:

	p1	p2 U p4	p3	p5
p1	0.00	?	0.41	0.35
p2 U p4	?	0.00	?	?
p3	0.41	?	0.00	0.85
p5	0.35	?	0.85	0.00

Example only!

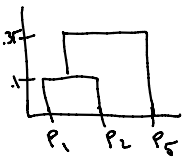


	$P_1 \cup P_2$	P_3	P_4	P_5
$P_1 \cup P_2$	—	.41	.47	.35
P_3	.41	—	.44	.85
P_4	.47	.44	—	.76

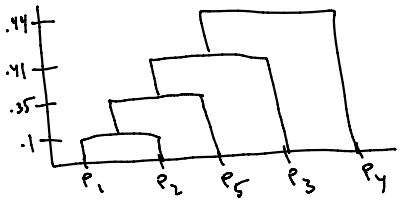


$P_1 \cup P_2$	—	.41	.47	.35
P_3	.41	—	.44	.85
P_4	.47	.44	—	.76
P_5	.35	.85	.76	—

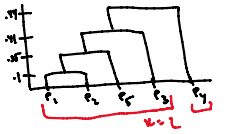
	$P_1 \cup P_2 \cup P_5$	P_3	P_4
$P_1 \cup P_2 \cup P_5$	—	.41	.47
P_3	.41	—	.44
P_4	.47	.44	—

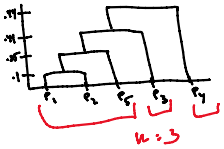


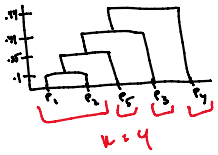
		P_4
$P_1 \cup P_2 \cup P_5 \cup P_3$	—	.44
P_4	.44	—



b. [6 points] Show the clusters when k = 2, k = 3, and k = 4.







3. https://github.com/NootCode/CS4210-Assignment-5/blob/master/Clustering/clustering.py

4. [10 points] The dataset below presents the user ratings on a 1-3 scale for 6 different rock bands.

	Bon Jovi	Metallica	Scorpions	AC/DC	Kiss	Guns n' Roses
Fred	1	3	1.5	3	1	3
Lillian	3	1.5	2	2	3	1
Cathy	2	2	2	3	1.5	2
John	3	2	2	2	?	?

a. [5 points] Apply **user-based** collaborative filtering on the dataset to decide about recommending the bands Kiss and Guns n' Roses to John. You should make a recommendation when the predicted rating is greater than or equal to 2.0. Use cosine similarity, a neutral value (1.5) for missing values, and the top 2 similar neighbors to build your model.

Cos(John, Fred) = 0.852

Cos (John, Lillian) = .995 > Closest 2

Cos (John, Cathy) = .982

Kiss

Avg John = 3+6 = 9/4 = 2.25

$$2.25 + \frac{[.995 * (3 - 2.125) + .982(1.5 - 2.25)]}{(.995 + .982)}$$

$2.25 + .068 = \boxed{2.32}$ ← Make the Recommendation

Guns N Roses

$$2.25 + \frac{[.995 * (1 - 2.125) + .982(2 - 2.25)]}{(.995 + .982)}$$

$2.25 - .69 = \boxed{1.56}$ ← Do not Recommend

b. [5 points] Now, apply **item-based** collaborative filtering to make the same decision. Use the same parameters defined before to build your model.

	Bon Jovi	Metallica	Scorpions	AC/DC	Kiss	Guns n' Roses
Fred	1	3	1.5	3	1	3
Lillian	3	1.5	2	2	3	1
Cathy	2	2	2	3	1.6	2
John	3	2	2	2	?	?

Kiss

Cos(Kiss, BJ) = .993
Cos(K, M) = .768
Cos(K, S) = .937
Cos(K, Ac) = .822
} 2 NN

Avg Kiss = 1.833

$$1.833 + \frac{[.993*(3-2) + .937(2-1.833)]}{(.993 + .937)}$$

$$= 1.833 + .597 = \boxed{2.43}$$

Guns N Roses

Cos(G, B) = .71
Cos(G, M) = .99
Cos(G, S) = .88
Cos(G, A) = .97
} 2 NN

Avg GNR = 2

$$2 + \frac{[.99*(2-2.17) + .97(2-2.67)]}{.99 + .97}$$

$$2 - .417$$

$$= \boxed{1.58}$$

5. [20 points] Consider the following transaction dataset.

Transaction ID	Items Bought
1	{a, b, d, e}
2	{b, c, d}
3	{a, b, d, e}
4	{a, c, d, e}
5	{b, c, d, e}
6	{b, d, e}
7	{c, d}
8	{a, b, c}
9	{a, d, e}
10	{b, d}

Suppose that minimum support is set to 30% (*minsup*) and minimum confidence is set to 60%.

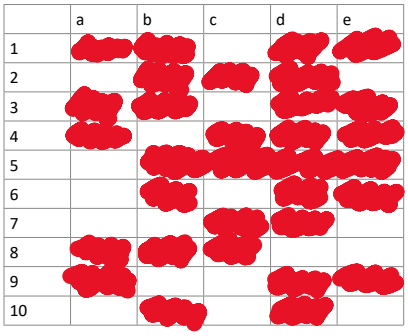
a. [5 points] Rank all frequent itemsets according to their support (list their support values).

{d} = .9
{b} = .7
{e} = .6
{a} = .5
{c} = .5
{a, e} = .5
{b, d} = .4
{a, d} = .4
{a, c} = .4
{b, e} = .4
{c, d} = .4
{a, d, e} = .4
{a, b} = .3
{b, c} = .3
{b, d, e} = .3

b. [5 points] For all frequent 3-itemsets, rank all association rules - according to their confidence values - which satisfy the requirements on minimum support and minimum confidence (list their confidence values).

{a, d, e}

{a, d} → {e} $c = \frac{\sigma(\{a,d,e\})}{\sigma(\{a,d\})} = \frac{4}{4} = 1$



{a, d, e}

{a, d} → {e} $c = \frac{\sigma(\{a,d,e\})}{\sigma(\{a,d\})} = \frac{4}{4} = 1$

{a, e} → {d} $c = \frac{\sigma(\{a,d,e\})}{\sigma(a,e)} = \frac{4}{4} = 1$

{d, e} → {a} $c = \frac{4}{5} = .8$

{a} → {d, e} $c = 4/5 = .8$

{e} → {a, d} $c = 4/6 = .667$

{b, d, e}

{b, d} → {e} $c = 3/4 = .75$

{b, e} → {d} $c = 3/4 = .75$

{d, e} → {b} $c = 3/5 = .6$

c. [5 points] Show how the 3-itemsets candidates can be generated by the $F_{k-1} \times F_{k-1}$ method and if these candidates will be pruned or not.

$F_2 \times F_2$

$F_2 = \{ab, ad, ae, bc, bd, be, cd, de\}$

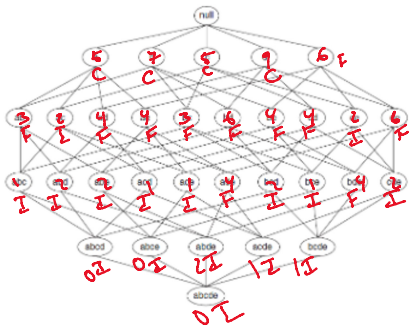
= abd
abe
ade
bcd
~~bce~~ ✓ Pruned
bde

ac
ce
→ infrequent

Only 2 taken

these candidates will be pruned or not.

d. [5 points] Consider the lattice structure given below. Label each node with the following letter(s): *M* if the node is a maximal frequent itemset, *C* if it is closed frequent itemset, *F* if it is frequent but neither maximal nor closed, and *I* if it is infrequent.



	a	b	c	d	e
1					
2					
3					
4					
5					
6					
7					
8					
9					
10					

Supp by Count (9)
{d}
Supp by Count (8)

6. https://github.com/NootCode/CS4210-Assignment-5/blob/master/ARM/association_rule_mining.py