ALDA Fall 2018
HW 4
11/12/2018

HW4 contains 6 questions. Please read and follow the instructions.

- **DUE DATE FOR SUBMISSION: 11/27/2018 11: 45 PM**
- **TOTAL NUMBER OF POINTS: 100**
- Make sure you clearly list each team member's **names and Unity IDs** at the top of your submission.
- Your submission should be a **single zip file** containing a PDF of your answers, your code, and a readme file with running instructions. Please follow the naming convention for your zip file: H(homework group number)_HW(homework number), e.g. H1_HW4.

---

1. (13 points) [**K-means Clustering**][**Xi Yang**] Use K-means clustering algorithm with *Euclidean Distance* to cluster the 10 data points in Figure 1 into 3 clusters. Suppose that the initial seeds are at points: B, D and J. Answer the following questions:
   (a) (4 points) Run 1 round of k-means algorithm. What are the coordinates of the new centroids? What are the new clusters? Show your work in Figure 1.

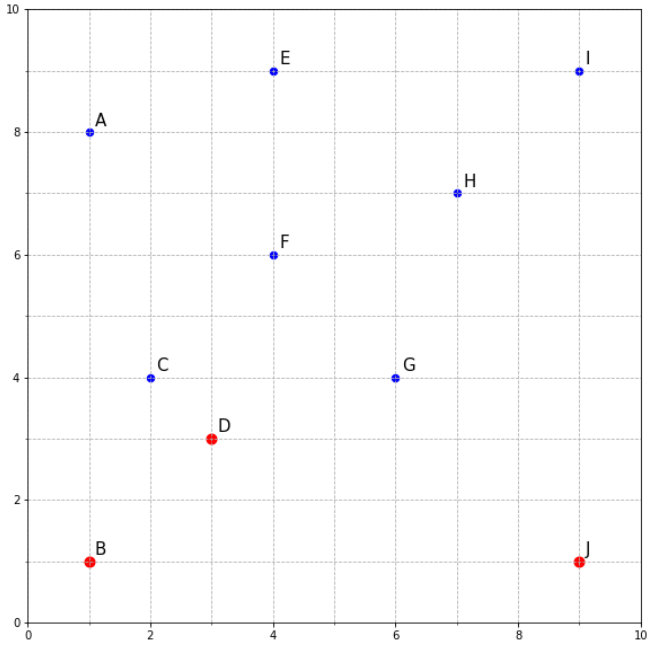

Figure 1: K-means Clustering (a)

(b) (9 points) How many rounds are needed for the K-means clustering algorithm to converge? Draw the result clusters and new centroid at the end of each round (including the first round) in the Figure 2. Indicate the coordinates along side corresponding centroids. **Add new graphics if needed; Stop when the algorithm converges and clearly label on the graph where the algorithm converges.**

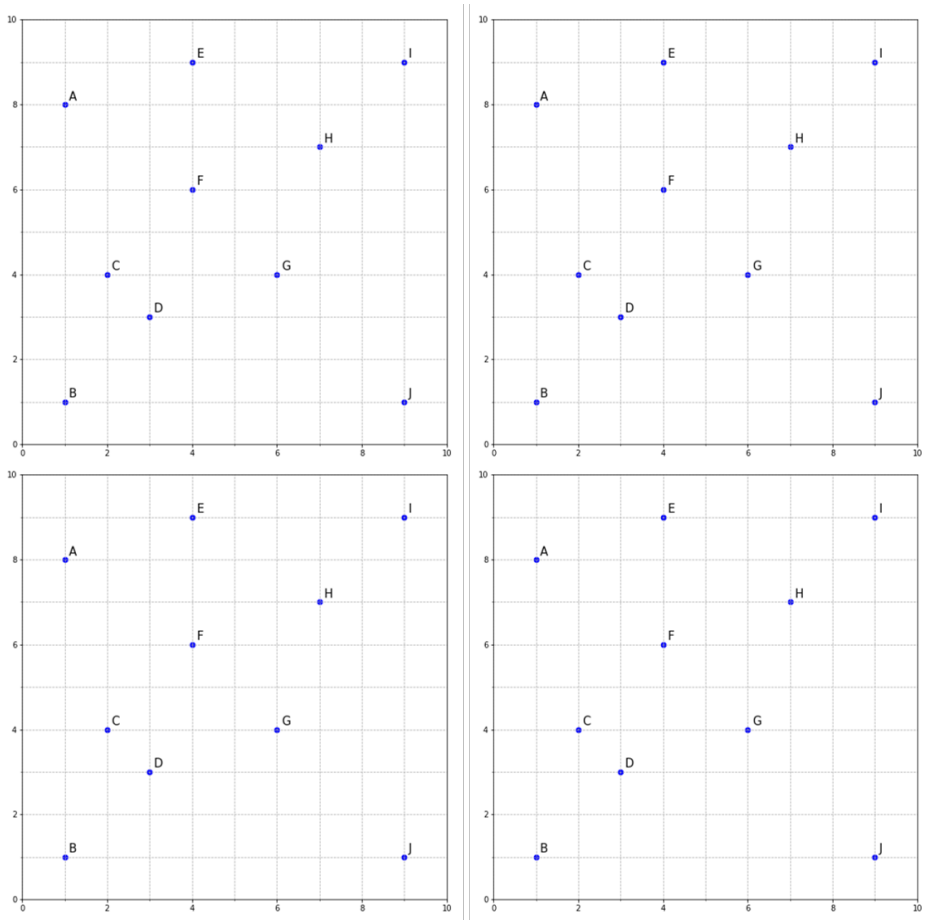

Figure 2: K-means Clustering (b)

2. (20 points) [**Hierarchical Clustering**] [**Ruth Okoilu**] We will use the same dataset A-J as in Question 1 for Hierarchical Clustering. The *Euclidean Distance* matrix between each pair of the datapoints are listed in the Table 3 below:

   (a) (10 points) Perform *single* and *complete* link hierarchical clustering. Show your results by drawing corresponding dendrogram. The dendrogram should clearly show the order and the height in which the clusters are merged.

   (b) (5 points) If we assume there are *three* clusters, which of the *single* and *complete* link hierarchical clustering will give better resulted clusters? Justify your answer.

   (c) (5 points) Compare your resulted clusters from 2(b) with the resulted clusters using K-means in Question 1 by calculating their corresponding Sum of Squared Error (SSE). Based on their SSE results, which resulted clusters, 1(b) or 2(b), are better?

	A	B	C	D	E	F	G	H	I	J
A	0.00	7.00	4.12	5.39	3.16	3.61	6.40	6.08	8.06	10.63
B	7.00	0.00	3.16	2.83	8.54	5.83	5.83	8.49	11.31	8.00
C	4.12	3.16	0.00	1.41	5.39	2.83	4.00	5.83	8.60	7.62
D	5.39	2.83	1.41	0.00	6.08	3.16	3.16	5.66	8.49	6.32
E	3.16	8.54	5.39	6.08	0.00	3.00	5.39	3.61	5.00	9.43
F	3.61	5.83	2.83	3.16	3.00	0.00	2.83	3.16	5.83	7.07
G	6.40	5.83	4.00	3.16	5.39	2.83	0.00	3.16	5.83	4.24
H	6.08	8.49	5.83	5.66	3.61	3.16	3.16	0.00	2.83	6.32
I	8.06	11.31	8.60	8.49	5.00	5.83	5.83	2.83	0.00	8.00
J	10.63	8.00	7.62	6.32	9.43	7.07	4.24	6.32	8.00	0.00

Figure 3: Hierarchical Clustering Dataset

3. (12 points) [**Song Ju**] For the transaction Table 1 given below, please answer the following questions:

TID	Items Bought
T1	{B,D,F,H}
T2	{C,D,F,G}
T3	{A,D,F,G}
T4	{A,B,C,D,H}
T5	{A,C,F,G}
T6	{D,H}
T7	{A,B,E,F}
T8	{A,D,F,G,H}
T9	{A,C,D,F,G}
T10	{(D,F,G,H}
T11	{A,C,D,E}
T12	{B,E,F,H}
T13	{D,F,G}
T14	{C,F,G,H}
T15	{A,C,D,F,H}

Table 1: Transactions Data

(a) (3 points) Explain what is frequent itemset and give an example of 2-itemset that is frequent itemset with support count = 8.

(b) (3 points) Explain what is closed frequent itemset and list all of them with support count = 8.

(c) (3 points) Explain what is maximal frequent and list all of maximal itemset with support count = 8.

(d) (3 points) Compute the support and confidence for association rule $\{D, F\} \rightarrow \{G\}$.

4. (13 points) [**Association Analysis**] [**Ruth Okoilu**] Consider the following market basket transactions shown in the Table 2 below.

Transaction ID	Items ordered
1	{Flour, Eggs, Bread}
2	{Soda, Coffee}
3	{Flour, Butter, Milk, Eggs}
4	{Bread, Eggs, Juice, Detergent}
5	{Bread, Milk, Eggs}
6	{Eggs, Bread}
7	{Detergent, Milk}
8	{Coffee, Soda, Juice}
9	{Butter, Juice, Bread}
10	{Milk, Bread, Detergent}

Table 2: Market Basket Transactions Data

For each of the following question, briefly explain your answers in 2-3 sentences.

(a) (2 points) How many items are in this data set? What is the maximum size of itemsets that can be extracted from this data set (only including itemsets that have ≥ 1 support)?

(b) (2 points) What is the maximum number of association rules that can be extracted from this data (including rules that have zero support)?

(c) (2 points) What is the *maximum number* of 2-itemsets that can be derived from this data set (including those have zero support)?

(d) (3 points) Find an itemset (of size 2 or larger) that has the largest support.

(e) (4 points) Given $minconf = 0.5$, find two pairs of items, a and b, such that the rules $\{a\} \rightarrow \{b\}$ and $\{b\} \rightarrow \{a\}$ have the same confidence, and their confidence is greater than or equal to the $minconf$ threshold.

5. (20 points) [**Apriori algorithm**][**Xi Yang**] Consider the data set shown in Table 3 and answer the following questions using apriori algorithm.

TID	Items
t_1	A,C,D,E
t_2	A,B,D,E
t_3	C,E
t_4	C,D
t_5	A,B,D
t_6	B,D,E
t_7	A,C,D
t_8	B,C,D,E

Table 3: Apriori algorithm

(a) (10 points) Show (compute) each step of frequent item set generation process using apriori algorithm, with support count of 3.

(b) (10 points) Show the lattice structure for the data given in table above, and mark the pruned branches if any. (Scanned hand-drawing is acceptable as long as it is clear.)

6. (22 points) [**Frequent Pattern Tree**][**Song Ju**] Consider the following data set shown in Table 4 and answer the following questions using FP-Tree.

TID	Items Bought
T1	{B,D,F,H}
T2	{C,D,F,G}
T3	{A,D,F,G}
T4	{A,B,C,D,H}
T5	{A,C,F,G}
T6	{D,H}
T7	{A,B,E,F}
T8	{A,D,F,G,H}
T9	{A,C,D,F,G}
T10	{(D,F,G,H}
T11	{A,C,D,E}
T12	{B,E,F,H}
T13	{D,F,G}
T14	{C,F,G,H}
T15	{A,C,D,F,H}

Table 4: Transactions Data Q6

(a) (12 points) Construct an FP-tree for the set of transactions in the table below as the first step towards identifying the itemsets with minimum support count of 2 (at least 2 occurrences). Do not forget to include the header table that locates the starts of the corresponding linked item lists through the FP-tree. For consistency,

please form your header table in the order of {F, D, G, H, A, C, B, E}

(b) (10 points) Using the FP-Tree constructed and support=3, generate all the frequent patterns with the base of item H step by step.