ALDA Fall 2021
HW 5
11/01/2021

HW5 contains 6 questions. Please read and follow the instructions.

- **DUE DATE FOR SUBMISSION: 11/17/2021 11:45 PM**
- **TOTAL NUMBER OF POINTS: 100**
- **NO PARTIAL CREDIT** will be given so provide concise answers.
- **Submissions and updates should be handled by the same person.**
- **You MUST manually add ALL team members in the submission portal when you submit through Gradescope.**
- Make sure you clearly list **your homework team ID, all team members' names and Unity IDs, for those who have contributed to the homework contribution** at the top of your submission.
- **[GradeScope and NCSU Github]:** Submit a PDF on GradeScope. **No code is required for this homework.**
- The materials on this course website are only for use of students enrolled in this course and **MUST NOT** be retained or disseminated to others.
- By uploading your submission, you agree that you have not violated any university policies related to the student code of conduct (https://policies.ncsu.edu/policy/pol-11-35-01/), and you are signing the Pack Pledge: **"I have neither given nor received unauthorized aid on this test or assignment"**.

---

1. (10 points) [**K-means Clustering**] [**John Wesley Hostetter (Designed) & Chengyuan Liu (Graded)**] Using K-means clustering and Euclidean distance, cluster the 11 data points in Figure 1 into *three* clusters. We assume that the initial seeds are at points E, F, and J (in yellow). Answer the following questions:

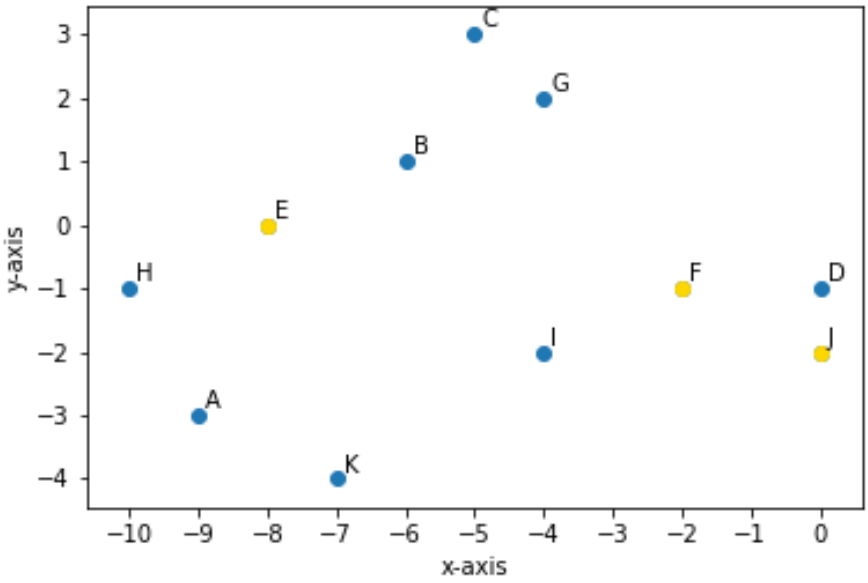

Figure 1: K-means Clustering (a)

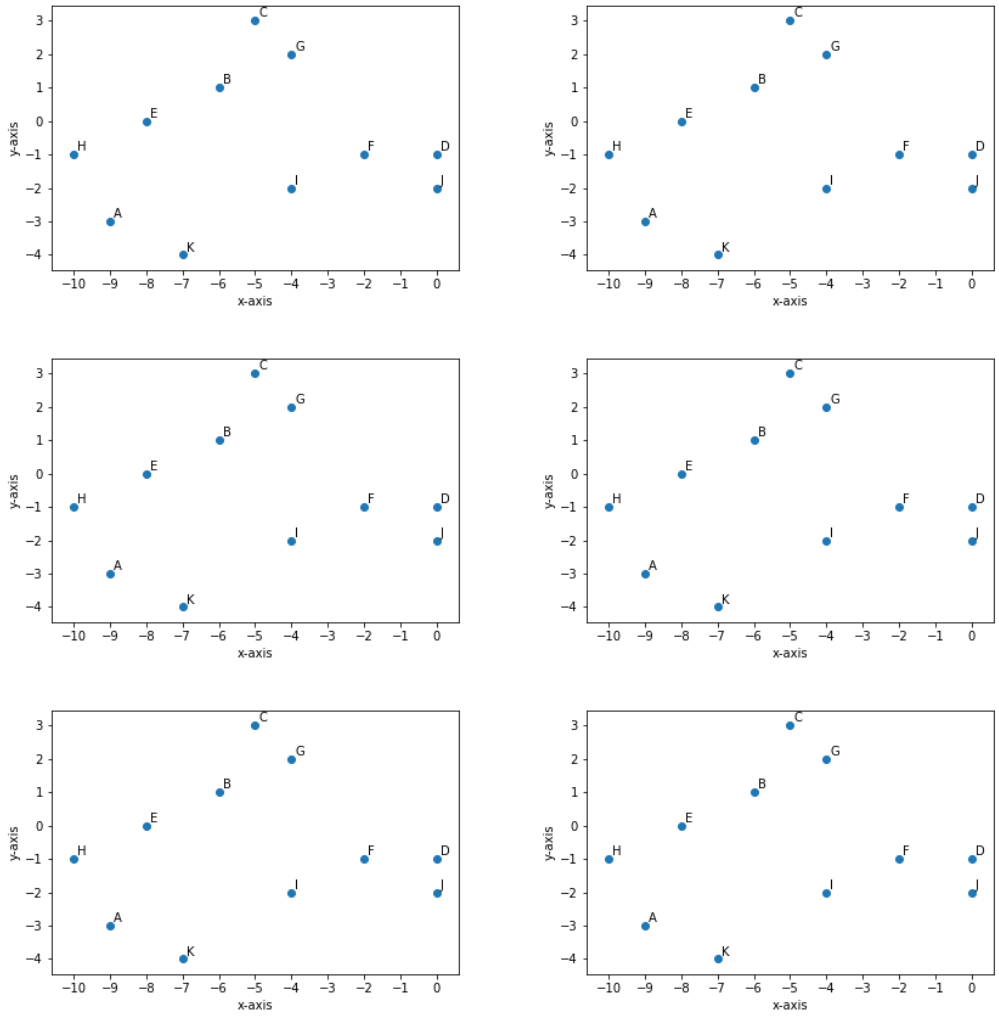

Figure 2: K-means Clustering (b)

(a) (4 points) Run the K-means algorithm for one round. Calculate the coordinates of the new centroids. What are the new clusters? Show your work in the first subgraph in Figure 2.

(b) (6 points) How many rounds are needed for the K-means clustering algorithm to converge? Draw the resulting 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.** NO PARTIAL CREDIT.

2. (15 points) [**Hierarchical Clustering**] [**John Wesley Hostetter(Designed) & Chengyuan Liu (Graded)**] We will use the same dataset as in Question 1 for the following problem. The *Euclidean Distance* matrix between each pair of the data points is listed in the figure below:

   (a) (8 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. **In case of a tie please resolve in alphabetical order of the points' labels**. NO PARTIAL CREDIT.

   (b) (4 points) Using Sum of Squared Error (SSE) and assuming there are three clusters, which of the *single link* and *complete link* hierarchical clustering will yield better results? Justify your answer.

   (c) (3 points) Compare the clusters from 2(b) with the clusters found using K-means in Question 1 by calculating their corresponding Sum of Squared Errors (SSE)s. According to their SSE results, which is better: K-means or hierarchical clustering?

	A	B	C	D	E	F	G	H	I	J	K
A	0.000	4.472	7.211	9.220	3.162	7.280	7.071	2.000	5.099	9.055	2.236
B	4.472	0.000	2.828	7.280	1.414	5.385	3.162	2.828	4.243	7.616	5.000
C	7.211	2.828	0.000	6.403	4.243	5.000	1.414	5.657	5.099	7.071	7.280
D	9.220	7.280	6.403	0.000	8.062	2.000	5.000	9.000	4.123	1.000	7.616
E	3.162	1.414	4.243	8.062	0.000	6.083	4.472	1.414	4.472	8.246	4.123
F	7.280	5.385	5.000	2.000	6.083	0.000	3.606	7.000	2.236	2.236	5.831
G	7.071	3.162	1.414	5.000	4.472	3.606	0.000	5.831	4.000	5.657	6.708
H	2.000	2.828	5.657	9.000	1.414	7.000	5.831	0.000	5.099	9.055	3.606
I	5.099	4.243	5.099	4.123	4.472	2.236	4.000	5.099	0.000	4.000	3.606
J	9.055	7.616	7.071	1.000	8.246	2.236	5.657	9.055	4.000	0.000	7.280
K	2.236	5.000	7.280	7.616	4.123	5.831	6.708	3.606	3.606	7.280	0.000

Figure 3: Euclidean Distance Matrix

3. (8 points) [**Frequent Itemset**] [**Angela Zhang (Designed) & Tyrone Wu (Graded)**] For the transaction Table 4 given below, please answer the following questions:

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

Table 1: Transactions Data

(a) (1 point) Explain what is frequent itemset and give an example of 2-itemset that is frequent itemset with minimal support count = 7.

(b) (3 points) Explain what is closed frequent itemset and list ALL of them with support count = 7 in *alphabetical order*. No partial credit.

(c) (3 points) Explain what is maximal frequent itemset and list ALL of maximal itemset with support count = 7 in *alphabetical order*. No partial credit.

(d) (1 point) Compute the support and confidence for the association rule $\{A, C\} \rightarrow \{F\}$.

4. (13 points) [**Association Analysis**] [**Ge Gao (Designed) & Tyrone Wu (Graded)**] Consider the following market basket transactions shown in the Table 2 below.

Transaction ID	Items ordered
1	{Beef, Butter, Chicken}
2	{Butter, Chicken, Milk}
3	{Butter, Cheese, Eggs, Soda}
4	{Beef, Eggs, Juice, Milk, Soda}
5	{Beef, Eggs, Soda}
6	{Beef, Bread, Cheese, Eggs}
7	{Bread, Soda}
8	{Beef, Cheese, Chicken}
9	{Chicken, Juice, Soda}
10	{Cheese, Milk, Soda}

Table 2: Market Basket Transactions Data

For each of the following question, briefly explain your answers in 2-3 sentences. NO PARTIAL CREDIT.

(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 count)?

(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 3-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) Find two pairs of items, a and b, such that the rules $\{a\} \rightarrow \{b\}$ and $\{b\} \rightarrow \{a\}$ have the same confidence, and the support for each rule is greater than or equal to 0.2.

5. (24 points) [**Apriori algorithm**] [**Ge Gao (Designed) & Pragna Bollam (Graded)**] Consider the data set shown in Table 3 and answer the following questions using Apriori algorithm.

TID	Items
t_1	A,B
t_2	A,B,E
t_3	A,B,D
t_4	B,E
t_5	A,B,D,E
t_6	C,D
t_7	B,C,D,F
t_8	A,D
t_9	A,B,D,F
t_{10}	A,B,C

Table 3: Apriori algorithm

(a) (10 points) Show (compute) each step of frequent itemset generation process using Apriori algorithm, with support count of 2.

(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.)

(c) (4 points) Mark closed and maximal frequent itemsets on the lattice structure from (b) if there's any.

6. (30 points) [**Frequent Pattern Tree**] [**Angela Zhang (Designed) & Chengyuan Liu (Graded)**] Consider the following data set shown in Table 4 and answer the following questions using FP-Tree.

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

Table 4: Transactions Data

(a) (15 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). Hint: Do not forget to include the header table that locates the starts of the corresponding linked item lists through the FP-Tree. NO PARTIAL CREDIT.

(b) (15 points) Using the FP-Tree constructed and support count = 2, generate all the frequent patterns with the base of item H step by step. List the frequent itemsets in alphabetical order. NO PARTIAL CREDIT.