CSC 522 HW 5

Group H29

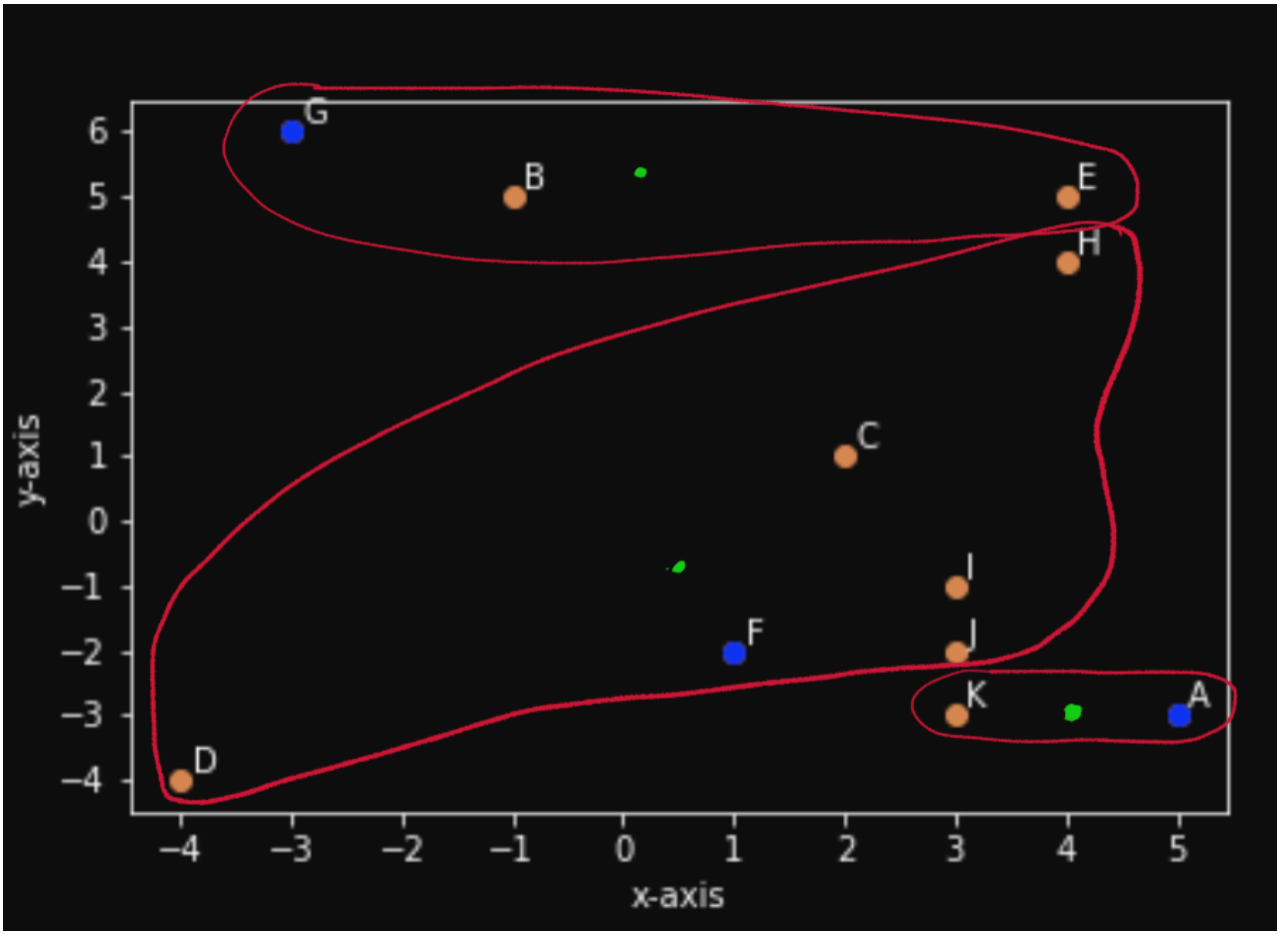
November 21, 2022

Ethan Purnell (efpurne2)
Corey Capooci (cvcapooc)
Jackson Ingraham (jtingrah)
Github Repository: engr-ALDA-Fall2022-H29
`https://github.ncsu.edu/efpurne2/engr-ALDA-Fall2022-H29`

Question 1:

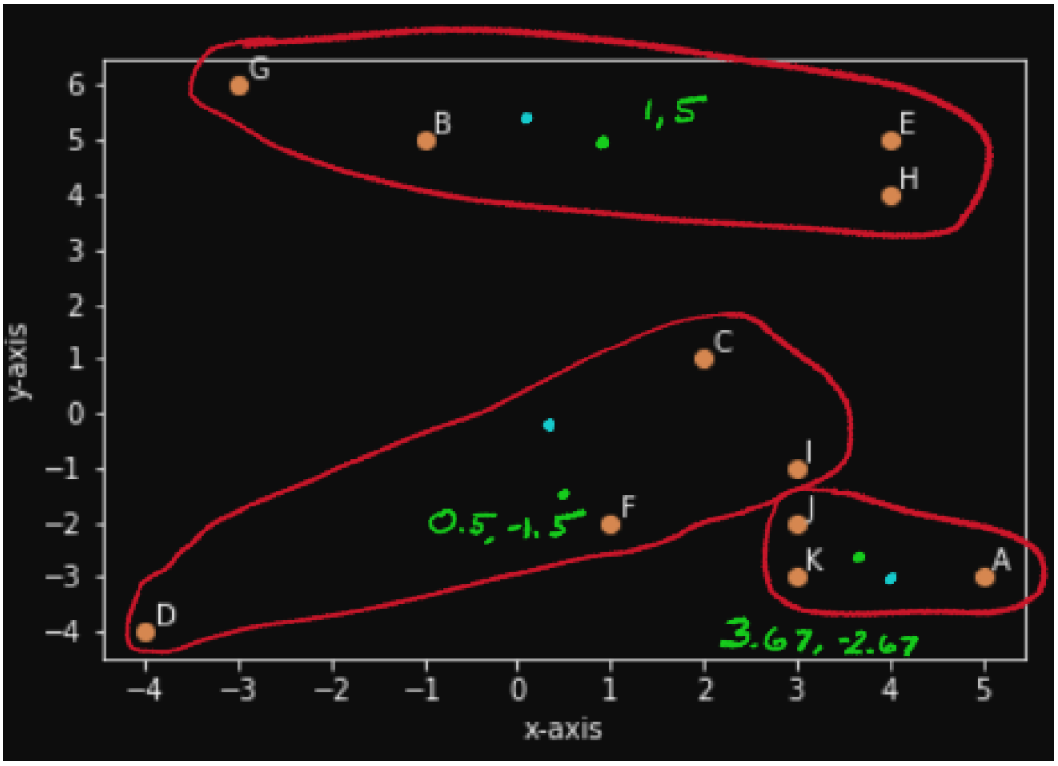
A:

The centroid coordinates after the first K-means clustering are (4, -3), (0, 5.333), and (1.5, -0.667).

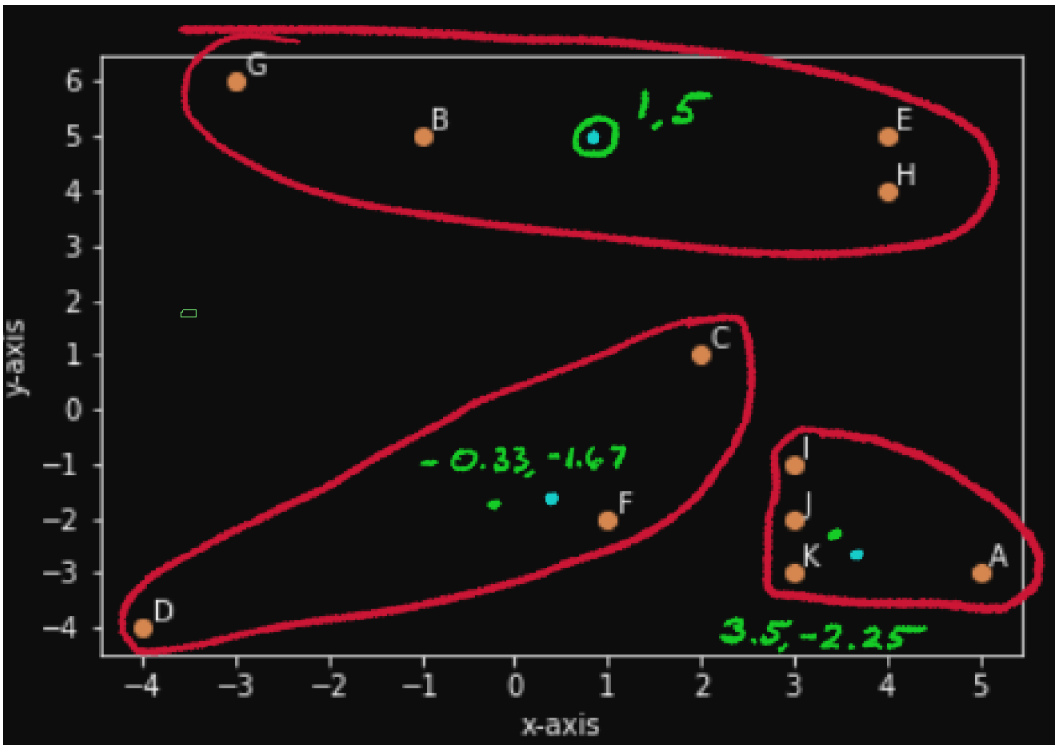


B:

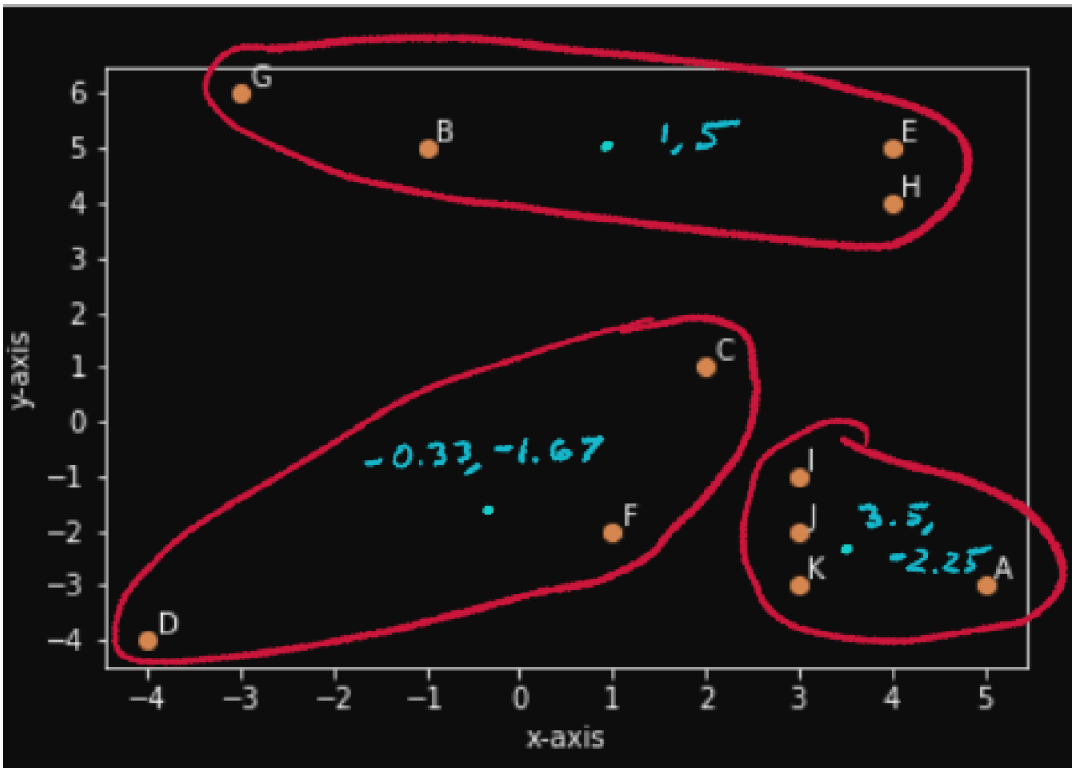
There are four total iterations for the clustering algorithm to converge - 1 iteration as shown in A and 3 more as shown below:



Coordinates: (1,5), (0.5, -1.5), (3.67, -2.67)



Coordinates: (1,5), (-0.33,-1.67), (3.5, -2.25)



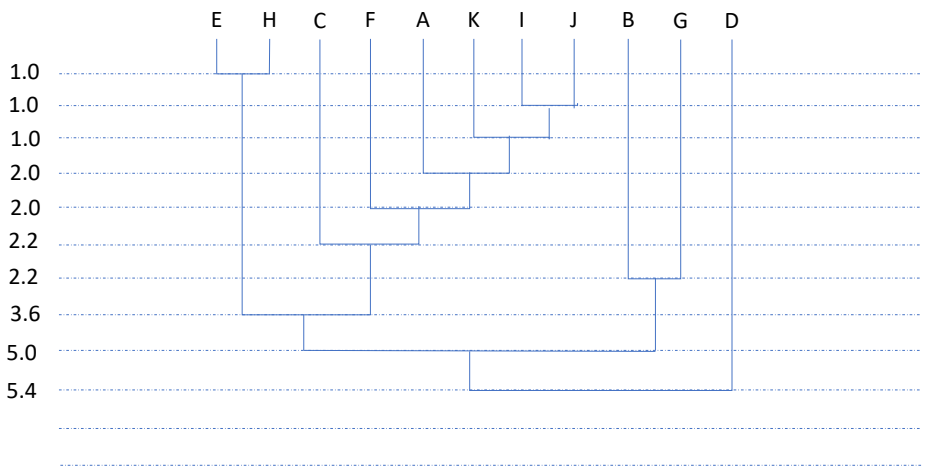
Coordinates remain the same, since this iteration just establishes convergence is reached

Question 2:

A:

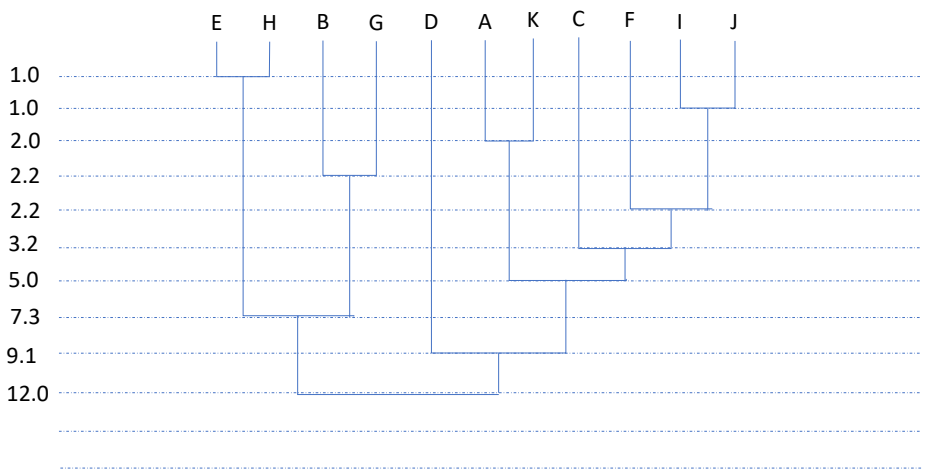
For both of the diagrams below, note that the heights are not to scale in order to show the order of which the itemsets were merged. The merge order is shown by the dotted lines with the highest line representing first merge and increments as the diagram goes down. Each merge then has a corresponding height shown at the beginning of the dotted line.

Single Link



Single Link

Complete Link



Complete Link

B:

Using the results from 2A to get three clusters, the clusters for single link are {D}, {B, G}, and {A, C, E, F, H, I, J, K}. The clusters for complete link are {D}, {B, G, E, H}, and {A, C, F, I, J, K}.

Based on these clusters, single link has a sum squared error of 82.25 and complete link has a sum square error of 60.167. Based on the results of the sum square error calculations, complete link yields better results.

C:

Using the center points of the three clusters calculated from 1, the k-means sum squared errors is 79.08. Comparing the sum square errors between k-means and hierarchical clustering, k-means performs better than single link clustering, but worse than complete link clustering. Therefore, there are no definitive conclusion of which performed better between k-means and hierarchical clustering.

Question 3:

A:

A frequent itemset is an itemset whose support count (number of occurrences, essentially), is greater than or equal to some minimal support count.

An example of a frequent 2-itemset with minimal support count = 7 is {G,H}.

B:

A closed frequent itemset is an itemset with frequency greater tan the minimal support count which has no immediate supersets with the same support count.

The only closed frequent itemset with support count = 7 we found was {G,H}.

C:

A maximal frequent itemset is an itemset with frequency greater than the minimal support count which has no supersets that are frequent.

Again, the only maximal frequent itemset we found satisfying the support count = 7 criterion was the itemset {G,H}.

D:

For the association rule {B,G} $\rightarrow$ {H}, we calculated the support to be s = 4/15 or s = 0.27. We calculated the confidence to be c = 4/6 or c = 0.67.

Question 4:

A:

After counting up all the individual items (i.e. Backpack, Eraser, Marker), the total amount of items found were 9. The maximum size of itemsets with support count ¿ 1 is equal to the largest transaction size, which is 5.

B:

The maximum number of association rules can be calculated by using $3^d - 2^d + 1$ where d is the items found in the dataset (9). Using this formula, the final output is 18660 possible association rules.

C:

The maximum number of 3-itemsets can be calculated using nCr, where n = 9 and r = 3. Once calculated, this is found to be 84 itemsets.

D:

The largest support directly correlates to the largest support count, so support counts can be compared once the possible combinations for 2-itemsets are found. The itemset which had the largest support/support count was a tie between {Glue, Marker} and {Glue, Pen}.

E:

The confidence between a pair of items is interchangeable if the number of times item a and item b aren't paired together are the same. The instance of this occurring is when a = {Eraser} and b = {Glue}.

Question 5:

A:

We will illustrate each step of the Apriori algorithm computations in a tabular format. First, we start with the individual items and their support counts.

Itemset	Support
{A}	3
{B}	2
{C}	8
{D}	7
{E}	6
{F}	3

In this case, all of the itemsets meet the support count = 2 criterion. Next, we build all of the possible 2-itemsets that are possible, scan the transactions for their support counts, and eliminate any itemsets with support count < 2. The rows containing itemsets to be eliminated will be highlighted in gray.

Itemset	Support
{AB}	1
{AC}	2
{AD}	1
{AE}	2
{AF}	0
{BC}	2
{BD}	1
{BE}	2
{BF}	0
{CD}	6
{CE}	4
{CF}	3
{DE}	4
{DF}	2
{EF}	1

Now, the remaining frequent 2-itemsets are:

Itemset	Support
{AC}	2
{AE}	2
{BC}	2
{BE}	2
{CD}	6
{CE}	4
{CF}	3
{DE}	4
{DF}	2

We will use these 2-itemsets to write the possible frequent 3-itemsets, scan the transactions again, and generate the table of 3-itemsets and their support with gray highlighted lines for the itemsets whose support count < 2.

Itemset	Support
{ACE}	1
{ABC}	1
{ACD}	1
{ACF}	0
{ADE}	0
{BCE}	2
{BCD}	1
{BCF}	0
{BDE}	1
{CDE}	3
{CDF}	2
{CEF}	1
{DEF}	1

Then, the remaining valid 3-itemsets are:

Itemset	Support
{BCE}	2
{CDE}	3
{CDF}	2

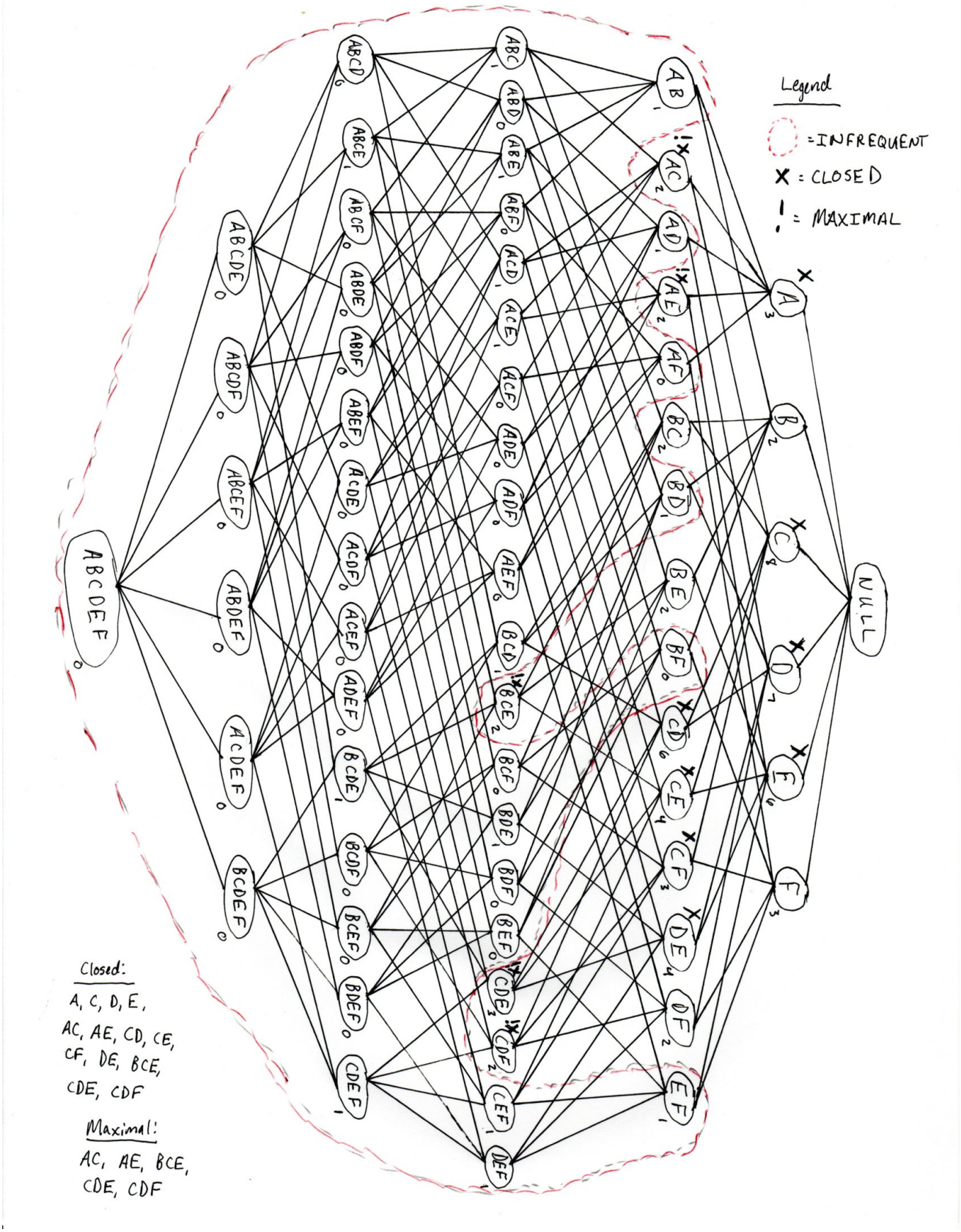
Now, we can combine these to see if there are any 4-itemsets with support count > 2:

Itemset	Support
{BCDE}	1
{BCDF}	0
{CDEF}	1

This time, there are no further supersets that could be created with a greater support count than those 4-itemsets above. So we're done!

B & C:

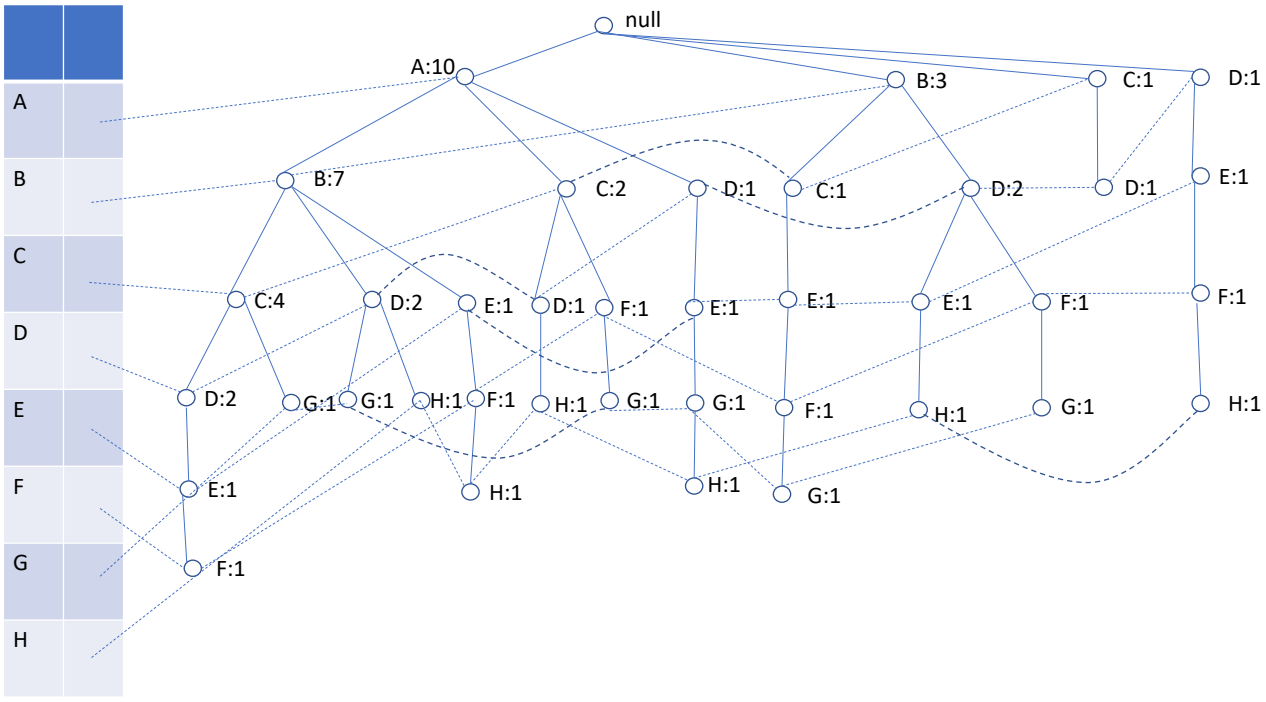
The lattice structure is shown on the following page. The branches to be pruned are circled in the red dashed line. Counts for each node are displayed at the bottom right of each node. The classification (maximal or closed) is given at the top left of the nodes that satisfy one or both of those classification criteria. In addition, the closed- and maximal-frequent itemsets are listed at the bottom left of the diagram.


Legend
= INFREQUENT
X = CLOSED
! = MAXIMAL
NULL
Closed:
A, C, D, E,
AC, AE, CD, CE,
CF, DE, BCE,
CDE, CDF
Maximal:
AC, AE, BCE,
CDE, CDF

Question 6:

A:

The initial construction for the FP-Tree based on the transactions is shown below.



B:

The frequent itemsets with support count = 2 and including H are,

{A, B, H} {A, D, H} {A, E, H} {A, H} {B, D, H} {B, E, H} {B, H} {D, E, H} {D, H} {E, F, H} {E, H} {F, H} {H}