Data Mining Practice Final Exam Solutions

Note: This practice exam only includes questions for material after midterm—midterm exam provides sample questions for earlier material. The final is comprehensive and covers material for the entire year.

True/False Questions:

1. [T] F — Our use of association analysis will yield the same frequent itemsets and strong association rules whether a specific item occurs once or three times in an individual transaction.

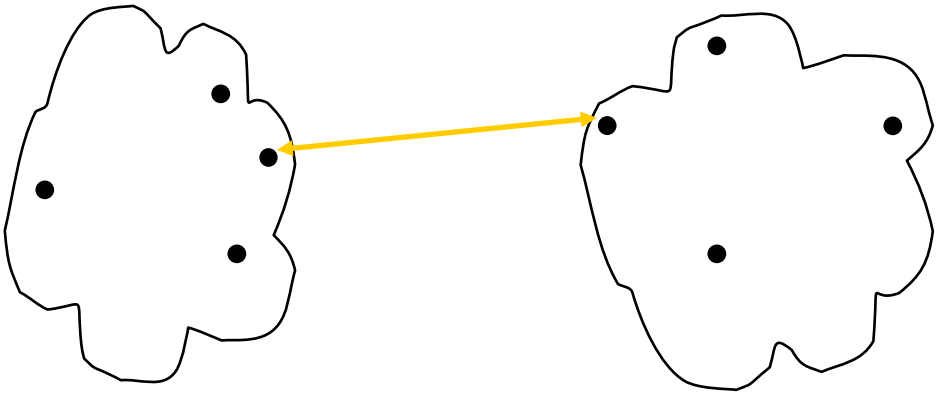
2. T [F] — The k-means clustering algorithm that we studied will automatically find the best value of *k* as part of its normal operation.

3. [T] F — A density-based clustering algorithm can generate non-globular clusters.

4. [T] F — In association rule mining the generation of the frequent itermsets is the computational intensive step.

Multiple Choice Questions

5. In the figure below, there are two clusters. They are connected by a line which represents the distance used to determine inter-cluster similarity.



Which inter-cluster similarity metric does this line represent (circle one)?

a. MIN
b. MAX
c. Group Average
d. Distance between centroids

Short Form Questions

6. (4 points) We generally will be more interested in association rules with high confidence. However, often we will not be interested in association rules that have a confidence of 100%. Why? Then specifically explain why association rules with 99% confidence may be interesting (i.e., what might they indicate)?

   While we generally prefer association rules with high confidence, a rule with 100% confidence most likely represents some already know fact or policy (e.g., checking account → savings account may just indicate that all customers are required to have a checking account if they have a savings account). Rules with 99% confidence are interesting not because of the 99% part but because of the 1% part. These are the exceptions to the rule. They may indicate, for example, that a policy is being violated. They might also indicate that there is a data entry error. Either way, it would be interesting to understand why the 1% do not follow the general pattern.

7. (4 points) The algorithm that we used to do association rule mining is the Apriori algorithm This algorithm is efficient because it relies on and exploits the Apriori property. What is the Apriori property?

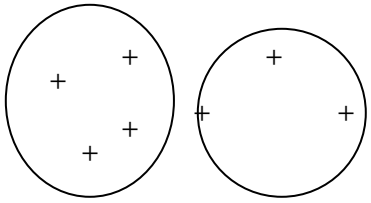
   The Apriori property state that if an itemset is frequent then all of its subsets must also be frequent.

8. (4 points) Discuss the basic difference between the agglomerative and divisive hierarchical clustering algorithms and mention which type of hierarchical clustering algorithm is more commonly used.

   Agglomerative methods start with each object as and individual cluster and then incrementally builds larger clusters by merging clusters. Divisive methods, on the other hand, start with all points belonging to one cluster and then splits apart a cluster each iteration. The agglomerative method is more common.

9. (4 points) Are the two clusters shown below well separated? Circle an answer: Yes [No]
   Now in one or two sentences justify your answer.

   It is not well separated because some points in each cluster are closer to points in another cluster than to points in the same cluster.



Long Problem (33 points)

1. You are given the transaction data shown in the Table below from a fast food restaurant. There are 9 distinct transactions (order:1 – order:9) and each transaction involves between 2 and 4 meal items. There are a total of 5 meal items that are involved in the transactions. For simplicity we assign the meal items short names (M1 – M5) rather than the full descriptive names (e.g., Big Mac).

Meal Item	List of Item IDs	Meal Item	List of Item IDs
Order:1	M1, M2, M5	Order:6	M2, M3
Order:2	M2, M4	Order:7	M1, M3
Order:3	M2, M3	Order:8	M1, M2, M3, M5
Order:4	M1, M2, M4	Order:9	M1, M2, M3
Order:5	M1, M3		

For all of the parts below the **minimum support is 2/9** (.222) and the **minimum confidence is 7/9** (.777). Note that you only need to achieve this level, not exceed it. Show your work for full credit (this mainly applies to part a).

a. Apply the Apriori algorithm to the dataset of transactions and identify *all* frequent k-itemsets. Show all of your work. You must show candidates but can cross them off to show the ones that pass the minimum support threshold. This question is a bit longer than the homework questions due to the number of transactions and items, so proceed carefully and neatly.

Note: if a candidate itemset is pruned because it violates the Apriori property, you must indicate that it fails for this reason and not just because it does not achieve the necessary support count (i.e., in these cases there is no need to actually compute the support count). So, explicitly tag the itemsets that are pruned due to violation of the Apriori property. This really did not come up on the homework because those problems were quite short. (If you do not know what the Apriori property is, do not panic. You will ultimately get the exact same answer but will just lose a few points).

b. Find all *strong* association rules of the form: $X \wedge Y \rightarrow Z$ and note their confidence values. Hint: the answer is not 0 so you should have at least one frequent 3-frequent itemset.

Show your work and solution for each part below and on the following two blank pages. I have copied the raw transactional data to each page so that you need not keep flipping back.

Solution Part a

C1/L1

Itemset	Support Count
M1	6
M2	7
M3	6
M4	2
M5	2

C2/L2

Itemset	Support Count
{M1, M2}	4
{M1, M3}	4
~~{M1, M4}~~	~~1~~
{M1, M5}	2
{M2, M3}	4
{M2, M4}	2
{M2, M5}	2
~~{M3, M4}~~	~~0~~
~~{M3, M5}~~	~~1~~
~~{M4, M5}~~	~~0~~

L2 (after pruning)

Itemset	Support Count
{M1, M2}	4
{M1, M3}	4
{M1, M5}	2
{M2, M3}	4
{M2, M4}	2
{M2, M5}	2

C3 initial

Itemset	Support Count
{M1, M2, M3}	
{M1, M2, M5}	
{M1, M3, M5}	
{M2, M3, M4}	
{M2, M3, M5}	
{M2, M4, M5}	

C3 after checking for Apriori Property

Itemset	Comment
{M1, M2, M3}	
{M1, M2, M5}	
~~{M1, M3, M5}~~	No {M3, M5}
~~{M2, M3, M4}~~	No {M3, M4}

~~{M2, M3, M5}~~	No {M3, M5}
~~{M2, M4, M5}~~	No {M4, M5}

C3 Final/L3

Itemset	Support Count
{M1, M2, M3}	2
{M1, M2, M5}	2

C4/L4

Itemset	Support Count
~~{M1, M2, I3, M5}~~	{No M3, M5}

Solutions Part b)

Rule	**Confidence**
~~M1 ∧ M2 → M3~~	~~2/4 = .50~~
~~M2 ∧ M3 → M1~~	~~2/4 = .50~~
~~M1 ∧ M3 → M2~~	~~2/4 = .50~~
~~M1 ∧ M2 → M5~~	~~2/4 = .5~~
M1 ∧ M5 → M2	2/2 = 1.0
M2 ∧ M5 → M1	2/2 = 1.0