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.

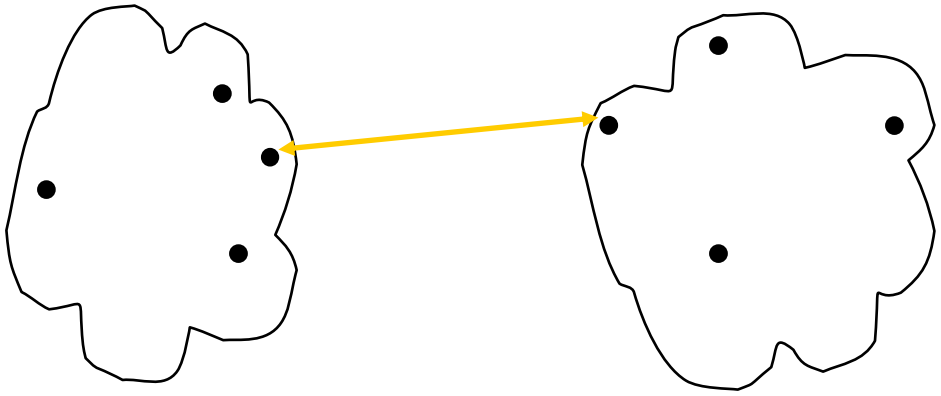
2. T [F] — The k-means clustering algorithm that we studied will automatically find the best value of *k* as part of its normal operation.

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?

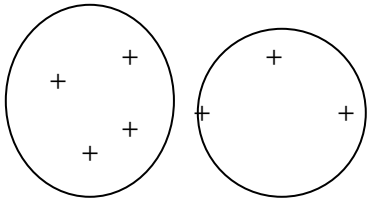
   The Apriori property state that if an itemset is frequent then all of its subsets must also be frequent.

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. A database has 4 transactions, shown below.

TID	Date	items_bought
T100	10/15/04	{K, A, D, B}
T200	10/15/04	{D, A, C, E, B}
T300	10/19/04	{C, A, B, E}
T400	10/22/04	{B, A, D}

Assuming a minimum level of support *min_sup* = 60% and a minimum level of confidence *min_conf* = 80%:

(a) Find *all* frequent itemsets (not just the ones with the maximum width/length) using the Apriori algorithm. Show your work—just showing the final answer is not acceptable. For each iteration show the candidate and acceptable frequent itemsets. You should show your work similar to the way the example was done in the PowerPoint slides.

Answer:

C1/L1

Itemset	Support Count
{A}	4
{B}	4
~~{C}~~	~~2~~
{D}	3
~~{E}~~	~~2~~
~~{K}~~	~~1~~

C2/L2

Itemset	Support Count
{A, B}	4
{A, D}	3
{B, D}	3

C3/L3

Itemset	Support Count
{A, B, D}	3

The final answer is: {{A}, {B}, {D}, {A, B}, {B, D}, {A, B, D}}

(include {A,D} above)

(b) List all of the strong association rules, along with their support and confidence values, which match the following metarule, where *X* is a variable representing customers and $item_i$ denotes variables representing items (e.g., "A", "B", etc.).

$\forall x \in transaction, buys(X, item_1) \wedge buys(X, item_2) \Rightarrow buys(X, item_3)$

Hint: don't worry about the fact that the statement above uses relations. The point of the metarule is to tell you to only worry about association rules of the form $X \wedge Y \Rightarrow Z$ (or $\{X, Y\} \Rightarrow Z$ if you prefer that notation). That is, you *don't* need to worry about rules of the form $X \Rightarrow Z$.

Grading: This part is worth 4 points. Each of the strong association rules is worth 2 points.

Answer:

$buys(X, A) \wedge buys(X, B) \rightarrow buys(X, D)$	(75%, 75%)	Not Strong
$buys(X, A) \wedge buys(X, D) \rightarrow buys(X, B)$	**(75%, 100%)**	**Strong**
$buys(X, B) \wedge buys(X, D) \rightarrow buys(X, A)$	**(75%, 100%)**	**Strong**