CSE-4502/5717 Big Data Analytics
Written Assignment 1

Data Processing; Rule Mining; Clustering

Due: 11:59 pm (23:59), 02/20

Requirements:

1. The written assignment will be graded based on correctness, accuracy and clarity.
2. Please prepare your answers using a **word** document, and submit the final assignment in a **pdf** file through HuskyCT.
3. Please explicitly indicate the ids of the questions you are answering for ease of grading.
4. Even if your answer is not correct, you may still get certain partial marks based on your calculation/analysis process. Please present the necessary calculation process if any.
5. Any submission of required assignments past the date they are due are subject to a grade reduction. In particular, late homework/assignments will be penalized by 10% (from the maximum points) each day and will not be accepted after 48 hours. For example, if the submission is 2 days late, 80 points (out of 100) will be the maximum score to obtain.

1. (total 10 marks) It is important to define or select similarity measures in data analysis. However, there is no commonly-accepted subjective similarity measure. Results can vary depending on the similarity measures used. Nonetheless, seemingly different similarity measures may be equivalent after some transformation.

Suppose we have the following two-dimensional data set:

Data Points	A_1	A_2
x_1	1.5	1.8
x_2	2.1	1.9
x_3	1.6	1.9
x_4	1.3	1.6
x_5	1.5	1.1

(a) (5 marks) Consider the data as two-dimensional data points. Given a new data point, x = (1.5, 1.3) as a query, rank the database points based on similarity with the query using Euclidean distance, Manhattan distance, and cosine similarity. Please provide a table to list all the similarity values calculated.

(b) (5 marks) Normalize the data set to make the norm of each data point equal to 1 (normalized into a vector (a,b) such that $a^2 + b^2 = 1$). Use Euclidean distance on the transformed data to rank the data points.

2. (total 15 marks) Consider the following set of frequent 2-itemsets:

{p, q}, {p, r}, {p, s}, {p, t}, {q, r}, {q, t}, {r, s}, {s, t}.

(a) (5 marks) List all the candidate 3-itemsets produced during the candidate generation step of the Apriori algorithm.

(b) (5 marks) List all the candidate 3-itemsets that survive the pruning step of the Apriori algorithm.

(c) (5 marks) Based on the list of candidate 3-itemsets given above, is it possible to generate at least one frequent 4-itemset? State your reason clearly.

3. (total 20 marks) Here we need to solve two problems related to association rule mining, one related to Apriori and one related to FPTree.

(1) (10 marks) A database has 7 transactions (TID: Transaction Index). Let the minimum support threshold min_sup is 0.5.

TID	Item Bought
T1	a, c, d, f, g
T2	a, b, d, e, g
T3	a, d, f, g
T4	b, d, f
T5	e, f, g
T6	a, b, c, d, g
T7	a, b, e, g

Use the Apriori algorithm to generate the frequent item sets. Please explain the process of the generation in details, including all the candidate item sets and frequent item sets. Please use the tables/diagrams shown in Association Rule Mining Part I (Slide 25) for the result demonstration.

(2) (10 marks) We are given the transaction database as:

TID	Item Bought
T1	A, C, D
T2	A, B, C
T3	A, B, E
T4	B, E

Please build the FPTree for the transaction database with the minimum support count 2. Please provide clear and readable figure or screenshot of the constructed FPTree (refer to Slide 9, FPTree, "Association Rule Mining Part II"). We assume alphabetical order for items with the same frequency.

4. (total 20 marks) Consider the closing prices for five stocks (A, B, C, D, and E) listed in the following table. Suppose you are interested in applying association rule mining to the data.

Figure 1 Example of Stock Market Data

Day	A	B	C	D	E
1	10.50	11.00	20.00	80.00	95.00
2	12.30	10.40	26.40	76.50	90.20
3	12.00	10.80	26.50	75.50	91.00
4	11.20	10.00	25.50	72.00	87.10
5	11.30	10.20	25.30	73.20	88.90
6	12.50	10.70	27.50	70.00	88.50
7	13.00	10.80	28.80	72.00	90.20
8	13.80	11.00	29.80	71.80	91.00
9	12.95	10.80	27.90	71.00	91.20
10	12.05	10.10	26.10	72.60	92.80
11	11.40	10.05	24.95	70.40	90.10

We first convert the stock market prices into transaction data. For each stock X on a trading day, compute the change in its closing price,

$$\Delta_X(t) = \frac{p_t(X) - p_{t-1}(X)}{p_{t-1}(X)}$$

which is the percentage of increase/decrease compared with the previous stock price. $p_t(X)$ is the price of stock X on day t. Next, create an "item" X-UP for a trading day if the increase is at least 5% ($\Delta_X(t)$ is greater than 0.05; if the closing price is up by at least 5%), or X-DOWN if decrease is at least 5% ($\Delta_X(t)$ is lower than -0.05; if the closing price is down by at least 5%). Assume each transaction corresponds to a trading day (starting from Day 2). Note that there are 10 possible items: A-UP, A-DOWN, B-UP, B-DOWN, …, E-UP, E-DOWN. Based on the original transactions, we can generate 10 transactions from above table as:

Transaction1: {A-UP, B-DOWN, C-UP, E-DOWN};
Transaction2: {};
Transaction3: {A-DOWN, B-DOWN};
Transaction4: {};
Transaction5: {A-UP, C-UP};
Transaction6: {};
Transaction7: {A-UP};
Transaction8: {A-DOWN, C-DOWN};
Transaction9: {A-DOWN, B-DOWN, C-DOWN};
Transaction10: {A-DOWN};

(a) (10 marks) Assuming the minimum support threshold is 20%, i.e., an itemset has to appear at least twice in the transaction data to be considered *frequent*, list all the frequent 1-itemsets, 2-itemsets, and so on (including their *support values*), that can be extracted from the data.

(b) (10 marks) Based on the frequent itemsets found in part (a), generate all the association rules with minsup = 20% and minconf = 60%. In your answer, for every rule generated you should list the support and confidence calculated. Please ignore the rules in which their left or right hand side correspond to an empty set.

5. (15 marks) Consider the following eight two-dimensional data points:

x_1: (23, 12), x_2: (6, 6), x_3: (15, 0), x_4: (15, 28), x_5:(20, 9), x_6: (8, 9), x_7: (20, 11), x_8: (8, 13),

Consider k-means algorithm to answer the following questions. You are required to show the information about each final cluster (including the mean of the cluster and all data points in this cluster). You can consider writing a program for this part but you are not required to submit the program.

(a) (5 marks) If k = 2 and the initial means are (20, 9) and (8, 9), what is the output of the algorithm? In the output, you are required to show the information about each final cluster (including the mean of the cluster and all data points in this cluster).

(b) (5 marks) If k = 2 and the initial means are (15, 0) and (15, 29), what is the output of the algorithm? In the output, you are required to show the information about each final cluster (including the mean of the cluster and all data points in this cluster).

(c) (5 marks) What are the advantages and the disadvantages of the k-means algorithm? For each disadvantage, please also give a suggestion to enhance the k-means algorithm.

6. (total 20 marks) Consider the following set of one-dimensional data points: {0.1, 0.2, 0.4, 0.5, 0.6, 0.8, 0.9}.

Index of Iteration	**Cluster assignment of data points** (put the label of cluster, either A, B or C for each data point; iteration 0 means initialization and no label is assigned)							**Centroid Locations** (calculate the updated coordinate for the centroid)		
	0.1	0.2	0.4	0.5	0.6	0.8	0.9	A	B	C
0	-	-	-	-	-	-	-	0.00	0.25	0.60
1										
2										
3										

(a) (15 marks) Suppose we apply kmeans clustering to obtain three clusters, A, B, and C. If the initial centroids are located at {0, 0.25, 0.6}, respectively, show the cluster assignments and locations of the centroids after the first **three** iterations (you can use a table as above and fill in the missing values).

(b) (5 marks) Compute the SSE (sum of squared errors) of the k-means solution (after **3** iterations; SSE is calculated after the centroid locations are updated). You can show the calculation process.