Assignment 1 - 10%

NAMES: Philip Cappello and Hassan Khan
STUDENT ID's: 216465098 and 216785099

1. **(5 points) Please answer the following questions:**
   (a) A financial institution would like to partition their investors into similar groups according to their demographic profiles. What data mining task is best suited to this problem?

   The data mining task best suited to this problem would be clustering the data based on demographic characteristics of the dataset. This is a descriptive data mining technique.

   From intro-2-2 slide 24: Basic Data Mining Tasks

   - **Predictive:** discover patterns on previous and current data in order to make predictions on future data
     - Classification
     - Regression
   - **Descriptive:** discover knowledge that characterizes general properties of data
     - Clustering
     - Concept characterization/summarization
     - Association analysis (frequent itemsets, association rules)
     - Sequential pattern mining
   - **Predictive or descriptive**
     - Time series analysis
     - Outlier detection

   (b) Suppose the above institution already knows for some of their investors whether or not they have bought a certain stock. Which data mining task would be suited to the problem of identifying the investors among the remaining investors, who might buy that particular stock?

   The data mining task best suited to this problem would be classification, a predictive data mining technique. This is because we are using data of investors that we know did or did not buy a certain stock, and we can use the past/current data to make predictions about the rest of the investors who we do not know if they bought a certain stock.

   (c) Suppose the institution has recorded all the transactions made by their investors. What data mining task would be best suited to finding sets of stocks that are often bought together on the same day with the Nvidia Stock?

We can use association analysis, a descriptive data mining technique. This is because association analysis is widely used for transactional data analysis which can help determine what products were purchased together filtered by a specific date.

(d) Suppose the institution finds the sets of stocks often bought together with the Nvidia Stock, how would this knowledge be used by the institution?

Given a set of stocks often bought with the Nvidia stock, this institution can suggest these stocks to its newer customers who buy Nvidia stocks, or to newer customers who are buying stocks that are a subset of the set of stocks often bought with the Nvidia stock.

(e) Suppose that a small number of investors lie about their demographic profile, and this results in a mismatch between the stock-buying behavior and the demographic profile, as suggested by comparison with the remaining data. Which data mining task would be best suited to finding such investors?

The data mining task best suited to finding such investors is Outlier/Anomaly Detection. This is because it is used frequently in rare events analysis. Methods of Outlier/Anomaly Detection include clustering, classification and regression analysis.

2. **(6 points) Suppose that you are employed as a data mining consultant for an online newspaper company. Describe how data mining can help the company by giving specific examples of how each of the techniques: classification, clustering, and association rule mining, can be applied.**

Classification - if you have a database of all the articles published by this online newspaper company and the traffic in which each article brings, you can use classification to determine what headlines brought the most views upon the articles. For example, if the newspapers that are about Politics have more reads than those about Sports, the online newspaper company should consider headlining their newspapers with political articles. If, from the previous data gathered, the online newspaper company has more reads of its sports articles during specific times of the year such as playoffs or finals of a sporting event, then the company should headline their newspaper with Sports headlines during these times of the year.

Clustering - perhaps using the age and geographic data of the readers of this online newspaper, along with the subject of the articles being read by customers with these characteristics, the company can tailor their articles to age groups and geographical areas where most reads come from. For example, if most people who read the comic section of the newspaper are between ages of 50-70 and live in Toronto, the online newspaper company should tailor the subject of the comic section to match the subjects read by consumers with these attributes. If most of these people who are aged

50-70 and live in Toronto read Political articles, then the company should use Political jokes referencing Toronto in its comic section. The company can also use this data to target people aged 50-70 in different cities and use political comic sections to draw new readers.

Association Rule Mining - association rule mining can be used for target advertising on the companies' online platform. For example if it is known that most of the online newspapers' readers are above the age of 50 and most people reading these articles click on ads pertaining to Investment Advice, then the online newspaper company can place Investment Advice ads on the articles most read by these consumers over the age of 50. You can also get support and confidence numbers values for specific articles based on the attributes of each article. These attributes can include age of reader, geographical area of reader, etc.

3. (8 points) Prove that the join step for generating C_{k+1} in the Apriori algorithm does not miss any frequent itemset in L_{k+1} (that is, it does not miss any frequent itemsets of length k+1).

3) Prove that the join step for generating C_{k+1} in the Apriori algorithm does not miss any frequent itemset in L_{k+1}.

Let C_k be set of candidate k-itemsets
Let L_k be set of frequent k-itemsets

We are given L_k and need the join step first to generate C_{k+1}
→ Join L_k with L_k by joining two k-itemsets in L_k. Two k-itemsets are joinable if their first k-1 items are the same and the last item in first itemset is smaller than the last item in the second itemset.

If $L_k \subseteq C_k$ then $L_{k+1} \subseteq C_{k+1}$ meaning all members in L_{k+1} are in C_{k+1} implying C_{k+1} does not miss any frequent itemsets in L_{k+1}

Proof: During join step, we extend each itemset in L_{k-1} with all possible items and remove all those whose (k-1)-subsets are not in L_{k-1}, being left with L_k.
The condition $p.item_{k-1} < q.item_{k-1}$ ensures NO duplicates.

Thus $L_k \subseteq C_k \Rightarrow L_{k+1} \subseteq C_{k+1}$

Therefore, the join step for generating C_{k+1} in the Apriori algorithm does not miss any frequent itemset in L_{k+1}.

4. (14 points) A transaction database contains ten transactions as shown below:

Let min_sup=30% and min-_conf=50%.

(a) Using FP-growth, find all frequent itemsets that contain item *m*. Show the steps.

(b) Find all the strong association rules whose antecedent is *m*.

(c) Are there any misleading rules in the result of (b)? If yes, identify them and explain why they are misleading.

(SEE BELOW ATTACHMENT FOR ANSWERS TO a-c)

4)

TID	Items Bought
1	{m, n, p, q}
2	{n, o, p}
3	{m, n, p, q}
4	{m, o, p, q}
5	{n, o, p, q}
6	{n, p, q}
7	{o, p}
8	{m, n, o}
9	{m, p, q}
10	{n, p}

min_sup = 30%.
min_conf = 50%.
min_sup_count: $0.3 \times 10 = 3$

Def. Assoc Rule: $X \Rightarrow Y$ [s, c]
X: antecedent, Y: consequent

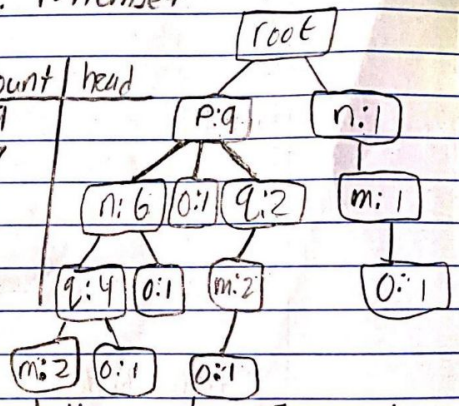
a) Using FP-Growth, find all frequent itemsets that contain item 'm'. Show the steps.

① Scan DB once, find freq. 1-itemset

Item	Count
m	5
n	7
o	5
p	9
q	6

⇒

Item	Count
p	9
n	7
q	6
m	5
o	5



TID	1	2	3	4	5
Freq Itemset	{p,n,q,m}	{p,n,o}	{p,n,q,m}	{p,q,m,o}	{p,n,q,o}

⋮

TID	6	7	8	9	10
Freq Itemset	{p,n,q}	{p,o}	{n,m,o}	{p,q,m}	{p,n}

For (m:5), it's Conditional Pattern base is: (pnq), (pq), (n)

Thus, freq patterns are: {(p,n,q:2), (p,q:2), (n:1)}

⇓

ANS: {(pm:2), (nm:3), (qm:4), (pnm:2), (pqm:4), (pnqm:2)}

b) Find all assoc rules whose antecedent is 'm'.

TID	Items Bought	Freq. Itemset	Support
1	{m,n,p,q}	{m}	50%
2	{n,o,p}	{n}	70%
3	{m,n,p,q}	{o}	50%
4	{m,o,p,q}	{p}	90%
5	{n,o,p,q} ⇒	{q}	60%
6	{n,p,q}	{m,n}	30%
7	{o,p}	{m,p}	40%
8	{m,n,o}	{m,q}	40%
9	{m,p,q}	{m,p,q}	40%
10	{n,p}		

Strong Rules: {m} → {n} [s=30%, c=60%]
{m} → {p} [s=40%, c=80%]
{m} → {q} [s=40%, c=80%]
{m} → {p,q} [s=40%, c=80%]

c) Are there any misleading rules in the result of (b)? If yes, identify them and explain why they are misleading.

* slide 84 (Assoc Rule 3-2) *

$$Lift\ (A \to B) = \frac{P(AB)}{P(A)P(B)}$$ for rule $A \to B$

if Lift of a rule ≤ 1, then it is misleading

we have
① $\{m\} \to \{n\}$ $P(m) = 50\%$, $P(n) = 70\%$, $P(mn) = 30\%$
② $\{m\} \to \{p\}$ $P(p) = 90\%$, $P(mp) = 40\%$.
③ $\{m\} \to \{q\}$ $P(q) = 60\%$, $P(mq) = 40\%$.
④ $\{m\} \to \{pq\}$ $P(pq) = 60\%$, $P(mpq) = 40\%$.

① $Lift\ (m \to n) = \frac{0.3}{0.5 \times 0.7} = 0.86$

Therefore the rule $\{m\} \to \{n\}$ is misleading because $P(n) > P(m)$

② $Lift\ (m \to p) = \frac{0.4}{0.5 \times 0.9} = 0.89$

Therefore the rule $\{m\} \to \{p\}$ is misleading because $P(p) > P(m)$

③ $Lift\ (m \to q) = \frac{0.4}{0.5 \times 0.6} = 1.33$

Therefore the rule $\{m\} \to \{q\}$ is NOT misleading.

④ $Lift\ (m \to pq) = \frac{0.4}{0.5 \times 0.6} = 1.33$

Therefore the rule $\{m\} \to \{pq\}$ is NOT misleading.

5. (7 points) This question is to get hands-on experience in finding association rules from a real data set by using an association rule mining library for Python. If you do not have Python installed, you can download it at https://www.python.org/downloads/

The libraries to be used in this question include:

Efficient_Apriori, a Python implementation of the Apriori algorithm for both frequent itemset and association rule mining. To install this library, use "pip install efficient_apriori" on a command line. For more information on this library, please see https://efficient-apriori.readthedocs.io/en/latest/

Pandas, one of the most widely used Python libraries in data science. It provides easy-to-use structures and data analysis tools. For this question, you can use this library for file output. To install Pandas, use "pip install pandas" on a command line.

The data set to be used in this question is a Walmart transaction data set, which can be downloaded here. The data set contains 12000 transactions. The transactions are described using item IDs. A dictionary that maps an item ID to its name can be found here.

For this question, you are given a Python program (association.py) that shows how you can read the input data file, how to use the apriori function in the Efficient_Apriori library, and how to write the rules into a csv file using the Pandas library. You can use this program and make changes to it to answer the following questions:

(a) Which itemset is the most frequent? What is its support (in percentage)?

The most frequent itemset is ('3873') : 1361

This is item with id 3873 and has a support of 1361/12000=11.3%

(b) What is the maximum length of the frequent itemsets that satisfies the minimum support 0.018? What is the most frequent itemset with the maximum length? What is the support (in percentage) of that itemset?

The maximum length of the frequent itemsets that satisfies the minimum support 0.018 are length-3 frequent itemsets

The most frequent itemset with that maximum length of 3 is ('2568','3623','3970')

The support of this itemset is 345/12000=2.8%

(c) Find the rules that satisfy minimum support 0.02 and minimum confidence 0.8. Interpret these rules using the meaningful names in the data dictionary provided above.

Strong Association Rules:

{3623} -> {2568} (conf: 0.829, supp: 0.033, lift: 14.426, conv: 5.527)

{4304} -> {4305} (conf: 0.817, supp: 0.022, lift: 25.392, conv: 5.282)

{3623, 3970} -> {2568} (conf: 0.922, supp: 0.029, lift: 16.043, conv: 12.155)

{2568, 3623} -> {3970} (conf: 0.876, supp: 0.029, lift: 14.758, conv: 7.564)

Interpretation using Data Dictionary:

{PINK REGENCY TEACUP AND SAUCER}->{GREEN REGENCY TEACUP AND SAUCER}
(conf: 0.829, supp: 0.033, lift: 14.426, conv: 5.527)

{SET6 RED SPOTTY PAPER CUPS}->{SET6 RED SPOTTY PAPER PLATE}
(conf: 0.817, supp: 0.022, lift: 25.392, conv: 5.282)

{PINK REGENCY TEACUP AND SAUCER,ROSES REGENCY TEACUP AND SAUCER}->
{GREEN REGENCY TEACUP AND SAUCER}
(conf: 0.922, supp: 0.029, lift: 16.043, conv: 12.155)

{GREEN REGENCY TEACUP AND SAUCER,PINK REGENCY TEACUP AND SAUCER}->
{ROSES REGENCY TEACUP AND SAUCER}
(conf: 0.876, supp: 0.029, lift: 14.758, conv: 7.564)

(d) Find the rules that satisfy minimum support 0.02, minimum confidence 0.6 and minimum lift 7 and write these rules into the walmart_rules.csv file, which will be used in the next question:

(answer is attached to assignment submission as walmart_rules.csv)