CSE 40647/60647 Data Science (Fall 2017)
Mid-term Exam

Instructor: Meng Jiang

(75 minutes, 100 marks, double sided reference, brief answers)

Name: NetID: Score:

1. [12] Introduction.
   Name at least **4** steps in "Data Science Research" or called "Knowledge Discovery from Data" (KDD).

   Answer: Any four of the following:

   (a) Task/problem definition
   (b) Data cleaning
   (c) Data integration
   (d) Task-relevant data selection
   (e) Data mining, Machine Learning
   (f) Pattern evaluation

□

2. [18] Data processing – Measures.

   (a) [9] (Distance measures) Given two data objects and four attributes/features, we have feature vectors of the two data objects as $(7, 4, -2, 1)$ and $(4, 5, -1, 6)$. Please **calcuate three** specific Minkowski distance measures between the two objects and **give the measures' names**. (Hint: $4^2 = 16$, $5^2 = 25$, $6^2 = 36$, $7^2 = 49$, $8^2 = 64$, $9^2 = 81$, $10^2 = 100$)

   Answer:

   i. Manhattan Distance (*i.e.*, L-1 norm): $3 + 1 + 1 + 5 = 10$.
   ii. Euclidean Distance (*i.e.*, L-2 norm): $\sqrt{3^2 + 1^2 + 1^2 + 5^2} = 6$.
   iii. Supremum Distance (*i.e.*, L-∞ norm): $\max\{3, 1, 1, 5\} = 5$.

□

(b) [9] (Correlation measures) Give one example wherein **Kulczynski measure** between two variables A and B is ***more appropriate*** than **Chi-square test** χ^2: You are asked to (1) explain your variables A and B, (2) give an equation to define the Kulczynski measure, (3) explain why Kulc measure is more appropriate in this example.

Answer: $Kulc$(A, B) = $\frac{1}{2}(\frac{s(A \cup B)}{s(A)} + \frac{s(A \cup B)}{s(B)})$, where $s(X)$ is the support of X.
Too many null transactions. Kulczynski is a null-invariant measure. □

3. [30] Data warehousing, OLAP, and data cube computation.
Suppose the base cuboid of a data cube contains two cells

- $(a_1, a_2, a_3, a_4, a_5, a_6) : 1$,
- $(a_1, \mathbf{b_2}, a_3, \mathbf{b_4}, a_5, \mathbf{b_6}) : 1$.

where $a_i \neq b_i$ for any dimension $i \in \{2, 4, 6\}$. Assume each dimension contains no concept hierarchy (i.e., has a single level). (Hint: $2^3 = 8$, $2^4 = 16$, $2^5 = 32$, $2^6 = 64$)

(a) [6] How many **nonempty cuboids** are there in this data cube?

Answer: 64. Since we have 6 dimensions with no concept hierarchy, there are 2^6 cuboids and all of them should not be empty. □

(b) [6] How many **nonempty closed cells** are there in this data cube?

Answer: 3. There are 3 closed cells, including the two base cells and $(a_1, *, a_3, *, a_5, *)$. □

(c) [6] How many **nonempty aggregated closed cells** are there in this data cube? What are they?

Answer: 1. There are 3 closed cells, including the two base cells and $(a_1, *, a_3, *, a_5, *)$. But only the latter one is an aggregated closed cell. □

(d) [6] How many **nonempty aggregated cells** are there in this data cube?

Answer: 118. For each base cell, there are $2^6 - 1$ aggregated cells. However, there are 2^3 cells that are counted twice since there are 3 common dimensions. Therefore, the total number of nonempty aggregated cells is $2 \cdot (2^6 - 1) - 2^3 = 118$. □

(e) [6] If we set **minimum support = 2**, how many **nonempty aggregated cells** are there in the corresponding **iceberg cube**?

Answer: 8. These two base cells have common value in 3 dimensions; therefore, there are 2^3 nonempty cells with support = 2 and all of them are aggregate cells. □

4. [40] Frequent pattern and association rule mining.
A data set shows 100 transactions in 5 days, each being summarized as a set of items associated with the number of transactions. Let *relative mininum support* to be $min\_sup = 0.5$ and *minimum confidence* to be $min\_conf = 0.6$. **Again, here we have 100 transactions, not just 5!!!**

date	items_bought	number of transactions
10/15	{a, b, c, m, p}	15
10/16	{b, e, f, p}	35
10/18	{a, c, k, p}	15
10/20	{b, e, p}	15
10/21	{a, e, g, p}	20

(a) [10] List the frequent 1-itemset associated with their absolute counts.

Answer: $p : 100, e : 70, b : 65, a : 50$ □

(b) [10] Draw the first frequent pattern tree (FP-tree) constructed and used in FP-Growth for the dataset. The tree is NOT for any conditional pattern base.

Answer: Any other trees will be ok if it is correct. □

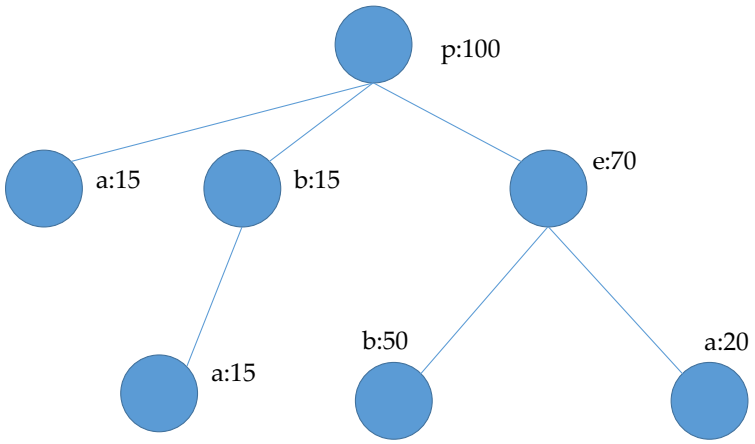

Figure 1: FP tree.

(c) [10] Present **all** the frequent k-itemsets for the **largest** k. Only list frequent itemsets of the largest size. The number of the largest frequent itemsets can be one, two, or many: Please list all of them.

Answer: $peb : 50$ □

(d) [10] Compute *relative* support and confidence on the following two rules. Are they good **association rules**? (Hint: compare with min_sup and min_conf.)

i. $pa \rightarrow b$, i.e., $\{p, a\} \rightarrow \{b\}$;
ii. $p \rightarrow e$, i.e., $\{p\} \rightarrow \{e\}$.

Answer: $pa \rightarrow b$ $(s : 15\%, c : 30\%)$, not an association rule;
$p \rightarrow e$ $(s : 70\%, c : 70\%)$, a good association rule. □