UCLA CS 249-80 Midterm Exam

Saturday, 1 May 2020

NAME: ________________________________

UID: ________________________________

Note

- This midterm exam consists of **6** pages, including **1** page of your name and your UID, and **6** pages of questions. All pages need to be returned at the end of the exam; otherwise, the answers will not be graded.
- For all questions ***including*** multiple-choice questions, please write down procedure and explanations for consideration of partial credits. There can be ***zero, one, or more than one*** correct answers for multiple-choice questions.
- You are ***not*** allowed to use or write any code in any form to solve any question. You must write down all your answers by hand, either on paper or on computer. You are allowed to use a calculator, but you must write down all the equations and computational steps by hand. If the use of any programming tool is detected, you will receive zero credit for that question.
- The exam is open note, but you must complete the exam by yourself. *It is imperative that you neither talk about nor reveal the contents or nature of the exam to anyone including social media, forums, or groups. In addition, if you gain knowledge of the exam from any source before taking it or during the exam, you are required to report the identity of the source and the extent of your knowledge to the instructor. Violators of these rules will be reported to the UCLA dean of students to open an investigation and to take appropriate disciplinary actions for the violation of the student conduct code.*
- A digital copy of your completed exam paper must be uploaded as a single file on CCLE. You can start any time between 9am and 9pm. Once started, the student will have 2 hours to finish the exam and upload their solution (So the latest finishing time is 11pm).

1. **Frequent Pattern Mining (50%)**
Consider the following transaction database and the profit list. Assume that the threshold `min_support` = 3.

TID	Items
T1	a,c,d,e,f
T2	c,d,e,f,g
T3	e,f,g
T4	b,c,f,g
T5	a,d,e,f,g

Item	a	b	c	d	e	f	g
Profit	10	5	40	30	-20	0	-10

(a) (10%) Find all the closed frequent itemsets and maximum frequent itemsets.

(b) (10%) Construct the FP-tree. What is the g-projected database?

(c) (10%) Which of the following pattern-space property (properties) does constraint *sum(profit)* > 40 have?

- (A) Succinct
- (B) Anti-monotonic
- (C) Convertible monotone
- (D) Strongly convertible
- (E) Monotonic

(d) (6%) What transaction(s) will be pruned in the d-projected database by the following constraint: *range(profit)* > 30?

(e) (6%) Now consider the PrefixSpan algorithm for sequential frequent pattern mining. What is the suffix generated by projecting $\langle a(ab)a(ab)a\rangle$ on $\langle aaa\rangle$? What is the suffix generated by projecting $\langle a(ab)a(ab)a\rangle$ on $\langle aaaa\rangle$?

(f) (8%) Suppose during the execution of the GSP algorithm on a sequence database, we find $\langle (bd)b(ba)d\rangle$ as the only length-6 frequent pattern. Can we claim there must be at least 6 length-5 frequent patterns found in the previous scan? Explain why.

2. **Clustering (50%)**
Consider the following two-dimensional points.

	x	y
p0	4	3
p1	5	1
p2	6	1
p3	7	3
p4	7	2
p5	6	4
p6	5	4
p7	4	2

(a) (10%) Simulate the K-means algorithm with 2 clusters for 2 iterations. Show the result after each iteration as a table of cluster assignment for each point. Pick p0 and p5 as your initial centroids. If there is a tie in cluster assignment, you can break the tie arbitrarily.

(b) (10%) Suppose the PAM algorithm is applied to the dataset with p0 and p5 as the initial medoids. At the end of the first iteration, suppose we would like to choose p1 and try swapping with p0. Would the swapping bring benefit?

(c) (10%) Which of the following statements about K-Means is (are) correct?

(A) K-Means sometimes cannot find the global optimal clustering.

(B) K-Means automatically determines the number of clusters.

(C) K-Means sometimes cannot converge.

(D) K-Means is sensitive to outliers but robust to data points with different densities.

(E) K-Means can deal with categorical features.

(d) (10%) Now suppose the DBSCAN algorithm is applied to the dataset. What of the following settings will make p1 a core point, but not density reachable from p7?

(A) Eps = 1, MinPts = 1

(B) Eps = 2, MinPts = 2

(C) Eps = 3, MinPts = 3

(D) Eps = 4, MinPts = 4

(e) (10%) Which of the following statements about clustering algorithms is (are) correct?

(A) The Manhattan distance between points $(-1, 2)$ and $(2, -1)$ is 5.

(B) Both K-Means and Agglomerative Hierarchical Clustering algorithms may suffer from convergence at local optima.

(C) Agglomerative Hierarchical Clustering and Divisive Hierarchical Clustering can have different time complexity.

(D) The K-Medoid algorithm is not suitable for clustering non-spherical (arbitrary shaped) groups of objects.

(E) The order of the data records inputted by the user affects the output of BIRCH.

(F) The order of the data records inputted by the user affects the output of OPTICS.

(G) If two points are density-connected, there exists a point p which is density-reachable from the two points.

(H) Grid-based methods for clustering include STING, CLIQUE, etc. whose results depend on the number of data objects.

3. **Bonus Question (10%)**
Consider the following sequence database. Assume that the threshold `min_support` = 2.

Sequence ID	Sequence
10	$\langle (acdf)bc(af) \rangle$
20	$\langle ae(efg)(df)cb \rangle$
30	$\langle fg(ac)d(cf) \rangle$
40	$\langle (ad)c(bc)(ae) \rangle$

Use the PrefixSpan algorithm to find all the sequential patterns having prefix $\langle d \rangle$.