Fall Semester 2019

KAIST EE412

Foundation of Big Data Analytics

Midterm

Name: _____

Student ID: _____

I agree to comply with the School of Electrical Engineering Honor Code.

Signature: _____

This exam is closed book. You may bring one sheet of paper containing notes. Read the questions carefully and focus your answers on what has been asked. You are allowed to ask the instructor/TAs for help only in understanding the questions, in case you find them not completely clear. Be concise and precise in your answers and state clearly any assumptions you may have made. You have 150 minutes (9:00 – 11:30 AM) to complete your exam. Be wise in managing your time. Good luck.

Question	Score
1	/10
2	/10
3	/10
4	/10
5	/10
6	/10
7	/10
Total	/70

1 (10 points) TRUE/FALSE

State if the following statements are true or false. Please write TRUE or FALSE in the space provided. Each problem is worth 1 point, but you will be penalized -1 point for an incorrect answer, so leave the answer blank if you don't know.

(a) In MapReduce, combiners push all the work of Reducers to the Map tasks.

Answer:_____

(b) In Spark, a function passed to a Reduce action is required to be associative and commutative.

Answer:_____

(c) An association rule can have a low confidence, but high interest.

Answer:_____

(d) When using the PCY algorithm, one can use the triangular-matrix method when at least 1/3 of the possible pairs appear in some basket.

Answer:_____

(e) The probability that the minhash values of two sets are the same is equal to the Jaccard distance.

Answer:_____

(f) Amplifying a $(d_1,\ d_2,\ p_1,\ p_2)$-sensitive family always results in a $(d_1,\ d_2,\ p_1',\ p_2')$-sensitive family such that $p_1' \geq p_1$ and $p_2' \leq p_2$.

Answer:_____

(g) The hierarchical clustering algorithm scales worse than the k-means algorithm on large datasets.

Answer:_____

(h) The GRGPF algorithm is designed to also work well on low-dimensional data.

Answer:_____

(i) The power iteration method finds the eigenpairs of any matrix of the form MM^T.

Answer:_____

(j) Compared to collaborative filtering, content-based systems are better at recommending items that few others are interested in.

Answer:_____

2 (10 points) MapReduce

For the following problems, describe how you would solve them using MapReduce by implementing the Map and Reduce functions. You may use pseudocode or explain the logic in words. Make sure you clearly specify the input and output of each function. If the problem cannot be solved with a single MapReduce pass, then describe how to solve it with multiple MapReduces where the output of each pass becomes the input of the next pass. However, we will penalize solutions that use passes unnecessarily.

(a) [4 points] Suppose we have as input a table with the schema (*address*, *zip*, *price*). Implement Map and Reduce functions such that the output is the average house price for each zip code.

(b) [6 points] Suppose we have as input the two tables R and S. R contains voter information as has the schema (*name*, *age*, *zip*). S contains disease information and has the schema (*age*, *zip*, *disease*). For each unique pair of age and zip values, output a list of names and a list of diseases of people with that age in that zip code. You may assume that any age-zip pair that occurs in one table also occurs in the other table.

3 (10 points) Frequent Itemsets

Suppose we want to find classes that are frequently taken by some KAIST students. The classes we consider are {CS101, CS206, EE201, EE209, EE210, EE211}, and the enrollment information is as follows:

Student	Enrolled classes
A	{CS101, CS206}
B	{CS206, EE209, EE210}
C	{CS101, CS206, EE201, EE210}
D	{EE201, EE210}
E	{CS206, EE210}
F	{CS101, CS206, EE210}
G	{EE209, EE210}
H	{CS101, CS206}
I	{EE209, EE211}

Given a support threshold of $s = 4$, answer the following questions.

(a) [1 point] What are the frequent singleton classes?

Answer:

(b) [3 points] Suppose we assign the following numerical values to classes:

Class	Value
CS101	1
CS206	2
EE201	3
EE209	4
EE210	5
EE211	6

Let us run the PCY algorithm using the hash function:

$$h(i, j) = (i + j) \bmod 8$$

that maps each pair of classes into one of eight buckets. Fill in the **bitmap** information below:

$$\text{bitmap} = \{\underset{0}{\_\_\_\_}, \underset{1}{\_\_\_\_}, \underset{2}{\_\_\_\_}, \underset{3}{\_\_\_\_}, \underset{4}{\_\_\_\_}, \underset{5}{\_\_\_\_}, \underset{6}{\_\_\_\_}, \underset{7}{\_\_\_\_}\}$$

(c) [2 points] Continuing from (b), what are the candidate pairs of classes in C_2?

Answer:

(d) [4 points] Now suppose we use the Multihash algorithm where, instead of using $h(i, j)$, we use the two hash functions below and two separate hash tables with 4 buckets each:

$$h_1(i, j) = (i + j) \bmod 4$$
$$h_2(i, j) = (i + 2 \times j) \bmod 4$$

When evaluating h_2, order the items so that $i < j$ to ensure that h_2 is symmetric.

What are the candidate pairs of classes in C_2?

Answer:

4 (10 points) Finding Similar Items

We briefly discussed in class that Uber uses minhashing and LSH in their application for finding similar driving routes. Each route is divided into smaller segments, and the idea is to find routes with many overlapping segments. Notice that this problem is more or less identical to finding common shingles among documents. Suppose we want to compute the minhash signature for two routes R_1 and R_2 using two pseudo-random permutation of columns using the following functions:

$$g_1(n) = 3n + 1 \bmod 7$$
$$g_2(n) = n - 1 \bmod 7$$

Here n is the segment number in the original ordering. We use the algorithm discussed in class where we sequentially read rows of the characteristic matrix and update the minhash values in the signature matrix. Complete the steps of the algorithm by filling in the blank spaces in the two matrices.

Segment	R_1	R_2	g_1	g_2
0	0	0		
1	1	0		
2	1	1		
3	0	0		
4	1	1		
5	0	1		
6	1	0		

Characteristic matrix

Hash fn	Sig(R_1)	Sig(R_2)
g_1		
g_2		

Signature matrix

5 (10 points) Clustering

(a) [6 points] Suppose we want to assign points to one of two cluster centroids, either $C_1 = (0, 0)$ or $C_2 = (100, 40)$. Depending on whether we use the L_1 or L_2 norms, a point (x, y) could be clustered with a different one of these two centroids.

Suppose (x, y) is (51, 15). When using each norm, which centroid is this point clustered with? Please choose C_1 or C_2.

Using the L_1 norm: _____

Using the L_2 norm: _____

Now suppose (x, y) is (55, 8) and answer the same questions:

Using the L_1 norm: _____

Using the L_2 norm: _____

(b) [4 points] We learned in class that a high-dimensional Euclidean space has a property called the curse of dimensionality. Without any information, we expect certain angles between the lines from a point Y to two other points X and Z to be approximately right angles. We denote the two lines as $X - Y - Z$ or simply XYZ. However, suppose we now have more information where the points A and B are in one cluster and points C, D, and E are in another cluster. The points in the same cluster are assumed to be close, while points in different clusters far away. Circle below the which of the following angles should **not** be approximately right angles.

ACB	ACE	ADC	ADE	AEB
BCD	BCE	BDC	BDE	CAE

6 (10 points) Dimensionality Reduction

Let us compute the SVD of a matrix $M = U\Sigma V^T$ as discussed in class.

$$M = \begin{bmatrix} 4 & 0 \\ 3 & -5 \end{bmatrix}$$

(a) [5 points] Derive V and Σ. For your answers, you may write expressions (e.g., $2+\sqrt{5}$) without computing the actual numbers. You must also show the derivation for full credit.

V :_____

Σ :_____

Derivation:

(b) [5 points] Derive U and Σ. Again, correct expressions are sufficient, and please show the derivation as well.

U :_____

Σ :_____

Derivation:

7 (10 points) Recommendation Systems

Suppose we have the following utility matrix of users A, B, C, and D rating the items a through e.

	a	b	c	d	e
A	4	5		5	1
B		3	4	3	1
C	2		1	3	
D	4	5	2	4	1

Let us cluster the items using hierarchical clustering. For the distance measure, we use the Jaccard distance between pairs of item ratings. To make the data more suitable for this measure, we round the ratings where we consider the ratings of 3, 4, and 5 as "1" and consider ratings 1 and 2 as unrated. We also take the distance between two clusters to be the minimum of the distances between any two items. When deciding which clusters to merge first, you may break ties arbitrarily. For your answer, show the complete grouping of the items using a tree/dendogram notation.

Answer:

[This can be used for scratch paper.]