Fall Semester 2021

KAIST EE412

Foundation of Big Data Analytics

Final Exam

Name: _____

Student ID: _____

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

Signature: _____

This exam is open book, notes, and computer. There are 20 pages in total. 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 assumption you may have made. You have 165 minutes (9:15 AM – 12:00 PM) to complete your exam. You may write your answers on a separate sheet (no need to copy the problems) or use this exam sheet. You may write in English or Korean. Upload your answers as a PDF file to KLMS by 12PM. **We will not accept PDFs after 12PM**, so make sure to leave enough time to generate and submit the PDF. Good luck.

Question	Score
1	/10
2	/8
3	/10
4	/8
5	/8
6	/10
7	/10
8	/14
9	/12
Total	/90

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) Spark is less likely to crash than MapReduce because it runs in memory.

Answer:_____

(b) The multistage algorithm for counting frequent pairs cannot use the triangular-matrix method.

Answer:_____

(c) The probability of minhash values being the same is equal to the Jaccard similarity of sets that are minhashed regardless of the hash functions.

Answer:_____

(d) When performing hierarchical clustering in a non-Euclidean space, we compare the centroids of clusters.

Answer:_____

(e) Performing dimensionality reduction using PCA or SVD gives the exact same result.

Answer:_____

(f) Collaborative filtering does not need to know item features to make recommendations.

Answer:_____

(g) Without taxation, a spider trap always drains all PageRank values to become 0.

Answer:_____

(h) Betweenness is used to find overlapping communities.

Answer:_____

(i) When using SVMs, setting the regularization parameter C to a large value reduces the margin.

Answer:_____

(j) One can delete elements from a Bloom filter without affecting other elements.

Answer:_____

2 (8 points) Spark

(a) [3 points] Suppose you are given the following Spark program along with its output. Please fill in *one missing line* so that the code prints the output. You may assume that the program already has imported the Python libraries and initialized the SparkContext. Any code that runs correctly will get full credit.

Spark program:

```
rdd1 = sc.parallelize([("1,2","a,b"), ("3,4","b,c"), ("5,6","d,e")])


rdd2 = _____________________________________________

print(rdd2.collect())
```

Output:

```
("1,2", "a"), ("1,2", "b"), ("3,4", "b"), ("3,4", "c"),
("5,6", "d"), ("5,6", "e")
```

(b) [5 points] Suppose you are given the following Spark program along with its output. Please fill in *three or fewer missing lines* so that the code prints the output. You may assume that the program already has imported the Python libraries and initialized the SparkContext. Any code that runs correctly will get full credit.

Spark program:

```
rdd1 = sc.parallelize(["a", "a", "b", "c", "d"])
rdd2 = sc.parallelize(["a", "a", "c", "d", "d"])
rdd3 = sc.parallelize(["b", "c", "d"])
rdd4 = rdd1.intersection(rdd2)

_____

_____

_____

print(rdd_final.collect())
```

Output:

```
("d", (0, 1)), ("c", (0, 1))
```

3 (10 points) Frequent Itemsets

Suppose there are 50 items numbered 1 to 50. Also say there are 50 baskets numbered 1 to 50. Item i is in basket b if and only if i divides b without a remainder. As a result, item 1 is in all baskets, item 2 is in the 25 even-number baskets, and so on. Basket 6 consists of the items (1, 2, 3, 6) as these are the integers that divide 6. Answer the following questions.

(a) [3 points] If the support threshold is 5, which *items* are frequent?

Answer:_____

(b) [4 points] If the support threshold is 5, which *pairs of items* are frequent? Make sure your answer is understandable.

Answer:

(c) [3 points] What is the sum of the sizes of all the baskets? You may write a mathematical expression (e.g., using $\sum$) or compute the sum exactly for full credit.

Answer:_____

4 (8 points) Locality Sensitive Hashing

Suppose we have an LSH family **F** of (d_1, d_2, 0.6, 0.4) hash functions.

(a) [2 points] Suppose we use two member functions from **F** and the AND-construction to form a (d_1, d_2, w, x) family. Calculate w and x. You may write a mathematical expression or compute the exact value for full credit.

w:_____

x:_____

(b) [2 points] Suppose we use two member functions from **F** and the OR-construction to form a (d_1, d_2, y, z) family. Calculate y and z. You may write a mathematical expression or compute the exact value for full credit.

y:_____

z:_____

(c) [4 points] Suppose we perform cascading where we do the OR-construction described in (b) and then the AND-construction described in (a). What is the fixedpoint p? You can either state the equation to compute the fixedpoint or compute the exact value for full credit.

Answer:_____

5 (8 points) Singular Value Decomposition

	M_1	M_2	M_3	M_4
U_1	1	1	0	0
U_2	2	2	0	0
U_3	0	0	3	0
U_4	0	0	0	4

Table 1: Ratings of movies by users

Suppose we have user ratings of movies as shown in Table 1. Each row represents a user's set of ratings, and each column represents a movie. The SVD decomposition of this utility matrix is computed as follows:

$$U\Sigma V^T = \begin{bmatrix} 0 & \frac{\sqrt{5}}{5} & 0 \\ 0 & \frac{2\sqrt{5}}{5} & 0 \\ 0 & 0 & 1 \\ 1 & 0 & 0 \end{bmatrix} \begin{bmatrix} 4 & 0 & 0 \\ 0 & \sqrt{10} & 0 \\ 0 & 0 & 3 \end{bmatrix} \begin{bmatrix} 0 & 0 & 0 & 1 \\ \frac{\sqrt{2}}{2} & \frac{\sqrt{2}}{2} & 0 & 0 \\ 0 & 0 & 1 & 0 \end{bmatrix}$$

Suppose the new user U_5 has the following reviews: 1 for M_1, 2 for M_3, and 3 for M_4. Thus, the respresentation of U_5 in the movie space is: [1 0 2 3].

(a) [2 points] What is the representation of U_5 in the concept space? You may write mathematical expressions or compute the exact values for full credit.

Answer:_____

(b) [3 points] Which user among U_1, U_2, U_3, and U_4 is the most similar to U_5 in terms of cosine distance in the concept space?

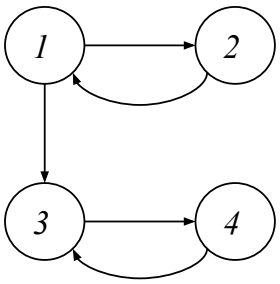
Answer:_____

(c) [3 points] Continuing from (b), calculate the cosine of the angle between the two users (i.e., U_5 and the closest user) in the concept space. You may write a mathematical expression or compute the exact value for full credit.

Answer:_____

6 (10 points) HITs Algorithm

Consider the link graph below.



Compute the hubbiness and authority scores using the HITs algorithm. First construct the link matrix L and do the following:

1. Start by assuming the hubbiness of each node is 1; that is, the vector $\mathbf{h}$ is the transpose of [1,1,1,1].
2. Compute an estimate of the authority vector $\mathbf{a} = L^T\mathbf{h}$.
3. Normalize $\mathbf{a}$ by dividing all values so the largest value is 1.
4. Compute an estimate of the hubbiness vector $\mathbf{h} = L\mathbf{a}$.
5. Normalize $\mathbf{h}$ by dividing all values so the largest value is 1.
6. Repeat Steps 2–5.

(a) [2 points] What is the link matrix L? Use the construction covered in class.

Answer:

(b) [4 points] After *two iterations* of Steps 2–5, what are the hubbiness scores of the four nodes?

Node 1:_____

Node 2:_____

Node 3:_____

Node 4:_____

(c) [4 points] Continuing from (b), what are the authority scores of the four nodes?

Node 1:_____

Node 2:_____

Node 3:_____

Node 4:_____

7 (10 points) Decision Trees

Suppose we are building a decision tree to decide whether a person is a vegetarian (V) or meat eater (M) based on his/her favorite food. A certain node of the tree is reached by six training examples as shown below where the examples are sorted by food name. Suppose we want to split these examples into two sets based on the food name.

Food	**V** or **M**
Apple	V
Beef	M
Duck	M
Grape	V
Juice	V
Pork	M

(a) [3 points] Compute the weighted GINI impurity when we split right after Apple. That is, Apple goes to the left child while the rest of the foods go to the right child.

Answer:_____

(b) [4 points] Where should we split to have the lowest weighted GINI impurity? If there are multiple places, please list all of them.

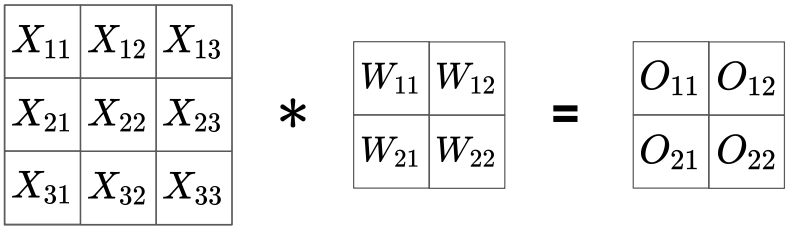
Answer: right after _____

(c) [3 points] Suppose we use the Accuracy impurity measure instead of Gini impurity, which was also covered in class. What is the lowest-possible weighted accuracy impurity?

Answer:_____

8 (14 points) Convolutional Neural Networks

(a) [3 points] Suppose we have a convolutional layer where the inputs, weights, and outputs are all variables as shown in the below figure. Here we use a stride of 1 and zero-padding of 0. Also, suppose we already computed the derivative of the loss function L with respect to O_{ij}. In other words, $\frac{\partial L}{\partial O_{ij}}$ is given, and you may use it in your answers.



Find the derivative of the loss function with respect to W_{11}, i.e., $\frac{\partial L}{\partial W_{11}}$.

Answer: _____

(b) [3 points] Continuing from (a), find the derivative of the loss function with respect to X_{22}, i.e., $\frac{\partial L}{\partial X_{22}}$.

Answer: _____

(c) [6 points] Suppose we have a convolutional neural network that consists of the layers in the table below. Fill in the output size, the number of weights, and the number of biases for each layer. For each output size, specify the dimensions, e.g., $3 \times 3 \times 10$.

Notation:

- CONV-F-N denotes a convolutional layer with N filters where each filter is of size $F \times F \times D$, and D is the number of channels in the input. We use a stride of 1 and zero-padding of 1, respectively.
- POOL-F denotes a $F \times F$ max-pooling layer with a stride of F.
- FC-N: denotes a fully-connected layer with N nodes

Layer	**Output size**	**Number of weights**	**Number of biases**
INPUT	$32 \times 32 \times 3$	0	0
CONV-3-4			
Leakly ReLU			
POOL-2			
FC-10			

(d) [2 points] Why is it important to use a non-linear activation function in a neural network instead of a linear function?

Answer:

9 (12 points) Mining Data Streams

Suppose we have a stream of positive integers as shown below. We assume each integer ranges from 0 to 7 and can thus be stored in 3 bits:

Stream: 1, 7, 3, 4, 4, 1, 4, 7

(a) [6 points] Let us extend the DGIM algorithm to estimate the sum of the last k integers for any $1 \leq k \leq N$ where N is the window size. For our stream, $N = 8$. We can treat the 3 bits of each integer as if they were from 3 separate bit streams. For example, the integer 4 is stored as 100, so the first bit stream has the value 1, the second stream has the value 0, and the third stream has the value 0. We then use the DGIM algorithm to count the 1's in each bit stream and combine the counts into the estimated sum of integers. What is the estimated sum for the last $k = 4$ integers? (The actual sum is 16.) Please explain your answer for any partial credit.

Answer:_____

(b) [3 points] Now suppose we use the AMS algorithm to estimate the fourth moment of this stream. What is the exact fourth moment of this stream?

Answer:_____

(c) [3 points] Continuing from (b), suppose we have variables for the 2nd and 4th positions of the stream. What is the average of the fourth moment estimates?

Answer:_____