Fall Semester 2019

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 closed book. You may bring two sheets 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 165 minutes (9:00 – 11:45 AM) to complete your exam. Be wise in managing your time. Good luck.

Question	Score
1	/10
2	/10
3	/15
4	/10
5	/15
6	/10
7	/10
8	/10
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) A directed graph that is strongly connected has no dead ends.

Answer:_____

(b) According to the affiliation-graph model, two people are always equally or more likely to be connected if they share more communities.

Answer:_____

(c) Given a graph of n nodes, a heavy-hitter triangle has a degree of at least $\sqrt{n}$.

Answer:_____

(d) A d-dimensional set of points can have d support vectors.

Answer:_____

(e) The SVM solution using gradient descent is always globally optimal.

Answer:_____

(f) The rectified linear unit (ReLU) does not have a saturation problem.

Answer:_____

(g) The Kullback-Leibler (KL) divergence is non-negative.

Answer:_____

(h) In a convolutional neural network, a pooling layer is a special case of a convolutional layer.

Answer:_____

(i) The DGIM algorithm can be extended to estimate the sum of negative integers.

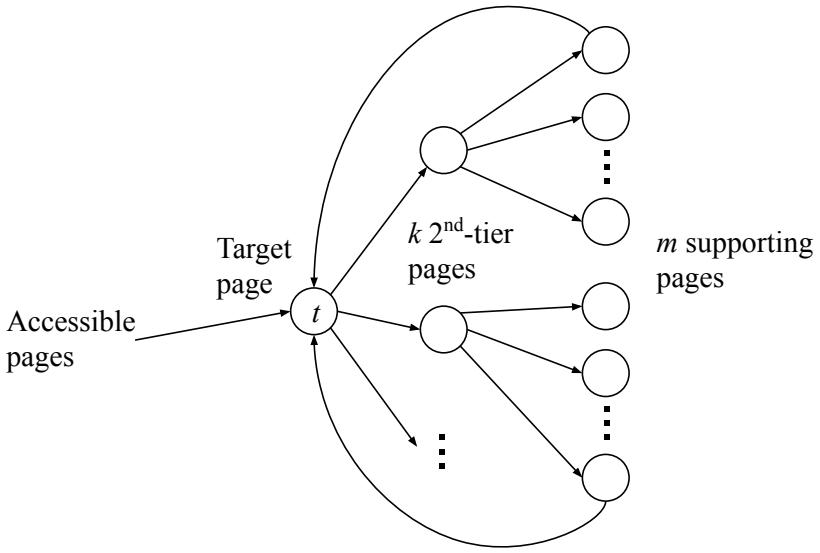
Answer:_____

(j) The competitive ratio of the Balance algorithm for 3 advertisers is lower than 3/4.

Answer:_____

2 (10 points) Link Analysis

The spam farm covered in class has a downside (for the spammer) that the target page needs to link to many supporting pages. Suppose the spammer wants to avoid this problem and designs the following spam farm instead:



The new spam farm has the following characteristics:

- The k second-tier nodes act as intermediate nodes.
- The target page t has one link to each of the k second-tier nodes.
- Each second-tier node has links to m/k nodes.
- Each supporting page only links to t (we only show two of the links in the figure to avoid clutter).

Also, we use the following parameters:

- n: the total number of pages in the Web.
- β: the taxation parameter.
- x: the amount of PageRank supplied from the outside (accessible pages) to t.
- y: the total PageRank of t.

Please answer the questions in the next page.

If we compute the formula for y in terms of x, k, m, and n, we get a formula of the form:

$$y = ax + bm/n + ck/n + d$$

Derive the expressions for a, b, c, and d. Use simple expressions to get full credit.

(2 points) a:_____

(3 points) b:_____

(3 points) c:_____

(2 points) d:_____

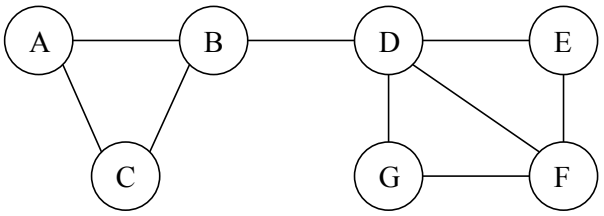
3 (15 points) Mining Social-Network Graphs

The edges of one graph G can be thought of as nodes of another graph G'. We construct G' from G using the following *dual construction*:

1. If (X, Y) is an edge of G, then XY, representing the unordered set of X and Y is a node of G'. Note that XY and YX represent the same node of G', not two different nodes.
2. If (X, Y) and (X, Z) are edges of G, then in G' there is an edge between XY and XZ. That is, nodes of G' have an edge between them if the edges of G that these nodes represent have a node (of G) in common.

Please answer the following questions:

(a) (3 points) Apply the dual construction to the following graph.



Dual construction graph G':

(b) (2 points) If we apply the dual construction to a network of friends G, what is the interpretation of the edges of the resulting graph G'?

(c) (4 points) How is the degree of a node XY in G' related to the degrees of X and Y in G?

(d) What we called the dual is not a true dual, because applying the construction to G' does not necessarily yield a graph isomorphic to G. Two graphs G and H are isomorphic if there exists an isomorphism (or equivalently, one-to-one and onto function) between the vertex sets of G and H.

$$f : V(G) \to V(H) \tag{1}$$

such that any two vertices u and v of G are adjacent in G if and only if $f(u)$ and $f(v)$ are adjacent in H.

Give an example graph G where the dual of G' is isomorphic to G and another example where the dual of G' is not isomorphic to G. Also briefly explain why.

- (3 points) Isomorphic example:

- (3 points) Non-isomorphic example:

4 (10 points) Large-scale Machine Learning

(a) Suppose we are learning using SVMs. The following training set obeys the rule that the positive examples all have vectors whose components sum to 10 or more, while the sum is less than 10 for the negative examples.

([2, 3, 7], +1)	([3, 4, 5], +1)	([4, 5, 6], +1)
([1, 2, 4], -1)	([3, 2, 1], -1)	([3, 2, 3], -1)

- (2 points) Which of these six vectors are the support vectors?

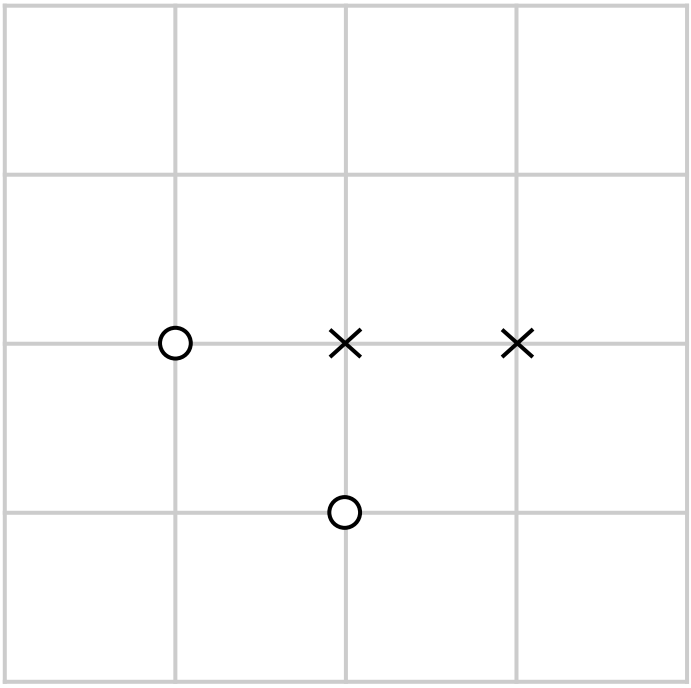
  Answer:

- (3 points) Suggest a vector $\mathbf{w}$ and constant b such that the hyperplane defined by $\mathbf{w.x} + b = 0$ is a good separator for the positive and negative examples. Make sure that the scale of $\mathbf{w}$ is such that all points are outside the margin; that is, for each training example (x, y), you have $y(\mathbf{w.x} + b) \geq +1$.

  Answer:

(b) Suppose we are learning using nearest neighbors where we look at the two nearest neighbors of a query point q. Say there are two labels O and X. Classify q with the common label if those two neighbors have the same label, and leave q unclassified if the labels of the two neighbors are different.

- (4 points) Sketch the boundaries of the regions for the following figure:

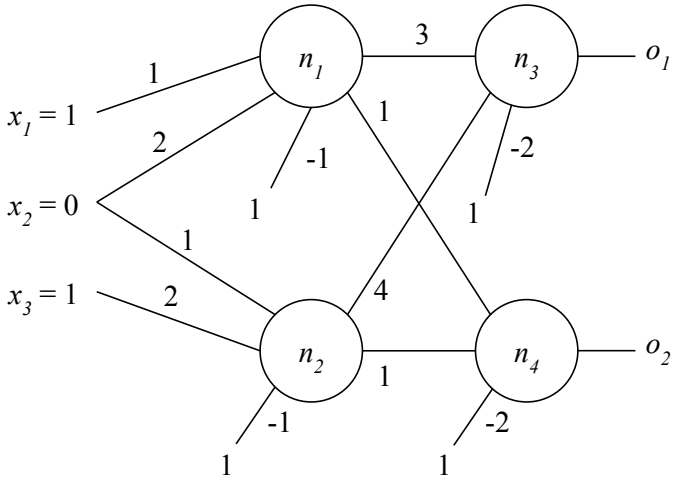


- (1 points) Would the boundaries always consist of straight line segments for any training data? Answer yes or no.

  Answer:

5 (15 points) Deep Learning I

Suppose we have the following neural network with particular values shown for all the weights and inputs. Say we use the sigmoid function to compute outputs of nodes at the hidden layer, and we use the softmax function to compute the outputs of nodes in the output layer.



Please answer the questions in the next pages.

(a) (8 points) Compute the values of the outputs for each of the four nodes. You may write expressions (e.g., e^2) without computing the actual numbers.

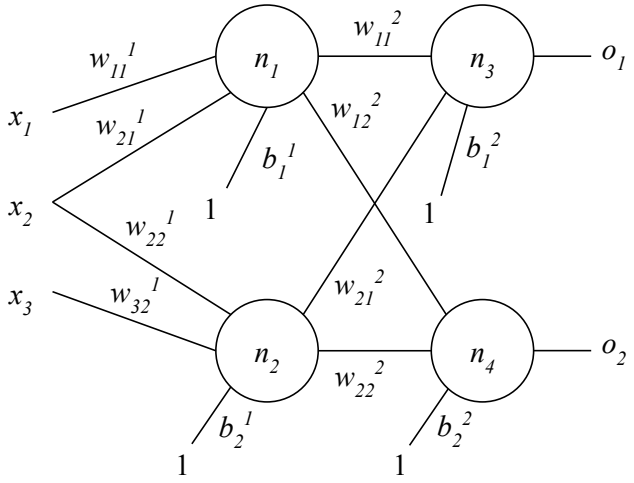
n_1:_____

n_2:_____

n_3:_____

n_4:_____

(b) (3 points) In this problem, let us assume that the inputs, weights, and biases are all variables as shown in the below figure. Express o_1 in terms of the variables. You may use intermediate variables in your expression.



Answer:

(c) (4 points) Continuing from (b), suppose we use cross entropy for the loss function, and the actual labels for the two outputs o_1 and o_2 are y_1 and y_2, respectively. Find the derivative of the loss function with respect to w_{11}^1. You may use intermediate variables in your expression.

Answer:

6 (10 points) Deep Learning II

(a) (4 points) Suppose we have a convolutional layer with a single filter:

1	0	-1
0	1	0
-1	0	1

We apply this filter to an array with the following values:

1	3	6	10
1	4	10	20
1	5	15	35
1	6	21	56

Assuming a zero padding of $p = 1$, apply the filter to the array to get another 4 x 4 array of values. Circle below the which of the following 8 values are in the responses.

-5 0 5 26 31 33 49 70

(b) (3 points) A layer of a convolutional neural network (CNN) has an array of inputs of size 50 x 50. It has 20 filters where each filter is of size 5 x 5. We use a stride of 1 and zero-padding of 2. How many output values does the layer have? You may assume that each filter produces only one output value.

Answer:

(c) (3 points) Suppose a pooling layer of a CNN has an input array of 100 x 100 pixels. The output consists of pools of size 4 x 4, and uses a stride of 3. How many output pixels are there?

Answer:

7 (10 points) Mining Data Streams

Suppose we have a stream of integers consisting of one 1's, two 2's, three 3's, and so on up to five 5's. That is, the stream is 122333444455555.

(a) (1 points) What is the first moment of this stream?

Answer:

(b) (4 points) Suppose we use the AMS algorithm to estimate the third moment of this stream where we have variables for the 3rd, 7th, and 12th positions of the stream. If we average the estimates given by each of these variables, what is the estimated third moment?

Answer:

(c) (3 points) Use the Flajolet-Martin algorithm to compute the estimated number of distinct values. Use the following two hash functions that both return 4-bit binary numbers. For example, $h_1(1) = 0010$.

- $h_1(x) = x + 1 \mod 16$
- $h_2(x) = 3x + 1 \mod 16$

Compute the average of the two estimates using h_1 and h_2.

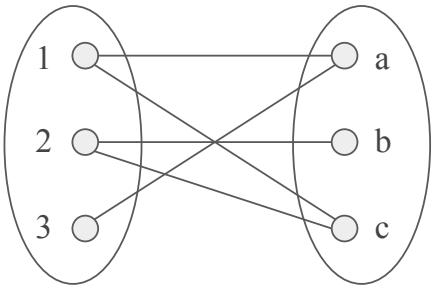
Answer:

(d) (2 points) In general, when using the Flajolet-Martin algorithm, is it problematic to count the number of leading zeros of a hash value instead of using the tail length? Answer Yes or No and briefly explain why.

Answer:

8 (10 points) Advertising on the Web

(a) (4 points) Whether or not the greedy algorithm gives us a perfect matching for the graph of the below graph depends on the order in which we consider the edges. Of the 5! possible orders of the five edges, how many give us a perfect matching?



Answer:

(b) (3 points) Suppose that there are three advertisers, A, B, and C. There are three queries, x, y, and z. There is one ad shown for each query. Each advertiser has a budget of 2. Advertiser A bids only on x; B bids on x and y, while C bids on x, y, and z. All bids are either 0 or 1, and all click-through rates are the same. Note that on the query sequence $xxyyzz$, the optimum off-line algorithm would yield a revenue of 6, since all queries can be assigned.

Will a greedy algorithm assign at least 4 of these 6 queries? Answer Yes or No and briefly explain.

Answer:

(c) (3 points) Continuing from (b), find another sequence of queries of any length such that the greedy algorithm may end up assigning only half the queries that the optimum off-line algorithm assigns on that sequence. Briefly explain why.

Answer:

[This can be used for scratch paper.]