CS246 Final Exam

March 16, 2016 8:30AM - 11:30AM

Name: _____

SUID: _____

I acknowledge and accept the Stanford Honor Code. I have neither given nor received unpermitted help on this examination.

(signed)_____

Directions: The exam is open book, open notes. Any inanimate materials may be used, including laptops or other computing devices. Access to the Internet is permitted. However, you must not communicate with any person. Answer all 24 questions in the spaces provided.

The total number of points is 180 (i.e., one point per minute of exam).

Numerical answers may be left as fractions, as decimals to an appropriate number of places, or as radicals, e.g., $\sqrt{2}$.

If you feel the need to make explanations, please do so on the reverse of the page, and indicate on the front that you are providing such an explanation (which we will read only in cases where there is apparent ambiguity).

Question 1 (6 points): Relation R(A,B) consists of the following four tuples: {(1,10), (2,10), (3,11), (4,10)}. Relation S(B,C) consists of the following five tuples: {(10,20), (11,21), (12,22), (10,23), (11,24)}. If we take the join of R and S using the MapReduce algorithm described in the text:

(a) How many key-value pairs are generated by all the mappers together? _____

(b) What is the minimum possible reducer size needed to execute this algorithm on this data? _____

(c) How many output tuples are produced?_____

Question 2 (6 points): Suppose a training set S with four output values has 80 examples with output A, 40 examples with output B, 20 examples with output C, and 20 examples with output D. Compute the following impurity measures for S:

(a) Accuracy _____

(b) GINI _____

(c) Entropy _____

Question 3 (6 points): In an attempt to speed up the computation of the transitive closure of a directed graph, we use the following recursion. **Basis**: at round 0: Path(U,V) = Arc(U,V); that is, we initialize the Path relation to be the same as the Arc relation. **Induction**: At subsequent rounds, we execute Path(U,V) += Path(U,X) o Path(X,Y) o Path(Y,V); that is, if there are paths from U to some node X, from X to some node Y, and from Y to V, then we add to the Path relation the fact that there is a path from U to V. Suppose we apply this algorithm to a directed graph that is a straight line with 8 nodes: 0→1→2→3→4→5→6→7. That is, there is an arc from node i to node i+1 for i = 0, 1,..., 6 and no other arcs.

(a) On the first round of the recursion, what new Path facts are discovered? You can just write (a,b) for Path(a,b).

_____

(b) On the second round of the recursion, what new Path facts are discovered?

_____

(c) Unfortunately, this algorithm does not correctly produce the transitive closure of this graph. Give an example of a path fact that is never discovered on any round. _____

Question 4 (10 points): Note: in this problem, to avoid your having to deal with non-round numbers, we'll accept anything that is within 1% of the true answer. Suppose we have 10,000 items and as many baskets as are needed so that (i) each basket has two items, and (ii) every pair of items is in exactly one basket. The support threshold is 1000. We wish to run the PCY algorithm on this data. On the first round, we use a hash function that distributes pairs into B buckets, randomly. We may assume that exactly half the buckets will receive more than an average number of pairs, and no bucket receives significantly more or less than the average.

(a) How many frequent items are there? _____

(b) What is the smallest possible value of B such that half the buckets are not frequent? _____

(c) Assume that on the second pass, we need just as much space to represent the frequent items as we did to count them on the first pass. Also assume that (i) buckets in both passes are 4-byte integers, (ii) B is as given in your answer to part (b), and (iii) main memory is no larger than what is needed on the first pass to hold the item counts and the buckets. How many bytes are available on the second pass to store counts of pairs? _____

(d) Would you store pair counts on the second pass using a table or a triangular array? _____
Why? _____

(e) Suppose that instead of PCY, we use the three-pass multistage algorithm. For the second pass, we use a new hash function that satisfies the same randomness assumption as for the first hash function. Then after the second pass of Multistage, how many candidate pairs would you expect to be generated (i.e., pairs that have to be counted on the third pass)? _____

Question 5 (6 points): In this problem, we are clustering words (character strings) using edit distance (allowing insertions and deletions only). Currently a cluster consists of the five words `he`, `she`, `her`, `they`, and `them`.

(a) If we define the clustroid to be the member with the smallest maximum distance to any member of the cluster, which word is the clustroid? _____

(b) What is the maximum distance between this clustroid and any other word? _____

(c) If we define the cohesion of the cluster to be the maximum distance between any two members of the cluster, what is the cohesion of the cluster? _____

Question 6 (10 points): Let our training set consist of positive points (1,1) and (2,3) and negative points (0,2) and (2,4), as:

(2,4) **-**
(2,3) **+**
(0,2) **-**
(1,1) **+**

(a) If we construct a perceptron to separate the positive and negative points, the separating hyperplane (a line in this two-dimensional case) can have a slope within some range. What are the lowest _____ and highest _____ possible slopes?

(b) Suppose we instead construct a support-vector machine to separate these points, using the positive point (2,3) and the negative points (0,2) and (2,4) as the support vectors. The separating hyperplane is defined by equation **w**.**x**+b=0. We want to maximize the margin, which means minimizing the length of vector **w** subject to the constraints that **w**.**x**+b$\geq$+1 for all positive points **x** and **w**.**x**+b$\leq$-1 for all negative points x. What are the values of **w** _____ and b _____? What is the margin? _____

Question 7 (6 points): This question concerns a MapReduce algorithm for duplicate elimination. The input is a set of key-value pairs, where the key is irrelevant. Our output is one copy of each value that appears one or more times in the input. The key for outputs is also irrelevant. Design the Map and Reduce functions for such an algorithm, and then answer the following questions about what your algorithm does. You may use * for an irrelevant key or value.

(a) Given an input (k,v), where v is the value and k is the (irrelevant) key, what key-value pair(s) does your Map function produce?

_____

(b) What does the input to the reducer for one intermediate key (output by the Map function) look like?

_____

(c) What output does your Reduce function produce from an input such as your answer to part (b)?

_____

Question 8 (6 points): Suppose there are six items A, B, C, D, E, and F. In a run of Toivonen's Algorithm, we find that in the sample, the following sets are the maximal frequent itemsets (that is, they and all their subsets are frequent, while no other itemsets are frequent in the sample): ABC, AD, BD, CD, and E. What is the negative border for these sets?

_____

Question 9 (6 points): This question concerns the minhash values for a column that contains six rows. Four of those rows hold 0 and the other two rows hold 1. There are 6! = 720 permutations of these six rows. When we choose a permutation of the rows and produce a minhash value for the column, we will use the number of the row, **in the permuted order**, that is the first in the permuted order to have a 1.

(a) For how many of the 720 permutations is the minhash value for the column 6? _____

(b) For how many of the 720 permutations is the minhash value for the column 5? _____

(c) For how many of the 720 permutations is the minhash value for the column 3? _____

Question 10 (6 points): Consider an undirected graph of 10 nodes named 0 through 9,arranged in a line, as 0-1-2-3-4-5-6-7-8-9. That is, there is an edge (i, i+1) for i = 0, 1,..., 8, and no other edges.

(a) What is the maximum betweenness of any edge? _____

(b) Which edge(s) has/have this betweenness? _____

(c) What is the minimum betweenness of any edge? _____

(d) Which edge(s) has/have this betweenness? _____

(e) Now, consider another graph with 10 nodes 0, 1,..., 9. This graph is complete; i.e., there is an edge between every pair of nodes. What is the betweenness of the edge (3,7)? _____

Question 11 (6 points): Suppose we are executing the DGIM algorithm for approximate counting of bits in a stream. Let the window size be 1000.

(a) What is the largest possible bucket size in the representation of this stream? _____

(b) Suppose each of the last 1000 bits is 1. What is the smallest possible size of the largest bucket in the representation of the current window? _____

(c) Describe a sequence of bucket sizes that justifies your answer to (b).
_____

Question 12: (8 points): A stream of integers consists of one 1, two 2's, three 3's, and so on, up to ten 10's.

(a) What is the zeroth moment of this stream? _____

(b) Suppose we apply the Flajolet-Martin algorithm with a single hash function h, to estimate the number of different elements in this stream. h(i) is simply i written as a 32-bit binary number (e.g., h(1) = 00...001, h(2) = 00...010). What estimate does h give as the number of distinct elements in this stream? _____

(c) What is the 2nd moment (surprise number) of this stream? _____

(d) Suppose we use the AMS method to estimate the second moment, and we have one variable for each position of the stream. If we average the estimates given by each of these variables, what is the resulting estimate of the second moment? _____

Question 13 (6 points): There are three advertisers A, B, and C, who bid on some of search queries P, Q, and R. Each advertiser has a budget of 2 search queries. A bids on P and Q only; B bids on Q and R only, and C bids on P and R only. We assign queries using the Balance Algorithm. When there are ties to be broken, use alphabetic order. That is, A is preferred to B or C in case of a tie in the remaining budgets, and B is preferred to C if they are tied. In what follows, represent the allocation of queries by a list of the advertiser given each query, in order, and use a dash (-) if no bidder gets a query. For example, if A gets the first query, B the second, and no one gets the third, then write AB-.

(a) Suppose six queries arrive in the order PQRPQR. what is the assignment of these queries to bidders according to the Balance Algorithm (ties broken alphabetically)? _____

(b) Give an example of a sequence in which the same six queries (two of each) could arrive, such that the Balance Algorithm fails to assign every query to an advertiser. _____ What is the sequence of assignments for this sequence of queries? _____
Hint: End your sequence with two of the same query.

(c) For your query sequence in (b), give an optimal assignment of these queries to advertisers. _____

Question 14 (8 points): Suppose we have a (.3, .7, .7, .3)-sensitive family *H*.

(a) Apply the (2, 2) AND-OR construction to H. The result is a (__, __, __, __)-sensitive family.
(b) Apply the (2, 2) OR-AND construction to H. The result is a (__, __, __, __)-sensitive family.

Question 15 (8 points):

(a) An association rule can have several items on the right as well as on the left. The *confidence* of an association rule S=>T is the fraction of baskets containing all the items in set S that also contain **all** the items in set T. Suppose we have 7 items numbered as 1,2,...,7. Which of the following association rules has a confidence that is certain to be at least as great as the confidence of 1,2 => 3,4,5,6,7 and no greater than the confidence of 1,2,3,4 => 5? Circle all and only those that do.

(i) 1,2,4 => 3,6,7 (ii) 1,2,3 => 4,5,7

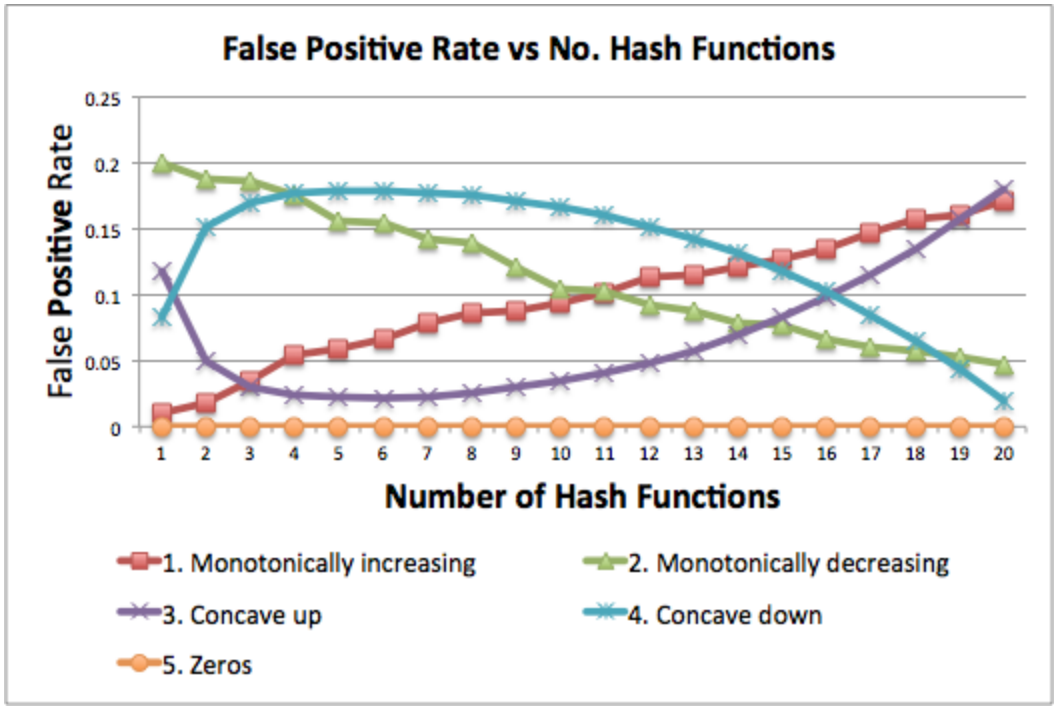
(iii) 2,3,4 => 1,5,7 (iv) 1,2,4 => 3,5,6

(b) Suppose ABC is a frequent itemset and BCDE is NOT a frequent itemset. Given this information, we can be sure that certain other itemsets are frequent and sure that certain itemsets are NOT frequent. Others may be either frequent or not. Circle all and only the correct statements.

(i) ACD is certainly not frequent. (ii) ABCD can be either frequent or not frequent.

(iii) BCE is certainly not frequent (iv) BC is certainly frequent

Question 16 (4 points): For this question, you are given five possible functions, below:



(a) Which of the following line best represents the Bloom Filter's false **positive** rate? (Circle the correct answer.)

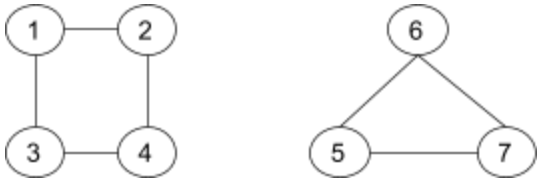
1. Monotonically increasing
2. Monotonically decreasing
3. Concave up
4. Concave down
5. Zeros

(b) Which of the following line best represents the Bloom Filter's false **negative** rate? (Circle the correct answer.)

1. Monotonically increasing
2. Monotonically decreasing
3. Concave up
4. Concave down
5. Zeros

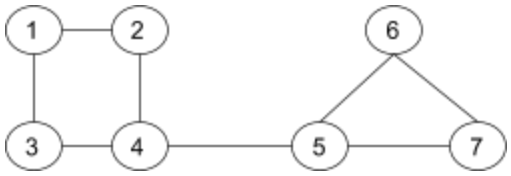
Question 17 (12 points):
For the following networks, define the most likely AGM that explains the data.
Note: You do not need to prove that your solution is the most likely model. Just "guess" the parameters and state them.

(a) State the most likely AGM with 2 communities:



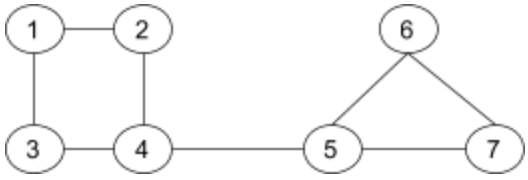
A = {_____}
B = {_____}
p(A) = _____
p(B) = _____

(b) State the most likely AGM with 3 communities:



A = {_____}
B = {_____}
C = {_____}
p(A) = _____
p(B) = _____
p(C) = _____

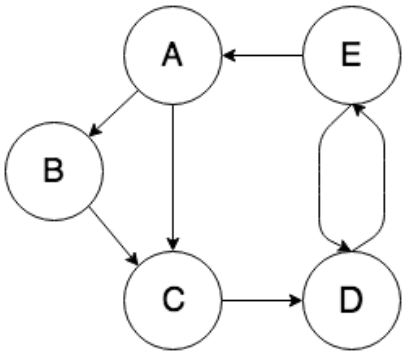
(c) State the most likely AGM *without any restriction on the number of communities*:



Describe all the communities as node sets (A,B, …) and state the respective edge probabilities (p(A), p(B),…).

_____
_____
_____
_____

Question 18 (10 points):



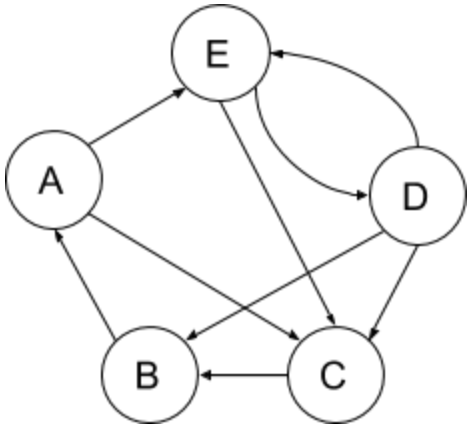
(a) Construct below the transition matrix for the network.

	A	B	C	D	E
A					
B					
C					
D					
E					

(b) Assuming no taxation (i.e., $\beta = 1$), and assuming we start with the PageRank vector $[1,1,1,1,1]^T$ for nodes A, B, C, D, and E in that order, give the PageRank vector at the next iteration (write in transposed form; no "T" needed): _____ and at the iteration after that: _____

(c) Now consider the teleport set {A, B} with teleport probability $1-\beta = 0.2$ For which nodes will the PageRank increase? _____

Question 19 (6 points): Consider the graph below:



In the HITS algorithm, let the initial hubbiness score of each node be 1. Calculate the authority and hubbiness scores for next three steps by filling in the following table.
Assume NO scaling is done here. i.e. $\lambda = \mu = 1$.

	A	**B**	**C**	**D**	**E**
hubbiness	1	1	1	1	1
authority					
hubbiness					
authority					

Question 20 (12 points): Answer each statement True (T) or False (F). Scoring: +2 for a correct answer, -2 for an incorrect answer, 0 for no answer; minimum total score is 0.

1. 1 is an eigenvalue of a column-stochastic matrix (just called "stochastic" in the slides on PageRank). _____

2. $f(S) = |S|^2$ (where $|S|$ is the number of elements in set S) is a submodular function. _____

3. Let M be an m-by-n matrix and E be an n-by-n orthonormal matrix. Then $||ME||_F = ||M||_F$ (F is the Frobenius norm). _____

4. The singular values of a matrix are unique (except for 0's). _____

5. If M is sparse, the SVD decomposition of M is guaranteed to be sparse as well. _____

6. Let M and N be column-stochastic matrices. MN Is also column stochastic. _____

Question 21 (12 points): Consider a modified version of the Bloom filter implemented as follows:

- Initialize an m-bit bitmap B with each bit set to 1 with probability p and 0 otherwise
- Choose N independent hash functions $h_1, h_2, ..., h_N$ and M independent hash functions $g_1, g_2, ..., g_M$, all of which hash uniformly to $\{1, 2, ..., m\}$
- For each x to be inserted, first set bits in B at positions $h_1(x), h_2(x), ..., h_N(x)$ to 1, then set bits at positions $g_1(x), g_2(x)..., g_M(x)$ to 0. Note that g may overwrite h for the same x.
- For each y to be checked for membership, return true if the bits in B at positions $h_1(y), h_2.(y), ..., h_N(y)$ are all 1 and the bits at positions $g_1(y), g_2(y), ..., g_M(y)$ are all 0. Return false if any bit is incorrect.

(a) In terms of M and N, what is the probability that a given bit is set to 1 when an element is inserted (regardless of whether it was 1 initially)?

_____

(b) In terms of p, M and N, what is the probability that a given bit is 1 **after the first** element is inserted?

_____

(c) For each of the following statements about the modified Bloom filter, answer True (T) or False (F). Scoring: +2 for a correct answer, -2 for an incorrect answer, 0 for no answer; minimum score on this part is 0.

(i) There can be false positives. _____

(ii) There can be false negatives. _____

(iii) The probability of at least one type of error (false positive or negative) increases as M increases. _____

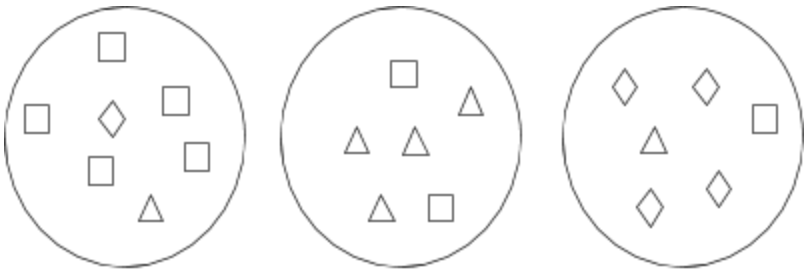
(iv) The probability of an error is independent of p. _____

Question 22 (8 points): Typically, we define "good clusters" as those that attain high intracluster similarity and low intercluster similarity. This is an internal criterion for measuring cluster quality. However, if we have a gold-standard clustering, we can use an external criterion to measure cluster quality by evaluating how well the clustering matches the gold standard. Given a gold-standard clustering of N data points into classes $C = \{C_1, C_2, \ldots, C_G\}$, and another clustering $L = \{L_1, L_2, \ldots, L_K\}$ of the same N data points, we define "purity" as follows: $\text{Purity}(C,L) = \Sigma_{1 \le k \le K} \max_{1 \le g \le G} |L_k \cap C_g|/N$. Answer the following questions:

(a) What is the maximum value of Purity(C,L), allowing any values of N, K, and G, any gold-standard clustering C (into G classes of size totaling N), and any clustering L (into K classes of size totaling N)? _____

(b) For each N and K, what is the minimum possible value of Purity(C,L), over all possible values of G, all possible gold-standard clusterings C (into G classes of size totaling N), and all possible clusterings L (into K classes of size totaling N)? _____

(c) Consider the clustering of a dataset of 19 data points consisting of 3 gold-standard classes (square, triangle, diamond) into 3 clusters (L1, L2, L3) as given in the figure below. Calculate the value of Purity for this clustering (you can leave your answer as a fraction). _____ Does this value decrease/increase/remain unchanged if each data point is duplicated (along with its class labels)? _____



Question 23 (6 points): Consider the following utility matrix:

	User 1	User 2	User 3	User 4	User 5
Item 1	4	4	4	1	1
Item 2	3	1		4	
Item 3	4	2		2	3
Item 4		2	3		1
Item 5			1	4	3
Item 6	1	1			2

(a) Assume that we use the Pearson correlation coefficient as the similarity measure and that we predict a rating by averaging the two nearest (most similar) neighbors. Which two users do we use to predict the rating of Item 4 by User 1: _____

(b) What is this predicted rating? _____

(c) What is the correlation coefficient for each? _____

Question 24 (6 points**)**: Suppose we have the following training set: ([1,2], +1), ([3,4], -1), ([3,1], -1), ([4,3], +1), and let us assume there is a new query point (3.5, 1.5) for which we are trying to obtain the class label. The distance between any two data points should be computed using the Euclidean distance metric. Given a query point x and its two nearest neighbors (x_1,y_1) and (x_2,y_2), the weighted average of the labels is computed as:
$[y_1/d(x_1,x) + y_2/(d(x_2,x)] / [1/d(x,x_1) + 1/d(x,x_2)]$, where d() is the Euclidean distance function. What would be the class label when the interpolation used is:

(a) The label of the nearest neighbor. _____

(b) The average of the labels of the two nearest neighbors. _____

(c) The average, weighted by distance, of the two nearest neighbors. _____

Note: when averaging labels of two nearest neighbors, you might end up with a noninteger value. Give us this value. In practice, all averages would be converted to +1 or -1, e.g., by taking the closest.