# CS246 Final Exam

March 21, 2017 3:30PM - 6:30PM

**Name**: _____

**SUID (digits)**: _____

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 22 questions in the spaces provided.

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

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).

Note: in some of these questions, it may appear that you are asked to perform some serious calculation, e.g., solving simultaneous linear equations. Be on the lookout for data that gives you a shortcut not available in the general case.

**Problem 1** (6 pts.) We wish to construct children for a node of a decision tree that receives the following 6 training-set examples. Each example is of the form (x,y), where x is a vector with two components called A and B. The criterion for impurity is GINI.

| **A** | **B** | **y** |
|---|---|---|
| 3 | Spades | +1 |
| 6 | Clubs | +1 |
| 9 | Diamonds | -1 |
| 2 | Clubs | +1 |
| 7 | Hearts | -1 |
| 5 | Spades | +1 |

a) We might want to split using a condition of the form A<i, where i is a particular integer. What value of i gives a split with the least average impurity of the children? _____
b) We might want to split using a condition of the form "B is in S," where S is a subset of {Clubs, Diamonds, Hearts, Spades}. Give a set S with the least average impurity of the children? _____
c) If we pick the split (based on either A or B) that gives the least average impurity, what is that impurity?_____

**Problem 2** (8 pts.) Suppose we have 12 minhash functions and apply LSH to the resulting signature matrix using b bands of r rows each. Let C1 and C2 be two columns of the original Boolean matrix (on which the minhash functions were computed), and let the Jaccard similarity of C1 and C2 be ½. You may assume that when we hash columns of the signature matrix, there are enough buckets that there are no accidental collisions, and columns go in the same bucket only if they are really identical in a band.

a) If r=3 and b=4, what is the probability of C1 and C2 NOT being a candidate pair; that is, the probability that C1 and C2 are never placed in the same bucket? _____
b) If r=4 and b=3, what is the probability of C1 and C2 NOT being a candidate pair? _____
c) Find the integer values of b and r such that we minimize the probability of C1 and C2 NOT being a candidate pair? b = _____; r = _____
d) Give the most important reason we would not want to choose these values of b and r?

_____

**Problem 3** (14 pts.) We have linearly separable data, and we wish to find a linear separator using SVM with positive support vectors (0,0) and (20,15) and negative support vector (20,0).

(20,15) +

(0,0) + (20,0) -

There are other members of the training set, but they are not important for this problem. We want to find the separating hyperplane (a line) based on these these three support vectors. The equation of the separating hyperplane can be written **w**.**x** + b = 0, while the upper and lower hyperplanes can be written **w**.**x** + b = +1 and **w**.**x** + b = -1, respectively.

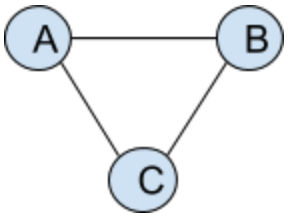
a) (6 pts.) What are the vector **w** _____ and the constant b _____?
b) What is the margin γ? _____
c) (6 pts.) Suppose we add a fourth point to the three in the diagram above. Several different things might happen, depending on the location of the point and its classification. Here are six possibilities:
   i) The data would no longer be separable without misclassified points.
   ii) The data would be separable without misclassified points, but only if one or more correctly classified points were between the new upper and lower hyperplanes.
   iii) The data would be separable without misclassified points, but the margin would be reduced.
   iv) The data would be separable without misclassified points, but the separating hyperplane would change.
   v) The data would be separable without misclassified points, and the margin would increase.
   vi) Nothing would change.

   For each of the following three points, indicate all those statements (i) - (vi) that are true when that point (and only that point) is added to the three above. Note that there may be more than one true statement; you must choose all true statements for each point.
   A negative point at (10,10). _____
   A positive point at (10,10). _____
   A negative point at (20, 5). _____

**Problem 4** (6 pts.) The hash function h(x) is computed by converting the integer x+5 to binary. We will use h as one of the hash functions in an application of the Flajolet-Martin algorithm, on a stream that happens to contain occurrences of the elements 1, 2, and 3, and no other elements.

a) What are all the different tail lengths computed when h is applied to the stream elements? _____
b) What is the estimate of the number of distinct elements provided by h? _____
c) Let g(x) be another hash function that produces the value x+a in binary, for some integer *a*. Give a value of *a* that minimizes the estimate of the number of distinct elements in the stream. _____ Note that there are many correct answers; pick any one.

**Problem 5** (10 pts.) For the triangle graph G:



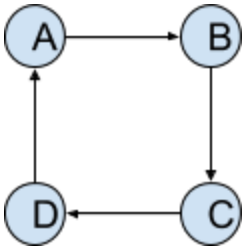

a) Write the Laplacian matrix of G in the space below:

b) What is the smallest eigenvalue of the Laplacian matrix of G? _____
c) What is the second-smallest eigenvalue of the Laplacian matrix of G? _____
d) There are several eigenvectors associated with the second-smallest eigenvalue. Give one of them (transposed to be a row). Make sure your vector is a unit vector, and the order of the elements corresponds to graph nodes A, B, and C in that order..
_____
e) What decomposition of G into two parts does your eigenvector suggest?
_____

**Problem 6** (6 pts.) Suppose we use a Bloom filter of N bits. We insert M different elements, and we use H independent hash functions (N is much larger than M or H).

(a) As a function of N, M, and H, write the expression that is the probability of a false positive (you should use the approximation that $(1- \varepsilon)^{1/\varepsilon}$ is $e^{-1}$ for small $\varepsilon$) .
_____
(b) Suppose N = 10 and M = 5. Either using your formula from (a) if you are certain it is correct, or reasoning directly for this special case, what is the probability of a false positive if H = 1? _____ In this and part (c), you may use sqrt(1/e) = .6066 and 1/e = .3679. An answer within 1% will be accepted as correct.
(c) If N = 10, M = 5, and H = 2, what is the probability of a false positive? _____

**Problem 7** (10 pts.) In the following problem, write all vectors (which are normally seen as columns) as row-vectors that are probability distributions (i.e., the sum of their components is 1). Also, their components should correspond to nodes A, B, C, and D, in this order. We are given the following Web graph:




a) Suppose the taxation rate is 50%, and only A is in the teleport set. What is the PageRank vector? _____

b) For the same taxation rate and teleport set, suppose that the teleport set represents the trusted set of pages; i.e., only A is trusted. What is the spam mass of node D? _____

c) Suppose we keep the taxation rate at 50% and A as the only member of the teleport set, but we delete the arc from D to A, and tax D 100% to account for the fact that it is now a dead-end, and we do not want to lose its PageRank. Which nodes have their PageRank **decrease**? _____

d) Now, consider the original Web graph, with the arc from D to A present, If we keep the taxation rate at 50% but change the teleport set to be {A, C}, which nodes have their PageRank **decrease**? _____

e) What is the hubbiness vector for this Web graph? (assume the largest component is 1)
_____

**Problem 8** (6 pts.) Suppose we have a large dataset of Sales tuples, of the form (Item, Buyer, Quantity, Date). We wish to take a 10% sample of the data and obtain an unbiased estimate of the following queries. For each, tell what the Key (attributes whose value determines the sample) should be when sampling.

a) On average, how many different Buyers bought each Item on each Date?
_____

b) On average, how many different Items does each Buyer buy?
_____

c) What fraction of the tuples are duplicates (i.e., there is at least one other identical tuple in the dataset)? _____

**Problem 9** (8 pts.) Here is a table of distances between four elements A, B, C, and D.

| | **A** | **B** | **C** | **D** |
|---|---|---|---|---|
| **A** | 0 | 2 | 3 | 4 |
| **B** | 2 | 0 | 1 | 5 |
| **C** | 3 | 1 | 0 | 6 |
| **D** | 4 | 5 | 6 | 0 |

a) Of the four triangles composed of three of these four elements, how many violate the triangle inequality? _____

b) If these four nodes form a cluster, and the clustroid is defined to be the node with the smallest average distance to the other nodes, which node is the clustroid? _____

c) If these four nodes form a cluster, and the clustroid is defined to be the node such that the square root of the sum of the squares of its distances to the other nodes is a minimum, which node is the clustroid? _____

d) Suppose now that these four nodes are in individual clusters, and we wish to cluster them. We always pick the clusters with the smallest distance between them, where the distance between two clusters X and Y is defined to be the minimum over all pairs of nodes x and y, with x in X and y in Y, of the distance between x and y. Draw below the dendrogram of the combination of the four clusters into a single cluster of four nodes.

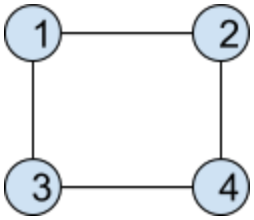
**Problem 10** (6 pts.) We apply Toivonen's Algorithm to a dataset with 10 items. In the sample, we find 7 of the items are frequent and the other three are not. We may also have found some other itemsets to be frequent in the sample.

a) What is the minimum number of singleton sets in the negative border? _____

b) What is the minimum number of pairs in the negative border? _____

c) What is the maximum number of pairs in the negative border? _____

**Problem 11** (10 pts.) We wish to find the Affiliation-Graph model of maximum likelihood for the following social graph, and now we conjecture that the best set of two communities are C = {1,2,3} and D = {3,4}.




Let the probabilities of "inspiring" an edge for the communities C and D be $p_C$ and $p_D$, respectively, and let ε be the probability of an edge that is part of no community.

a) In terms of $p_C$, $p_D$, and ε, write the expression for the likelihood of seeing the graph above. _____
b) Find the values of $p_C$ and $p_D$ that maximize the likelihood: $p_C$ = _____, $p_D$ = _____
c) What is this maximum likelihood? _____
d) Suppose we modify our guess regarding the communities to include node 2 in D (no change to C). Would the maximum likelihood of the graph
Increase Decrease Stay the Same (circle the correct answer).

**Problem 12** (6 pts.) In Smart Transitive Closure, round 0 initializes Q to be the arcs of the graph and P to consist of all and only the pairs (u,u), where u is a node (i.e., all the paths of length 0). On rounds 1, 2,… we:

1. First add to P the paths that are the composition of a path in Q followed by a path in the previous value of P.
2. Then compute Q to be those paths that are the composition of two paths in the previous value of Q, and subtract those paths in the value of P computed at step (1).

Suppose our directed graph is a straight line of length 8, that is, 0->1->2->...->7->8.

a) After round 1, how many pairs are in P? _____
b) After round 1, how many pairs are in Q? _____
c) After round 3, what are all the pairs in Q?
_____

**Problem 13** (6 pts.) Let the graph G have 2N nodes, where N is a large number, say N ≥ 100. The nodes of G are divided into two groups, the *left* and the *right*, each with N nodes. There is an edge between every left node and every right node. There is also an edge between every pair of left nodes, but no edge between any two right nodes. Each of the following questions has an answer that is an expression involving N and is based on the notion of "heavy hitter" used in the optimal algorithm for finding triangles.

a) How many heavy-hitter nodes are there? _____
b) How many heavy-hitter triangles are there? _____
c) How many other triangles are there? _____

**Problem 14** (8 pts.) Let the matrix M be:

$$M = \begin{matrix} 1 & 2 & 3 \\ 4 & 5 & 6 \\ 7 & 8 & 9 \end{matrix}$$

We want to decompose M into the product UV, where U has one column and three rows, while V has one row and three columns.

a) Suppose

$$U = \begin{matrix} 1 \\ 2 \\ 3 \end{matrix} \quad \text{and} \quad V = [1 \;\; 2 \;\; 3]$$

What is the product UV? (write your answer below)

b) What is the Frobenius norm (square root of the sum of the squares of the errors in each element) of the error matrix M - UV? _____

c) Suppose we allow the first element of V to vary. That is, replace V by [x 2 3]. What value of x minimizes the Frobenius norm of the error M - UV? _____

d) Suppose we want to perform a CUR decomposition of M. When selecting a row for the matrix R, what is the probability that we will pick the first row of M? _____

**Problem 15** (6 pts.) There are three advertisers, A, B, and C, and three search queries P, Q, and R. Each advertiser has a budget of 2 queries. A bids on all of P, Q, and R; B bids on only P and Q, and C bids only on P. Queries are allocated according to the Balance Algorithm, and when there is a tie (several bidders with the same remaining budget) you may allocate the query to any of them. You are asked to give sequences of queries and their assignment to advertisers with certain properties, and your answers should use the following notation. X->Y means query X is assigned to advertiser Y. If Y=*, then X is assigned to no advertiser. For example, P->A, Q->B, R->*, means first query P arrives and is assigned to A. Then query Q arrives and is assigned to B. Then query R arrives and is not assigned to any advertiser.

a) Give a sequence of two P's, two Q's, and two R's (in any order) and an assignment of queries to advertisers, following the Balance algorithm, that results in all six queries being assigned to advertisers.

_____

b) Give a sequence of two P's, two Q's, and two R's (in any order) and an assignment of queries to advertisers, following the Balance algorithm, that results in exactly five of the six queries being assigned to advertisers.

_____

c) Give a sequence of two P's, two Q's, and two R's (in any order) and an assignment of queries to advertisers, following the Balance algorithm, that results in the minimum possible number of queries being assigned to advertisers.

_____

**Problem 16** (10 pts.) We wish to multiply a matrix M with 10 rows and 20 columns, times a matrix N with 20 rows and 30 columns.

a) If we use the two-pass method described in the text, how many key-value pairs are generated (by all the mappers together) on the first pass? _____
b) Again using the two-pass method, on the first pass, what is the minimum reducer size for any key? _____
c) On the second pass of the two-pass method, how many different keys (reducers) are there? _____
d) If instead we use the one-pass method described in the text, how many key-value pairs are generated (by all the mappers together)? _____
e) Again using the one-pass method, what is the minimum reducer size for any key? _____

**Problem 17** (8 pts.) We wish to construct the sets of k-shingles, for some $k \geq 1$ for certain documents, with the goal of identifying documents (strings of letters) whose shingle sets have high Jaccard similarity.

a) Let $S_1$ = aaabbaaaa and $S_2$ = bbbabaaba. What is the smallest $k \geq 1$ such that the Jaccard similarity between the k-shingle representations of $S_1$ and $S_2$ is strictly less than 1? _____

b) Let $S_1$ = aaabaaaaa and $S_2$ = aaaaaaaaa. What is the largest $k \leq 9$ such that the Jaccard similarity between the k-shingle representations of $S_1$ and $S_2$ is strictly greater than 0? _____

c) Let $S_1$ = bbabbaaaa and $S_2$ = bbbabaaba. Let k = 3 and the universal set be the set of eight 3-shingles consisting only of *a*'s and *b*'s. What are the minhash values of $S_1$ and $S_2$ for the permutation corresponding to the alphabetical order, i.e., (aaa, aab, aba, abb, …)? For $S_1$: _____ For $S_2$: _____ (Use the shingle itself as the minhash value.)

d) What is the probability that the minhash values of $S_1$ and $S_2$ are equal when the permutation in part (c) is chosen uniformly from the 8! possibilities? _____

**Problem 18** (10 pts.) Rajan is running a research project with n people, and wants to determine the order of authorship. Since he does not have time to supervise all of the workers, he decided to implement a peer-assessment system, where each worker is given 1 unit of credit to divide among his fellow teammates in any way they want. At the end of the project, the distribution of credit looked like:

| | Alice | Bob | Carol | David |
|---|---|---|---|---|
| Alice | | 1/2 | 1/4 | 1/4 |
| Bob | 3/4 | | | 1/4 |
| Carol | | | | 1 |
| David | | | 1 | |

In this situation, Bob (for example) gives 3/4 of his credit to Alice, and 1/4 of his credit to David.

In order to assign authorship, Rajan builds an appropriate graph using the credit distribution information, and runs a modified version of PageRank on the graph. Specifically, we define the "share of authorship" $a_p$ for a participant p as

$$a_p = \Sigma_{\text{all participants q}}\, a_q * (\text{fraction of q's vote given to p})$$

All authorship shares are normalized so that

$$\Sigma_{\text{all participants p}}\, a_p = 1$$

a) Using the data in the table above, what are the modified PageRank equations relating the authorship shares of the four participants? Use A, B, C, and D for the authorship shares of Alice, Bob, Carol, and David, respectively. Write your equations below:

b) What is the solution to these equations? A = _____; B = _____; C = _____; D = _____

c) We might suppose we could solve the equations from (a) using relaxation (iterative approximation to the solution as for PageRank). However, for the given data, the iteration may or may not converge to the solution of the equations (b). Give an example of initial values of A, B, C, and D, summing to 1, such that iteration **does** converge to the solution (b). A = _____; B = _____; C = _____; D = _____

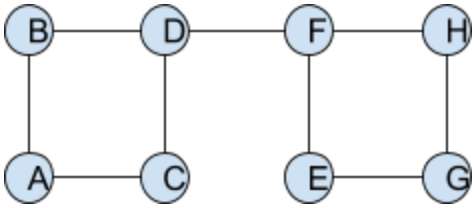
d) Now give an example of initial values for A, B, C, and D, summing to 1, such that iteration **does not** converge to the solution (b). A = ____; B = ____; C = ____; D = ____

(Continues on next page)

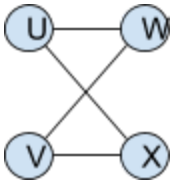
e) Bob is unhappy with his position in this order and asks Rajan to implement a taxation scheme, with a tax rate of 1-β. Now, the equations for the authorship shares change to $a_p = \beta \Sigma_{\text{all participants q}} a_q$ * (fraction of q's vote given to p) + (1-β)/(number of participants)
$\Sigma_{\text{all participants p}} a_p = 1$
where $0 \leq \beta \leq 1$. What value of β maximizes Bob's share of authorship? _____

(For more information about Rajan and his research group, you may visit http://crowdresearch.stanford.edu/ after the exam is over. When supervising the Stanford Crowd Research Collective, Rajan did in fact use this PageRank-like strategy to assign authorship credit.)

**Problem 19** (10 pts.) For the following social graph:




a) Calculate the betweenness of each of the following edges:
(D,F) _____ (B,D) _____ (A,B) _____




b) Now consider a complete bipartite graph with two nodes on each side, as above. Each edge has the same betweenness. What is that betweenness? _____
c) Next, generalize (b) to a complete bipartite graph with N nodes on each side. Write as a function of N, the betweenness of each edge. _____

**Problem 20** (8 pts.) Suppose we have 10,000 items, of which K are frequent. You may assume:

1) Each item is represented by a unique integer.
2) All integers require 4 bytes.
3) There are 1,000,000 baskets.
4) Each basket contains exactly 4 items.
5) The support threshold is s = 2000.

We run the PCY Algorithm on this data.

a) What is the maximum possible number K of frequent items? _____
b) Suppose that on pass 1 we use B = 5000 buckets. What is the minimum number of bytes needed for all the work done on pass 1 of the PCY Algorithm? _____
c) What is the maximum number of these B = 5000 buckets that can be frequent? _____
d) Having used 5000 buckets on pass 1, what is the maximum number of pairs we might have to count on pass 2? _____

**Problem 21** (12 pts.) Here is a matrix M:

| | |
|---|---|
| 1 | -1 |
| -1 | 1 |

a) (2 pts.) What is the rank of M? _____
b) (2 pts.) If we wish to find the Singular-Value Decomposition of M, we need first to find the eigenvalues and eigenvectors of another matrix A. Write the matrix A below:

c) (4 pts.) What are the eigenpairs for the matrix A? Write them as (e, [f,g]), where e is the eigenvalue and [f,g] is the eigenvector, transposed to be a row.
_____
d) (4 pts.) Give the exact decomposition of M as the product of $U\Sigma V^T$, using as few singular values as possible, in the space below.

**Problem 22** (6 pts.) Consider the table of ratings below.

| | Item 1 | Item 2 | Item 3 | Item 4 | Item 5 |
|---|---|---|---|---|---|
| User 1 | 5 | | 3 | | 4 |
| User 2 | 3 | 3 | 1 | 5 | 2 |
| User 3 | | 5 | 5 | 2 | |
| User 4 | 3 | 5 | 1 | | 1 |
| User 5 | | | 5 | 2 | 5 |
| User 6 | 5 | 3 | ? | 5 | |

a) You wish to predict the rating of Item 3 by User 6. Suppose we use item-item collaborative filtering and use Pearson correlation as the similarity metric. What is the Pearson correlation coefficient between Item 1 and Item 3? (In this and later parts of the problem, 2 decimal places are sufficient, or you may give exact expressions.)

_____

b) Which other Items could be used to help predict the rating of Item 3 by User 6?

_____

c) One problem with pure Collaborative Filtering is newcomers who have not rated anything. This is where the content-based approach comes in handy. Suppose we have a new user 7 with profile features [1,3,0,2,-2]. We also know that the profile features for Item 3 are [-1,0,2,2,0], what is the cosine similarity (you should give the value of the cosine itself, rather than the angle with that cosine) between User 7 and Item 3?

_____