CS246 Final Exam, Winter 2019

- Your Name: _____
- Your SUNetID (e.g., pirroh): _____
- Your SUID (e.g., 01234567): _____

I acknowledge and accept the Stanford Honor Code.

Signature: _____

1. These questions require thought, but do not require long answers. Please be as concise as possible.
2. Please write all answers in the space provided. You can use scratch paper for anything, but you may **not** attach any scratch paper with this exam.
3. The duration of this exam is **3 hours**.
4. This exam is open-book and open-notes. You may use notes (digitally created notes are allowed) and/or lecture slides and/or any reference material. You may **not** use the Internet during the exam.
5. Acceptable uses of computer:
   - You may **not** access the Internet or communicate with any other person.
   - You may **not** use your computer to write code, only to do arithmetic calculations.
   - You may only use features that would be present in a standard scientific calculator, such as addition, subtraction, multiplication, division, logarithms, exponents, etc.
   - You can use your computer as a calculator or an e-reader.
6. Numerical answers may be left as fractions, as decimals to an appropriate number of places or as radicals (e.g., $\sqrt{2}$).
7. There are **17** questions on this exam; the maximum score that you can obtain is **180** points.
8. Each question is worth **10** points with the exception of the last two questions, which are worth **15** points each.

1 Dimensionality Reduction [10 points]

Consider the following 3×3 matrix:

$$M = \begin{bmatrix} 3 & 1 & 0 \\ 1 & 3 & 0 \\ 0 & 0 & 1 \end{bmatrix}$$

The SVD of M is given as follows:

$$M = U\Sigma V^T = \begin{bmatrix} 1/\sqrt{2} & 1/\sqrt{2} & 0 \\ 1/\sqrt{2} & -1/\sqrt{2} & 0 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} 4 & 0 & 0 \\ 0 & 2 & 0 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} 1/\sqrt{2} & 1/\sqrt{2} & 0 \\ 1/\sqrt{2} & -1/\sqrt{2} & 0 \\ 0 & 0 & 1 \end{bmatrix}$$

(a) **Singular Values**: What are the singular values of M?

(b) **Low-rank Approximation**: We wish to obtain N, which is the **best possible rank 2 approximation** of M (in terms of reconstruction error based on the Frobenius norm). Remember that N must also be a 3×3 matrix. Calculate N and its singular value decomposition. **Note**: Frobenius norm of a matrix A is defined as follows: $||A_F|| = \sqrt{\sum(A_{ij}^2)}$

(c) **Reconstruction Error**: The reconstruction error between two matrices is defined as the Frobenius norm of their difference. What is the reconstruction error between M and N?

2 MapReduce [10 points]

In Gradiance quizzes and homework you have seen matrix-vector multiplication in MapReduce. This time you will work on a slightly different task. Given two sets

$$R = \{a, b, c\}, S = \{b, e, f\},$$

we want to find the difference of $R \setminus S$, i.e., the set of elements that exists in R but not in S. Design Map, Group by Key and Reduce functions to compute the set difference, and write your answers below in terms of a, b, c, e, f, R, S.

(a) What key-value pairs does your Map function produce for R, S?

(b) What does the Group by Key function produce?

(c) What does the Reduce function produce for the set $R \setminus S$?

3 Frequent Item Set Mining [10 points]

Suppose we are given some documents of different domains or topics, and we would like to categorize them according to the words in the documents. Specifically, we treat documents as baskets, and words as items. Your goal is to find the frequent item sets, and then categorize the item sets, so that documents can be assigned to different categories.

In this problem we focus on the frequent item set mining part. As a toy example, assume the whole item set is $S = \{$banana, apple, basket, friend, atmosphere, learning$\}$. (An example of document can be "The apple is in the basket".) And we have the following table of baskets and items:

Sentence index	words in the sentence and S
1	banana, apple, basket
2	basket, learning
3	apple, friend, learning
4	basket, friend, atmosphere
5	banana, friend, atmosphere, learning
6	basket, friend, atmosphere

Consider **support threshold** $s = 3$ in this case. Apply the A-priori algorithm to find the frequent item sets.

(a) Find the frequent items: $L_1 = \{$_____$\}$

(b) Then, construct the candidate pairs using items in L_1. The set of candidate pairs is:

$C_2 = \{$_____

_____$\}$.

(c) Filter on C_2 to obtain frequent item tuples: $L_2 = \{$_____$\}$

(d) Now suppose instead of A-priori, you are using the PCY algorithm to further optimize the process. During the first pass, you also hash each pair of items into a bucket, and maintain the count of pairs for each bucket.

For a bucket b with count c, what can you say about the pairs that hash to b if $c \geq s$? (Possible answers: They must be frequent, they cannot be frequent, or not sure.)

_____

What can you say if $c < s$? (They must be frequent, they cannot be frequent, or not sure.)

_____

How are these counts stored when doing the second pass? (one line answer is sufficient)

_____

4 Learning Through Experimentation [10 points]

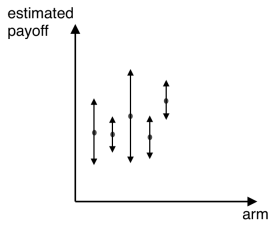


Figure 1: Estimated payoff with confidence interval plot.

(a) Consider Fig. 1 about a bandit with arms 1-5, reporting the current estimated payoff with 99% confidence intervals. Which arm will be chosen next based on pure exploitation? Which arm will be chosen next based on the UCB (Upper Confidence Bounds) algorithm?

(b) Suppose you are facing the following scenario where the estimated payoffs for all the arms are the same, and each arm is played for the same number of times. What action will be chosen based on ϵ-greedy algorithm when $\epsilon = 0.3$? What action will be chosen based on UCB?

(c) UCB-TUNED is a slight modification of the UCB algorithm, with the policy defined as below:

$$j(t) = \arg\max_a \hat{\mu_a} + \sqrt{\frac{\ln(t)}{m_a} \min\left(\frac{1}{4}, V(m_a)\right)}$$

where

$$V = \hat{\sigma_a^2} + \sqrt{\frac{2\ln(t)}{m_a}}$$

j(t) is the arm to be chosen at time t. $\hat{\mu_a}$ is our estimate of payoff of arm a, and m_a is the number of pulls of arm a so far. $\hat{\sigma_a^2}$ is the estimate of the variance of arm a.
Supposing you are facing the same scenario in (b), which arm will be chosen?

5 Machine Learning Memes for Decision Trees [10 points]

Feature selection in regression

Like most people, you are fond of memes. You decide to build a decision tree to predict the number of likes y_i for a given post i on the Stanford meme page. Your training dataset has 1000 posts. The variance of y_i in your training dataset is 500. You have been given a few candidate features and want to figure out which is the best one to split on. Let $|D_L|$ and $|D_R|$ represent the size of the left and right child datasets after splitting. Let $Var(L)$ and $Var(R)$ represent the variance of y_i in the left and right child datasets after splitting.

(a) Which of these features would you choose for splitting at the top level, and why?

- Word count: $|D_L| = 800$, $|D_R| = 200$, $Var(L) = 600$, $Var(R) = 100$
- Content related to machine learning: $|D_L| = 300$, $|D_R| = 700$, $Var(L) = 100$, $Var(R) = 600$
- Number of prior posts by user: $|D_L| = 400$, $|D_R| = 600$, $Var(L) = 100$, $Var(R) = 700$

Bagging: Now consider a smaller dataset D_1 with only 3 examples:

Word Count	ML Content	Prior Posts	Likes
5	Yes	4	500
10	Yes	6	400
15	No	8	100

Your friends Leland and Stanford want to use bagging to improve model performance. They use D_1 to generate synthetic datasets for bagging. Leland's dataset D_1' looks like this (table below):

Word Count	ML Content	Prior Posts	Likes
15	No	8	100
5	Yes	4	500
15	Yes	10	900

Stanford's dataset D_1'' looks like this (table below):

Word Count	ML Content	Prior Posts	Likes
15	No	8	100
5	Yes	4	500
15	No	8	100

(b) Which of the two has an incorrect implementation of bagging, and why?

6 Clustering [10 points]

(a) In this question, you will perform a simple K-means calculation.

Points	x	y
1	1.0	1.0
2	1.5	2.0
3	3.0	4.0
4	5.0	7.0
5	3.5	5.0
6	4.5	5.0
7	3.5	4.5

If we set $k = 2$, the initial centroids be **P1** and **P4**, and we are using Euclidean distance, what are the final outputs of K-means?

Centroids coordinates	members

(b) *True/False* question about K-means.
If we use **Euclidean distance**, the cost over iterations always decreases: _____

(c) Explain two different types of 2-dimensional data distributions where K-means might fail to produce accurate clusters. Provide sketches of the data in the 2D Euclidean space, and briefly explain.

7 Advertising [10 points]

During the lecture, you were given $\psi_i(q) = x_i(1 - e^{-f_i})$, where x_i is the bid and f_i is the fraction of left over budget for bidder i. In reality, sometime you would probably want to compute $\psi_i(q) = x_i CTR_i(1 - e^{-f_i})$, where CTR_i is the click through rate for bidder i.

(a) How do you interpret the new formula for ψ_i against the original form (short answer in 1 sentence)?

(b) Suppose you have the following table for 3 advitisers. From now on, use the new ψ formula as your metric to choose advertisers.

Advertisers	Bid	CTR	Budget	Spent so far
A	50	2%	1000	100
B	80	1.5%	2000	250
C	50	2.5%	1500	300

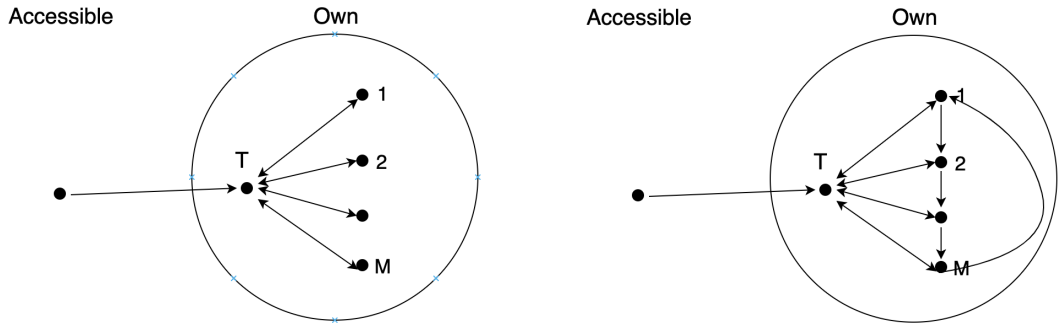
A new query targeted for all three advertiser arrived. Who is the winner?

(c) Following up from (b), the system received a new query targeted for all three advertiser. Who is the winner at this round?

(d) Give a reasonable bound for the algorithm used in this problem (answers like ≤ 1 will not be awarded). Why is this form preferable compared to the one given in the lecture?

8 Link Spam [10 points]

Consider two spam farm structures shown below. Spam farm 1 is the example you studied in class, and spam farm 2 is a modified version of the given example. For both structures, node T is the target page, and the spammer hopes to maximize its page rank y. The spammer owns M farm pages and has 1 accessible page with only one out link linking to T. The accessible page has page rank a. There are N pages in the entire web. The teleportation parameter is β. The only difference between the two structures is that in spam farm 2, each farm page has one additional out link pointing to another farm page, and these out links form a directed cycle as shown in the graph.



(a) Spam Farm 1 (b) Spam Farm 2

Figure 2: Two Spam Farms

(a) For spam farm 2, write down the PageRank equations to compute the PageRank score of the farm page (f) and the PageRank score of the target page (y) (in terms of f, y, a, β, M, and N)

(b) Which spam farm out of the two do you think will yield higher page rank for T? Please explain your reasoning.

9 Collaborative Filtering [10 points]

Consider the table of ratings below, which shows the ratings for three different books by five different readers. In this question, you want to figure out the rating of Reader1 for Book1 using item-item and user-user collaborative filtering methods. Notice that some of the ratings are unknown.

	Book1	Book2	Book3
Reader1	?	2.0	1.0
Reader2	3.0	1.0	
Reader3	1.0		
Reader4	2.0	1.0	
Reader5	0.0		3.0

(a) Use the user-user collaborative filtering method and the **cosine similarity** measure to calculate the rating of Reader1 for Book1 based on the **two** most similar readers to Reader1. Show your steps and highlight your final rating. **You do not need to subtract row mean to normalize the table**.

(b) Use the item-item collaborative filtering method and the **cosine similarity** measure to calculate the rating of Reader1 for Book1 based on the **most** similar book to Book1. Show your steps and highlight your final rating. **You do not need to subtract row mean to normalize the table**.

10 Latent Factors [10 points]

In this question, rather than using collaborative filtering, you will use a basic latent factor model for making recommendations. For this model, the predicted rating $\hat{r}_{xi}$ for user x on item i is computed as follows:

$$\hat{r}_{xi} = q_i \cdot p_x$$

where q_i is row i of matrix Q, and p_x is row x of matrix P. Consider the incomplete ratings matrix R in the table below, where r_{xi} is the true rating for user x on item i, along with the partially completed latent factor matrices Q and P^T.

	User1	User2	User3	User4
Item1	1.0		4.5	
Item2		5.0		2.0
Item3	1.5			3.0
Item4		2.5	1.5	

$$Q = \begin{bmatrix} 2 & \text{---} \\ \text{---} & 2 \\ 1 & \text{---} \\ \text{---} & 1 \end{bmatrix}$$

$$P^T = \begin{bmatrix} -0.5 & 1 & \text{---} & \text{---} \\ 2.0 & \text{---} & 1.5 & 1 \end{bmatrix}$$

(a) Fill in the missing entries of Q and P^T (denoted by ---) such that $r_{xi} - \hat{r}_{xi} = 0$ for all observed ratings. Please rewrite both matrices fully in the space given below.

(b) Using the completed matrices from part (a), fill in the unobserved ratings in the ratings matrix for User 4 with predictions generated from the model.

11 Bloom Filter [10 points]

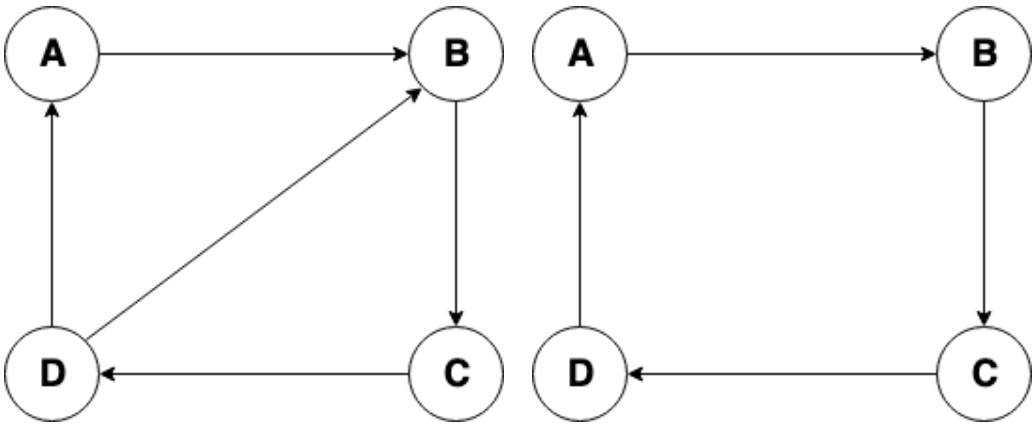
Suppose you are building a Youtube recommendation system. Each video on Youtube is represented by a unique 64-bit index. You have already developed recommendation algorithm which would give us a set R of videos a particular user might like. But you certainly don't want to recommend a video already watched by that user. So you want to test whether a video returned by your recommendation algorithm is in the watch history of that user or not. Clearly, a user might have seen so many videos that you can't afford storing all the video identifiers in memory. Therefore, you decide to use a bloom filter data structure to achieve the goal.

(a) Suppose the user has watched a total of 1000 videos. You want to construct a bloom filter with a bit array B of m bits, using a hash function $h : \{0, 1, 2, ..., 2^{64}-1\} \rightarrow \{0, 1, 2, ..., m-1\}$. The minimum value of m to achieve a false positive probability of 0.1 for your bloom filter will be _____. (Write down your answer as an integer)

(b) Now, suppose you are not satisfied with the 0.1 false positive probability for your bloom filter, and want to achieve lower false positive probability. Suggest two different ways to modify the design of the bloom filter in part (a) to achieve your goal. How will your approaches affect the false negative probability? (For each approach, give 1-2 sentence(s) description and 1-2 sentence(s) explanation of why it would decrease false positive probability, and how it would affect false negative probability.)

12 PageRank [10 points]

Figure 3 gives two directed graphs G1=(V, E1), G2=(V, E2) where:
V = $\{A, B, C, D\}$
E1 = $\{(A, B), (B, C), (C, D), (D, A), (D, B)\}$
E2 = $\{(A, B), (B, C), (C, D), (D, A)\}$
Based on Figure 3, answer the following questions:



(a) Directed Graph 1 (b) Directed Graph 2

Figure 3: Two Directed Graphs

(a) Suppose the teleportation parameter is $\beta = 0.8$, and the teleport set is V. After running PageRank on **graph G1**, which node has the lowest page rank and which node has the highest page rank?
Node with the smallest page rank: _____, Node with the highest page rank: _____

(b) Suppose the teleportation parameter is $\beta = 0.2$, and the teleport set is V. After running PageRank on **graph G1** and **comparing with question (a)**, give one node whose page rank will increase, and one node whose page rank will decrease.
Node whose page rank increases: _____, Node whose page rank decreases: _____

(c) Suppose the teleportation parameter is $\beta = 0.8$, and the teleport set is V. After running PageRank on **graph G2** and **comparing with question (a)**, give one node whose page rank will increase, and one node whose page rank will decrease.
Node whose page rank increases: _____, Node whose page rank decreases: _____

(d) Give the rank of nodes A, B, C, D in **question (c)** above. Please make sure the total rank of the four nodes sum to 1.
Rank of node A: _____, Rank of node B: _____
Rank of node C: _____, Rank of node D: _____

(e) Suppose the teleportation parameter is $\beta = 0.8$, and the teleport set is $\{A, B\}$ only. After running PageRank on **graph G2** and **comparing with question (c)**, give one node whose page rank will increase, and one node whose page rank will decrease.
Node whose page rank increases: _____, Node whose page rank decreases: _____

13 Data Streams [10 points]

DGIM

Suppose you are using the DGIM method to maintain a count of the most recent 1s. Represent each bucket as (i, t), where i is the number of $1s$ in the bucket and t is the timestamp of its end (i.e., time of the most recent 1 in the bucket). The window size is 50. Suppose the list of buckets at t = 169 is as follows:

$$(16, 120)(8, 128)(8, 148)(4, 156)(2, 160)(1, 164)(1, 169)$$

(a) Write down how the buckets-list gets modified if the incoming stream for the next 10 timestamps (t = 170 to 179) is: 1010100000 . Note that here 0 is the last bit in the stream (corresponding to $t = 179$)

Flajolet-Martin

Suppose you are using the Flajolet-Martin algorithm to count the number of distinct elements using the hash function $h_a(x) = (x + a) \bmod 128$
Let the stream you saw so far be 1,3,1,1,6,3,6,1,1

(b) What is the estimated number of distinct elements on using the hash function h_1?

(c) What is the estimated number of distinct elements on using the hash function h_2?

(d) Give a value of a such that using h_a gives the smallest possible estimate of the number of distinct elements for this stream. Explain in a few words why this is the least possible estimate.

14 Community Detection [10 points]

Consider the undirected weighted graph below. You are using the Louvain algorithm to perform community detection. Suppose in a certain iteration, after considering nodes 1, 2, and 3, you end up with the following configuration where nodes {1, 2, 3} form a community and nodes {4, 5, 6, 7, 8, 9} form another community.

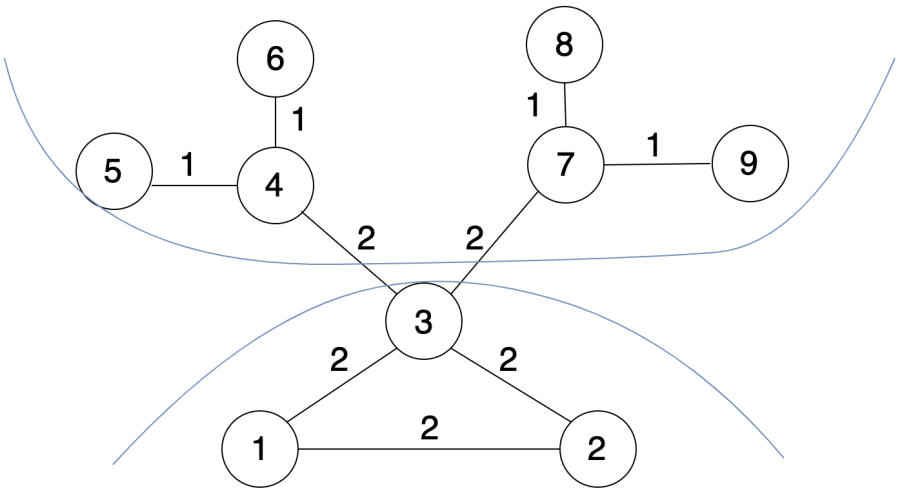


Figure 4: Current Configuration

(a) Now consider node 4. Will node 4 change its community assignment and join node 3's community? Please show your reasoning / calculation.

(b) After considering node 4, process also nodes 5, 6, 7, 8, and, 9 sequentially. After performing possible local changes for these nodes, please circle the resulting communities in the graph below. You don't need to show calculation. (Tip: You can use symmetry to help you reduce the amount of calculations needed.)

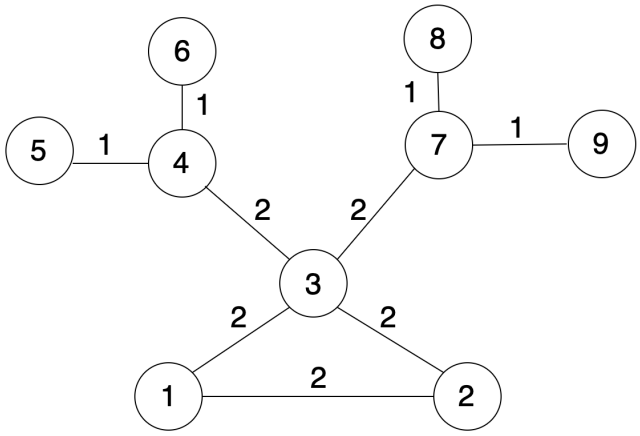


Figure 5: Communities to Detection

(c) Do you think the community assignments (in particular for nodes 4 through 9) you obtained are desirable? Please explain your rationale. If you don't think they are desirable, please give an idea on how to modify the Louvain algorithm in order to overcome this problem.

15 Graph Representation Learning [10 points]

In this question, we will explore an algorithm called *struct2vec* which captures the structural information of nodes in a graph, and compare it with *node2vec*.

Figure 6: An example of two nodes (u and v) that are structurally similar (respectively degrees 5 and 4, connected to 3 and 2 triangles, and connected to the rest of the network by two nodes), but very far apart in the network.

Here is how *struct2vec* works. Given a graph $G(V, E)$, it defines K functions $g_k(u, v)$, $k = 1, 2, .., K$, which measure the structural similarity between nodes. The parameter k means that only the local structures within distance k of the node are taken into account.

With all the nodes in G, regardless of the existing edges, it forms a new clique graph where any two nodes are connected by an edge whose weight is equal to the structural similarity between them. Since *struct2vec* defines K structural similarity functions, each edge has a set of possible weights corresponding to $g_1, g_2, ...g_K$.

The random walks are then performed on the clique. During each step, weights are assigned according to different g_k's selected by some rule (omitted here for simplification). Then, the algorithm chooses the next node with probability proportional to the edge weights.

(a) Characterize the vector representations of the 10-node cliques after running the *node2vec* algorithm on the graph in Fig. 7. Suppose to set the parameters of the random walk such that nodes that are close to each other have similar embeddings. Do you think the node embeddings will reflect the structural similarity? Justify your answer.

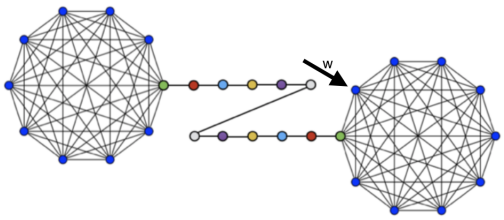


Figure 7: Graph with two 10-node cliques

(b) In Fig. 7, suppose that you arrive at node w. What are the nodes that you can reach after taking one step further with the *node2vec* algorithm? What about with the *struct2vec* algorithm (suppose that for this graph, $g_k(u, v) > 0$ for any u, v, k)?

(c) Why is there a need to consider different g_k's during the random walk?

(d) Characterize the vector representations of the two 10-node cliques after running the *struct2vec* algorithm on the graph in Fig. 7.

16 Locality Sensitive Hashing [15 points]

16.1 LSH Application: Finding Similar Documents

In this question, you will apply Locality-Sensitive Hashing to efficiently find candidate document pairs that likely have high similarity.

(a) Assume to use single words as tokens, and to convert documents into sets of 2-shingles. Recall that you can use Jaccard similarity as a similarity measure for shingled documents.
Consider the following two documents (pre-processed to remove punctuation and converting all letters to lower-case):

$D_1 =$ the quick brown fox jumps over the lazy dog

$D_2 =$ jeff typed the quick brown dog jumps over the lazy fox by mistake

The Jaccard similarity of D_1 and D_2 on 2-shingles is: $sim(D_1, D_2) =$_____.

(b) Recall the min-hashing algorithm covered in the lecture.
Assume you are given the following input matrix A:

$$A = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 1 & 0 & 1 \\ 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 \end{bmatrix}$$

where each row i corresponds to a shingle, each column j corresponds to a document. An element A_{ij} is set to 1 if document j includes shingle i, or 0 if the document does NOT include shingle i (i.e., same as defined in the lecture).

Assume to generate 2 permutations of the shingles, represented as a permutation matrix π:

$$\pi = \begin{bmatrix} 3 & 1 \\ 6 & 3 \\ 1 & 5 \\ 2 & 6 \\ 4 & 2 \\ 5 & 4 \end{bmatrix}$$

where each column corresponds to a permutation of shingles, specifying the order in which to iterate through the rows of the input matrix (i.e., same as defined in the lecture).

Fill out the signature matrix M (2 rows, 4 columns) resulting from running Min-Hashing using A and π:

$$M = \begin{bmatrix} -- & -- & -- & -- \\ -- & -- & -- & -- \end{bmatrix}$$

(c) Assume to use the LSH algorithm with $b = 10$ bands and $r = 5$ rows per band.
If a pair of documents D_3, D_4 have Jaccard similarity $sim(D_3, D_4) = 0.5$, then the probability that they have identical hash values in **at least** 1 band is: _____.
If a pair of documents D_5, D_6 have Jaccard similarity $sim(D_5, D_6) = 0.8$, then the probability that they have identical hash values in **at least** 1 band is: _____.
(You can leave exponents in your expression without calculating them.)

(d) Assume Emily and Pierre are using LSH with inputs from M different Min-hash functions. If Emily cares more about higher precision (i.e., less false positives) while Pierre cares more about higher recall (i.e., less false negatives), and their target similarity threshold s are the same, who should divide the M Min-Hash functions into **more** bands (by setting the number of bands b higher, with each band containing fewer rows)?
Circle exactly one option below.

A. Emily, who cares more about higher precision, should set the number of bands b higher
B. Pierre, who cares more about higher recall, should set the number of bands b higher
C. They should choose the same number of bands
D. Changing the number of bands b does not affect precision and recall

(e) Given a fixed number M of input Min-Hash functions, what will happen if the number of bands b is set too high, and each band contains very few rows?

16.2 Theory of LSH

(f) For a given hash family H, we have $\forall h \in H, h(x = y) = Similarity(x, y)$. Suppose you construct a $(3, 4, 5)$ way OR-AND-OR hash family G from H. Given similarity s, what is the probability of two candidate pairs get hashed into the same bucket for each $g \in G$?

(g) How many hash function from H do you need to construct a hash family composed of $(3, 4, 5)$ OR-AND-OR construction followed by $(6, 7)$ AND-OR construction?

17 Machine Learning with Gradient Descent [15 points]

(a) In some cases, it is possible to show that gradient descent with sufficiently small step size η will converge to the globally optimal solution. Let X be a matrix of the training data with data as columns, and consider linear regression (with no intercept term), where you try to fit the following system:

$$X^T\mathbf{w} = \mathbf{y}$$

with the squared error loss function:

$$L(\mathbf{x}_i, y_i) = \frac{1}{2}(\mathbf{x}_i^T\mathbf{w} - y_i)^2.$$

Then you want to minimize:

$$f(\mathbf{w}) = \frac{1}{2}(X^T\mathbf{w} - \mathbf{y})^T(X^T\mathbf{w} - \mathbf{y}),$$

and you can show that:

$$\nabla f = XX^T\mathbf{w} - X\mathbf{y}.$$

Now linear regression has a well-known closed form solution of $\mathbf{w} = (XX^T)^{-1}Xy$. Let $\mathbf{w}^{(t+1)}$ be the weight values after t update steps and let $\mathbf{w}_0$ be the initial weights. Using the fact that $(\eta A)^{-1} = \sum_{i=0}^{\infty}(I - \eta A)^i$ for small η, show that as $t \to \infty$ you have $\mathbf{w}^{(t+1)} \to (XX^T)^{-1}X\mathbf{y}$.

(b) The proof above only holds when η is sufficiently small. As you have seen on the homework, the ability of Gradient Descent to converge depends on whether your η is selected correctly. Suppose you train a model with gradient descent with a variety of training rates. You plot the train and test error for each epoch and see the output in Figure 8. The plotted accuracy on the training dataset is shown as a solid line and test accuracy as a dashed line. Given these results, what is the best learning rate for your model and data? Please justify your answer.

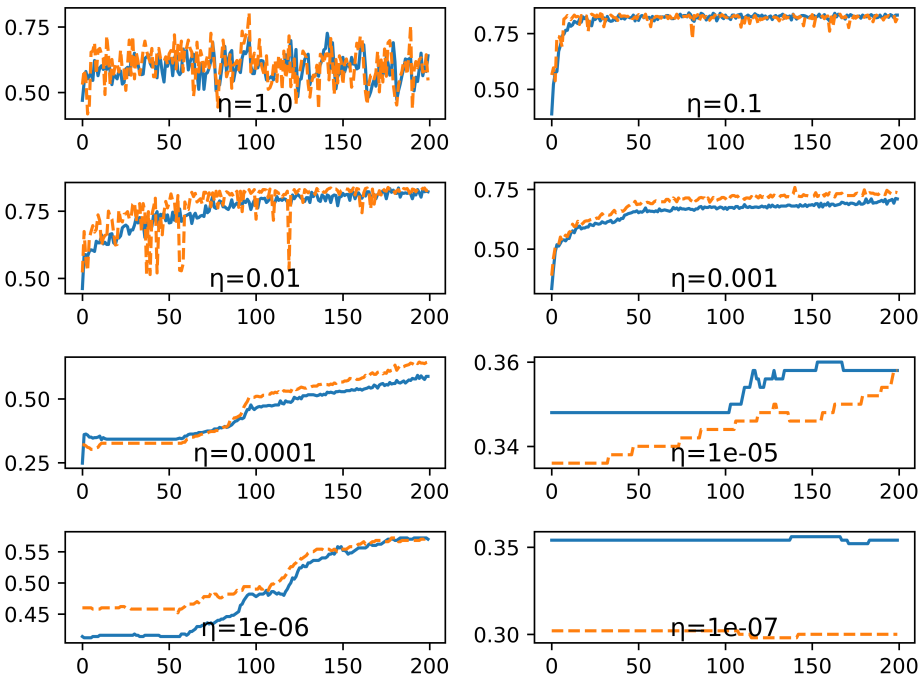


Figure 8: Model accuracy vs. epoch with varying learning rates, η. (*Pay attention to accuracy plotted on the Y-axis!*)

As discussed in class, Stochastic Gradient Descent is often used in place of Gradient Descent when dealing with large data sets that cannot fit in memory. In this part of the question, you will devise a way to run SGD efficiently when dealing with sparse data, i.e., when your data $\mathbf{x} \in \mathbb{R}^d$ but the average number of non-zero entries of $\mathbf{x}$, s, is much smaller than d ($s \ll d$).

(c) Consider the same objective function as before except now with ℓ^2 regularization in your objective:

$$f(\mathbf{w}) = \frac{1}{2}(X^T\mathbf{w} - y)^T(X^T\mathbf{w} - y) + \frac{\lambda}{2}\sum_{i=1}^{d}\mathbf{w}_i^2.$$

Write down the SGD update rule for $\mathbf{w}_i$.

(d) Suppose you can represent the vectors with two different kinds of data structures: dense and sparse. In a dense representation, you can iterate over the elements of a vector $v \in \mathbb{R}^d$ in $O(d)$

time, and in a sparse representation you can iterate over the s non-zero elements of $v \in \mathbb{R}^d$ in $O(s)$ time. Suppose $\lambda = 0$. If you use a dense data structure, what is the runtime of one update to $\mathbf{w}_i$? What is it if you use a sparse data structure?

(e) Suppose you have a sequence of examples $\mathbf{x}^{(t)}, \cdots, \mathbf{x}^{(t+k)}$ where a given component i is always 0, i.e., $\mathbf{x}_i^{(t)} = \mathbf{x}_i^{(t+1)} = \cdots = \mathbf{x}_i^{(t+k)} = 0$. Provide an expression for $\mathbf{w}_i^{(t+k)}$ in terms of $\mathbf{w}_i^{(t)}$, k, η, and λ.

(f) Using the observation from part (e), propose an $O(s)$ update algorithm for SGD for when $\lambda > 0$.