CS 246 Final Exam, Winter 2018[1]

- Your Name: _____
- Your SUNetID (e.g. jtysu): _____
- Your numerical SUID (e.g. 01234567): _____

I acknowledge and accept the Stanford Honor Code, and promise that **I will not discuss the exam with anyone until Wednesday, March 21, 2018, 8:00 am Pacific Time.**

Signature: _____

1. Please bring a calculator or a computer.
2. Acceptable uses of your computer:
   - You may not access the Internet (except to view the course website, lecture slides, and digital version of the textbook), or communicate with any other person.
   - You may not use your computer to write code, only do arithmetic calculations. You may only use computational features that would be present in a standard scientific calculator, such as addition, subtraction, multiplication, division, logarithms, exponents, etc.
   - You are intended to use your computer as a calculator and an e-reader.
3. SCPD students may call Jessica at 561-543-1855 if they have questions during the final exam. However, the answer to most questions will be "use your best judgment." If Jessica doesn't pick up the phone you should send an email to cs246-win1718-staff@lists.stanford.edu.
4. The final exam will be open-book and open-notes. You may only use notes that you have written yourself (digitally-created notes are okay), and/or lecture slides/the textbook from the course website.
5. Numerical answers may be left as fractions, as decimals to an appropriate number of places, or as radicals, e.g., $\sqrt{2}$.
6. There are 17 questions on this final; the maximum score you can attain is 137 points.

[1]Special thanks to Yutian Li, Ansh Shukla, Heather Blundell, Chang Yue, Qijia Jiang, and Kush Goyal for taking early versions of this exam, Jessica Su for reviewing everyone's problems and compiling them into a polished document, Jure Leskovec and Jeff Ullman for offering continuous feedback throughout the exam design process, and Jeff Ullman, Sanyam Mehra, Dylan Liu, Praty Sharma, Wanzi Zhou, Hiroto Udagawa, and Sen Wu (as well as the other people previously listed) for each contributing and peer-reviewing a couple of problems.

1 Computing Degree Distributions on Networkly (5 points)

Alice is a data scientist at Networkly, a large social networking website where users can follow each other. Alice has a deadline in three hours, and needs your help running an analysis.

Suppose you are given the Networkly follow graph. Assumptions you may make:

- The graph has 330 million nodes and 100 billion directed edges.
- The most followed user on Networkly has 100 million followers.
- Your computing cluster has 1,000 computers, each with 64 GB of RAM.

Using Spark, write a parallel program to compute the in-degree distribution of the graph. That is, for each k, your program should compute the number of users who have k followers (unless this number is zero, in which case you do not need to include that count in the final output).

Your program should effectively take advantage of parallel processing and not attempt to read 100 billion directed edges on a single machine.

Your input is an RDD, where each entry is an edge: (u, v) is an entry in the RDD if u follows v. Your function should return another RDD, where the entries are of the form (k, number of users who have k followers).

What to submit: Please fill in the code stubs provided for you on the next page; you may choose to fill in the stub for **either** Python, Java, or Scala.

You should write real code (rather than pseudocode). However, because you do not have access to a compiler/interpreter, it will be possible to get full credit even if your code does not compile. Specifically, the grading rubric will be

- (3 points) A correct pseudocode description of the algorithm, without any code.
- (4 points) Real code that partially works, and corresponds to a correct high-level algorithm.
- (5 points) Real code that completely works, minus one or two missing parentheses.

Python version

```
import sys
from pyspark import SparkConf, SparkContext

# Input: An RDD containing entries of the form
#   (source node, destination node)
# Output: An RDD containing entries of the form
#   (k, number of users who have k followers)
# TODO: Write your answer code in this function.
def degree_distribution(edges):









    return distribution

if __name__ == '__main__':
    conf = SparkConf()
    sc = SparkContext(conf=conf)
    sc.setLogLevel("WARN")

    # Reads input and converts all node ids to integers
    data = sc.textFile(sys.argv[1]).map(lambda line: map(int, line.split()))

    # Computes the degree distribution
    distribution = degree_distribution(data)

    # Writes the output to a file
    distribution.sortByKey().saveAsTextFile(sys.argv[2])

    sc.stop()
```

Java version

```
package edu.stanford.cs246;

import java.util.Arrays;
import org.apache.spark.SparkConf;
import org.apache.spark.api.java.JavaPairRDD;
import org.apache.spark.api.java.JavaRDD;
import org.apache.spark.api.java.JavaSparkContext;
import scala.Tuple2;

public class DegreeDistributionJava {
    // TODO: Fill in this method
    // Input: An RDD containing entries of the form
    //      (source node, destination node)
    // Output: An RDD containing entries of the form
    //      (k, number of users who have k followers)
    public static JavaPairRDD<Integer, Integer>
        degreeDistribution(JavaPairRDD<Integer, Integer> edges) {









        return distribution;
    }

    public static void main(String[] args) throws Exception {
        SparkConf conf = new SparkConf();
        JavaSparkContext sc = new JavaSparkContext(conf);

        JavaRDD<String> lines = sc.textFile(args[0]);

        JavaPairRDD<Integer, Integer> data = lines.mapToPair(l -> {
            String[] nodes = l.split(" ");
            return new Tuple2<Integer, Integer>(Integer.parseInt(nodes[0]),
                                                Integer.parseInt(nodes[1]));
        });
```

```
        JavaPairRDD<Integer, Integer> distribution = degreeDistribution(data);

        distribution.sortByKey().saveAsTextFile(args[1]);
    }
}
```

Scala version

```
package edu.stanford.cs246

import org.apache.spark.rdd.RDD
import org.apache.spark.{SparkConf,SparkContext}
import org.apache.spark.SparkContext._

object DegreeDistribution {
    // TODO: Fill in this method
    // Input: An RDD containing entries of the form
    //      (source node, destination node)
    // Output: An RDD containing entries of the form
    //      (k, number of users who have k followers)
    def degreeDistribution(edges: RDD[(Int, Int)]): RDD[(Int, Int)] = {









        return distribution
    }

    def main(args: Array[String]) = {
        val conf = new SparkConf()
        val sc = new SparkContext(conf)
        val data = sc.textFile(args(0))
            .map(line => line.split(" "))
            .map(array => (array(0).toInt, array(1).toInt))

        val distribution = degreeDistribution(data)

        distribution.sortByKey().saveAsTextFile(args(1))

        sc.stop()
    }
}
```

★ **SOLUTION:** Idea:

1. Map (user, friend) to (friend, 1)
2. Group by key, sum all the values; output is (user, number of followers)
3. Map (user, number of followers) to (number of followers, 1)
4. Group by key, sum all the values; output is the degree distribution.

The code is provided in a separate file.

2 Frequent Itemsets (10 points)

Alice, a proud Stanford alumnus, is the developer of `carta.stanford.edu`, and wants to get insight into which classes Stanford students most frequently take. She can't find a bug in her code, so she created a toy dataset and wants your help generating unit tests.

The classes in her dataset are {CS103, CS140, CS161, CS224W, CS341}, and the toy student records are

Student	Classes they took
Chang	{CS103, CS140}
Heather	{CS140, CS161, CS224W}
Jessica	{CS140, CS341}
Kush	{CS103, CS140, CS161}
Praty	{CS103, CS140, CS161, CS341}
Qijia	{CS140, CS224W, CS341}
Sanyam	{CS224W, CS341}
Wanzi	{CS103, CS140, CS341}

Given a **support threshold** of $s = 4$, please answer the following questions:

(a) (1 point) What are the frequent individual classes? _____

(b) (4 points) Suppose we assign numerical values to every class:

Class	Value
CS103	1
CS140	2
CS161	3
CS224W	4
CS341	5

Now we run the PCY algorithm, defining a hash function

$$h(i, j) = (i + j) \mod 6$$

that maps pairs of items into six buckets using their numerical values. What does the bitmap look like?

$$\text{bitmap} = \{\underset{0}{\_\_}, \underset{1}{\_\_}, \underset{2}{\_\_}, \underset{3}{\_\_}, \underset{4}{\_\_}, \underset{5}{\_\_}\}$$

(c) (2 points) What are the frequent pairs of classes? _____

(d) (3 points) What are the closed items and item pairs? _____

_____

What to submit: Please fill in the blanks for parts (a), (b), (c), and (d).

★ **SOLUTION:**

(a) {CS103, CS140, CS341}

(b) bitmap = {1, 1, 0, 1, 0, 0}

(c) {CS103, CS140}, {CS140, CS341}

(d) {CS140}, {CS224W}, {CS341}, {CS103, CS140}, {CS140, CS161}, {CS140, CS224W}, {CS140, CS341}, {CS224W, CS341}

3 Locality-Sensitive Hashing (4 points)

Suppose we have a $(0.4, 0.6, 0.8, 0.2)$-sensitive family $\boldsymbol{F}$ of hash functions.

(a) (2 points) The family constructed by a 2-way AND followed by a 2-way OR using $\boldsymbol{F}$ is a (_____, _____, _____, _____)-sensitive family.

★ **SOLUTION:** $(\underline{0.4}, \underline{0.6}, \underline{0.8704}, \underline{0.0784})$-sensitive family.

(b) (2 points) The family constructed by a 2-way OR followed by a 2-way AND using $\boldsymbol{F}$ is a (_____, _____, _____, _____)-sensitive family.

★ **SOLUTION:** $(\underline{0.4}, \underline{0.6}, \underline{0.9216}, \underline{0.1296})$-sensitive family.

What to submit: Fill in the blanks for parts (a) and (b).

4 Clustering (18 points)

In this question, we will compare different clustering strategies for points in the two-dimensional Euclidean plane.

(a) (8 points) Suppose that we want to cluster the following 5 points $\boldsymbol{p}_i$, $i = 1, \ldots, 5$ into 2 clusters:

point	x	y
$\boldsymbol{p}_1$	6	3
$\boldsymbol{p}_2$	2	2
$\boldsymbol{p}_3$	0	2
$\boldsymbol{p}_4$	3	0
$\boldsymbol{p}_5$	6	6

(i) (4 points) Suppose that we use the K-means algorithm ($K = 2$) with Euclidean distance and $\boldsymbol{p}_3$ and $\boldsymbol{p}_5$ as the initial centroids. What are the coordinates of the final cluster centroids and cluster assignments at convergence?

What to submit: Please fill in the table.

Centroid	Points assigned to that cluster

★ **SOLUTION:** Centroid $(5/3, 4/3)$ has assigned points $\boldsymbol{p}_2$, $\boldsymbol{p}_3$, and $\boldsymbol{p}_4$ and centroid $(6, 9/2)$ has assigned points $\boldsymbol{p}_1$ and $\boldsymbol{p}_5$. (Note: converges in 1 iteration)

(ii) (4 points) Instead suppose that we use hierarchical agglomerative clustering and we merge clusters using *"single-linkage"* Euclidean distance. That is, we define the distance between two clusters to be the minimum of the distances between any two points, one chosen from each cluster. Which points are in each of the two clusters at convergence?

What to submit: Please fill in the blanks.

- Points assigned to one cluster: _____
- Points assigned to the other cluster: _____

★ **SOLUTION:** Same as k-means.

(b) (5 points) Now consider clustering n points $\boldsymbol{p}_i$, $i = 1, \ldots, n$ in the 2-dimensional Euclidean plane in general. Explain a scenario in which single-linkage hierarchical clustering performs worse than k-means clustering (i.e. a scenario where k-means clustering assigns the points to more intuitively correct clusters than single-linkage hierarchical clustering does). Sketch the structure of the resulting dendogram produced by single-linkage hierarchical clustering in this scenario. Also, support your explanation with an example set of points in the Euclidean plane.

What to submit:

- A 1-3 sentence explanation of a scenario where single-linkage hierarchical clustering performs worse than k-means.
- A dendogram sketch of this scenario.
- An example set of points where this scenario applies.

★ **SOLUTION:** Hierarchical clustering with single-linkage can exhibit a chaining phenomenon when points are close together in a line in the Euclidean plane. So dissimilar points may be placed in the same cluster. Example set of points: $(1, 0)$, $(2, 0)$, $(3, 0)$, $(4, 0)$, ..., $(10, 0)$ would be in one cluster and the points $(1, 2)$ and $(3, 2)$ would be in the other cluster. Dendogram sketch of chaining:

(c) (5 points) Consider the application of clustering grocery store customers into segments based on purchase habits. Suppose that the x-coordinate represents money spent on vegetables (in dollars) and the y-coordinate represents money spent on candy (in dollars). The store wants to cluster customers into a "healthy" cluster and an "unhealthy" cluster.

(i) (3 points) Explain a potential issue with the Euclidean distance metric in this scenario.

What to submit: A 1-paragraph explanation.

★ **SOLUTION:** We might want to consider customers as similar based on the ratio of their spending on each category instead of the absolute amounts of spending. So a customer who spent x dollars on vegetables and y dollars on candy should be considered as more similar to a customer who spent $5x$ dollars on vegetables and $5y$ dollars on candy than a customer who spent x dollars on vegetables and 0 dollars on candy. The larger amounts of spending could just be due to a different budget of spending.

(ii) (2 points) What would be a better choice of distance metric in this case and why?

What to submit:

- Your choice of distance metric.
- A 1-3 sentence explanation of why your choice is better.

★ **SOLUTION:** Cosine distance or correlation distance. These metrics consider the distance between vectors based on angle between them or correlation and are not affected by differing lengths of the vectors.

5 Singular-Value Decomposition on Networkly (5 points)

Alice is a data scientist at Networkly, and wants to use the Networkly follow graph to understand the different characteristics of Networkly users. Specifically, she thinks we can infer things about Networkly users based on the set of users that follow them. Alice has the idea of representing each user u by an n-element vector (where n is the number of users on Networkly); element i of the vector is 1 if user i follows user u, and 0 otherwise.

Since there are hundreds of millions of users on Networkly, this method results in absurdly long vectors. Alice decides to use **singular-value decomposition** to encode approximately the same information using fewer numbers, and keeps the top 100 features. To measure the performance of her approach, she calculates the **reconstruction error** between the original Networkly follow matrix and the product of the matrices in the decomposition.

Alice's friend Bob is preparing for an interview at Networkly, and wants to practice using the same technique. However, since he does not work at Networkly (yet), he doesn't have a copy of the Networkly follow graph, and uses a **random graph** instead.

The random graph has the same number of nodes as the Networkly follow graph. Edges between each pair of nodes are independently formed with probability p. Bob wants to use a realistic value for the probability p, so he asks Alice what the edge density of the Networkly follow graph is (i.e. the number of edges divided by the number of pairs of nodes), and uses that for his value of p.

(1) (1 point) How would you expect Bob's reconstruction error to compare to Alice's? Select one of the following:

   (a) Bob's reconstruction error should be lower than Alice's.

   (b) Bob's reconstruction error should be higher than Alice's.

   (c) Bob's reconstruction error should be similar to Alice's.

(2) (4 points) Please explain your answer.

What to submit: Your answer choice, and a brief (1-3 sentence) explanation.

★ **SOLUTION:** A randomly generated matrix would be expected to naturally vary along many more dimensions than a social network matrix. So if we limit the number of dimensions we get rid of a lot of the signal, and we would expect Bob's reconstruction error to be higher.

6 Recommender Systems (7 points)

You work as a software engineer for Streamingly, a video streaming service popular for streaming East-Asian media including anime, manga and music. Customer feedback suggests that they want Streamingly to provide them with recommendations. You are the lead engineer in this project.

(a) (3 points) You start by considering different approaches to build the recommender system. Streamingly has a lot of anime titles in their database. However, they only recently started recording user ratings and have very few ratings in their database. In case you need to implement the system before being able to collect a lot of data, which of the following would you consider to be a better recommender system?

(a) User-user collaborative filtering

(b) Item-item collaborative filtering

(c) Content-based recommendation

Give a brief justification of your answer.

What to submit: Your answer choice (1 point), and a 1-5 sentence justification (2 points).

★ **SOLUTION:** c. The only system from the above choices that does not depend on the ratings from other users is option (c) Content-based recommendation system.

(b) (4 points) A few years later, Streamingly has collected lots of data, and the situation in the previous part of the question no longer applies. You are working on the team that decides if two anime shows are similar to each other. To do this for two shows A and B, you compute the Pearson correlation coefficient between the ratings given to show A, and the ratings given to show B.

Your friend Alice works at Networkly, and proposes to use a similar strategy to determine if users are similar to each other, based on whether other users have followed them or not. Whenever a user u follows another user f, Alice gives the (user, user) pair a "rating" of $r_{uf} = 1$. Because there are (330 million choose 2) pairs of users on Networkly, Alice does not store any ratings for pairs of users who don't follow each other. Alice again proposes to compute the Pearson correlation coefficient between two sets of ratings.

Is Alice's strategy a good strategy? Why or why not?

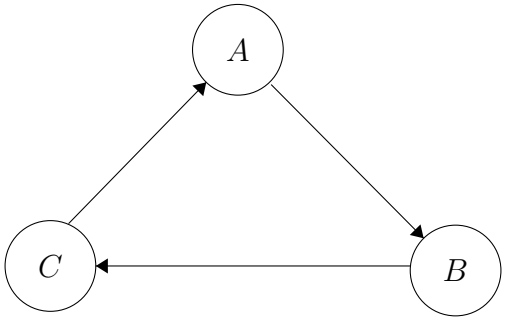
What to submit: Your answer (yes or no), followed by a 1-paragraph explanation of why Alice's strategy is good or bad. Please write your answer on the next page.

★ **SOLUTION:** No. To calculate Pearson's correlation coefficient, we subtract the average nonzero rating from all the nonzero entries of the vector. Since all the ratings are the same, this leaves us with a vector that is entirely composed of zeros, which would not give us an accurate result.

[more space for part (b)]

7 Web Spam (9 points)

Here is a tiny Web graph:



You may assume that the PageRank of each node is 1/3. (Note: this is an example of a pathological graph where PageRank without taxation does not converge unless you initialize the estimate to have all ranks the same to begin with.)

(a) (3 points) Supposing there is a 50% taxation rate, and only A is a trusted node, write down the equations for the TrustRank vector. Use a, b, and c as the variables that represent the TrustRank of nodes A, B, and C, respectively.

What to submit: Please write the equations in the space below.

★ **SOLUTION:** $a = a/2 + b/2 + c$, $b = a/2$, $c = b/2$

(b) (3 points) Assuming $a + b + c = 1$, what is the solution to these equations?

$a =$ _____

$b =$ _____

$c =$ _____

What to submit: Please fill in the blanks.

★ **SOLUTION:** $a = 4/7$, $b = 2/7$, $c = 1/7$.

(c) (3 points) What is the spam mass of each node?

Node	Spam mass
a	
b	
c	

What to submit: Please fill in the table.

★ **SOLUTION:** In general, if a node has PageRank p and TrustRank t, its spam mass is $(p - t)/p$. So for node A, the spam mass is $(1/3–4/7)/(1/3) = -5/7$. For B, it is $(1/3–2/7)/(1/3) = 1/7$, and for C it is $(1/3–1/7)/(1/3) = 4/7$.

8 Social Networks (6 points)

Suppose there are two clubs A and B, where A contains nodes $\{1, 2, 3\}$ and B contains nodes $\{4, 5, 6, 7\}$. Their relationships are represented in a graph shown in Figure 1, where an edge between two nodes means that the two nodes are good friends with each other.

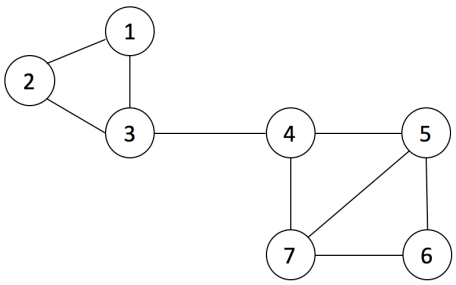


Figure 1: Original Graph

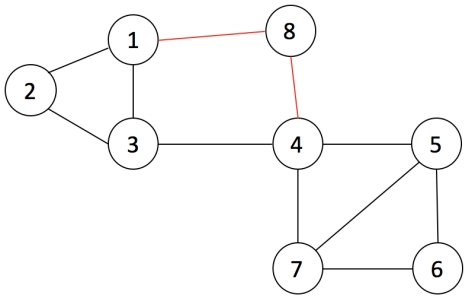


Figure 2: After node 8 joins

Now the two clubs have a severe conflict, and the only edge between them, i.e edge 3-4, is about to break. At this time, a newcomer node 8 joins. Node 8 has good relationships with both node 1 and node 4, but he can only join one club and thus is forced to break his edge with either 1 or 4. He does not know what to do so he wants your advice. The current graph is shown in Figure 2.

We want to partition the graph into good clusters based on the measurement of conductance score.

- (2 points) What is the conductance of the cut $((1, 2, 3), (4, 5, 6, 7, 8))$? _____
- (2 points) What is the conductance of the cut $((1, 2, 3, 8), (4, 5, 6, 7))$? _____
- (2 points) Which club (A or B) should we suggest that node 8 join? _____

What to submit: Please fill in the blanks.

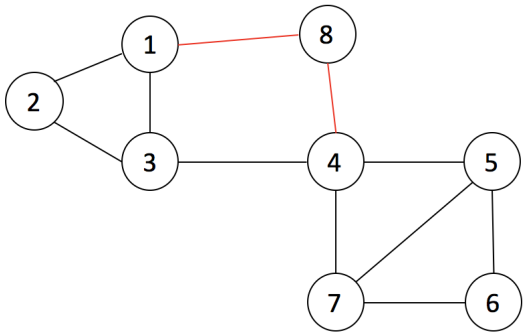
★ **SOLUTION:** For cutting edges 3-4 and 1-8

$$\text{conductance} = \frac{2}{\text{degree}(1) + \text{degree}(2) + \text{degree}(3)} = \frac{2}{3+2+3} = \frac{1}{4}$$

For cutting edges 3-4 and 4-8

$$\text{conductance} = \frac{2}{\text{degree}(1) + \text{degree}(2) + \text{degree}(3) + \text{degree}(8)} = \frac{1}{3+2+3+2} = \frac{1}{5} < \frac{1}{4}$$

Thus we should choose to partition the graph into $\{1, 2, 3, 8\}$ and $\{4, 5, 6, 7\}$ and ask node 8 to join club A.


1
8
2
3
4
5
7
6

9 Algorithms on Large Graphs (15 points)

9.1 AGM (8 points)

Consider a set of 9 nodes: $v_1, v_2, \ldots, v_9$. Consider an AGM model on this graph where nodes $v_1, \ldots, v_5$ belong to a community A with parameter p_A and nodes $v_5, \ldots, v_9$ belong to community B with parameter p_B. Let ϵ be the probability that two nodes are connected if they share no communities. Using this AGM model we generate a graph G. Answer the following questions:

(a) (2 points) What is the probability that nodes $v_1, \ldots, v_5$ form a clique in graph G? ____

(b) (3 points) What is the probability that nodes $v_1, \ldots, v_5$ form a 3-2 complete bipartite graph in graph G? (A 3-2 complete bipartite graph is a five-node graph whose vertices can be divided into two groups, one of size 3 and one of size 2; each vertex in one group has an edge to every vertex in the other group, but there are no edges between vertices that are part of the same group.) _____

_____

(c) (3 points) What is the probability that nodes v_1, v_2, v_5, and v_6 form a clique in graph G? _____

_____

What to submit: Fill in the blanks for parts (1-3).

9.2 BigCLAM (7 points)

Consider the graph given below:

Using the relaxed AGM model, find the values of x_A, x_B which maximise the log-likelihood of this graph. Assume that

- There are two communities, A and B.
- $x_A + x_B = 1$.
- $0 \leq x_A \leq 1$ and $0 \leq x_B \leq 1$.
- For this part of the question only, you may assume $\epsilon = 0$.
- All the other membership strengths are fixed for the other nodes, and are set to the values indicated in the graph.

After computing the values of x_A, x_B, give the expression for the log-likelihood of the graph.

Your answer:

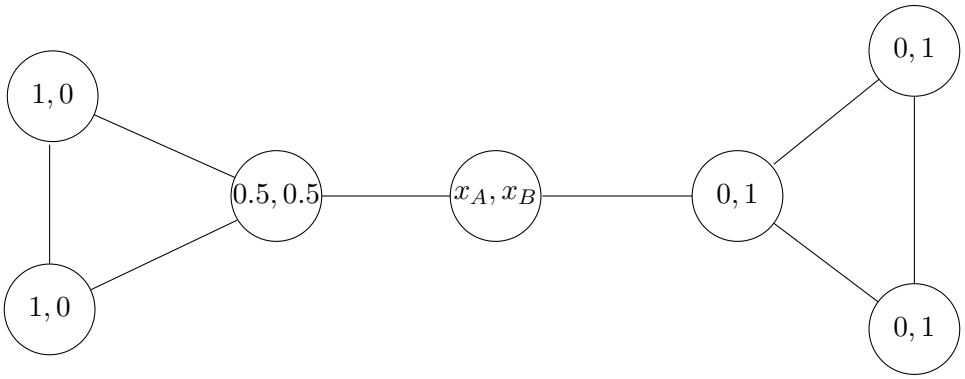


Figure 3: The graph, with membership strengths. For each node in the graph, the first number is the membership strength of that node with community A, and the second number is the membership strength of that node with community B.

(a) (2 points) $x_A =$ _____

(b) (2 points) $x_B =$ _____

(c) (3 points) Log-likelihood = _____

What to submit: Fill in the blanks for x_A, x_B, and the log-likelihood.

★ SOLUTION:

1. p_A^{10}
2. $10 * p_A^6 * (1 - p_A)^4$
3. $p_A^3 * p_B * \epsilon^2$
4. $x_B = 1$, $x_A = 0$.
   Log-Likelihood $= log((1 - e^{-1})^5 * (1 - e^{-0.5})^3) - 3.5$

10 Large-Scale Machine Learning (10 points)

Consider a dataset of positively labeled points $x_1 = (0, 1)$ and $x_2 = (2, 3)$, along with negatively labeled points $x_3 = (1, 0)$ and $x_4 = (3, 3)$.

(a) (3 points) Express the range of possible slopes of a separating hyperplane as an interval.

- Your answer: _____

★ **SOLUTION:** $(\frac{2}{3}, 3)$

(b) (3 points) Follow the convention that data in d-dimensional space have $d + 1$ support vectors. What are the possible sets of support vectors?

- Your answer: _____

★ **SOLUTION:** $\{x_1, x_2, x_4\}, \{x_2, x_3, x_4\}$.

(c) (4 points) Choose one of the sets of support vectors from part b. We wish to maximize the margin, which entails minimizing $\|w\|$ subject to the constraints that $w \cdot x + b \geq 1$ for all positively labeled points and $w \cdot x + b \leq -1$ for all negatively labeled points. State your choice of support vectors. What are the values of w, b, and the margin?

- Your choice of support vectors: _____
- w: _____
- b: _____
- Margin: _____

★ **SOLUTION:** For $\{x_1, x_2, x_4\}$, $w = (-2, 2)$, $b = -1$, and the margin is $\frac{1}{2\sqrt{2}}$. For $\{x_2, x_3, x_4\}$, $w = (-2, \frac{4}{3})$, $b = 1$, and the margin is $\frac{3}{2\sqrt{13}}$.

What to submit: Please fill in the blanks.

11 Decision Trees (11 points)

11.1 Predictions on Regression Trees (3 points)

Suppose we are given N data points. Let x_i be the features associated with the ith data point, and y_i be the label. In this case, we are solving a regression problem, so y_i is a real number, and does not have to be 0 or 1.

Let's say that we have built out a regression tree and are now determining what values to assign the leaf nodes as our prediction.

There are K leaf nodes and N points. Let's call S_j the set of points associated with leaf $j \in \{1, ..., K\}$. We must assign a value r_j for $j \in \{1, ..., K\}$ that minimizes the error:

$$\text{error} = \sum_{i=1}^{N} (y_i - f(x_i))^2$$

where

$$f(x_i) = \sum_{j=1}^{K} r_j I(x_i \in S_j)$$

that is, the predicted value associated with the leaf that x_i is located in. Here $I(x_i \in S_j)$ is an indicator variable denoting whether or not x_i is part of the set S_j.

What value of r_j minimizes the error?

What to submit: Either a mathematical formula or 1-2 sentences describing the correct value of r_j.

★ **SOLUTION:**

$$r_j = \sum_{i=1}^{N} y_i I(x_i \in S_j) / \sum_{i=1}^{N} I(x_i \in S_j)$$

that is, the average output value for the data in leaf j.

11.2 Calculating Best Splits on Classification Tree (8 points)

Consider this training set of 9 examples described in terms of three attributes a_1, a_2 and a_3 in addition to the class label.

a_1	a_2	a_3	Class Label
True	True	1.0	+
True	True	6.0	+
True	False	5.0	−
False	False	4.0	+
False	True	7.0	−
False	True	3.0	−
False	False	8.0	−
True	False	7.0	+
False	True	5.0	−

Recall that we defined the entropy of a set of data as:

$$H(X) = -\sum_{j=1}^{m} p_j \log_2(p_j)$$

and the information gain:

$$IG(Y|X) = H(Y) - H(Y|X)$$

(a) (2 points) What is the entropy of this dataset? _____

★ **SOLUTION:** 0.9911

(b) (4 points) What are the information gains of a_1 and a_2? _____

★ **SOLUTION:**

$a_1 : 0.2294$

$a_2 : 0.0072$

(c) (2 points) What is the best split among a_1 and a_2 according to the information gain? _

★ **SOLUTION:** a_1 produces the best split.

What to submit: Fill in the blanks for parts (a)-(c).

12 Decision Trees on Networkly (4 points)

Alice, a data scientist at Networkly, is trying to get more people to click on Networkly ads. She tells her manager that fewer than 5% of ad views result in a click. Her manager tells her to build a decision tree to predict which users will click on Networkly ads, so that she can optimize the ad delivery.

Every time a user sees an ad on Networkly, a new entry (X_i, y_i) is added to the dataset. X_i is a vector that contains the features corresponding to the user. y_i is 1 if the user clicked on the ad, and 0 otherwise.

Alice's manager told her that it is very important to know when to stop building the decision tree, because otherwise you might overfit to the training set. She decides to stop building the decision tree when her leaves are "pure," i.e. when $Var(y_i) < 0.05$ for the data in each leaf. In a lot of classification problems, this stopping criterion might produce a good decision tree, but on this dataset it might fail.

What might be the problem? (1-3 sentences)

What to submit: A 1-3 sentence description of what the problem is.

★ **SOLUTION:** The problem is that the classes are highly imbalanced, because very few people click on ads. Even with a click-through rate of 5%, which is unreasonably high for online advertising, the variance at the root node is only $1^2 \cdot 0.05 - (1 \cdot 0.05)^2 = 0.0475 < 0.05$.

13 Data Streams (8 points)

There is a stream of "registration records" with the format (Name, Course, Department, Units) giving the name of a student, a course in which the student is enrolled, the department teaching that course, and the number of units for which they are enrolled in this course. As at Stanford, students can enroll in the same course for different units.

(a) (4 points) Suppose we want to use a sample consisting of approximately 20% of the stream's records and get an unbiased estimate of the total number of units for each course (i.e., for each course, the sum of the number of units awarded to each of the students enrolled in that course). Explain how you would select the sample and how you would compute the estimates for each course?

What to submit: A 1-paragraph explanation of how you would select the sample, and how you would compute the estimates for each course.

★ **SOLUTION:** Hash only the student name to five buckets, and take one of the buckets as a sample. Within the sample, compute the sum of units for each course. Then, multiply each sum by 5. To be precise:

SELECT Course, 5*SUM(Units) FROM Sample GROUP BY Course;

Note: It is also correct to hash on any set of the attributes that includes Student, although I doubt anyone will do that.

Also note: it may be ambiguous whether a course is uniquely defined by the value of Course, or whether you need Department as well (i.e., is this course just "246" or "CS246"?). it doesn't matter regarding the essence of the solution, but if you need Department, then that attribute would have to be added to the SELECT and GROUP-BY clauses in the above.

(b) (2 points) Suppose that all is as in (a), but now we want, for each course, the minimum number of units for which any student in that course is registered. Would your approach of (a) still work? Explain briefly.

What to submit:

- A yes/no answer.
- An explanation of why your approach in (a) would or would not work.

★ **SOLUTION:** No. If we dropped even one record, it might be the unique student in a course with the MINIMUM number of units. Thus, the estimate could be too high, but never too low.

(c) (2 points) Suppose that all is as in (a), except that for each course we want the AVERAGE number of units for which any student in that course is registered. Would your approach of (a) still work? Explain briefly.

What to submit:

- A yes/no answer.
- An explanation of why your approach in (a) would or would not work.

★ **SOLUTION:** Yes. The expected number of records in the sample for any course is 1/5 of the total number of records for that course, and their average of units will have the same expected value as the average for the entire class.

14 Mining Data Streams (8 points)

In this problem, we will walk through the DGIM algorithm from lecture.

Consider the bucketized stream: 0 0 0 1 0 1 0 1 0, where new elements are added on the right.

The current state is: 0 0 0 $\boxed{1\ 0\ 1}$ 0 $\boxed{1}$ 0.

Assume the maximum number of buckets with the same size is 2.

(a) (1 point) Draw the state after the bit of value 1 arrives: 0 0 1 0 1 0 1 0 1.

★ **SOLUTION:** 0 0 $\boxed{1\ 0\ 1}$ 0 $\boxed{1}$ 0 $\boxed{1}$

(b) (1 point) Draw the state after another bit of value 1 arrives: 0 1 0 1 0 1 0 1 1.

★ **SOLUTION:** 0 $\boxed{1\ 0\ 1}$ 0 $\boxed{1\ 0\ 1}$ $\boxed{1}$

(c) (1 point) Draw the state after the bit of value 0 arrives: 1 0 1 0 1 0 1 1 0.

★ **SOLUTION:** $\boxed{1\ 0\ 1}$ 0 $\boxed{1\ 0\ 1}$ $\boxed{1}$ 0

(d) (3 points) Based on the state in part (c), what is your estimation of the number of 1s of the last 3 bits? Of the last 4? Of the last 8?

Bits	Estimated number of 1s
3	
4	
8	

★ **SOLUTION:** 2; 2; 4

(e) (2 points) What is the maximum error rate if the maximum number of buckets with the same size is 4? _____

★ **SOLUTION:** 25%

What to submit: For parts (a)-(c), please draw boxes around the numbers to indicate the state. For part (d) you should fill in the table, and for part (e) you should fill in the blank.

15 Data Streams on Networkly (5 points)

Alice is presented with a stream of randomly chosen user ids on Networkly (which are stored as 64-bit integers), and wants to use the Flajolet-Martin algorithm to count the number of distinct user ids in the stream. Alice hashes each element in the stream to at least 64 bits, using the hash function

$$h(a) = \text{base-2 representation of } a$$

However, Alice decides to count the number of leading zeroes in each hashed element, instead of the number of trailing zeros.

Describe a common way of assigning user ids to users that would cause this approach to fail. Explain why this algorithm would produce inaccurate counts in this situation.

What to submit: A one-paragraph explanation that outlines (1) the approach for assigning user ids to users, and (2) why this algorithm would produce inaccurate counts for this user id generation scheme.

★ **SOLUTION:** If user ids were contiguous integers starting from 0, it is likely that small numbers would be overrepresented compared to large numbers, and therefore many more than $1/2$ of all ids would have 1 leading zero, and more than $1/4$ of ids would have 2 leading zeros, etc. In general this would tend to overestimate the counts of users.

16 Generalized BALANCE Algorithm (8 points)

Recall the Generalized BALANCE Algorithm where for each bidder we have

- Bid = x_i
- Budget = b_i
- Amount spent so far = m_i
- Fraction of budget remaining $f_i = 1 - \frac{m_i}{b_i}$

We then define $\psi_i(q) = x_i(1 - \exp(-f_i))$, and allocate query q to the bidder i with the largest value of $\psi_i(q)$. Ties should be broken alphabetically; if bidders A and B both have the same value of $\psi_i(q)$, we should choose bidder A.

Also recall the regular BALANCE Algorithm where for each query, we pick the advertiser with the largest unspent budget. Ties should be broken alphabetically; if bidders A and B both have the same unspent budget, we should choose bidder A.

Now let's say we have the following advertisers:

- A bids \$4 each on queries x and y
- B bids \$5 each on queries y and z
- C bids \$3 each on queries x, y, and z

Each advertiser begins with the same budget of \$12.

We have the query stream '**z x y y z**'.

(a) (4 points) What is the revenue yield for the Generalized BALANCE algorithm? _____

★ **SOLUTION:** Generalized: 21 (B A B A C)

(b) (2 points) What is the revenue yield for the regular BALANCE algorithm? _____

★ **SOLUTION:** Regular: 20 (B A C C B)

(c) (2 points) What is the optimal yield? _____

★ **SOLUTION:** Optimal: 22 (B A A A B)

What to submit: Please fill in the blanks.

17 Optimizing Submodular Functions (4 points)

Let f, g be submodular functions. Which of the following functions must also be submodular?

(a) $f + g$

(b) $f - g$

(c) $\max(f, g)$

(d) $\min(f, g)$

What to submit: Please circle one of the answers (a)-(d) on the page.

★ **SOLUTION:** Only a.

[This space may be used for scratch paper.]

[This space may be used for scratch paper.]

[This space may be used for scratch paper.]

[This space may be used for scratch paper.]