Course Home Page

Quick Questions 15

COURSE

Announcements

Video Lectures

Video Errata

EXERCISES

Quizzes (Homeworks)

Surveys

Final (Basic)

Help

The **due date** for this exam is **Mon 1 Dec 2014 12:01 AM PST**.

This is the Basic Final Exam for the MMDS Course. All students are expected to take this exam. You must submit your work within 3 hours of opening. The exam is open-book; any inanimate source may be used. There is no penalty for a wrong answer (compared with no answer), so feel free to guess.

☐ **In accordance with the Coursera Honor Code, I (Tommi Hovi) certify that the answers here are my own work.**

Question 1

How many distinct 3-shingles are there in the string "hello world"? (Note: the quotes are not part of the string.)

○ 9

ABOUT THE COURSE

Syllabus

Grading and Logistics

COMMUNITY

Discussion Forums

Course Wiki

Join a Meetup

Help Articles

Course Materials Errors
Technical Issues

- ○ 4
- ○ 10
- ○ 6

Question 2

Here is a column representing a set, whose minhash value we wish to compute. The hash function we shall use to determine the order of rows is h(x) = (3x+2) modulo 11.

Row Number	Column Value
1	0
2	1
3	0
4	1
5	0

What is the minhash value for this column? Note: take the minhash value to be the row number, NOT the hash value of the row number.

- ○ 1
- ○ 2
- ○ 3
- ○ 4

- ○ 5

Question 3

This question involves three different Bloom-filter-like scenarios. Each scenario involves setting to 1 certain bits of a 10-bit array, each bit of which is initially 0.

Scenario A: we use one hash function that randomly, and with equal probability, selects one of the ten bits of the array. We apply this hash function to four different inputs and set to 1 each of the selected bits.

Scenario B: We use two hash functions, each of which randomly, with equal probability, and independently of the other hash function selects one of the of 10 bits of the array. We apply both hash functions to each of two inputs and set to 1 each of the selected bits.

Scenario C: We use one hash function that randomly and with equal probability selects two **different** bits among the ten in the array. We apply this hash function to two inputs and set to 1 each of the selected bits.

Let a, b, and c be the expected number of bits set to 1 under scenarios A, B, and C, respectively. Which of the following correctly describes the relationships among a, b, and c?

- ○ a = b = c
- ○ a < b = c
- ○ a < b < c

- a = b < c

Question 4

In this market-basket problem, there are 99 items, numbered 2 to 100. There is a basket for each prime number between 2 and 100. The basket for p contains all and only the items whose numbers are a multiple of p. For example, the basket for 17 contains the following items: {17, 34, 51, 68, 85}. What is the support of the pair of items {12, 30}?

- 2
- 3
- 4
- 5

Question 5

To two decimal places, what is the cosine of the angle between the vectors [2,1,1] and [10,-7,1]?

- 0.84
- 0.65
- 0.47
- -0.38

Question 6

In this question we use six minhash functions, organized as three bands of two rows each, to identify sets of high Jaccard similarity. If two sets have Jaccard similarity 0.6, what is the probability (to two decimal places) that this pair will become a candidate pair?

- ○ 0.64
- ○ 0.26
- ○ 0.74
- ○ 0.36

Question 7

Suppose we have a (.4, .6, .9, .1)-sensitive family of functions. If we apply a 3-way OR construction to this family, we get a new family of functions whose sensitivity is:

- ○ (.4, .6, .973, .729)
- ○ (.4, .6, .973, .271)
- ○ (.4, .6, .999, .729)
- ○ (.4, .6, .999, .271)

Question 8

Suppose we have a database of (Class, Student, Grade) facts, each giving the grade the student got in the class. We want to estimate the fraction of students who have gotten A's in at least 10 classes, but we do not want to examine the entire relation, just a sample of 10% of the tuples. We shall hash tuples to 10 buckets, and take only those tuples in the first bucket. But to get a valid estimate of the fraction of students with at least 10 A's, we need to pick our hash key judiciously. To which Attribute(s) of the relation should we apply the hash function?

- ○ Student only
- ○ Class only
- ○ Student and Class
- ○ Class and Grade

Question 9

Suppose the Web consists of four pages A, B, C, and D, that form a chain

A-->B-->C-->D

We wish to compute the PageRank of each of these pages, but since D is a "dead end," we will "teleport" from D with probability 1 to one of the four pages, each with equal probability. We do not teleport from pages A, B, or C. Assuming the sum of the PageRanks of the four pages is 1, what is the PageRank of page B, correct to two decimal places?

- 0.40
- 0.33
- 0.25
- 0.20

Question 10

Suppose in the AGM model we have four individuals (A,B,C,D} and two communities. Community 1 consists of {A,B,C} and Community 2 consists of {B,C,D}. For Community 1 there is a 30% chance it will cause an edge between any two of its members. For Community 2 there is a 40% chance it will cause an edge between any two of its members. To the nearest two decimal places, what is the probability that there is an edge between B and C?

- 0.58
- 0.40
- 0.70
- 0.42

Question 11

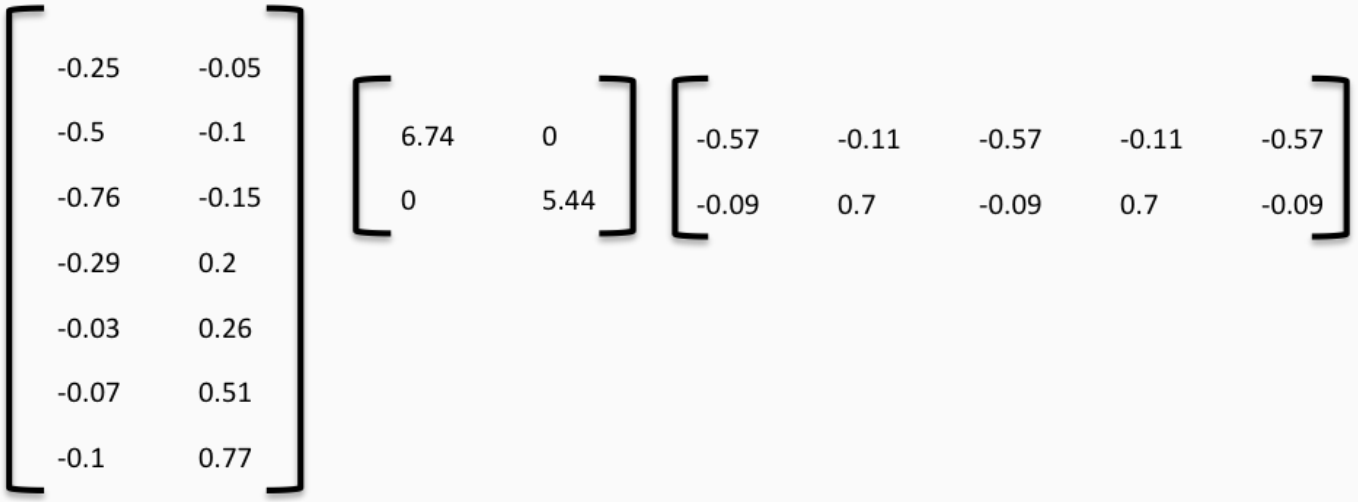
X is a dataset of n columns for which we train a supervised Machine Learning algorithm. e is the error of the model measured against a validation dataset. Unfortunately, e is too high because

model has overfitted on the training data X and it doesn't generalize well. We now decide to reduce the model variance by reducing the dimensionality of X, using a Singular Value Decomposition, and using the resulting dataset to train our model. If i is the number of singular values used in the SVD reduction, how does e change as a function of i, for i $\in$ {1, 2,...,n}?

- ○ e starts low, then increases, then decreases.
- ○ e starts high, then decreases.
- ○ e starts high, then decreases, then increases.
- ○ e starts low, then decreases
- ○ e doesn't change.

Question 12

A is a users times movie-ratings matrix like the one seen in class. Each column in A represents a movie, and there are 5 movies in total. Recall that a Singular Value Decomposition of a matrix is a multiplication of three matrices: U, Σ and V. The following is such a decomposition for matrix A:



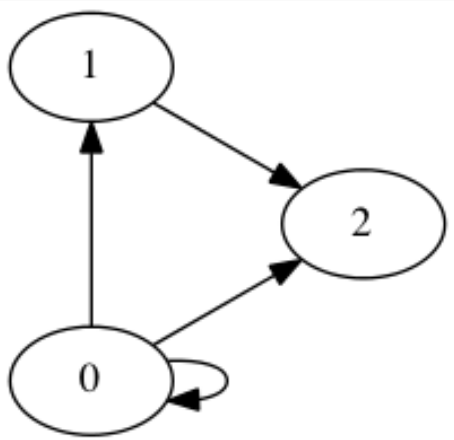
What is the cosine similarity between a user with ratings [5,0,0,0,0] and a user with ratings [0,2,0,0,4] using their concept space vectors? (round your answer to two decimals. For example, if your answer is 0.345 the rounded answer is 0.35. If your answer is 0.344, the rounded answer is 0.34.)

Question 13

Recall that the power iteration does r=X·r until converging, where X is a nxn matrix and n is the number of nodes in the graph.

Using the power iteration notation above, what is matrix X value when solving topic sensitive Pagerank with teleport set {0,1} for the following graph? Use beta=0.8. (Recall that the teleport

set contains the destination nodes used when teleporting).



○

11/30	1/10	1/10
11/30	1/10	1/10
4/15	8/10	0

○

1/3	0	0
1/3	0	0
1/3	1	0

○

8/30	0	0
8/30	0	0
8/30	8/10	0

○

1/6	1/2	1/2
1/6	1/2	1/2
1/3	1	0

Question 14

Here are two sets of integers S = {1,2,3,4} and T = {1,2,5,6,x}, where x stands for some integer. For how many different integer values of x are the Jaccard similarity and the Jaccard distance of S and T the same? (Note: x can be one of 1, 2, 5, or 6, but in that case T, being a set, will contain x only once and thus have four members, not five.)

- ○ 6
- ○ 2
- ○ An infinite number
- ○ 4

Question 15

Which of the following are advantages of using decision trees? (check all correct options)

- ☐ It avoids overfitting
- ☐ It can handle multiple output easily
- ☐ It can handle categorical input data without any special preprocessing
- ☐ The resulting model is easy to interpret
- ☐ The training is easy to parallelize

Question 16

The hard margin SVM optimization problem is:

$$\begin{aligned} &\textit{minimize } \tfrac{1}{2}\|w\|^2 \\ &s.t.\ y_i \cdot (x_i \cdot w + b) \geq 1, \quad \forall i = 1, \ldots, n \end{aligned}$$

and the soft margin SVM optimization problem is:

$$\begin{aligned} &\textit{minimize } \tfrac{1}{2}\|w\|^2 + C \sum_{i=1}^{n} \xi_i \\ &s.t.\ y_i \cdot (x_i \cdot w + b) \geq 1 - \xi_i, \quad \forall i = 1, \ldots, n \\ &\xi_i \geq 0, \ \forall i = 1, \ldots, n \end{aligned}$$

Consider a dataset of points $x_i, \ldots x_n$ with labels $y_i, \ldots, y_i \in \{-1, 1\}$, such that the data is separable. We run a soft-margin SVM and a hard-margin SVM, and in each case we obtain parameters w and b. Check the option that is true:

- ○ The resulting w and b can be different, and the boundaries can be different.
- ○ The resulting w and b are the same in the two cases, hence boundaries are the same.
- ○ The resulting w and b can be different in the two cases, but the boundaries are the same.
- ○ None of the above.

Question 17

Consider the following MapReduce algorithm. The input is a collection of positive integers. Given integer X, the Map function produces a tuple with key Y and value X for each prime divisor Y of X. For example, if X = 20, there are two key-value pairs: (2,20) and (5,20). The Reduce function, given a key K and list L of values, produces a tuple with key K and value sum(L) i.e., the sum of the values in the list. Given the input 9, 15, 16, 23, 25, 27, 28, 56 which of the following tuples appears in the final output?

- ○ (3, 51)
- ○ (5, 51)
- ○ (4, 51)
- ○ (7, 51)

Question 18

Suppose we run K-means clustering over the following set of points in 2-d space using the L_1 distance metric: (1,1), (2,1) (2,2), (3,3), (4,2), (2,4), (4,4). We pick k=2 and the initial centroids are (1,1) and (4,4). Which of these is the centroid of the cluster containing the point (3,3) when the algorithm terminates?

Recall that the L_1 distance between two points is the sum of their distances along each dimension, e.g. the L_1 distance between (1, 2) and (-1, 3) is 3.

- ○ (4, 4)

- ○ (13/4, 13/4)
- ○ (5/3, 4/3)
- ○ (3, 3)

Question 19

In an implementation of the Bradley-Fayyad-Reina (BFR) algorithm over a 3-dimensional data set, the discard set for a cluster is summarized by the following parameters:

N = 1000

SUM = (-323, 1066, 1776)

SUMSQ = (412, 1500, 3500)

Which of the following choices is closest to the Mahalanobis distance of the point (0,0,0) from the centroid of this cluster?

- ○ 2.55
- ○ 1.55
- ○ 3.55
- ○ 4.55

Question 20

Consider an execution of the BALANCE algorithm with 4 advertisers, A1, A2, A3, A4, and 4 kinds of queries, Q1, Q2, Q3, Q4. Advertiser A1 bids on queries Q1 and Q2; A2 bids on queries Q2 and Q3; A3 on queries Q3 and Q4; and A4 on queries Q1 and Q4. All bids are equal to 1, and all clickthrough rates are equal. All advertisers have a budget of 3, and ties are broken in favor of the advertiser with the lower index (e.g., A1 beats A2). Queries appear in the following order:

Q1, Q2, Q3, Q3, Q1, Q2, Q3, Q1, Q4, Q1

Which advertiser's budget is exhausted first?

- ○ A4
- ○ A2
- ○ A1
- ○ A3

Question 21

Consider the bipartite graph with the following edges (you might want to draw a picture):

(a,1), (a,3), (b,1), (b,2), (b,4), (c,2), (d,1), (d,4)

Which of the following edges appears in NO perfect matching?

- ○ (a, 1)
- ○ (d, 4)
- ○ (b, 4)
- ○ (c, 2)

Question 22

The Utility Matrix below captures the ratings of 5 users (A,B,C,D,E) for 5 movies (P,Q,R,S,T). Each known rating is a number between 1 and 5, and blanks represent unknown ratings. What is the Pearson Correlation (also known as the Centered Cosine) between users B and D?

	P	Q	R	S	T
A	2		4		
B		3	1	2	
C	5			5	
D		4	3		2
E	4			5	1

- ○ 0.74
- ○ 0.23
- ○ 0.96

○ 0.5

Question 23

The Utility Matrix below captures the ratings of 5 users (A,B,C,D,E) for 5 movies (P,Q,R,S,T). Each known rating is a number between 1 and 5, and blanks represent unknown ratings. Let (U,M) denote the rating of movie M by user U. We evaluate a Recommender System by withholding the ratings (A,P), (B,Q), and (C,S). The recommender system estimates (A,P)=1, (B,Q)=4, and (C,S)=5. What is the RMSE of the Recommender System, rounded to 2 decimal places?

	P	Q	R	S	T
A	2		4		
B		3	1	2	
C	5			5	
D		4	3		2
E	4			5	1

○ 0.82

○ 2.36

○ 0.0

○ 1.44

Question 24

We are going to perform a hierarchical (agglomerative) clustering on the four strings {he, she, her, their}, using edit distance (just insertions and deletions; no mutations of characters). Initially, each string is in a cluster by itself. The distance between two clusters is the **minimum** edit distance between two strings, one chosen from each of the two clusters. When we complete the hierarchical clustering, there is one cluster containing all four strings, and we performed three mergers of clusters to get to that point. For each of the three mergers there was a distance between the merged clusters. What is the sum of those three distances?

- ○ 3
- ○ 4
- ○ 5
- ○ It depends on how we break ties when there are two pairs of clusters at the same distance.

☐ **In accordance with the Coursera Honor Code, I (Tommi Hovi) certify that the answers here are my own work.**

Submit Answers Save Answers

You cannot submit your work until you agree to the Honor Code. Thanks!

Time remaining
2:57:35