NAME	NIA	GRADE

Mining of Massive Datasets (sample)

FINAL EXAM

WRITE YOUR ANSWERS CLEARLY IN THE BLANK SPACES. Please read carefully each question. If you do not understand the question, ask for a clarification. Please underline your final answer (numerical or symbolic questions) or important key words or phrases (text questions). You can attach an extra sheet to your exam. In this case, indicate clearly that the solution can be found in the extra sheet.

Problem 1 *0.4 point*

What is the difference between *explicit* and *implicit* preferences in an utility matrix? Give examples of each one.

The difference is (give examples):

Problem 2 *0.6 point*

About the *cold-start problem* of recommender systems:

Define the cold-start problem in recommender systems:

Name one possible way to tackle this problem:

Problem 3 *1 point*

Consider the utility matrix below, which includes the preferences of two users (u, v) on three items, based on two attributes (a1, a2). Suppose we build a content-based recommender system in which the rating is a linear function on the attributes.

	Attributes		User ratings	
Item	a1	a2	u	v
A	1	0	$+1$	-2
B	0	1	-1	$+1$
C	1	1	X	Y

What is your guess of the model of ratings of user u given a1, a2?

What is your guess of the value of X (the rating u would give to item C)?

What is your guess of the model of ratings of user v given a1, a2?

What is your guess of the value of Y (the rating v would give to item C)?

Problem 4 *2 points*

Consider a recommender system based on user similarity, such as the one we saw in class, for which the similarity between users u and v is given by:

$$\text{sim}(u,v) = \frac{\sum_{i \in I_{u,v}} (u_i - \hat{u}) \cdot (v_i - \hat{v})}{\sqrt{\sum_{i \in I_{u,v}} (u_i - \hat{u})^2 \cdot \sum_{i \in I_{u,v}} (v_i - \hat{v})^2}}$$

where u_i and v_i are the ratings given respectively by users u and v to item i, $\hat{u}$ and $\hat{v}$ are the average scores given by users u and v, and $I_{u,v}$ is the set of items rated by users u and v.

Consider that the estimated rating of user u on item i is given by:

$$\text{score}(u,i) = \hat{u} + \frac{\sum_{v: v_i \neq \text{NULL}} \text{sim}(v,u) \cdot (v_i - \hat{v})}{\sum_{v: I_{u,v} \neq \emptyset} |\,\text{sim}(v,u)|}$$

Consider the utility matrix of 3 users (u1, u2, u3) on 4 movies (m1, m2, m3, m4) as shown on the next table.

	m1	m2	m3	m4
u1	1	1	1	-2
u2	-1		-1	2
u3	2	-1		

Please draw a box around your final values when answering the following.

What are the values of $\widehat{u1}$, $\widehat{u2}$, $\widehat{u3}$*?*

$\widehat{u1} =$ $\widehat{u2} =$ $\widehat{u3} =$

What is $I_{u1,u3}$*? What is* $sim(u1,u3)$*?*

$I_{u1,u3} = \{\quad\}$ $\text{sim}(u1,u3) =$

What is $I_{u2,u3}$*? What is* $sim(u2,u3)$*?*

$I_{u2,u3} = \{\quad\}$ $\text{sim}(u2,u3) =$

What is $v : v_{m3} \neq \text{NULL}$*? What is* $v : I_{u3,v} \neq \emptyset$*? What is* $score(u3,m3)$*?*

$v : v_{m3} \neq \text{NULL} = \{\quad\}$ $v : I_{u3,v} \neq \emptyset = \{\quad\}$

$\text{score}(u3,m3) =$

What is $v : v_{m4} \neq \text{NULL}$*? What is* $v : I_{u3,v} \neq \emptyset$*? What is* $score(u3,m4)$*?*

$v : v_{m4} \neq \text{NULL} = \{\quad\}$ $v : I_{u3,v} \neq \emptyset = \{\quad\}$

$\text{score}(u3,m4) =$

Problem 5 *1 point*

When we use non-negative matrix factorization for recommender systems, we want to factorize the utility matrix D into two matrices U and V such that $D \approx UV^T$.

What is the objective function that the factorization tries to minimize? Define all variables that you use.

What are the constraints of this minimization?

Once U and V are obtained, how do you estimate the score that user i will give to item j?

Problem 6 *1 point*

Consider the following four items with two attributes:

	a1	a2	z-score of a1	z-score of a2
1	1	100		
2	2	50		
3	0	150		
4	1	30		
μ				
σ				

Use the method based on z-scores seen in class to do the following. Remember $\mu = (\sum x_i)/N$ and $\sigma = \sqrt{\sum (x_i - \mu)^2 / N}$:

Complete μ, σ, and the z-scores of a1 and a2 in the above table, with 4 decimal digits

Which item is the outlier, and because of which attribute?

Problem 7 *1 point*

Suppose you do reservoir sampling with a reservoir of size 2 to store a stream whose first 5 elements are $\{a, b, c, d, e\}$.

Suppose you use a function $r(i, j)$ which is supposed to return a random integer between i and j, both inclusive. Whenever you want to decide if an element has to be inserted in the reservoir with probability x/y, you insert the element if $r(1, y) \leq x$, and whenever you want to decide which element to eject from the reservoir, you eject element in position $r(1, 2)$.

Now, suppose there is an implementation error, and $r(i, j)$ always returns i.

Which elements are in the reservoir after $\{a, b, c\}$ have been read? Explain.

Which elements are in the reservoir after $\{a, b, c, d\}$ have been read? Explain.

Which elements are in the reservoir after $\{a, b, c, d, e\}$ have been read? Explain.

Problem 8 *1 point*

Given the following series: $1.00, 5.00, 67.00, 10.00, 23.00$, and using only two decimals, answer the following:

The smoothed series using moving averages and a window size of 2 is:

The smoothed series using moving averages and a window size of 3 is:

Problem 9 *2 points*

Consider the series $X = 1, 1, 2, 4, 3$ and $Y = 1, 3, 3, 2$. Apply Dynamic Time Warping (DTW) using the following space for your computations:

	X1=1	X2=1	X3=2	X4=4	X5=3
Y1=1					
Y2=3					
Y3=3					
Y4=2					

Complete the table above, marking with circles or with color the alignment path, as seen in class

Indicate to which element(s) of the Y series the following should be aligned.

- Item X1 should be aligned with:
- Item X2 should be aligned with:
- Item X3 should be aligned with:
- Item X4 should be aligned with:
- Item X5 should be aligned with: