

<b>NAME</b>	<b>NIS (uXXXXX) or NIA</b>	<b>GRADE</b>
-------------	----------------------------	--------------

## MINING OF MASSIVE DATASETS (2022-2023)

### ————— FINAL EXAM —————

**WRITE YOUR ANSWERS BRIEFLY and CLEARLY IN THE BLANK SPACES.** PLEASE UNDERLINE KEY WORDS IN YOUR ANSWERS. PLEASE IF YOU INCLUDE INTERMEDIATE CALCULATIONS, CIRCLE THE FINAL RESULT. IF NEEDED, 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

*1 point*

*Describe an advantage of content-based recommender systems over interaction-based recommender systems:*

*Describe an advantage of interaction-based recommender systems over content-based recommender systems:*

#### Problem 2

*1 point*

Consider the utility matrix below, which includes the preferences of two users ( $u$ ,  $v$ ), on a series of items (A, B, C, ...) described by three attributes ( $a_1$ ,  $a_2$ ,  $a_3$ ). Suppose we build a content-based recommender system in which the rating is a linear function on the attributes.

Item	Attributes			User ratings	
	$a_1$	$a_2$	$a_3$	$u$	$v$
A	1	0	0	+1	-2
B	0	1	0	-1	0
C	0	0	1	0	+1
D	1	0	1	+1	-1
E	0	1	1	-1	+1
F	1	1	1	X	Y

*What is your guess of the model of ratings of user  $u$  given  $a_1$ ,  $a_2$ ,  $a_3$ ?*

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

*What is your guess of the model of ratings of user  $v$  given  $a_1$ ,  $a_2$ ,  $a_3$ ?*

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

**Problem 3**

2 points

Consider the following utility matrix  $V$ , that we use as input to build recommendations based on **latent factors**.

<b>V</b>	Julia	Emma	Pol	Leo
Sweets	0	2	0	1
Chips	1	1	0	0
Veggies	1	0	1	2

We apply non-negative matrix factorization, and obtain matrices  $W$ ,  $H$ , where A and B are two latent factors.

<b>W</b>	Julia	Emma	Pol	Leo
A	1.38	0.00	1.38	2.75
B	0.33	1.67	0.00	0.67

<b>H</b>	A	B
Sweets	0.00	1.19
Chips	0.00	0.60
Veggies	0.73	0.00

For which two users will the recommendations be more accurate, and for which two users will the recommendations be less accurate? Justify your answer in terms of reconstruction error.

*Recommendations would be **more accurate** for users:*

*Recommendations would be **less accurate** for users:*

*Reconstruction error calculations:*

**Problem 4**

2 points

Consider a dataset of two attributes  $X$ ,  $Y$ , composed of the five data points A, B, C, D, E.

	X	Y
A	0.2	0.3
B	0.1	0.4
C	0.2	0.6
D	1.2	0.9
E	1.0	1.0

We have clustered this dataset and determined that cluster 1 contains points A, B, C, and cluster 2 contains points D, E. Use the method seen in class for clustering-based outlier detection, and indicate which element is the outlier.

*The outlier is point:*

*The centroid of cluster 1 is:*

*The centroid of cluster 2 is:*

*Distance calculations:*

**Problem 5**

1 point

We want to create a bloom filter using  $h$  hashes in a table of  $M$  bits, to be initialized with  $N$  elements.

*What is the probability that the first bit of the table remains 0 after the initialization? Justify.*

**Problem 6**

2 points

Consider the following temperature readings:

Day	Time	Temperature [°C]
Monday	03:00	6
Monday	06:00	8
Tuesday	06:00	12
Wednesday	18:00	18

Complete the following table with linear interpolations.

Use the space on the right for calculations.

Day	Time	Temperature [°C]
Monday	12:00	
Tuesday	12:00	
Wednesday	12:00	

**Problem 7**

1 point

Assume we have created an autoregressive model for a time series  $x_t = 3 \times x_{t-1} - x_{t-2}$ . When performing **multi-step** forecasting we assume that the prediction will be perfect, and use predicted points as if they were actual inputs.

Perform multi-step forecasting in the following series to predict points  $x_3$ ,  $x_4$ , and  $x_5$ .

$x_1$	$x_2$	$x_3$	$x_4$	$x_5$
2	4			

Justify your answer by providing calculations: