NAME	**NIS (uXXXXX) or NIA**	**GRADE**

MINING OF MASSIVE DATASETS (2023-2024)

———— *FINAL EXAM* ————

WRITE YOUR ANSWERS <u>BRIEFLY</u> and <u>CLEARLY</u> 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*

What is white noise?

Give an example of a dataset containing an outlier that is not an extreme value:

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 (a1, a2, a3). 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	a3	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 a1, a2, a3?

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 a1, a2, a3?

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**.

V	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.

W	Julia	Emma	Pol	Leo
A	1.38	0.00	1.38	2.75
B	0.33	1.67	0.00	0.67

H	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	0.3	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.

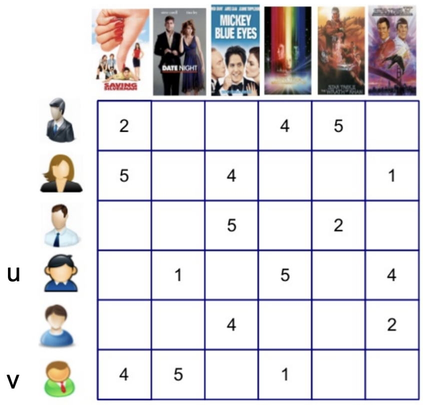
The outlier is point:

The centroid of cluster 1 is:

The centroid of cluster 2 is:

Distance calculations:

Problem 5 *1 point*



	2			4	5	
	5		4			1
			5		2	
u		1		5		4
			4			2
v	4	5		1		

a) Given the above utility matrix, compute the similarity between users u and v:
sim(u,v) =

The similarity between users u and v is given by:

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

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.

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

b) Suppose you have computed all similarities of users to u. Explain how do you recommend movies to user u using the formula above:

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	15:00	
Tuesday	03:00	
Wednesday	09:00	

Problem 7 *1 point*

Assume we have created an autoregressive model for a time series $x_t = 2 \times x_{t-1} - x_{t-2}$. 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: