NAME	uXXXX	GRADE

MINING OF MASSIVE DATASETS (2024-2025)

FINAL EXAM

WRITE YOUR ANSWERS BRIEFLY and CLEARLY IN THE BLANK SPACES. If you do not know the answer to a question, leave it blank. No points are awarded for what does not answer the question being asked. Please underline key words in your answers, and when including intermediate calculations, circle the final result. If needed, you can attach an extra sheet. In this case, indicate clearly that the solution is in the extra sheet.

Problem 1 *2 points*

Consider the following partially defined utility matrix describing the ratings of 4 users on 3 items:

D	I_1	I_2	I_3
U_1	5.00	4.00	1.00
U_2	1.00		4.00
U_3		2.00	5.00
U_4		1.00	

We have performed non-negative matrix factorization to approximate $D \approx UV^T$. The resulting matrices are:

U	x_1	x_2
U_1	0.34	2.18
U_2	1.68	0.08
U_3	2.09	0.55
U_4	0.18	0.52

V	x_1	x_2
I_1	0.44	1.96
I_2	0.45	1.54
I_3	2.08	0.10

where x_1 and x_2 are the latent factors. For the following computations, use just two digits of precision:

(1p) Compute the reconstruction error for U_2 and U_3 as Root Mean Square Average (RMSE). RMSE is the square root of the average of the square differences. Indicate for which of the two users the approximation is better.

(1p) Compute the predicted ratings for U_4 on the items that user has not rated. Indicate which of those items you would recommend.

Problem 2 *2 points*

Regarding outliers.

(1p) What is an outlier? Provide a definition, not an example.

(0.5p) Some outliers are not extreme values. Provide an example of an outlier that is not an extreme value.

(0.5p) In the isolation forest method, we claim that a point is an outlier if it is close to the root of many trees. Explain briefly and precisely why. Do not give an example, and do not draw an isolation forest.

Problem 3 *2 points*

Consider the reservoir sampling method with a reservoir of size $s = 3$, and the input sequence $\langle a, b, c, d, e, f, \dots \rangle$.

After processing the fifth element, e in the input, compute the following probabilities:

$p_5(e)$, *the probability that element e is in the sample:*

$p_5(d)$, *the probability that element d is in the sample:*

$p_5(c)$, *the probability that element c is in the sample:*

$p_5(b)$, *the probability that element b is in the sample:*

$p_5(a)$, *the probability that element a is in the sample:*

Problem 4 *1 point*

Consider a Bloom filter of size n, in which we will use k hash functions.

Write the expression for the number of cells that are "on" (i.e., set to 1) after inserting m keys. Justify briefly your answer. You can compute directly if you find it easier, or consider that n is large and use the approximation $(1-\epsilon)^{(1/\epsilon)} \approx 1/e$ for small ϵ.

Problem 5 *1 point*

Remove the linear trend in this series:

t	1	2	3	4	5	6	7
$x(t)$	1.6	2.9	4.6	5.9	7.6	8.9	10.6

(0.5p) Write the expression for the underlying trend:

(0.5p) Write the series with the trend removed:

Problem 6 *2 points*

Consider a bivariate time series $\langle x, y \rangle$ in which we want to perform an autoregressive forecast with a single lag: $x_{t+1} = \alpha x_t + \beta y_t + \gamma$, $y_{t+1} = \kappa x_t + \lambda y_t + \mu$. Consider that all the coefficients are drawn from the set $\{-1, 1\}$. Given that there are three coefficients per model, that means there are $2^3 = 8$ possible equations for each forecasting model.

t	x_t	y_t
1	3	2
2	4	0
3	3	-3
4	-1	-5
5	-7	-3
6	-11	5

(0.5p) Guess the model for the first variable: $x_{t+1} =$

(0.5p) Guess the model for the second variable: $y_{t+1} =$

(1p) Perform multi-step forecasting (i.e., assuming the predictions are real and use them to further predict) to determine x_7, y_7, x_8, y_8. Write your calculations.

t	x_t	y_t
7		
8		