# Mining of Massive Datasets [Final Exam]

Please use your @upf.edu e-mail address. Remember to submit frequently to avoid losing your work.

* Required

1. Email *

2. Q1 [1 point]. What is the difference between explicit and implicit utility matrices in recommender systems?

3. Q2 [1 point]. In a regression-based approach to recommendations in which there are N items that are described by K features, we need as input ...

   *Mark only one oval.*

   - At least K ratings in total
   - At least N ratings in total
   - At least K ratings per user
   - At least N ratings per user

Figure 1: User ratings for a regression-based recommender system

| | Is it an action movie? | Is it an adventure movie? | Rating by user *u* |
|---|---|---|---|
| Movie 1 | y | n | Liked |
| Movie 2 | y | y | Disliked |
| Movie 3 | y | y | Not rated |
| Movie 4 | n | n | Not rated |
| Movie 5 | n | n | Liked |
| Movie 6 | n | y | Disliked |

4. Q3 [1 point]. Suppose you are given the user ratings on Figure 1 and want to use the regression-based approach to recommender systems. Suppose you encode "action movie" and "adventure movie" with a binary numerical variable so that 1=yes and 0=no. The coefficients will be ...

*Mark only one oval per row.*

| | A negative number | Zero | A positive number |
|---|---|---|---|
| Action movie coefficient | ◯ | ◯ | ◯ |
| Adventure movie coefficient | ◯ | ◯ | ◯ |

Figure 2. Utility matrix V, which is factorized into matrices W and H

Matrix V

| | Jordi | Xavier | Paula | Aisha |
|---|---|---|---|---|
| **Broccoli** | 5 | 2 | 1 | 1 |
| **Tofu** | 3 | 5 | 1 | 1 |
| **Potatoes** | 3 | 2 | 0 | 2 |
| **Fish** | 0 | 1 | 1 | 2 |
| **Pork** | 0 | 0 | 4 | 0 |
| **Chicken** | 0 | 0 | 3 | 2 |

Matrix W

| | X1 | X2 |
|---|---|---|
| **Broccoli** | 1.74 | 0.20 |
| **Tofu** | 1.89 | 0.17 |
| **Potatoes** | 1.33 | 0.00 |
| **Fish** | 0.34 | 0.64 |
| **Pork** | 0.00 | 1.71 |
| **Chicken** | 0.07 | 1.55 |

Matrix H

| | Jordi | Xavier | Paula | Aisha |
|---|---|---|---|---|
| **X1** | 2.16 | 1.88 | 0.24 | 0.76 |
| **X2** | 0.00 | 0.00 | 2.11 | 0.69 |

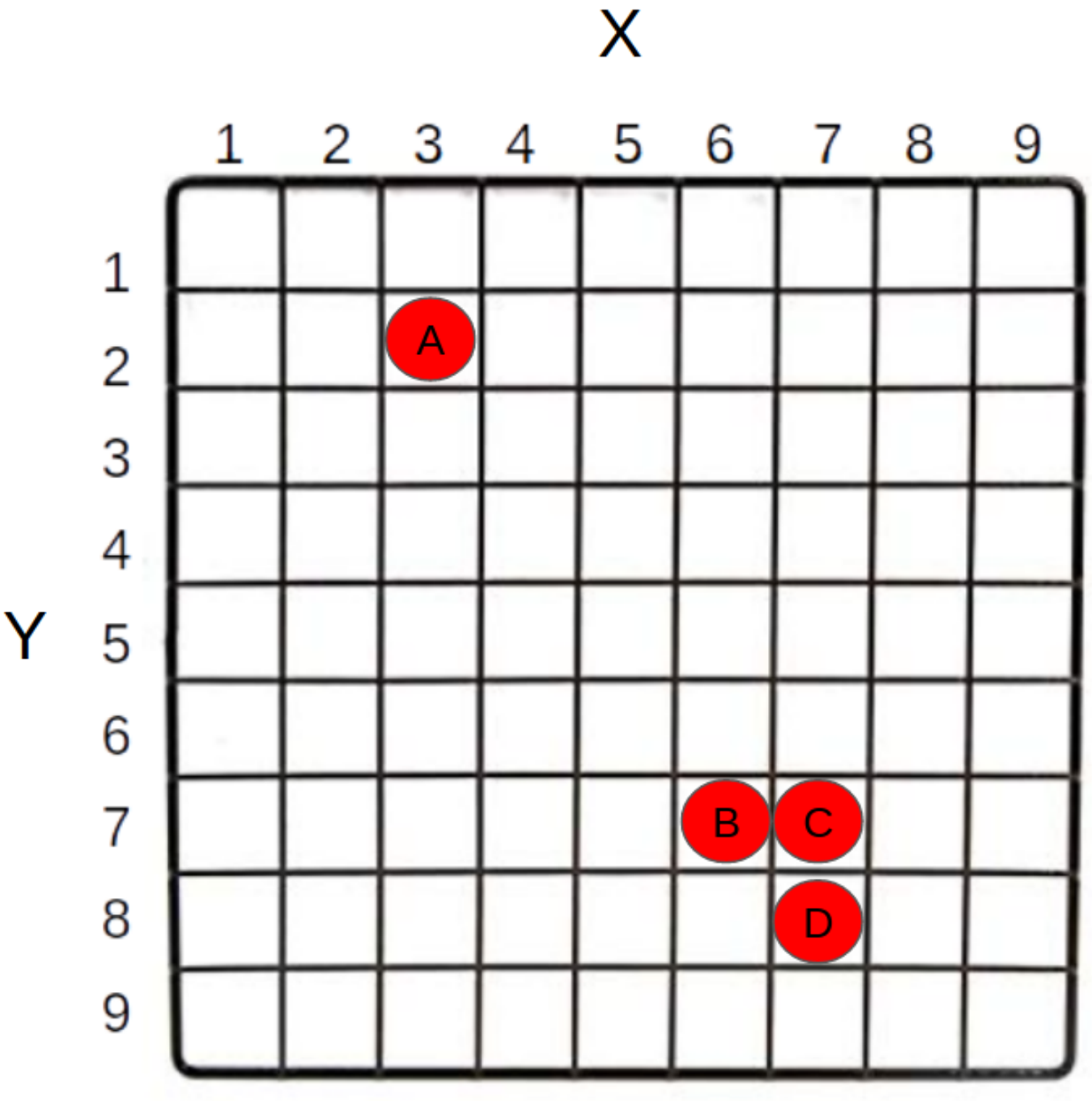
5. Q4 [1 point]. Suppose in a supermarket you are given user preferences (matrix V) factorized into two non-negative matrices (W and H), as expressed in Figure 2. How do we call X1 and X2, in general (not specifically in this example)?

6. Q5 [2 points]. Suppose in a supermarket you are given user preferences (matrix V) factorized into two non-negative matrices (W and H), as expressed in Figure 2. What would be the score of a recommendation of Fish to Xavier? What would be the score of a recommendation of Chicken to Xavier? Which of these two would you recommend to Xavier?

   Your answer should contain two numbers.

7. Q6 [1 point]. Suppose we have measured the wingspan of different individual birds of the same species, converted them into a z-score by standardization, and observe that for the largest bird in our sample the z-score of the wingspan is 5.0. What does that mean, precisely? Under which circumstances that would be enough to mark this bird as an outlier?

Figure 3. Data for isolation tree.



8. Q7 [2 points]. Upload a photo of the "cuts" you would make to create one isolation tree of maximum depth 2 on the data of Figure 3. Use a random number generator of your calculator, or simply make up the random numbers when you need them. (Note that a tree of three nodes in which one node is the root and the other two nodes are its children has depth 1.)

   Files submitted:

9. Q8 [1 point]. Indicate the depth of each of the elements in the photo. Your answer should have the form "depth(A) = number, depth(B) = number, ..., depth(D) = number"

10. Q9 [1 point]. What is the difference between a standing query and an ad-hoc query in a stream processing system?

11. Q10 [1 point]. You receive a stream of data about traffic tickets in a country. Each traffic ticket is a record of the form <timestamp, plate, description>. Plate numbers for cars are composed of a series of 4 digits followed by 3 letters. How would you randomly sample 1% of all vehicles and all of their tickets using a streaming sampling algorithm?

12. Q11 [2 points]. Suppose you want to do reservoir sampling and have a reservoir of 5 elements, which currently are r1, r2, r3, r4, r5. Suppose that element number 500 of the stream arrives. What is the probability that element r1 currently in the reservoir is evicted?

13. Q12 [2 points]. Why is the probability of false positives of a Bloom filter a convex function of the number of hash functions used?

14. Q13 [1 point]. A thermometer marked 10 degrees at 10:00 and 20 degrees at 13:00. What temperature would we have at 12:00 if we apply linear interpolation?

    Indicate a number with 2 decimals.

15. Q14 [1 point]. Given the series x = (2.0, 4.0, 10.0, 5.0, 3.0, 2.0), compute a new series y that should be the moving average of x with a window of size 2. Use one decimal for your numbers.

16. Q15 [2 points]. Suppose we did dynamic time warping of two series X=(x(1),x(2),...,x(n)) and Y=(y(1),y(2),...,y(m)) using distance function d(a,b) = |a-b|, and found that the best mapping was that x(i) should be aligned with y(p(i)) where p(i) is a function from 1...n into 1...m. To obtain this mapping, we built a matrix in a certain ordering. Let's say we started by the bottom-left corner and end in the top-right corner. Now there is a number M in the top-right corner. What is M, exactly (as a function of the variables we have defined)?

17. Q16 [2 points]. Let's assume you're given the input series x(t) = (1, 4, 4, 7, 10, 16). You want to build a linear autoregressive (AR) model for this series with lags 1 and 2, which we will represent as y(t). We would like this AR model to be such that y(t) = a * x(t-1) + b * x(t-2) + c, with a, b, and c coefficients of the model. Write the 4 approximate equalities that you would like this model to satisfy in this specific case. Guess the parameters a, b, and c, and indicate the error of the model.

    Files submitted:

---