# 2020 Mining of Massive Datasets / Midterm exam

To avoid losing work, note that you can submit and resubmit as many times as you want. Answer each question, then click submit, then click on edit your answer so come back to this screen.

Answer the questions that require a photo AT THE END -- do them on paper first, then upload them during the last 10 minutes of the exam. Those answers cannot be edited. Everything else can be edited.

* Required

1. Email *

2. Q1. Regarding nondependency vs dependency data [1 point]

   *Mark only one oval.*

   - Graphs are dependency oriented, tables are nondependency oriented
   - Graphs are nondependency oriented, tables are dependency oriented
   - Both graphs and tables are dependency oriented
   - Both graphs and tables are nondependency oriented

3. Q2. "Ordinal attributes are .... attributes" [1 point]

   Complete with a single word.

4. Q3. Suppose in a table you have 3 attributes: name, age, civil status. Suppose civil status can be single, married, divorced, widowed. If you use one-hot encoding for civil status, how many attributes would you have? [1 point]

5. Q4. If an attribute has values 5, 30, 50, 1000, and you want to discretize it into a categorical attribute with two values "low" and "high" using EQUI-WIDTH ranges, in which category will each of these go? [1 point]

   *Mark only one oval per row.*

| | Low | High |
|---|---|---|
| 5 | ◯ | ◯ |
| 30 | ◯ | ◯ |
| 50 | ◯ | ◯ |
| 1000 | ◯ | ◯ |

6. Q5. Indicate what is the range for Low and the range for High that you used in the previous question [1 point]

   Answer: Low=[___, ___], High=[___, ___]

7. Q6. Suppose in a database of purchases containing timestamp, customer-id, product-id, price you are missing one attribute. Suppose you want to run association rules mining to find products co-purchased together. What should you do in each case? [1 point]

   Timestamp is the moment the purchase was made, customer-id identifies the customer, product-id identifies the product, price is the price paid.

   *Mark only one oval per row.*

| | Impute the value | Delete the record |
|---|---|---|
| Missing timestamp | | |
| Missing product-id | | |
| Missing customer-id | | |
| Missing price | | |

8. Q7. Suppose you have customers A, B, C, D, E, F, in which A, B, C are men, C, D, F are women. Suppose you want to do a STRATIFIED SAMPLE by gender in which the sample will contain only four elements. Indicate a valid sample from this set. [1 point]

9. Q8. Suppose you have cars with an ordinal attribute indicating fuel efficiency, which from most efficient to least efficient is: A+, A, B+, B, C+, C. What is the ordinal SIMILARITY between two cars having fuel efficiency A+ and B? What is the ordinal SIMILARITY between two cars having fuel efficiency C+ and C? [1 point]

   Answer sim(A+,B) = _____; sim(C+, C) = _____; both should be numbers between 0 and 1.

10. Q9. Suppose you are given the following points: A=(1,4); B=(2,3); C=(3,3); D=(2,2); E=(3,2); F=(4,1). According to the Mahalanobis distances, which points are closer? [1 point]

    *Mark only one oval.*

    - Points B and E are closer
    - Points C and D are closer

11. Q10. What is the Jaccard DISTANCE between vectors x = (1, 2, 3, 0, 0) and y = (0, 1, 2, 0, 3)? [1 point]

12. Q11. What is the Tanimoto SIMILARITY between vectors x = (1, 2, 3, 0, 0) and y = (0, 1, 2, 0, 3)? [1 point]

13. Q12. What are the shingles (word 3-grams) in the phrases D1 = "A happy hippo hopped and hiccuped"; D2 = "The kangaroo hopped and hiccuped"; D3 = "The kangaroo hopped over a happy hippo" [1 point]

    Answer S(D1) = { ...}; S(D2) = {....}, S(D3) = {...}

14. Q13. Draw the binary shingles/document matrix for the exercise above. Indicate for each row the shingle it represents, and for each column the document it represents. Upload a photo of your matrix. [1 point]

    Remember to upload the photo at the END of the exam, in the last 10 minutes.

    Files submitted:

15. Q14. Consider now this other document matrix and these permutations (Image here: https://drive.google.com/file/d/15NiNu4l1SLKQBuObQ6_QFyCSbfz8MTs8/view?usp=sharing). Upload a photo of your signature matrix [2 points]

    Remember to upload the photo at the END of the exam, in the last 10 minutes.

    Files submitted:

16. Q15. What is the main difference between locality sensitive hashing (LSH) and normal hashing (i.e., the kind of hash functions used in cryptography)? [1 point]

17. Q16. What is the support monotonicity property? [1 point]

18. Q17. What is a closed itemset? [1 point]

19. Q18. Given transactions {A, B, C}, {A, C, D}, {B, C, D}, {A, B} what is the confidence of the rule A => D? [1 point]

20. Q19. Given the following transactions (image: https://drive.google.com/file/d/1VRyq5eGRzTZE1U09V4Q3LdkUUEZWwkyV/view?usp=sharing) ... draw (a) all 1-itemsets and their support, marking with an X the itemsets that do not satisfy the minsup criteria (b) all 2-itemsets and their support, also marking with an X the itemsets that do not satisfy the minsup criteria (c) two rules that satisfy the minsup criteria, indicating the confidence of each rule [3 points]
Remember to upload the photo at the END of the exam, in the last 10 minutes.

Files submitted:

21. Q20. Given the following hash tree for verifying the given transaction (image: https://drive.google.com/file/d/1WiFZOvmdAdNtyloavquDQXoDi5EvwZaO/view?usp=sharing) ... indicate in a drawing which leaf nodes are examined for matching candidates; indicate intermediate steps as we did in class [2 points]
Remember to upload the photo at the END of the exam, in the last 10 minutes.

Files submitted: