NAME	**NIA**	**GRADE**

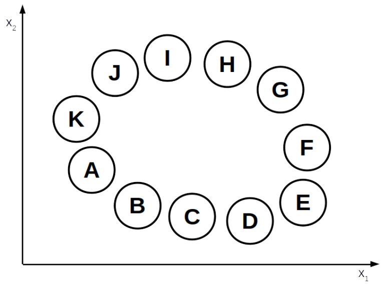
Mining of Massive Datasets (2019-2020)

FIRST MID-TERM (TT01-TT04)

WRITE YOUR ANSWERS <u>CLEARLY</u> IN THE BLANK SPACES. Write as if you were trying to communicate something in written to another person who is going to evaluate what you write. If for some reason (for example, if after you have written the solution you realize that there is some mistake that you would like to correct) you can attach an extra sheet to your exam. In this case, indicate clearly that the solution can be found in the extra sheet. Also, you may use other sheets to perform your calculations.

Problem 1 *1 point*

Consider the two-dimensional dataset on the right, where we will apply an Isomap-like distance calculation of distances considering two nearest neighbors and without the graph projection step. According to this method, indicate the relationship (less than, equal, greater than) between distance(A,H) and distance(H, D), justifying your answer.



Problem 2 *2 points*

Consider a dataset of three cars as follows:

- X1: year=2014, km=40,000, maker=Ford, color=Red, engine=Diesel
- X2: year=2017, km=20,000, maker=Seat, color=Red, engine=Diesel
- X3: year=2018, km=10,000, maker=Ford, color=Blue, engine=Gasoline

1. Indicate the distance between each pair of cars, using L_2 distance for numerical variables, and Jaccard distance for categorical variables. Use a linear combination with a weight of 0.5 for each group of variables.
   - distance(X1, X2) =
   - distance(X1, X3) =
   - distance(X2, X3) =
2. Perform the same calculation using L_2 distance after min-max normalization of each numerical variable, and Jaccard distance for categorical variables, also with a weight of 0.5 for each group of variables.
   - distance(X1, X2) =
   - distance(X1, X3) =
   - distance(X2, X3) =

Problem 3 *2 points*

1. Describe two cases in which a missing value should make us drop the item from the dataset.

   -

   -

2. Describe two cases in which a missing value should be imputed, and how.

   -

   -

Problem 4 *5 points*

Consider the following documents:

- D1: "hello this is a test"
- D2: "bye this is a test"
- D3: "hello this is nice"
- D4: "hi this is nice"

1. Enumerate all **six** distinct shingles in this dataset, indicating their number (start from 1) and the text of the shingle. Use word trigrams as shingles:

- Shingle 1 =
- Shingle 2 =
- Shingle 3 =
- Shingle 4 =
- Shingle 5 =
- Shingle 6 =

2. Write a binary document/shingle matrix in which each column is a document and each row is a shingle.

3. Indicate the similarity between all pairs of documents, using the document/shingle matrix (D_i is the column corresponding to document i).

- $\text{sim}(D_1, D_2) =$
- $\text{sim}(D_1, D_3) =$
- $\text{sim}(D_1, D_4) =$
- $\text{sim}(D_2, D_3) =$
- $\text{sim}(D_2, D_4) =$
- $\text{sim}(D_3, D_4) =$

4. Considering the following permutations: $\pi_1 = (2, 3, 1, 4, 6, 5)$, $\pi_2 = (5, 2, 4, 6, 1, 3)$, and $\pi_3 = (4, 5, 1, 2, 6, 3)$, write the document/signature matrix in which each row is a permutation and each column is a document.

5. Indicate the similarity between all pairs of documents, using the document/signature matrix (S_i is the signature of document i).

- $\text{sim}(S_1, S_2) =$
- $\text{sim}(S_1, S_3) =$
- $\text{sim}(S_1, S_4) =$
- $\text{sim}(S_2, S_3) =$
- $\text{sim}(S_2, S_4) =$
- $\text{sim}(S_3, S_4) =$