NAME	NIA	GRADE

Mining of Massive Datasets (2021-2022)

MID-TERM EXAM

WRITE YOUR ANSWERS CLEARLY IN THE BLANK SPACES. Please 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 *0.5 point*

What is the difference between *ordinal* and *categorical* data?

The difference is:

Problem 2 *0.5 point*

Convert in the following table the columns **Continent** and **Source** to 1-hot encoding, deleting the original columns. Assume there are no other continents or sources beyond the ones appearing here. Provide the resulting list of columns.

Country	Continent	Population	Source
China	Asia	1,411,778,724	Census
India	Asia	1,383,524,897	Estimate
United States	Americas	332,593,407	Census
Indonesia	Asia	271,350,000	Estimate
Pakistan	Asia	225,200,000	Estimate
Brazil	Americas	213,856,536	Census
Nigeria	Africa	211,401,000	Estimate

The resulting list of columns is:

Problem 3 *1 point*

For the table in the previous problem, divide the countries into 3 equi-depth and 3 equi-log (base 10 logarithms) bins by population. Bins should be named "low", "medium" and "high". Your answer should be "Low: [list of countries], Medium: [list of countries], High: [list of countries]". It is OK in equi-depth that bins do not have exactly the same members, and it is OK in equi-log bins that some bins are empty.

Countries in equi-depth bins:

Countries in equi-log bins:

Boundaries of each equi-log bin:

Problem 4 *1 point*

Consider the following sets: $X = \{a, b, c\}$, $Y = \{b, d, e\}$, $Z = \{a, b, c, d, e, f, g\}$, $W = \{b, d, e, f\}$.

Compute Jaccard similarity between all pairs of sets, and draw a graph in which you include only the edges of similarity larger or equal to 0.3, with the Jaccard similarity drawn on the edge.

Your graph:

Problem 5 *0.5 point*

Describe a *unique*, a *range*, a *set membership*, and a *data type* constraint for the table of the country populations.

Unique constraint:

Range constraint:

Set membership constraint:

Data type constraint:

Problem 6 *1 point*

In the following situations where there is missing data, indicate (i) whether it is a case of missing not at random (MNAR), missing at random (MAR) or missing completely at random (MCAR), then (ii) whether you would drop the row, impute the value, or use a data mining method that can deal with null values (iii) If you decide to impute the value, which value would you use, and (iv) justify your decision.

We are doing a study on survival probability of patients upon entering an emergency room, and the age of patients arriving into an emergency room is missing for about half of the patients who were unconscious at time of admission:

We are doing a study on the amount spent per person in a store that sells hearing aids, and the information on whether the customer had an additional disability (in addition to needing a hearing aid) is missing for about 3% of the people. People for which this information is missing do not seem to have anything in common:

Problem 7 *0.5 point*

Given the following series: $1, 5, 67, 10, 23, 41$

The min-max scaled series is:

The standardized series is:

Problem 8 *0.5 point*

Explain how would you do a sampling stratified by day of the week of the purchases of 100 people entering a store that is open Monday through Friday. The entire sample should contain 100 people.

To do this stratified sampling by day of week:

Problem 9 *1 point*

Suppose in a meeting 20% of people are from Italy, 30% from Germany, and 50% from France. Suppose there are only two attributes: age and nationality. Compute the distance between the following pairs of people using L2 distance for age and Goodall measure for nationality, combining them in equal parts.

What would be the distance between a 30-year old French and a 31-year old German?

What would be the distance between a 30-year old German and a 31-year old German?

Problem 10 *1 point*

Extract 3-word shingles and compute the Jaccard similarity between the sets of shingles of the following sentences (remove punctuation and lowercase first):

- S1: The rice is good but the food is tasteless.
- S2: The food is tasteless, but the rice is good.

3-word shingles of sentence S1:

3-word shingles of sentence S2:

Jaccard similarity:

Problem 11 *0.5 point*

Consider the permutations $\pi_1 = \{3, 2, 4, 1, 5\}, \pi_2 = \{2, 5, 4, 1, 3\}$. Consider the shingles matrix below in which each row is a shingle and each column a document. Indicate the signatures of the two documents. Remember that each signature is a vector.

D1	**D2**
0	0
1	0
0	1
1	0
0	0

Signature for D1:

Signature for D2:

Problem 12 *1 point*

Consider the following database of transactions.

t1	a, b, c, d
t2	b, c, d
t3	a, b, c, e
t4	b, c, d, e
t5	a, d, e

Indicate one closed 2-itemset, or explain why there is not any:

Indicate the support of all frequent 3-itemsets in this database, with minimum support 0.4:

Compute the confidence of the following rule: $b, c \rightarrow d$ (write the formula and apply it, show each step):

Can the confidence of the rule $b, c \rightarrow d$ be smaller than the one of the rule $b \rightarrow c, d$? Why or why not?

Problem 13 *1 point*

Execute the candidate generation part of the a priori algorithm on the database given above. Consider min support 0.4. Notice you only need to execute the candidate generation part (itemsets), you do not need to provide the rules, just the itemsets.

Indicate up to four tables containing 1-itemsets, 2-itemsets, ... here. Strikethrough the itemsets that are not frequent.