Homework#1

ACS577 Knowledge Discovery and Data Mining, Summer 2023

Due: July 4

- For your homework1 submission, prepare a single file, *YourLastName_FirstName*_ACS575_HW1.zip
- Organize the submission file with Part I and Part II and clearly number your answer with the question number.
- For each problem solving question, give your answer and also show the steps of computation to get the answer, if any.

Part I. Problem Solving

1. Classify the attribute type of given attributes with each of four category below:

Category 1: *categorical, numeric*
Category 2: *discrete, continues*
Category 3: *qualitative, quantitative*
Category 4: *nominal, binary, ordinal, interval, or ratio*

Example: Age in years. *Answer***:** numeric**,** discrete, quantitative, ratio
Some cases may have more than one interpretation in a category, so briefly indicate your reasoning if you think there may be some ambiguity.

(1) Height above sea level

(2) Bronze, Silver, and Gold medals as awarded at the Olympics.

2. Answer the following questions.

The (score) values for a subject (*subject1*) tuples are in increasing order.

13, 15, 16, 16, 19, 20, 20, 21, 22, 22, 25, 25, 25, 25, 30, 33, 33, 35, 35, 35, 35, 36, 40, 45, 46, 52, 70

(1) Normalize **a value 20** based on ***min-max normalization.***

(2) Normalize **a value 20** based on ***z-score normalization***.

(3) Explain why often normalize attribute values before analysis.

3. Answer the following questions.

The (score) values for a subject (*subject2*) tuples are in increasing order.

11, 12, 13, 15, 17, 20, 20, 21, 21, 22, 22, 23, 23, 25, 30, 31, 31, 32, 35, 35, 35, 36, 40, 45, 45, 53, 55

(1) Partition them into three bins by ***equal-width partitioning***. List values in each bin.

(2) Use ***smoothing by bin means*** to smooth the data from Q3 (1) result. List values in each bin.

4. Consider the time-series (−3,−1, 1, 3, 5, 7, ∗). Here, a missing entry is denoted by ∗. What would be the estimated value of the missing entry using ***linear interpolation,*** $y = a \cdot x + b$, on a window of size 3 ?

[*Hint*] When you derive the relationship $y = a \cdot x + b$, x is the time-stamp. So, you can present the dataset with (x, y)s, where y is a value measured at a time point x (any ordered value x), e.g., (1, -3), (2, -1) , (3, 1) , (4, 3), (5, 5) , (6, 7), (7,*).
In $y = a \cdot x + b$, the values a and b are typically solved using methods such as least squares regression, $a = \frac{N\sum xy - \sum x \sum y}{N\sum x^2 - (\sum x)^2}$, $b = \frac{\sum y - a\sum x}{N}$.
For linear interpolation of ∗ on a window of size 3, you need to consider its previous three values, i.e., 3, 5, and 7.

5. Consider the following datasets with 2 attributes A_1 and A_2 :

	A_1	A_2
X_1	1.5	1.7
X_2	2	1.9
X_3	1.6	1.8
X_4	1.2	1.5
X_5	1.5	1.0

Given a new data point, X = (1.4, 1.6) as a query, rank the data points based on similarity with the query using (1) ***Euclidean distance*** and (2) ***Manhattan distance***.

6. Compute the (1) ***match-based similarity*** and (2) ***Jaccard coefficient*** between the two sets {A, B, C} and {A, C, D, E}.

7. Consider the two sentences below.

S1: "The sly fox jumped over the lazy dog."
S2: "The dog jumped at the intruder."

(1) Convert S1 and S2 to ***frequency term vectors***.
Assume the lexicon here is {the sly fox jumped over lazy dog at intruder}.

(2) Compute the ***cosine similarity*** measure of S1 and S2:

8. Consider a matrix representing four sets as following.

Element	S_1	S_2	S_3	S_4
0	0	1	0	1
1	0	1	0	0
2	1	0	0	1
3	0	0	1	0
4	0	0	1	1
5	1	0	0	0

The columns of the matrix correspond to the sets, and the rows correspond to elements of the universal set from which elements of the sets are drawn. There is a 1 in row *r* and column *c* if the element for row *r* is a member of the set for column *c*, otherwise 0.

(1) Compute the ***minhash signature*** for each column when we use a hash function: $h(x) = 2x+1 \bmod 6$. Show the final **signature matrix**.

(2) Estimate the ***Jaccard similarities*** of the underlying sets $\boldsymbol{S_2}$ and $\boldsymbol{S_4}$ from the signature matrix from (1).

Part II. Hands-on practice

1. The purpose of this practice is to get familiar with R scripts used in data exploration.
Download the **Iris** data set from https://archive.ics.uci.edu/dataset/53/iris . Follow "Ch 3. Data Exploration" (from Zhao, "R and Data Mining") provided. Submit your practice on all scripts from Ch 3.1 to Ch 3.5 with showing the results.

Alternatively, you can show the same results using Python or any programming language you choose.

Deliverables: (1) Your program codes, e.g., the script codes in Ch 3.1 – Ch 3.5 provided.
(2) A proof to show the successful execution of your program, e.g., screen shots on running and the program output.

2. Download the **Arrythmia** data set from https://archive.ics.uci.edu/ml/datasets/Arrhythmia. Using a programming language or data mining tool you chose, normalize all the numeric attributes in the dataset with a mean of 0 and a standard deviation of 1.

Deliverables: (1) A list of first 10 records transformed with normalized values, i.e., z-scores.
(2) Your program codes, if any
(3) A proof to show the program's successful execution,

3. Again download the **Arrythmia** data set. Discretize each numerical attribute into 10 equiwidth ranges (bins), and then transform each attribute's value to the corresponding bin name. Assume the bin names of each attribute are given as *attribute name*_Bin1, *attribute name*_Bin2, …, *attribute name*_Bin10.

Deliverables: (1) A list of first 10 records transformed using the discretization.
(2) Your program codes, if any
(3) A proof to show the program's successful execution,