Homework 2

Due: 9/22

Note: You must show all intermediate calculations/results. You can do manual calculations or use any software (e.g., Weka, Excel, JMP, R, Python) to answer the questions unless otherwise noted. In any case, you need to attach the relevant file(s) or screenshot(s) that shows how you obtained your answers.

Problem 1 (20 points) Consider the dataset *a2-p1.csv* which is posted along with this assignment. It has 100 instances and one attribute.

(1). Calculate the mean, median, and standard deviation (sample) of the attribute *A*1.
(2). Determine Q1, Q2, and Q3 of *A*1.
(3). Detect outliers using the IQR method, which we discussed in the class, and show the *A*1 values of the detected outliers.
(4). Plot the boxplot of the attribute *A*1. In your boxplot, you need to show outliers separately.

Note: You may use any tool to determine mean, median, standard deviation, Q1, Q2, and Q3. When you detect outliers, you must do it yourself using the method we discussed in the class.

Problem 2 (10 points). This problem uses *a2-p2.csv* dataset, which has 4 attributes and 100 tuples. Plot the scatterplots of all possible pairs of attributes (you need to show 6 scatterplots), and determine, by visual observation, which two attributes have the strongest correlation.

Problem 3 (10 points). Consider the following dataset that has some information about 10 people.

ID	job	marital	education	default	housing	loan	contact
P1	unemployed	married	primary	no	no	no	cellular
P2	services	married	secondary	no	yes	yes	cellular
P3	management	single	tertiary	no	no	no	cellular
P4	management	married	tertiary	no	yes	yes	unknown
P5	blue-collar	single	secondary	no	yes	no	unknown
P6	management	single	tertiary	no	no	yes	cellular
P7	self-employed	married	tertiary	no	yes	no	cellular
P8	services	married	secondary	yes	no	no	cellular
P9	entrepreneur	married	tertiary	no	yes	no	unknown
P10	services	single	primary	no	yes	yes	cellular

Calculate the distance between P8 and P9, *d*(P8, P59), and the distance between P8 and P10, *d*(P8, P10). Is P8 closer to P9 or P10? Here, all attributes are nominal attributes.

Problem 4 (10 points). Consider the following dataset with two objects.

Object	A1	A2	A3	A4
O1	2	second	gold	large
O2	5	third	silver	small

Here, all attributes are ordinal attributes and ranks of their values are shown below (lowest rank on the left):

A1: {1, 2, 3, 4, 5}
A2: {first, second, third}
A3: {bronze, silver, gold}
A4: {small, medium, large}

Calculate the distance between O1 and O2 using the method discussed in the class. Use the Euclidean distance measure.

Problem 5 (10 points). Consider the following dataset, which has attributes of mixed data types:

Tuple ID	A1	A2	A3	A4	A5	A6	A7
1	37	0	No	No	Yes	Low	Large
2	26	1	No	No	No	High	Medium
3	29	0	No	Yes	Yes	Low	Large
4	35	0	Yes	No	No	Middle	Large
5	73	1	No	No	No	High	Medium
6	**27**	**0**	**Yes**	**No**	**No**	**High**	**Small**
7	**52**	**1**	**No**	**Yes**	**No**	**Low**	**Medium**
8	36	1	No	No	Yes	High	Medium
9	12	0	Yes	No	Yes	High	Large
10	24	1	Yes	No	Yes	Low	Medium

Here,

- A1 is a numeric attribute
- A2 A3 are symmetric binary attributes
- A4 and A5 are asymmetric binary attributes, where Yes is more important than No
- A6 is a categorical (nominal) attribute
- A7 is an ordinal attribute. *Small* has the lowest rank, *Medium* has the next lowest rank, and *Large* has the highest rank.

Calculate the distance between **Tuple 6** and **Tuple 7** using the method we discussed in the class.

Problem 6 (10 points). Consider the following dataset.

Document	apple	orange	banana	pear	lemon	tomato	grape	berry	pineapple	mango
D1	3	3	2	1	2	2	4	2	2	1
D2	1	1	0	3	4	2	2	3	1	1
D3	2	2	1	4	0	1	0	1	2	2
D4	1	3	4	3	0	3	5	0	4	0

Calculate the similarity between D1 and D2, *cos*(D1, D2), and the similarity between D1 and D3, *cos*(D1, D3), using the cosine similarity measure. Is D1 closer to D2 or D3? You must calculate the cosine similarity yourself (i.e., you must not use a built-in function of a software).

Submission:

Submit the solutions in a single Word or PDF document and upload it to Blackboard. Use *LastName_FirstName_hw2.docx* or *LastName_FirstName_hw2.pdf* as the file name. If necessary, you may submit an additional file that shows how you obtained your answers. Make sure that this additional file also has your last name and first name as part of the file name. If you have multiple files, then combine them into a single archive file, name it *LastName_FirstName_hw2.EXT*, where *EXT* is an appropriate file extension (such as *zip* or *rar*), and upload it to Blackboard.