

Assignment 1

Note: Show all your work. You can do manual calculations, use R, use Python, or use any other software (e.g., Weka, Excel, JMP) 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. Install Weka (10 points)

- Complete WEKA installation using the attached installation instruction.
- Start Weka and start Explorer.
- Open the *iris.arff* dataset. Provide a snapshot of the resulting screen in your submission.

Problem 2. Install JMP Pro (10 points)

- Complete JMP installation using the attached installation instruction.
- Follow the instruction and capture the last screen and include it in your submission file.

Problem 3 (20 points) Consider the dataset *a1.csv* which is posted along with this assignment. It has 100 instances and 5 attributes.

- (1). Calculate the mean, median, and standard deviation (sample) of the attribute A5.
- (2). Determine Q1, Q2, and Q3 of A5.
- (3). Detect outliers using the IQR method, which we discussed in the class, and show the A5 values of the detected outliers. When detecting outliers, use only the A5 values.
- (4). Plot the boxplot of the attribute A5. 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. However, when you detect outliers, you must do it manually using the method we discussed in the class.

Problem 4 (20 points). This problem also uses *a1.csv* dataset.

- (1). Calculate the following four Pearson's correlation coefficients:

```
correl(A1, A5)
correl(A2, A5)
correl(A3, A5)
correl(A4, A5)
```

(2). Show the four scatterplots.

(3). Among A1, A2, A3, and A4, which one has the strongest correlation with A5?

Problem 5 (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	technician	married	secondary	no	yes	no	cellular
P9	entrepreneur	married	tertiary	no	yes	no	unknown
P10	services	married	primary	no	yes	yes	cellular

Calculate the distance between P4 and P5, $d(P4, P5)$, and the distance between P4 and P6, $d(P4, P6)$. Is P4 closer to P5 or P6? Here, all attributes are nominal attributes.

Problem 6 (10 points). Consider the following dataset with two objects.

Object	A1	A2	A3	A4
O1	43	14	32	21
O2	25	20	15	14

(1). Calculate the distance between O1 and O2 using the Manhattan distance.

(2). Calculate the distance between O1 and O2 using the Euclidean distance.

Problem 7 (10 points). Consider the following dataset with two objects.

Object	A1	A2	A3	A4
O1	2	second	silver	small
O2	4	first	gold	large

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

A1: {1, 2, 3, 4}

A2: {first, second, third}

A3: {bronze, silver, gold}

A4: {small, medium, large}

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

Problem 8 (10 points). Consider the following dataset.

Document	apple	orange	banana	pear	lemon	tomato	grape	berry	pineapple	mango
D1	5	1	2	1	3	2	4	2	3	1
D2	2	1	0	0	4	2	3	3	1	1
D3	2	0	1	0	3	1	0	2	3	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_hw1.docx* or *LastName_FirstName_hw1.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_hw1.EXT*, where *EXT* is an appropriate file extension (such as zip or rar), and upload it to Blackboard.