Homework 2

Due: 2/5

Note: Show all your work. You can do manual calculations, use R, or use any 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 (20 points) Consider the dataset *a2.csv* which is posted along with this assignment. It has 190 instances and 5 attributes.

(1). Calculate the mean, median, and standard deviation of the attribute *A*5.
(2). Determine Q1, Q2, and Q3 of *A*5.
(3). Detect outliers using the IQR method, which we discussed in the class, and show the *A*5 values of the detected outliers. When detecting outliers, use only the *A*5 values.
(4). Plot the boxplot of the attribute *A*5. 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 2 (10 points). This problem also uses *a2.csv* dataset. Using JMP Pro (or any other tool), plot the scatterplots of all possible pairs of attributes (you need to show 10 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	yes	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	no	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 P9 and P8, *d*(P9, P8), and the distance between P9 and P10, *d*(P9, P10). Is P9 closer to P8 or P10? Here, all attributes are nominal attributes.

Problem 4 (10 points). Consider the following dataset.

Document	apple	orange	banana	pear	lemon	tomato	grape	berry	pineapple	mango
D1	3	1	2	1	2	2	4	2	2	1
D2	0	1	0	0	6	2	3	6	1	0
D3	2	3	1	0	0	1	0	4	3	2
D4	1	2	4	3	5	3	0	0	1	0

Calculate the similarity between D2 and D1, *cos*(D2, D1), and the similarity between D2 and D3, *cos*(D2, D3), using the cosine similarity measure. Is D2 closer to D1 or D3?

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.