Homework 2

Due: 2/6

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-p1.arff* which is posted along with this assignment. It has 190 instances and 7 attributes. The same dataset in Excel format is also posted.

(1). Calculate the mean, median, and standard deviation of the attribute *A*5.
(2). Determine Q1, Q2, and Q3, and plot the boxplot of the attribute *A*5. In your boxplot, you don't need to show outliers separately.
(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.

You can convert an *arff* file to a *csv* file (for easy manipulation) by opening it in Weka Explorer and saving it as a *csv* file.

Problem 2 (10 points). Consider the following dataset that has income and age information of 10 people.

ID	Income	Age
P1	60000	23
P2	70000	27
P3	75000	49
P4	60000	52
P5	95000	25
P6	90000	65
P7	100000	63
P8	120000	38
P9	27000	47
P10	63000	72

(1). Calculate the distance between P2 and P1, *d*(P2, P1), and the distance between P2 and P3, *d*(P2, P3), using the Euclidean distance measure. Is P2 closer to P1 or P3?
(2). First, standardize *Income* and *Age* using the z-score method that uses the standard deviation as the denominator. Then, calculate *d*(P2, P1) and *d*(P2, P3). Is P2 closer to P1 or P3?
(3). What conclusion can you draw from the above calculations?

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	married	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 P8 and P7, *d*(P8, P7), and the distance between P8 and P9, *d*(P8, P9). Is P8 closer to P7 or P9? Here, all attributes are nominal attributes.

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

Document	*team*	*coach*	*hockey*	*baseball*	*soccer*	*penalty*	*score*	*win*	*loss*	*season*
Document1	5	0	3	0	2	0	0	2	0	0
Document2	3	0	2	0	1	1	0	1	0	1
Document3	0	7	0	2	1	0	0	3	0	0
Document4	0	1	0	0	1	2	2	0	3	0

Calculate the similarity between Document2 and Document 3, *d*(D2, D3), and the distance between Document2 and Document4, *d*(D2, D4), using the cosine similarity measure. Is Document 2 closer to Document3 or Document4?

Submission:

Submit the solutions in a single Word or PDF document and upload it to Blackboard. Please make sure that there are no spaces in the file name. 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 (such as a zip file or a rar file), name it as *LastName_FirstName_hw2.EXT*, where *EXT* is an appropriate file extension (such as zip or rar), and upload it to Blackboard.