

Assignment 6

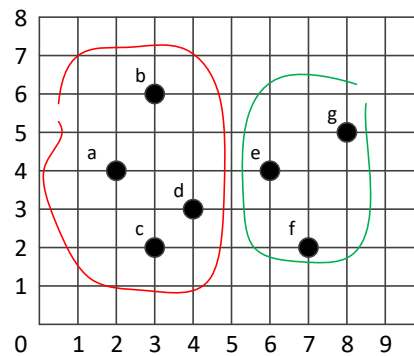
Note: Show all your work.

Problem 1 (20 points). The k-means algorithm, which we discussed in the class, is being run on a small two-dimensional dataset. After a certain number of iterations, you have two clusters as shown below:

ID	x	y	Cluster
1	3	4	Cluster 1
2	5	3	Cluster 1
3	6	4	Cluster 1
4	4	5	Cluster 2
5	4	7	Cluster 2
6	7	6	Cluster 2
7	8	4	Cluster 2

Run one more iteration of the k-Means clustering algorithm and show the two clusters at the end of the iteration. Use Manhattan distance when calculating the distance between objects.

Problem 2 (20 points). Consider the following two clusters:



Compute the distance between the two clusters (1) using the maximum distance method and (2) using the mean distance method. Use the Manhattan distance measure when calculating the distance between objects.

Problem 3 (20 points). Consider the following dataset, which has eight 2-dimensional objects.

Object	x	y
a	1	7
b	2	3
c	2	4
d	5	1
e	5	8
f	6	7
g	7	2
h	8	8

Using the agglomerative hierarchical clustering approach that we discussed in the class, create two clusters. Use the *minimum distance* method with the Manhattan distance measure. You need to show, at each step, which two clusters are merged. You must also show, at each step, all clusters and their objects at the step. You must decide which two clusters are merged yourself and you must not use any software to do that.

Problem 4 (20 points).

(1). Use the *a6-p4-1.csv* dataset for this problem. It has 2 attributes and 150 tuples. Run the *SimpleKMeans* algorithm of Weka on this dataset with $k = 2, 3, 4, 5, 6$, and 7. For each k , record the value of *within cluster sum of squared errors* (which you can find in Weka's cluster output window) and plot a graph where the x-axis is k and y-axis is *within cluster sum of squared errors*. Then, determine an optimal number of clusters using the *elbow method* which we discussed in the class.

(2). Use the *a6-p4-2.csv* dataset for this problem, which has 4 attributes and 77 tuples. This dataset was downloaded from JMP sample dataset library and modified for this assignment. The dataset has nutritional information for 77 breakfast cereals. Each tuple in the dataset represents a cereal product and the four attributes are *calories*, *fiber*, *sugars*, and *potassium*.

(2)-a. Run the *SimpleKMeans* algorithm of Weka on this dataset with $k = 4$ and prepare the following table that summarizes the profiles of four clusters:

		Cluster			
		1	2	3	4
Calories	mean				
	stddev				
	max				
	min				
Fiber	mean				
	stddev				
	max				
	min				

S u g a r s	m e a n				
	s t d d e v				
	m a x				
	m i n				
P o t a s s i u m	m e a n				
	s t d d e v				
	m a x				
	m i n				

(2)-b. In general, a cereal is considered healthy if it has low calories, high fiber, low sugars, and high potassium. Using this standard, which cluster has the healthiest cereals?

Note: You may use a k-means algorithm implemented on other tools, including JMP Pro, R, or Python, for both Problem 4-(1) and Problem 4-(2). If you use JMP Pro you must include appropriate screenshots in your submission. If you use R or Python, you must submit script files as well as your answers.

Problem 5 (20 points). Follow the instructions in *JMP-clustering-assignment.pdf* file. Include the required screenshots and your answers to some questions in your submission.

Submission:

Include all answers in a single file and name it *LastName_FirstName_HW6.EXT*. Here, “EXT” is an appropriate file extension (e.g., docx or pdf). If you have multiple files, then combine all files into a single archive file and name it *LastName_FirstName_HW6.EXT*. Here, “EXT” is an appropriate archive file extension (e.g., zip or rar). Upload the file to Blackboard.