Assignment 6

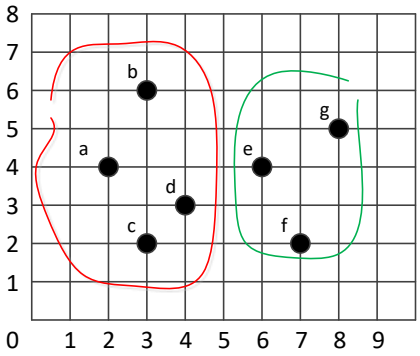
Note: Show all your work as you have on previous assignments.

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, show how the individual objects are aggregated into clusters. Continue this process until no further aggregation is possible. Use the *minimum distance* method with the Manhattan distance measure. You need to show, at each step, which two clusters are merged. At each step you should list which objects are in which cluster (example: cluster 1 contains objects w and y and cluster 2 contains objects x and z… though this will be different for you since your objects are labeled a – h). 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				
Sugars	mean				
	stddev				
	max				
	min				
Potassium	mean				
	stddev				
	max				
	min				

(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 Guidelines:

- **Submit the solutions in a single Word or PDF document and upload it to Blackboard. Make sure the filename includes your name.**
- **Your submission should be organized well and easy to follow. Clearly list which question you are answering and provide the answer requested (one or more numbers, a screenshot, a plot, an explanation, whatever the question asks you to do). If a question contains multiple parts, you should clearly provide your answer for each part.**
- **If a problem requires you to work out formulas or calculations, you should provide this detailed work either in an Appendix in your Word/PDF or as a separate file. All separate files should include your name in the filename.**
- **If you need more than 3-4 separate files, you might ZIP/RAR these files. Please still upload the Word/PDF separately. The separate Word/PDF greatly simplifies the facilitator's ability to comment on your assignment.**
- **If your facilitator tells you to submit the files differently than the above guidelines, you are expected to respect your facilitator's wishes starting on the next assignment.**
- **Facilitators can deduct up to 20% if you fail to follow these requirements (more if the questions are not actually answered).**
- **Facilitators can deduct 5% for each day the assignment is late.**
- **Unless your facilitator or the professor agrees, your assignment will not be graded if it is more than 3 days late (e.g., no credit will be given after Friday**

at 6 AM Boston time). The professor will usually ask the facilitator to make the decision but in rare cases (<1% of the time) has overridden a facilitator. Do not expect the professor to override in most cases.