CS 699
Assignment 6

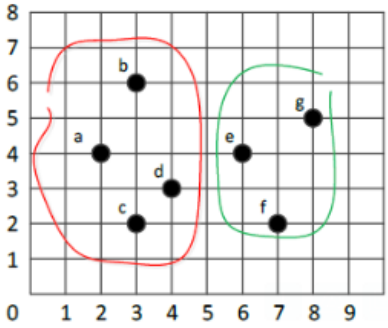
Note: R is recommended since that is what is taught in class, but you may use any software you choose. As explained in submission guidelines on the next page, please provide your "final answers" in a document and provide a separate file containing your code or other detail that explains how you obtained your answers.

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 (15 points). Consider the following two clusters plotted to the right. Compute the distance between the two clusters (1) using the maximum distance method and (2) using the centroid distance method. Use the Manhattan distance measure when calculating the distance between objects.



Problem 3 (15 points). Use the dataset built into R called *swiss*. This dataset has 47 observations with 6 features.

(1). Apply k-means clustering to the data. Which value of k do you recommend?
(2). Which cluster or clusters seem to be associated with high Examination and high Education scores? Which Swiss provinces are associated with these clusters?

(Assignment continues on the next page)

Problem 4 (20 points). Consider the following dataset with 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 5 (30 points). Use the dataset built into R called *AirPassengers.* This is a 12-year monthly time series.

(1). Plot the time series.
(2). Create either a seasonal subseries plot or a seasonal boxplot. Comment on the patterns you see.
(3). Hold out the last year of data. Attempt at least two regression-based forecasts. Measure each model's performance using one or more appropriate metrics of your choice.
(4). Report the best model. Explain the trend.

Submission Guidelines:

- **Submit the solutions in a single Word or PDF document and upload it to Blackboard. You should separately upload the code or another document that explains how you arrived at the answers. Make sure each filename includes your name.**
    - **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.**
- **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 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.**