Emory University
Department of Computer Science
Dr. Andreas Züfle

CS 485/584: Spatial Computing

Fall 2024

Assignment 5: Spatial Clustering

Submit all assignments until **Monday, October 28th at 2:30pm**

Assignment 5-1 *k-means clustering* **(6 Points)**

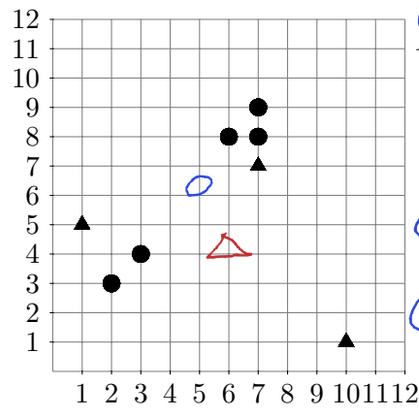
Given the following data set with 8 objects (in $\mathbb{R}^2$):

△ centroid:

$\triangle_1 = (1,5)$ $\triangle_2 = (7,7)$

$\triangle_3 = (10,1)$

$\triangle = \left(\frac{1+7+10}{3}, \frac{5+7+1}{3}\right) = \left(6, \frac{13}{3}\right)$



○ centroid

$\bigcirc_1 = (2,3)$ $\bigcirc_2 = (3,4)$

$\bigcirc_3 = (6,8)$ $\bigcirc_4 = (7,8)$

$\bigcirc_5 = (7,9)$

$\bigcirc = \left(\frac{2+3+6+7+7}{5}, \frac{3+4+8+8+9}{5}\right) = \left(5, \frac{32}{5}\right)$

In the following, as distance function use the Manhattan distance (L_1 norm), which is defined on two objects x, y as $L_1(x,y) = \sum_{i=1}^{d} |x_i - y_i|$

(a) Compute a partitioning into $k = 2$ clusters using the k-means algorithm *as introduced in the lecture*. The initial assignment of objects is given using the triangle and circle markers. That is, initially assume that cluster 1 has the three triangles and cluster 2 has the five circles.

Instead of starting with random centroids (as k-means would normally do, but which may trivialize this question) start with computing the initial centroids (mark centroids as large triangle and large circle). Then, draw the new cluster assignments and new centroids after each step/iteration. Explain each step.

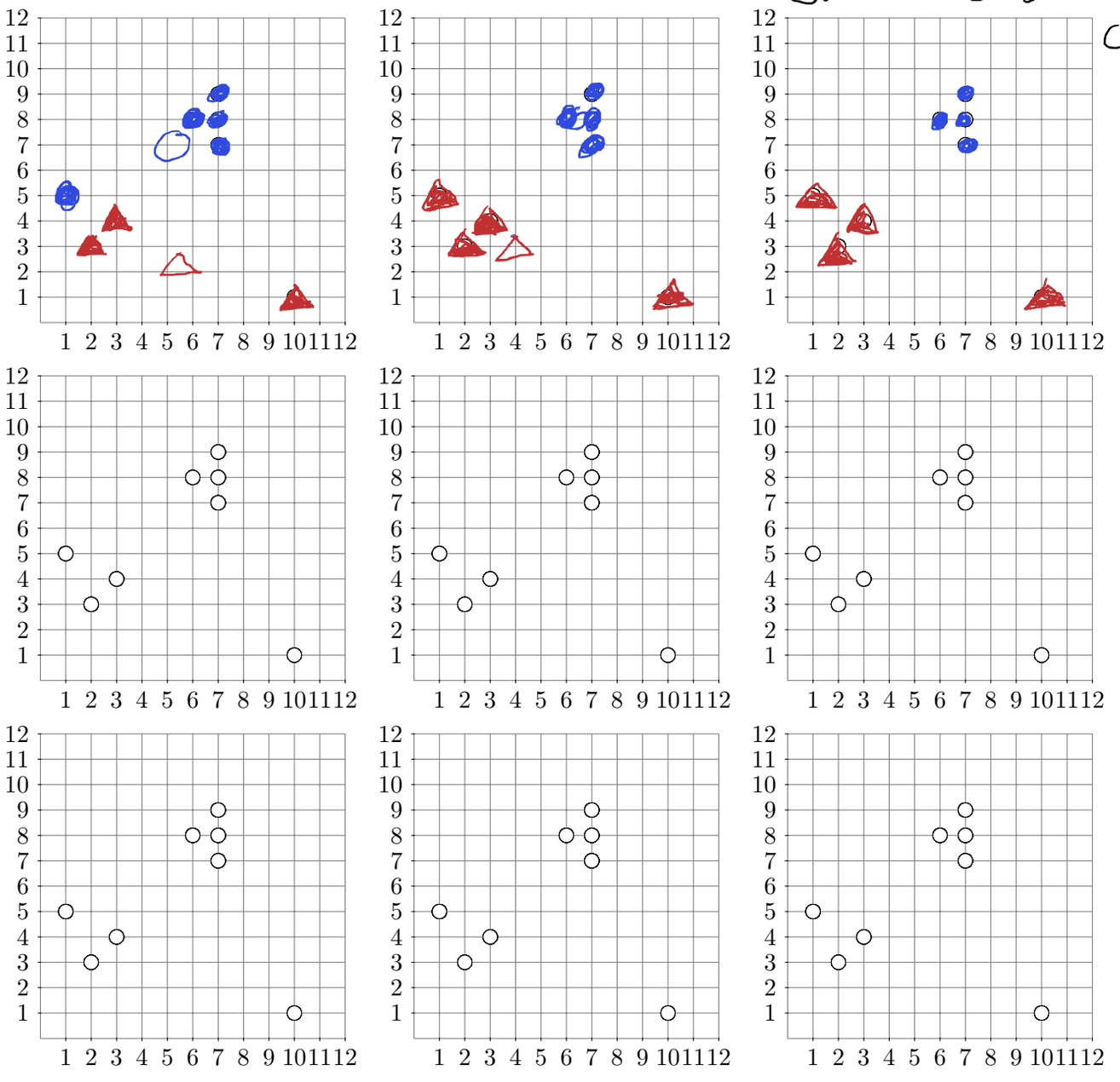
You may also use different colors instead of triangles and circles.

Remember to use the L_1 norm for computing distances!

You can copy the next page of this assignment multiple times if you need more space for sketching.

(b) Optional (No Points): also try k-medoids instead.

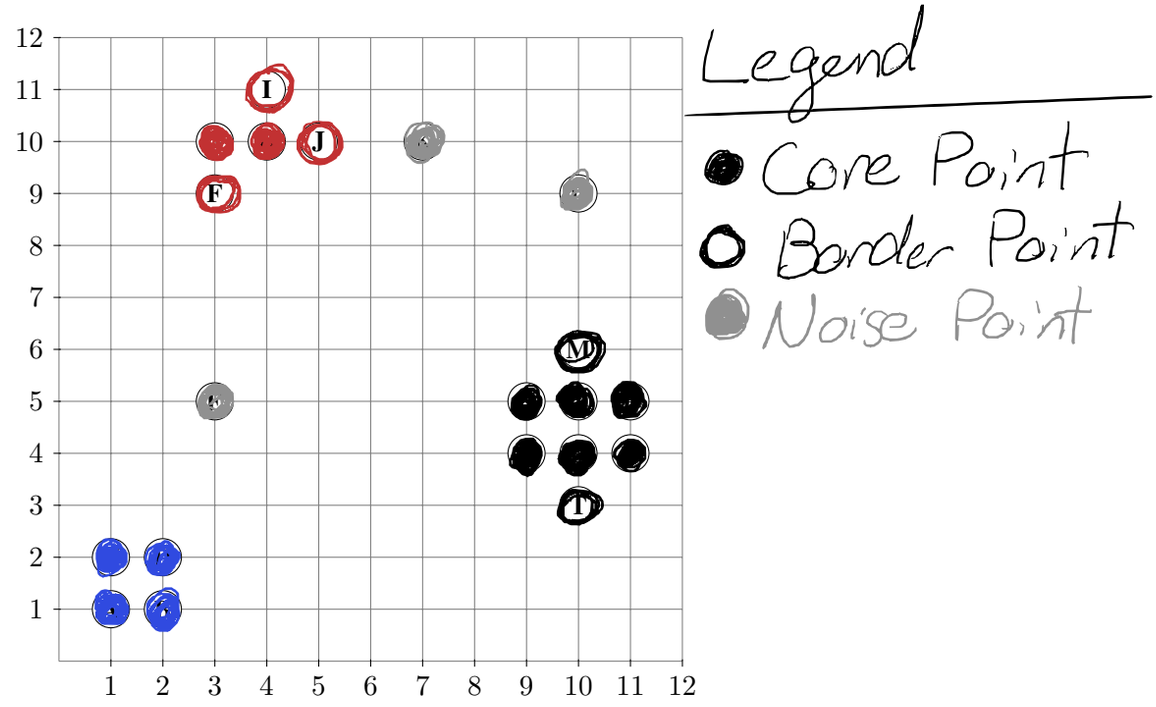
Clusters constant ⇒ converge!



Assignment 5-2 *DBSCAN* **(8 Points)**

Given the following data set:

radius $\varepsilon = 1.1$
min Pts $= 3$



As distance function, use Manhattan Distance:

$$L_1(x, y) = |x_1 - y_1| + |x_2 - y_2|$$

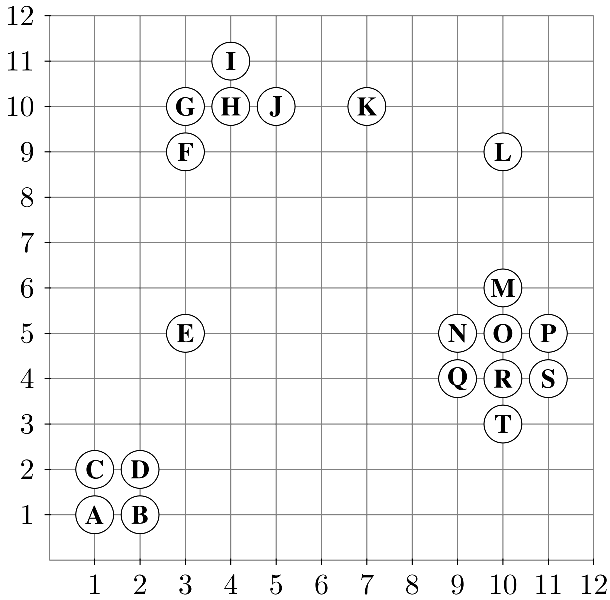
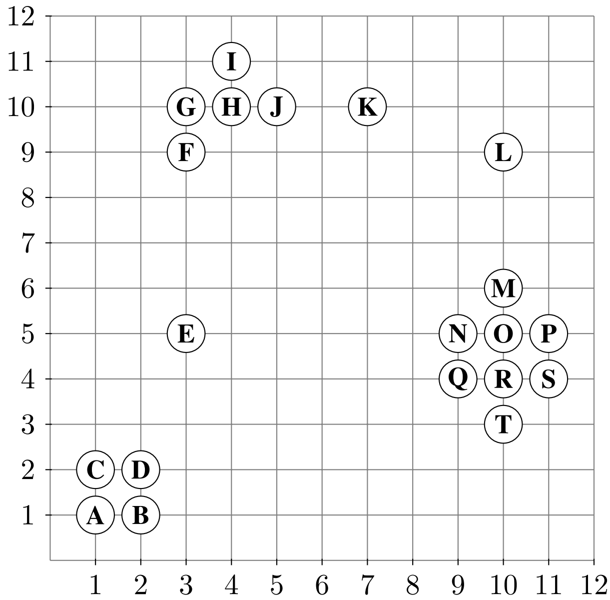
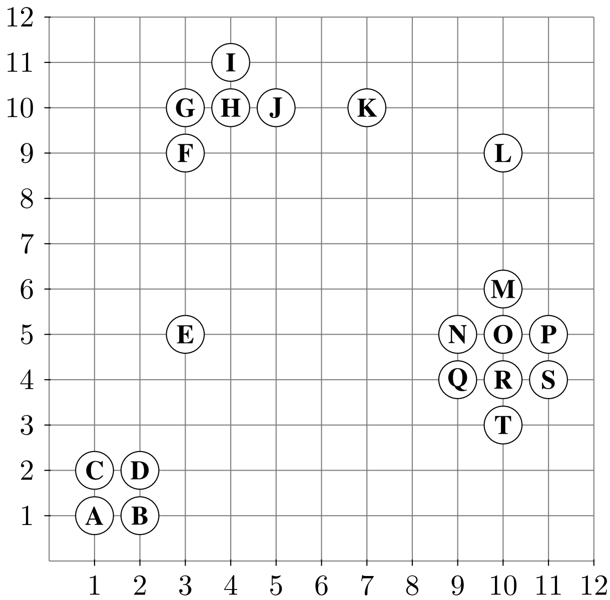
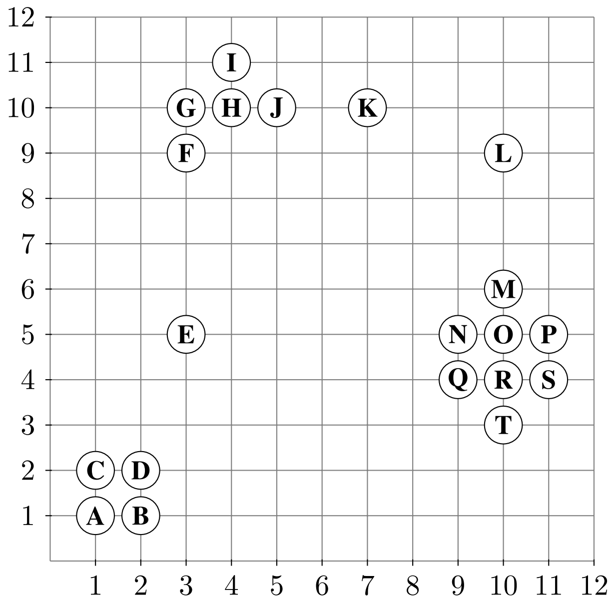
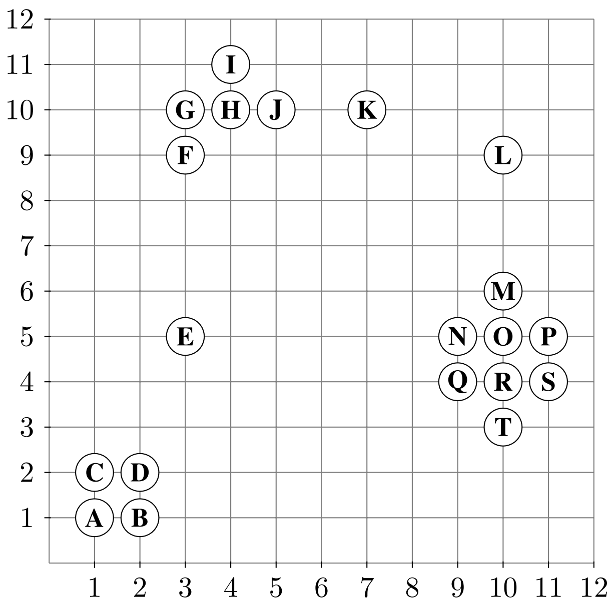
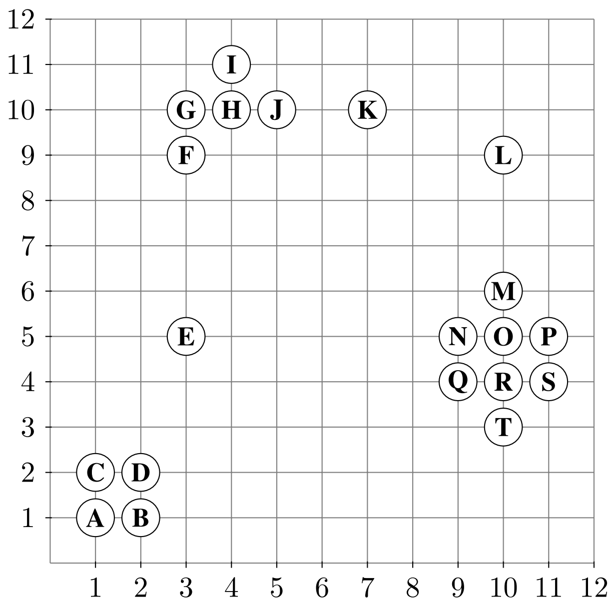
Compute DBSCAN and indicate which points are core points, border points and noise points.

Use the following parameter settings:

- Radius $\varepsilon = 1.1$ and *minPts* $= 2$ (optional)
- Radius $\varepsilon = 1.1$ and *minPts* $= 3$
- Radius $\varepsilon = 1.1$ and *minPts* $= 4$ (optional)
- Radius $\varepsilon = 2.1$ and *minPts* $= 4$ (optional)
- Radius $\varepsilon = 4.1$ and *minPts* $= 5$ (optional)
- Radius $\varepsilon = 4.1$ and *minPts* $= 4$ (optional)

When *minPts* $= 2$, what happens to border points?

The following page has copies of the dataset above for sketching. You can copy this page multiple times if you need more space for sketching.

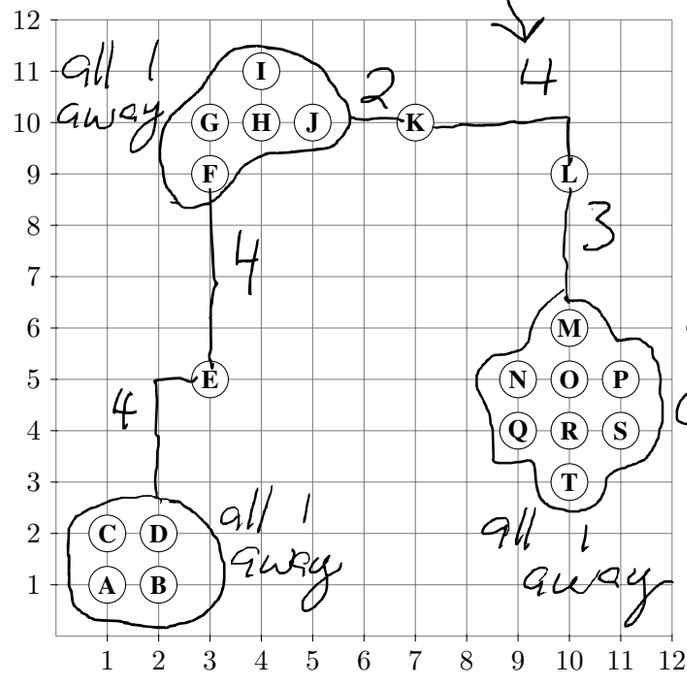

A B C D E F G H I J K L M N O P Q R S T
1 2 3 4 5 6 7 8 9 10 11 12

A B C D E F G H I J K L M N O P Q R S T
1 2 3 4 5 6 7 8 9 10 11 12

A B C D E F G H I J K L M N O P Q R S T
1 2 3 4 5 6 7 8 9 10 11 12

A B C D E F G H I J K L M N O P Q R S T
1 2 3 4 5 6 7 8 9 10 11 12

A B C D E F G H I J K L M N O P Q R S T
1 2 3 4 5 6 7 8 9 10 11 12

A B C D E F G H I J K L M N O P Q R S T
1 2 3 4 5 6 7 8 9 10 11 12

Assignment 5-3 *Hierarchical Clustering* **(4 Points)**

Given the following data set:

single-link: these are shortest distances between clusters

Looking at ABCD cluster, we see AC, CD, DB, and AB are all 1 away so even though AD and BC are 2 away they are included in the 1 cluster.



Further agglomerative clustering is on the last page.

As distance function, use Manhattan Distance:

$$L_1(x, y) = |x_1 - y_1| + |x_2 - y_2|$$

Compute a dendrograms for this data set using agglomerative hierarchical clustering. To compute the distance of sets of objects, use

- the single-link method
- the average-link method (optional - no assignment points)

Hint: with discrete distance values, nodes may have more than two children. more space for sketching. You do not need to materialize the entire distance matrix - instead, you may identify the pair of clusters having the lowest distance simply by looking at the figure above.

Assignment 5-4 *Implementing DBSCAN* **(6 Points (CS 584 Only))**

Implement the DBSCAN algorithm (in a program language of your choice)!

Hint: You will need to compute many ε-range queries (to decide if points are core points and in the expand-cluster function). You may implement these ε-range queries naively using a linear scan (that computes distances to all other points). What I'm trying to say here: You do not need to implement a spatial index structure to efficiently support the range queries that DBSCAN requires.

Hint: You may use data structures are heaps and queues using libraries are you see fit. You may also use other libraries that you may find helpful, for example for visualization. Just don't use an existing DBSCAN algorithm.

Didn't due - no time :(

Question 5-1: K-means continued

K-means after 1 iteration

$\Delta_1 = (2,3)$ $\Delta_2 = (3,4)$
$\Delta_3 = (10,1)$

$\Delta = \left(\frac{2+3+10}{3}, \frac{3+4+1}{3}\right)$
$= \left(5, \frac{7}{3}\right)$

$O_1 = (1,5)$ $O_2 = (6,8)$
$O_3 = (7,7)$ $O_4 = (7,8)$ $O_5 = (7,9)$

$O = \left(\frac{1+6+7+7+7}{5}, \frac{5+8+7+8+9}{5}\right)$
$= \left(\frac{28}{5}, \frac{37}{5}\right)$

K-means after 2 iterations

$\Delta_1 = (1,5)$ $\Delta_2 = (2,3)$
$\Delta_3 = (3,4)$ $\Delta_4 = (10,1)$

$\Delta = \left(\frac{1+2+3+10}{4}, \frac{5+3+4+1}{4}\right)$
$= \left(4, \frac{13}{4}\right)$

$O_1 = (6,8)$ $O_2 = (7,7)$
$O_3 = (7,8)$ $O_4 = (7,9)$

$O = \left(\frac{6+7+7+7}{4}, \frac{8+7+8+9}{4}\right)$
$= \left(\frac{27}{4}, 8\right)$

Question 5-3: Hierarchical Clustering

- step 1:
  Identify clusters 1 away:
  A-D, F-J, and M-T
- step 2:
  Add K (dist 2) to F-J cluster.
- step 3:
  Add L (dist 3) to M-T cluster
- step 4:
  E is dist 4 away from
  - A-D
  - F-K
  - L-T

  } so we combine all clusters