CS251 HW 7 | Mon Apr 8, 2019 | Week 10

Name:

Notes

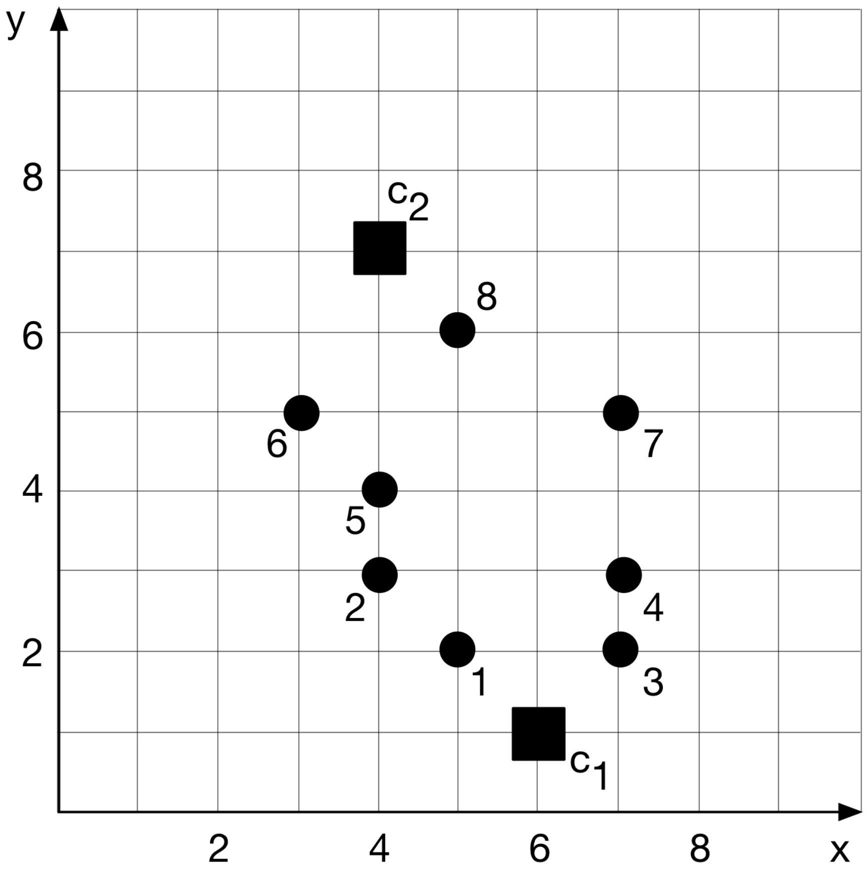
- **Use city-block distance (L^1 norm) throughout this homework.**
- Do this HW by hand (no need to write code).

Question 1: K-means

The plot shows 8 2D data points that will be clustered with K-means, with two clusters ($K = 2$).

a) If c_1 and c_2 represent the initial cluster centers, use K-means to assign each data point to the appropriate cluster.

Note: Assume that we loop through centroids in order to do the assignment (i.e. we assign points to Cluster 1 in the case of distance ties).



b) Compute (**and draw** in the plot) the new position of the centroids after we assign the data points to the clusters according to your answer from (a).

c) Rather than minimize the sum of squared error (SSE), city-block K-means minimizes the **cost**, which we compute as:

- The total city-block distance between each data point within a cluster and its center
- Sum these totals across all clusters.

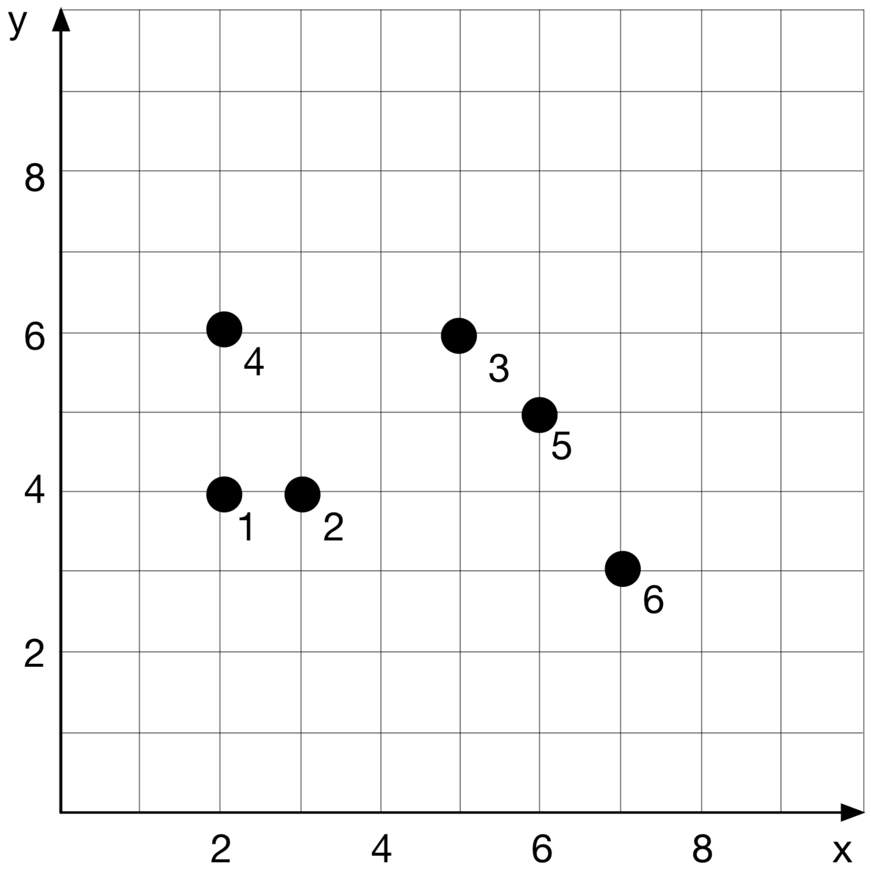
Compute the cost of the clustering with the new centroids from (b). Is it an improvement over the original centroids?

Note: Point assignments do not change with the new centroids.

d) Explain why we generally should run K-means multiple times with the same K and dataset, but different initial conditions.

e) Using a uniform random initialization scheme, the centroids were selected as $c_1 = c_2 = (8, 8)$. How would K-means assign points to clusters?

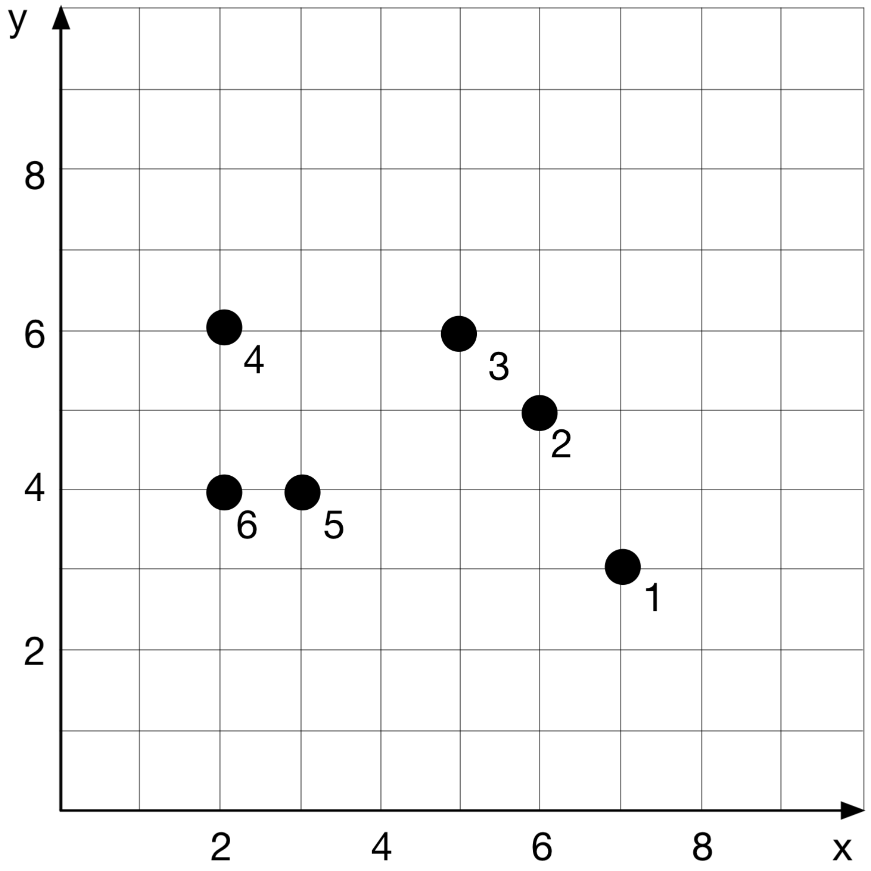
Question 2: Leader algorithm



a) Determine the cluster assignments for each data point using the leader algorithm with distance threshold $T = 3$. A point belongs to a leader's cluster if the distance $\leq T$. **Remember to use city-block distance.**

- Run the algorithm in this order: data points $1, 2, 3, \ldots, 6$.
- Label the leaders in the plot (e.g. L_i)
- Draw the cluster boundaries around each leader.

b) Repeat (a) and find the leader algorithm cluster assignments for the relabeled data in the order: $1, 2, 3, \ldots, 6$.


y
8
6
4
2
2
4
6
8
x
4
3
2
6
5
1