k-Means Clustering

Problem: The following table gives the locations of data points (P_i) on a flat, 2-D coordinate surface. It also provides the initial random locations of two centroids (C_k).

Name	x	y
P_1	2	2
P_2	4	3
P_3	8	8
P_4	9	4
P_5	10	6
C_1	6	7
C_2	7	2

The goal is to find *k*=2 clusters, using the Manhattan distance metric.

<u>*k*-Means Algorithm:</u>

Initialize: select *k* random centroids
Repeat

1. **Assign all points to nearest centroid *m* using the error metric *b*:**

$$b_i^t \leftarrow \begin{cases} 1 \text{ if } \|x^t - m_i\| = \min_j \|x^t - m_j\| \\ 0 \text{ otherwise} \end{cases}$$

2. **Re-compute centroid *m* of each cluster:**

$$m_i \leftarrow \frac{\sum_t b_i^t x^t}{\sum_t b_i^t}$$

Until centroids are stable

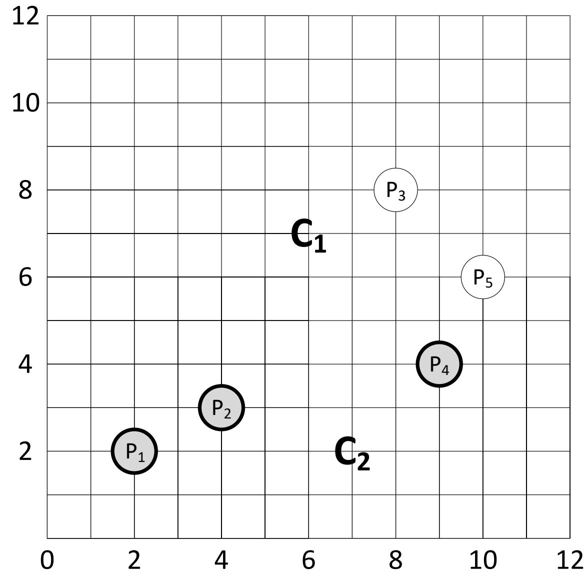
where the ***x*** are data points, and the ***m*** are centroids.

<u>Manhattan Distance metric</u>

$$d(\mathrm{P}_1, \mathrm{P}_2) = |x_1 - x_2| + |y_1 - y_2|$$

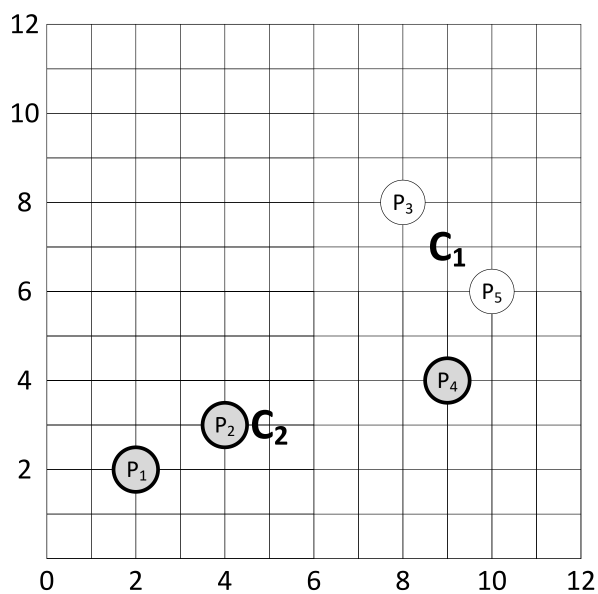
1a) Assign points to nearest centroid. Points P_1, P_2, and P_4 (shaded circles) are closer to centroid C_2 (7, 2). Points P_3 and P_5 (clear circles) are closer to centroid C_1 (6, 7):

Name	$d(P_i, C_1)$	$d(P_i, C_2)$
P_1	9	5
P_2	6	4
P_3	3	7
P_4	6	4
P_5	5	7



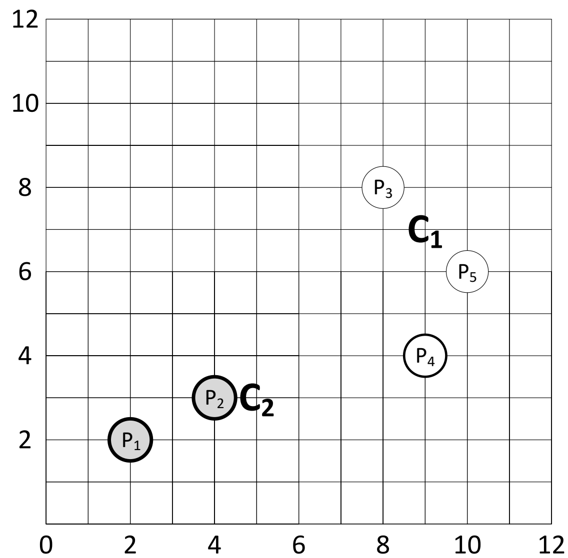
1b) Re-compute centroid locations based on center of mass of the two clusters:

C_1: (6, 7) → (9, 7) C_2: (7, 2) → (5, 3)

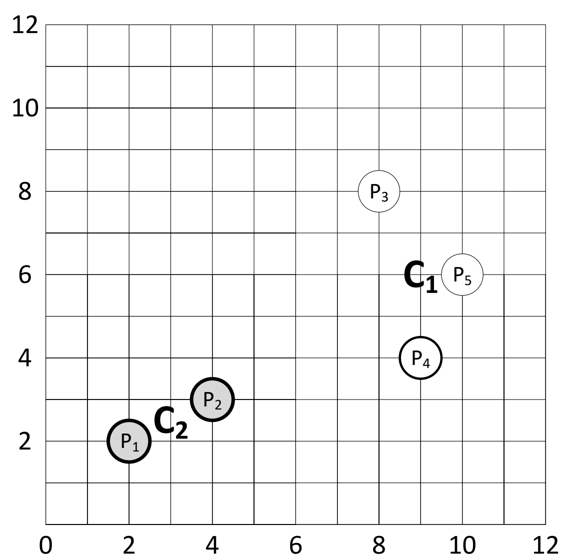


2a) Re-assign points to nearest centroid. Points $P_1$and P_2 (shaded circles) are still closer to centroid C_2; points P_3, P_5 and now P_4 (clear circles) are closer to centroid C_1:

Name	**d(P_i, C_1)**	**d(P_i,C_2)**
P_1	12	4
P_2	9	1
P_3	2	8
P_4	3	5
P_5	2	8



2b) Re-compute centroid locations based on center of mass of the two clusters:
C_1: (9, 7) → (9, 6) C_2: (5, 3) → (3, 2.5)



No points will change clusters ⇨ the centroids are stable.