Seminar 9

Algorithm 1 K-means($\{\vec{x}_1, \ldots, \vec{x}_N\}, K$, stopping criterion)

$(\vec{s}_1, \ldots, \vec{s}_K) \leftarrow$ SelectRandomSeeds($\{\vec{x}_1, \ldots, \vec{x}_N\}, K$)
for $k \leftarrow 1$ to K **do**
    $\vec{\mu}_k \leftarrow \vec{s}_k$
end for
repeat
    for $k \leftarrow 1$ to K **do**
        $\omega_k \leftarrow \{\}$
    end for
    for $n \leftarrow 1$ to N **do**
        $j \leftarrow \mathrm{argmin}_{j'} |\vec{\mu}_{j'} - \vec{x}_n|$
        $\omega_j \leftarrow \omega_j \cup \{\vec{x}_n\}$ ▷ reassigning vectors
    end for
    for $k \leftarrow 1$ to K **do**
        $\vec{\mu}_k \leftarrow \frac{1}{|\omega_k|} \sum_{\vec{x} \in \omega_k} \vec{x}$ ▷ recomputing centroids
    end for
until a stopping criterion has been met
return $\{\vec{\mu}_1, \ldots, \vec{\mu}_K\}$

Exercise 9/1

Use the K-means algorithm with Euclidean distance to cluster the following $N = 8$ examples into $K = 3$ clusters: $A_1 = (2, 10)$, $A_2 = (2, 5)$, $A_3 = (8, 4)$, $A_4 = (5, 8)$, $A_5 = (7, 5)$, $A_6 = (6, 4)$, $A_7 = (1, 2)$, $A_8 = (4, 9)$. Suppose that the initial seeds (centers of each cluster) are A_1, A_4 and A_7. Run the K-means algorithm for 3 epochs. After each epoch, draw a 10×10 space with all the 8 points and show the clusters with the new centroids.

$d(A, B)$ denotes the Euclidean distance between $A = (a_1, a_2)$ and $B = (b_1, b_2)$. It is calculated as $d(A, B) = \sqrt{(a_1 - b_1)^2 + (a_2 - b_2)^2}$.

Take seeds $\vec{s}_1 = A_1 = (2, 10)$, $\vec{s}_2 = A_4 = (5, 8)$, $\vec{s}_3 = A_7 = (1, 2)$.

By 1 we count the alignment for epoch 1: $A_1 \in \omega_1$, $A_2 \in \omega_3$, $A_3 \in \omega_2$, $A_4 \in \omega_2$, $A_5 \in \omega_2$, $A_6 \in \omega_2$, $A_7 \in \omega_3$, $A_8 \in \omega_2$; and we get the clusters: $\omega_1 = \{A_1\}$, $\omega_2 = \{A_3, A_4, A_5, A_6, A_8\}$, $\omega_3 = \{A_2, A_7\}$.

Centroids of the clusters: $\vec{\mu}_1 = (2, 10)$, $\vec{\mu}_2 = ((8+5+7+6+4)/5, (4+8+5+4+9)/5) = (6, 6)$, $\vec{\mu}_3 = ((2+1)/2, (5+2)/2) = (1.5, 3.5)$.

After epoch 2 the clusters are $\omega_1 = \{A_1, A_8\}$, $\omega_2 = \{A_3, A_4, A_5, A_6\}$, $\omega_3 = \{A_2, A_7\}$ with centroids $\vec{\mu}_1 = (3, 9.5)$, $\vec{\mu}_2 = (6.5, 5.25)$ and $\vec{\mu}_3 = (1.5, 3.5)$. And finally after epoch 3, the clusters are $\omega_1 = \{A_1, A_4, A_8\}$, $\omega_2 = \{A_3, A_5, A_6\}$, $\omega_3 = \{A_2, A_7\}$ with centroids $\vec{\mu}_1 = (3.66, 9)$, $\vec{\mu}_2 = (7, 4.33)$ and $\vec{\mu}_3 = (1.5, 3.5)$.

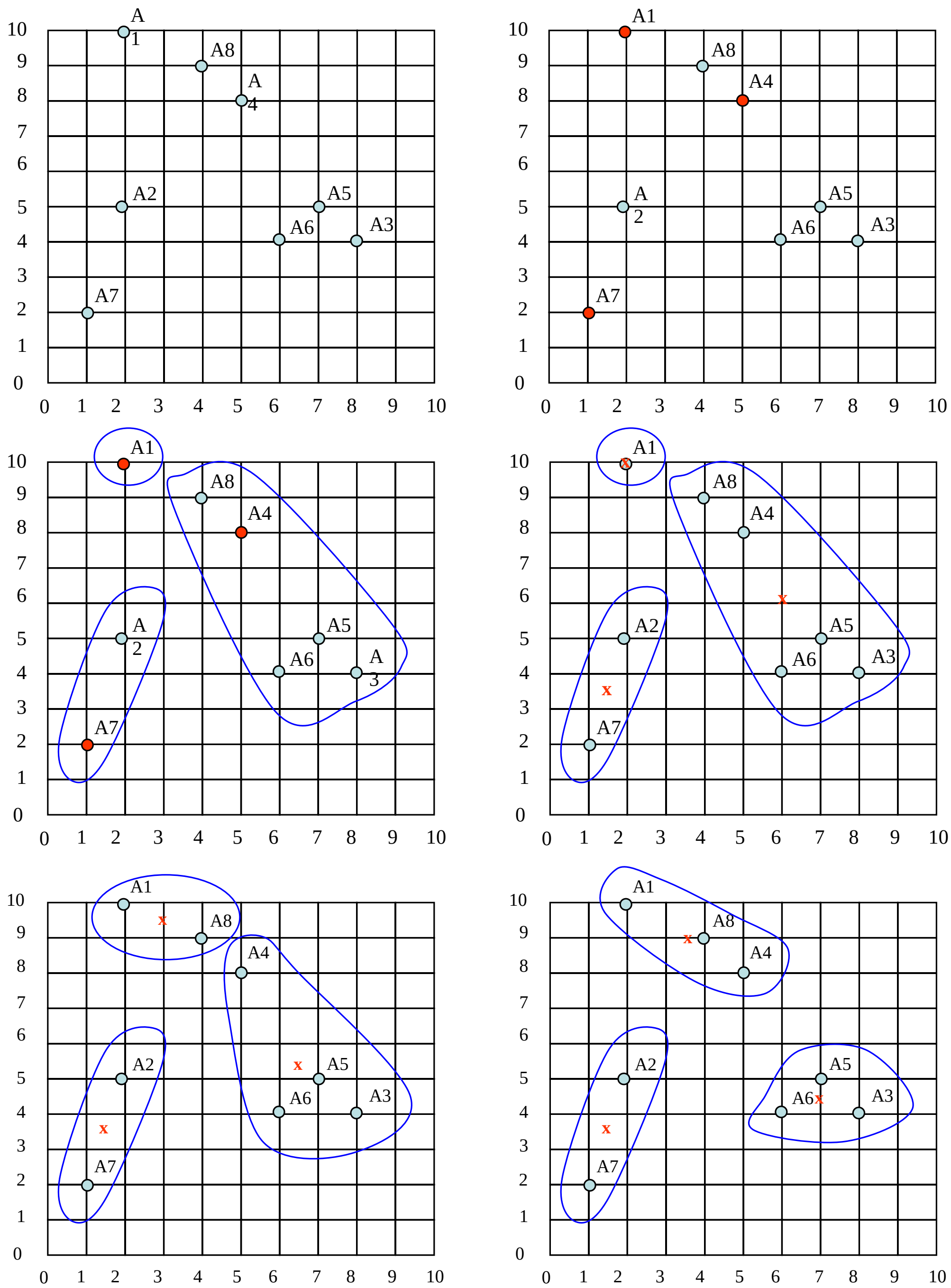

Figure 1: Visualization of K-means clustering algorithm.

Exercise 9/2

Consider three points: $A_1 = [1, 1]$, $A_2 = [3, 1]$, $A_3 = [6, 1]$. Give an example of a point A_4 such that the K-means clustering algorithm with seeds $\{A_2, A_4\}$ and the agglomerative hierarchical clustering algorithm result in different clusterings of $\{A_1, A_2, A_3, A_4\}$ into 2 classes.

For example, if $A_4 = [2, 1]$, then K-means results in $\{\{A_1, A_4\}, \{A_2, A_3\}\}$ and agglomerative in $\{\{A_1, A_2, A_4\}, \{A_3\}\}$.

Exercise 9/3

What makes a good clustering? Give some clustering evaluation metrics.

Answers can vary. For official definition refer to the Manning book.