

# W3WI DS304.1 Applied Machine Learning Fundamentals

## Exercise Sheet # 9 - Clustering

### Question 1 EX 2021 (*k*-means algorithm)

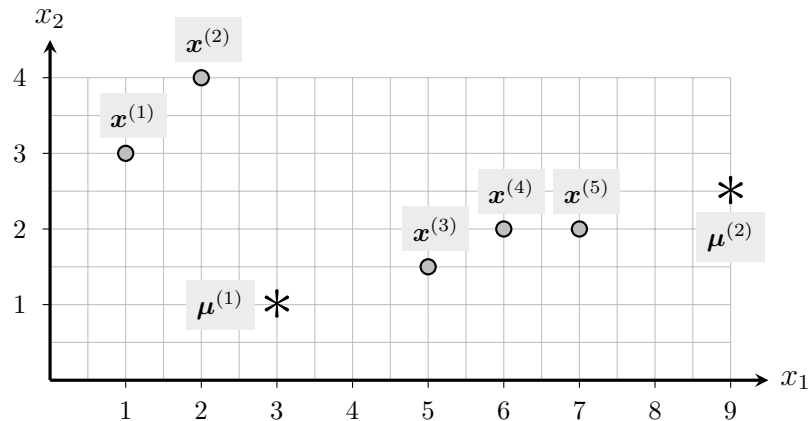
Briefly describe the *k*-means algorithm in your own words. Which steps have to be executed? Does the algorithm always find the optimal solution?

### Question 2 EX 2020 (*k*-means algorithm by hand)

Figure 1 below plots a small set of data points  $\mathbf{x}^{(i)} \in \mathbb{R}^2$  ( $1 \leq i \leq 5$ ) as well as two randomly initialized cluster centroids  $\boldsymbol{\mu}^{(1)}$  and  $\boldsymbol{\mu}^{(2)}$  denoted by  $*$ . Perform one iteration of the *k*-means algorithm using the Euclidean distance metric given by

$$d_2(\mathbf{x}, \boldsymbol{\mu}) := \sqrt{\sum_{j=1}^m (x_j - \mu_j)^2}.$$

Fill in the tables 1 and 2 below with your results and mark the updated cluster means in the plot. Has the algorithm already converged?



**Figure 1:** Plot of the dataset and the initialized cluster centroids.

$i$	$\mathbf{x}^{(i)}$	$d_2(\mathbf{x}^{(i)}, \boldsymbol{\mu}^{(1)})$	$d_2(\mathbf{x}^{(i)}, \boldsymbol{\mu}^{(2)})$	Cluster assignment (1 or 2)
1	(1.0, 3.0)			
2	(2.0, 4.0)			
3	(5.0, 1.5)			
4	(6.0, 2.0)			
5	(7.0, 2.0)			

**Table 1:** Cluster assignments: Add your results to this table!

$\ell$	Before update: $\boldsymbol{\mu}^{(\ell)}$	After update: $\boldsymbol{\mu}_{\text{new}}^{(\ell)}$
1	(3.0, 1.0)	
2	(9.0, 2.5)	

**Table 2:** Update of cluster centroids: Add your results to this table!

**Question 3** EX 2022 (Choice of  $k$ )

How can you choose a suitable value for the hyper-parameter  $k$  in the  $k$ -means algorithm?