Artificial Intelligence and Machine Learning

Exercises – Clustering

Question 1 (KMeans algorithm) ⊛

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

Question 2 (KMeans algorithm by hand) ⊛

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

$$d_2(\boldsymbol{x}, \boldsymbol{\mu}) := \sqrt{\sum_{m=1}^{M} (x_m - \mu_m)^2}.$$

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 after one iteration?

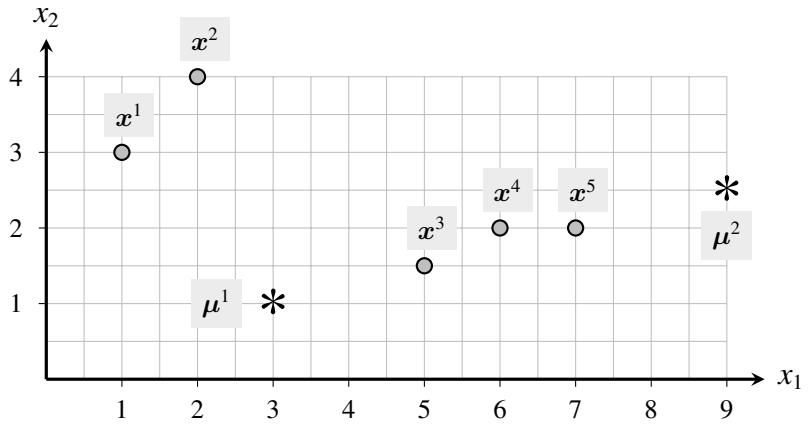

Figure 1: Plot of the dataset and the initialized cluster centroids.

n	$\boldsymbol{x}^n$	$d_2(\boldsymbol{x}^n, \boldsymbol{\mu}^1)$	$d_2(\boldsymbol{x}^n, \boldsymbol{\mu}^2)$	**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!

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

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

Question 3 (Choice of K) ⊛

How can you choose a suitable value for the hyperparameter K in the KMeans algorithm?

Question 4 (Implementing KMeans)

Implement the KMeans algorithm to cluster the dataset generated by the following code snippet.

```
from sklearn.datasets import make_blobs

X, _ = make_blobs(
        n_samples=250,
        n_features=2,
        cluster_std=5.50,
        centers=3,
        shuffle=False,
        random_state=42
)
```

Compress an image of your choice. The `imageio` library may be helpful.