Deadline: Th. January 23th , 14:00 Drop your printed or legible handwritten submissions into the boxes at Samelsonplatz. Alternatively upload a `.pdf` file via LearnWeb. (e.g. exported Jupyter notebook)

1. K Means Clustering (10 points)

A. [2p] Compute the (squared) **distance matrix** $D_{ij} = \text{dist}_{\text{eucl.}}(x_i, x_j)^2$, given the data from Table 1.

B. [4p] Perform K-means clustering on the dataset from Table 1. Use the first and last datapoints as initial centers ($K = 2$). Given the final parameters, which cluster would $x^* = \binom{1}{1}$ belong to?

x_1	x_2
0	0
0	1
-1	2
2	0
3	0
4	-1

Table 1

C. [1p] For a set of points $(x_n)_{n=1:N}$ in $\mathbb{R}^m$, show that the **mean** $\hat{\mu} = \frac{1}{N}\sum_{n=1}^{N} x_n$ is the solution to the optimization problem

$$\hat{\mu} = \underset{\mu \in \mathbb{R}^m}{\text{argmin}} \sum_{i=1}^{N} \text{dist}_{\text{eucl.}}(x_n, \mu)^2 \tag{1}$$

I.e. for a set of points, their mean can be characterized as the point which is, on average, closest to all the other points with respect to the **squared euclidean distance**.

D. [3p] For a set of points $(x_n)_{n=1:N}$ in $\mathbb{R}^m$, the **geometric median** is defined as the point

$$\hat{\mu} = \underset{\mu \in \mathbb{R}^m}{\text{argmin}} \sum_{n=1}^{N} \text{dist}_{\text{eucl.}}(x_n, \mu) \tag{2}$$

Note that in contrast to the mean, (2) does not have a closed form solution. However, the minimum can be found numerically by a fixed point iteration scheme (algorithm 1). Given the dataset from Table 2 (rows are datapoints!), compute both the mean and the geometric median. What happens to both if we change the last datapoint to $\binom{100}{0}$?

x_1	x_2
-1	-1
-1	1
1	-1
1	1
10	0

Table 2

Algorithm 1: Weiszfeld's algorithm

$\mu^{(0)} = \frac{1}{N}\sum_{n=1}^{N} x_n$;
for $t = 0, 1, 2 \ldots,$ *max_iter* **do**
  $\mu^{(t+1)} = \left(\sum_{n=1}^{N} x_n \|x_n - \mu^{(t)}\|^{-1}\right) \Big/ \left(\sum_{n=1}^{N} \|x_n - \mu^{(t)}\|^{-1}\right)$;
  if *converged* **then**
    return μ

2. Gaussian Mixture Models (GMMs) (8 points)

Two datasets ("MOONS" and "STRIPES") were each clustered by 3 different methods: K-means clustering, Gaussian-Mixture-Models and Hierarchical Clustering (single link). The results are shown in Table 3.

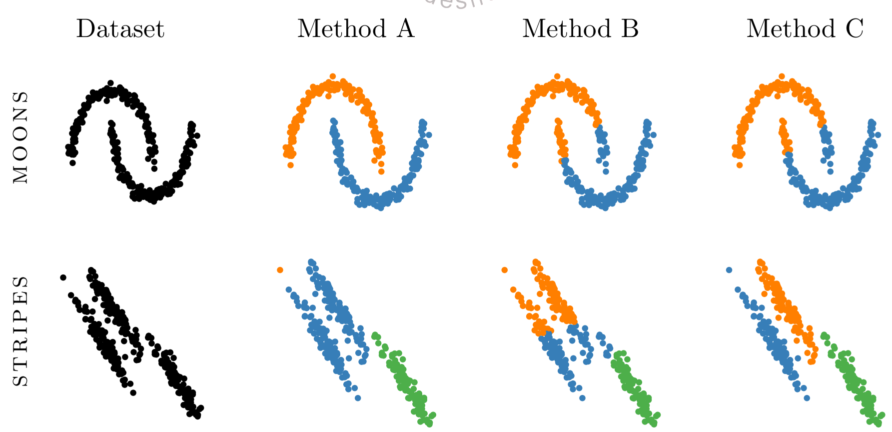

Table 3: Different Clustering Methods

A. [2p] Decide which method corresponds to A, B and C. Explain your decision.

B. [6p] Given the data from Table 1, and the initial configuration $\pi_1, \pi_2 = \frac{1}{2}$, $\mu_1 = \left(\begin{smallmatrix}0\\1\end{smallmatrix}\right)$, $\mu_2 = \left(\begin{smallmatrix}3\\0\end{smallmatrix}\right)$, $\Sigma_1, \Sigma_2 = \mathbb{I}$, perform 1 iteration of the (soft partition) EM algorithm to fit a GMM. Which cluster would $x^* = \left(\begin{smallmatrix}1\\1\end{smallmatrix}\right)$ belong to according the initial/final parameters?

3. Hierarchical Clustering (6 points)

A. [2p] Compute the **distance matrix** $D_{ij} = \text{dist}(x_i, x_j)$, using the **Manhatten distance** (i.e. L^1), given the data from Table 4.

B. [4p] Perform **agglomerative Hierarchical Clustering** using **single linkage** as the cluster distance measure. Draw the associated tree (as in slides 26/27).

x_1	x_2
0	0
1	0
2	0
-0.5	-1
0.5	-1
0	-1.5

Table 4