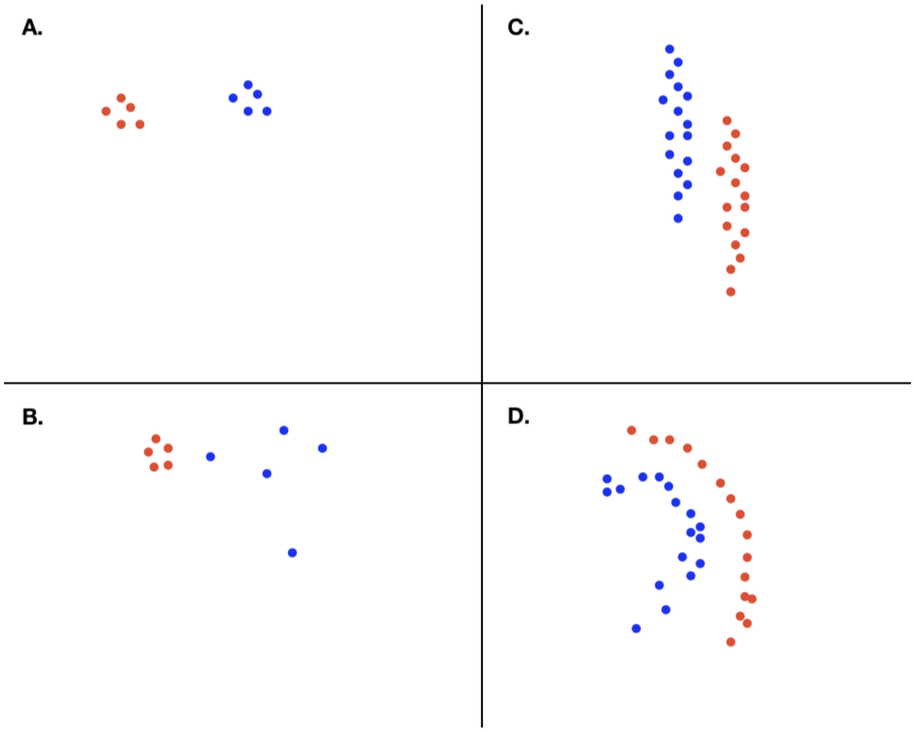
Unsupervised Learning: Clustering

Question 1:

In the plot below, which of the following options could have produced each clustering (multiple answers are possible): *K-means, Single linkage (hierarchical clustering), Gaussian Mixture Models*.



Question 2: Hierarchical Clustering

For four branches of a supermarket chain, the following values are obtained for the characteristics turnover and sales area, each measured in suitable units:

branch	1	2	3	4
turnover	8	5	10	4
sales area	24	22	25	21

Using the squared Euclidean distance as the distance between individual objects both times,

a) Perform a hierarchical clustering with the *Single Linkage* method

b) Perform a hierarchical clustering with the *Zentroid* method.

c) Draw the complete dendrograms for both methods.

Question 3:

a) For a set of points $(x_i)_{i=1}^m$ in $\mathbb{R}^m$, show that the arithmetic mean $\hat{\mu} = \frac{1}{m}\sum_{i=1}^m x_i$ is the solution to the optimization problem

$$\hat{\mu} = \underset{\mu \in \mathbb{R}^m}{\operatorname{argmin}} \sum_{i=1}^{n} \|x_i - \mu\|^2$$

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.

b) Consider the following six points in $\mathbb{R}^2$:

$$x_1 = \begin{pmatrix} 0 \\ 0 \end{pmatrix}; \; x_2 = \begin{pmatrix} 0 \\ 1 \end{pmatrix}; \; x_3 = \begin{pmatrix} -1 \\ 2 \end{pmatrix}; \; x_4 = \begin{pmatrix} 2 \\ 0 \end{pmatrix}; \; x_5 = \begin{pmatrix} 3 \\ 0 \end{pmatrix}; \; x_6 = \begin{pmatrix} 4 \\ -1 \end{pmatrix}.$$

Use Lloyd's algorithm and "random" initialization $\{x_1; x_6\}$ to perform **both** *k-means* and *k-medoids* (also with squared euclidean distance) clustering for $K = 2$.

Question 4:

a) Outline the model assumptions used in the Gaussian Mixed Models (GMMs). How can a GMM be fit?

b) Consider a one-dimensional Gaussian Mixture Model with 2 clusters and parameters $(\mu_1, \sigma_1^2, \mu_2, \sigma_2^2, \pi_1, \pi_2)$. Here (π_1, π_2) are the mixing weights, and (μ_1, σ_1^2), (μ_2, σ_2^2) are the centers and variances of the clusters. We are given a dataset $\mathcal{D} = \{x_1, x_2, x_3\} \subset \mathbb{R}$, and apply the EM-algorithm to find the parameters of the Gaussian mixture model. What is the complete log-likelihood that is being optimized for this problem?

c) Assume that the dataset $\mathcal{D}$ consists of the following three points, $x_1 = 1, x_2 = 10, x_3 = 20$. At some step in the EM-algorithm, we compute the expectation step which results in the following matrix: $\mathrm{T} = \begin{pmatrix} 1 & 0 \\ 0.4 & 0.6 \\ 0 & 1 \end{pmatrix}$, where τ_{ij} denotes the probability of x_i belonging to cluster j.

Given the above T for the expectation step, write the result of the following maximization step, specifically the

- mixing weights π_1, π_2
- centers μ_1, μ_2
- variance values σ_1^2, σ_2^2