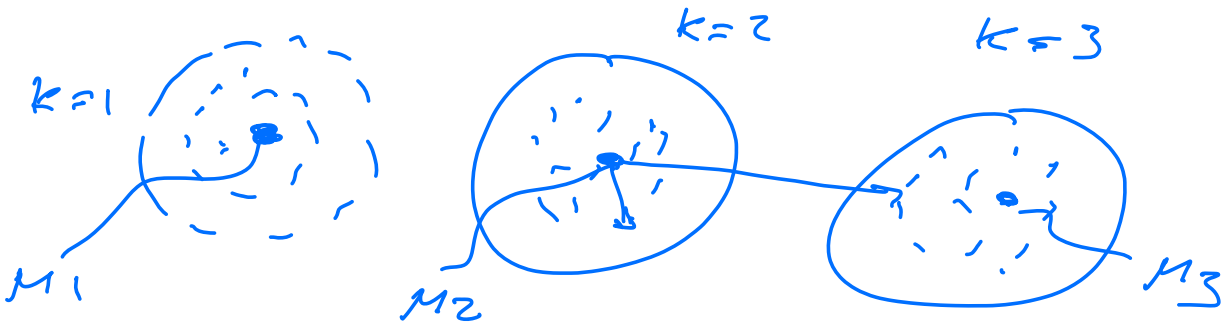
FYS-STK 4155, NOV 25, 2022

K-means clustering

Suppose we have a data set $\{x_0, x_1, \ldots, x_{n-1}\}$ of random D-dimensional variable x.

The goal is to find a partition of the data into some number of clusters

- K -



introduce a variable μ_K (center of each cluster)

The goal is to find the assignment of points

belonging to a given cluster – k –

Define cost function

$$C = \sum_{i=0}^{n-1} \sum_{k=0}^{K-1} r_{ik} \|x_i - \mu_k\|^2$$

$r_{ik} = \{0, 1\}$

1 if inside cluster
0 else

optimization:

- First select initial values for μ_k
  - optimize C wrt r_{ik} keeping μ_k fixed
- Second stage
  - optimize C wrt μ_k keeping r_{ik} fixed.
- Continue till convergence criterion has been reached.

$$r_{ik} = \begin{cases} 1 & \text{if } k = \arg\min_j \|x_i - \mu_j\|^2 \\ 0 & \text{else} \end{cases}$$

μ_k Derivative wrt μ_k

$$2 \sum_{i=0}^{n-1} r_{ik}(x_i - \mu_k) = 0$$

$$\Rightarrow \quad \mu_k = \frac{\sum_i r_{ik} x_i}{\sum_i r_{ik}}$$

$\sum_i r_{ik}$ ↑ # points in a cluster $-k-$