$K \equiv$ # of clusters (hyperparameter)

$\theta_j \equiv$ cluster centroid for cluster j

$u_{ij} \equiv$ membership of point x_i in cluster w/ centroid θ_j

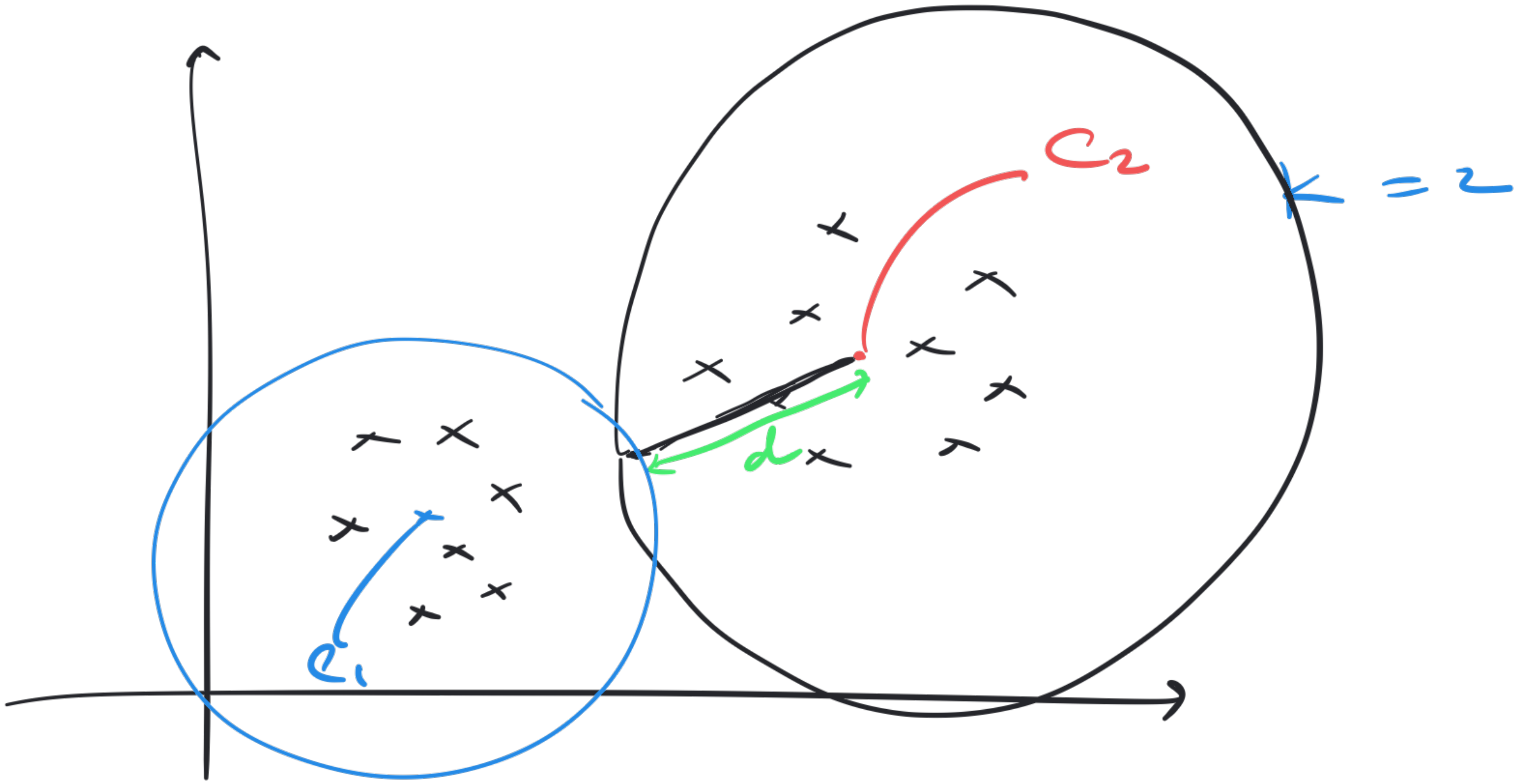
$$\boxed{J(\theta, U) = \sum_{i=1}^{N} \sum_{j=1}^{K} u_{ij}\, d^2(x_i, \theta_j) \qquad u_{ij} \in \{0,1\} \text{ and } \sum_{j=1}^{K} u_{ij} = 1 \quad \forall i}$$

$d(x_i, \theta_j) \equiv$ distance (with any metric) between x_i and θ_j

$$d_E(x_i, \theta_j) = \|x_i - \theta_j\|_2$$

① If using Euclidean distance

$$J(\theta, U) = \sum_{i=1}^{N} \sum_{j=1}^{K} u_{ij} \cdot \|x_i - \theta_j\|_2^2$$



$$J(\theta, U) = \sum_{i=1}^{N} \sum_{j=1}^{K} u_{ij} \cdot \|x_i - \theta_j\|_2^2$$

optimize for the membership U: $\boxed{\text{E-STEP}}$

θ_j are fixed

$$\frac{\partial J}{\partial u_{ij}} = 0 \iff \sum_{i=1}^{N} \|x_i - \theta_j\|_2^2 = 0$$

$$\Rightarrow u_{ij} = \begin{cases} 1, & \text{if } j = \arg\min \|x_i - \theta_j\|_2^2 \\ 0, & \text{otherwise} \end{cases}$$

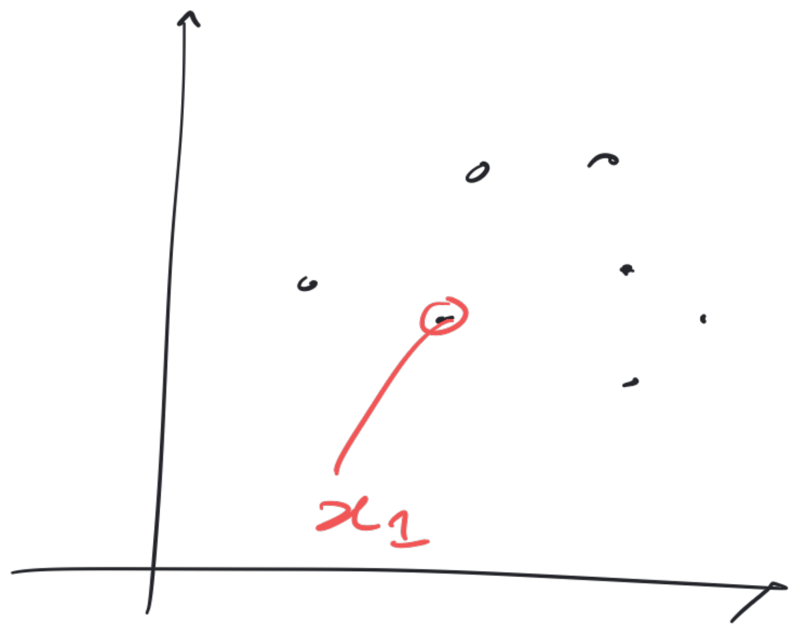
$$U_{N \times K} = \begin{array}{c} \\ 1 \\ 2 \\ 3 \\ \vdots \\ N \end{array} \begin{array}{c} \begin{array}{cccc} 1 & 2 & \cdots & K \end{array} \\ \left[\begin{array}{cccc} 0 & 1 & \cdots & 0 \\ & & & \\ & & & \\ & & & \\ & & & \end{array} \right] \end{array}$$

$$\sum_{j=1}^{K} u_{1j} = 1$$

② optimize for cluster centroids:

$$J(\theta, U) = \sum_{i=1}^{N} \sum_{j=1}^{K} u_{ij} \, \|x_i - \theta_j\|_2^2$$

$$= \sum_{i=1}^{N} \sum_{j=1}^{K} u_{ij} \cdot (x_i - \theta_j)^T (x_i - \theta_j)$$



$$x_1 = \begin{bmatrix} x_{11} \\ x_{12} \end{bmatrix}$$

$$\frac{\partial J(\theta, U)}{\partial \theta_j} = 0$$

$$\Leftrightarrow \sum_{i=1}^{N} u_{ij} (-2) \cdot (x_i - \theta_j) = 0$$

$$\Leftrightarrow \sum_{i=1}^{N} u_{ij} \cdot x_i - \sum_{i=1}^{N} u_{ij} \cdot \theta_j = 0$$

$$\Leftrightarrow \theta_j = \frac{\sum_{i=1}^{N} u_{ij} \cdot x_i}{\sum_{i=1}^{N} u_{ij}} \underset{j=1}{=} \frac{\sum_{i=1}^{N} u_{i1} \cdot x_i}{\sum_{i=1}^{N} u_{i1}}$$

$$\Leftrightarrow \boxed{\theta_j = \sum_{x_i \in C_j} x_i \Big/ N_j}$$

= AVERAGE of points assigned to cluster with centroid θ_j

Observations:

① K-means is sensitive to feature scaling.

Must scale the data prior to run K-means.