CS4641B Machine Learning

Focus video: K-Means

Rodrigo Borela ▸ rborelav@gatech.edu

Formal statement of the clustering problem

- Given N data points, $\{\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_N\} \in \mathbb{R}^{N \times D}$ $\rightarrow x_j^T = [\ ,\ ,\ ]$
- Find k cluster centers $\{\boldsymbol{\mu}_1, \boldsymbol{\mu}_2, \dots, \boldsymbol{\mu}_K\} \in \mathbb{R}^{K \times D}$
- And assign each data point $\mathbf{x}_n$ to one cluster k such that $r_{nk} = 1$ and $r_{nj} = 0$ for $j \neq k$ (1-of-K encoding)
- Such that the average square distances from each data point to its respective cluster center (distortion measure) is small:

$$\min_{\boldsymbol{\mu}_k, r_{nk}} \sum_{n=1}^{N} \sum_{k=1}^{K} r_{nk} \|\mathbf{x}_n - \boldsymbol{\mu}_k\|_2^2$$

Formal statement of the clustering problem

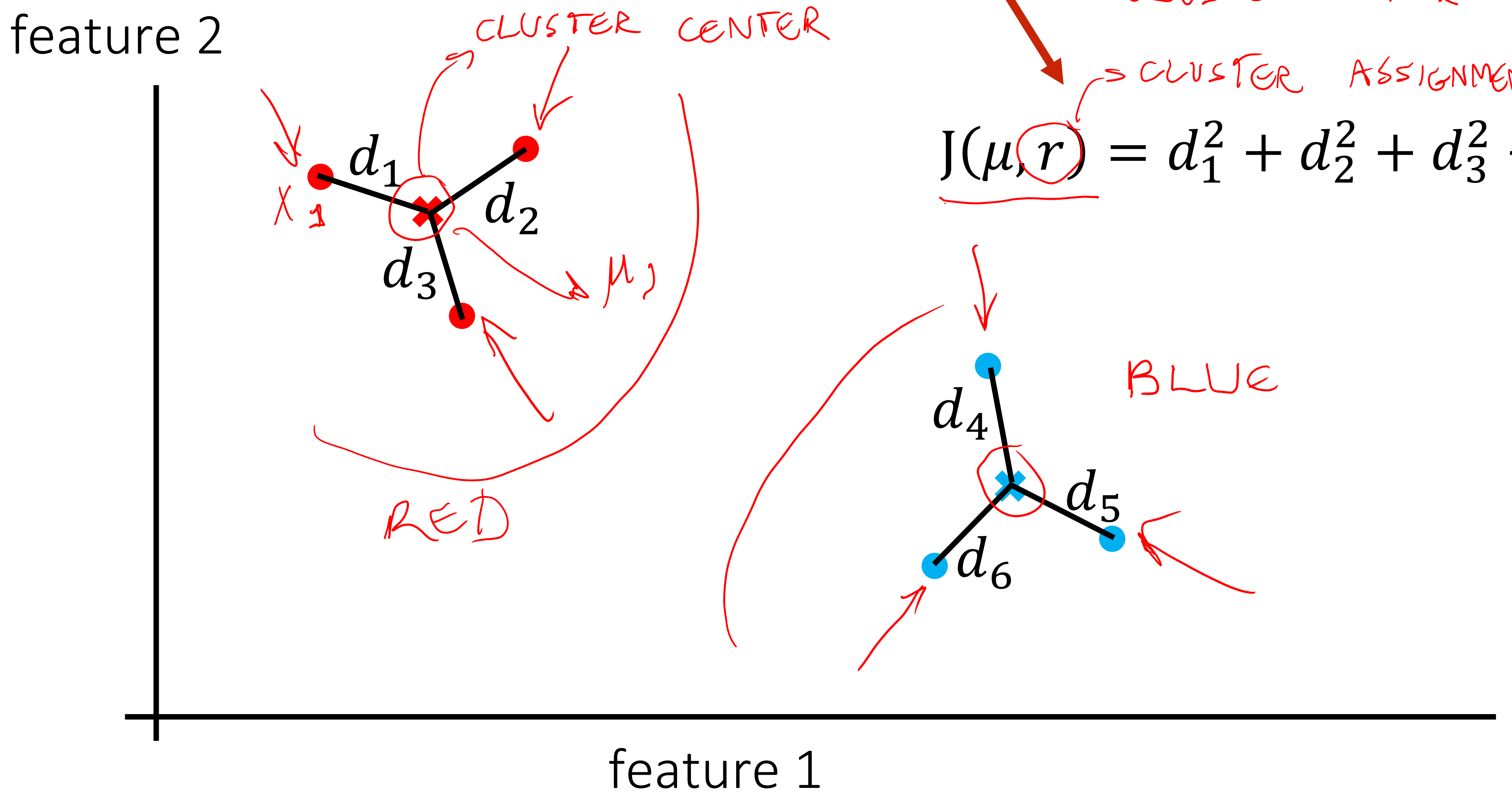
POINTS

$$\boxed{\min_{\mu_k, r_{nk}} \sum_{n=1}^{N} \sum_{k=1}^{K} r_{nk} \|\mathbf{x}_n - \boldsymbol{\mu}_k\|_2^2}$$

CLUSTER CENTER

CLUSTER ASSIGNMENT

$$J(\mu, r) = d_1^2 + d_2^2 + d_3^2 + d_4^2 + d_5^2 + d_6^2$$



K-means algorithm revisited

- Step 1: Keeping $\boldsymbol{\mu}_k$ and computing the squared distances between $\mathbf{x}_n$ and $\boldsymbol{\mu}_k$, we can optimize the objective simply by assigning $\mathbf{x}_n$ to the nearest cluster center

$$r_{nk} = \begin{cases} 1 & \text{if } k = \arg\min_j \|\mathbf{x}_n - \boldsymbol{\mu}_k\|_2^2 \\ 0 & \text{otherwise} \end{cases}$$

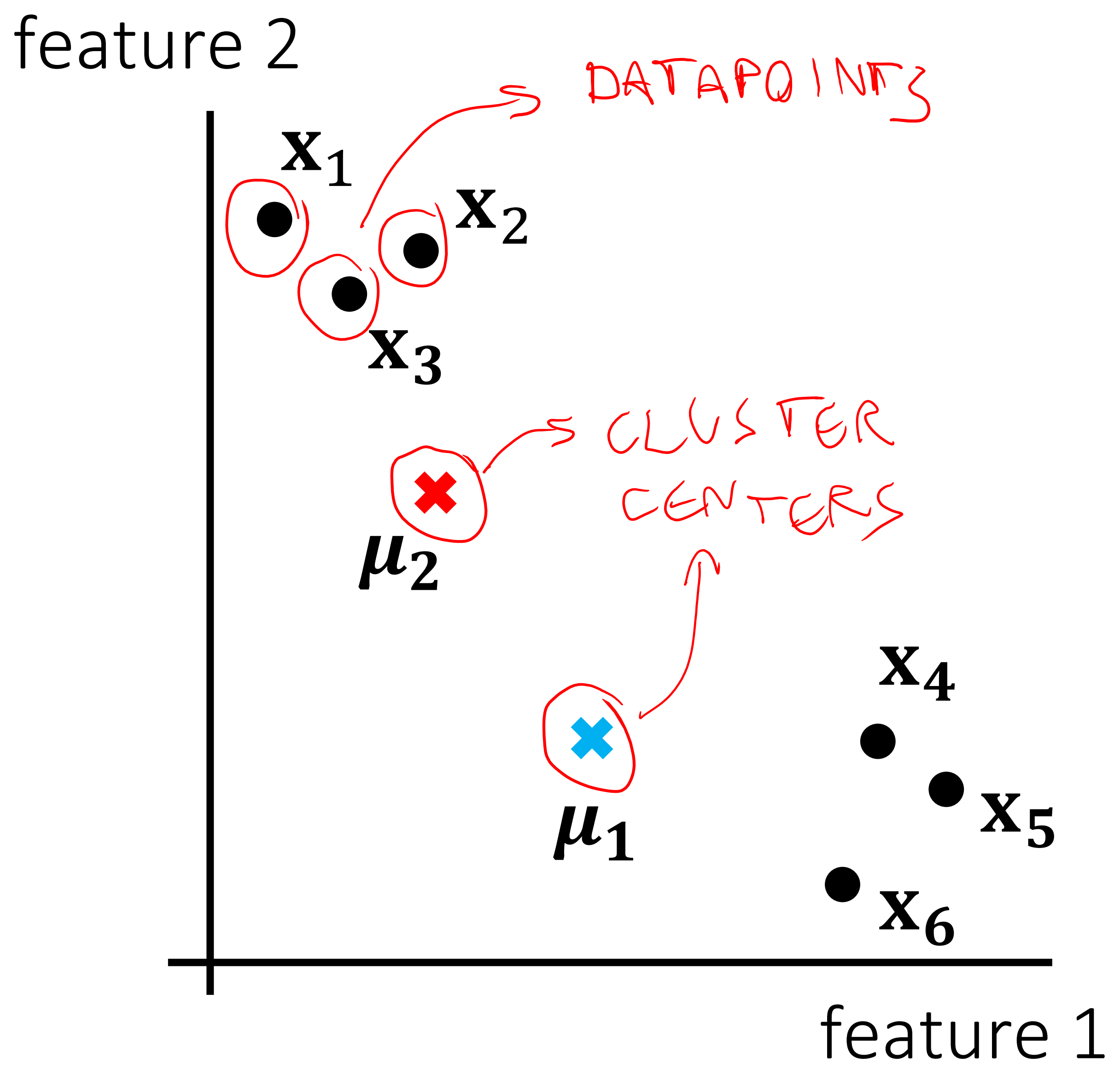
CLUSTER ASSIGNMENT

- **Step 2:** Keeping r_{nk} fixed we can optimize the objective with respect to $\boldsymbol{\mu}_k$ by setting the derivative wrt to $\boldsymbol{\mu}_k$ to zero

$$\frac{\partial}{\partial \boldsymbol{\mu}_k} \sum_{n=1}^{N} r_{nk}(\mathbf{x}_n - \boldsymbol{\mu}_k)^T(\mathbf{x}_n - \boldsymbol{\mu}_k) = 2\sum_{n=1}^{N} r_{nk}(\mathbf{x}_n - \boldsymbol{\mu}_k) = 0 \rightarrow \boldsymbol{\mu}_k = \frac{\sum_n r_{nk}\mathbf{x}_n}{\sum_n r_{nk}}$$

COMPUTING NEW CLUSTER CENTER

K-means algorithm example



x_1^T

$$\text{Dataset: } \mathbf{X} = \begin{bmatrix} \mathbf{x}_1^T \\ \mathbf{x}_2^T \\ \mathbf{x}_3^T \\ \mathbf{x}_4^T \\ \mathbf{x}_5^T \\ \mathbf{x}_6^T \end{bmatrix}_{N \times D} = \begin{bmatrix} 1.0 & 8.0 \\ 2.5 & 7.5 \\ 2.0 & 7.0 \\ 8.5 & 2.5 \\ 9.0 & 2.0 \\ 8.0 & 1.0 \end{bmatrix}_{N \times D = 6 \times 2}$$

$$\text{Cluster assignment: } \mathbf{R} = \begin{bmatrix} ? & ? \\ ? & ? \\ ? & ? \\ ? & ? \\ ? & ? \\ ? & ? \end{bmatrix}_{N \times K = 6 \times 2}$$

μ_1^T

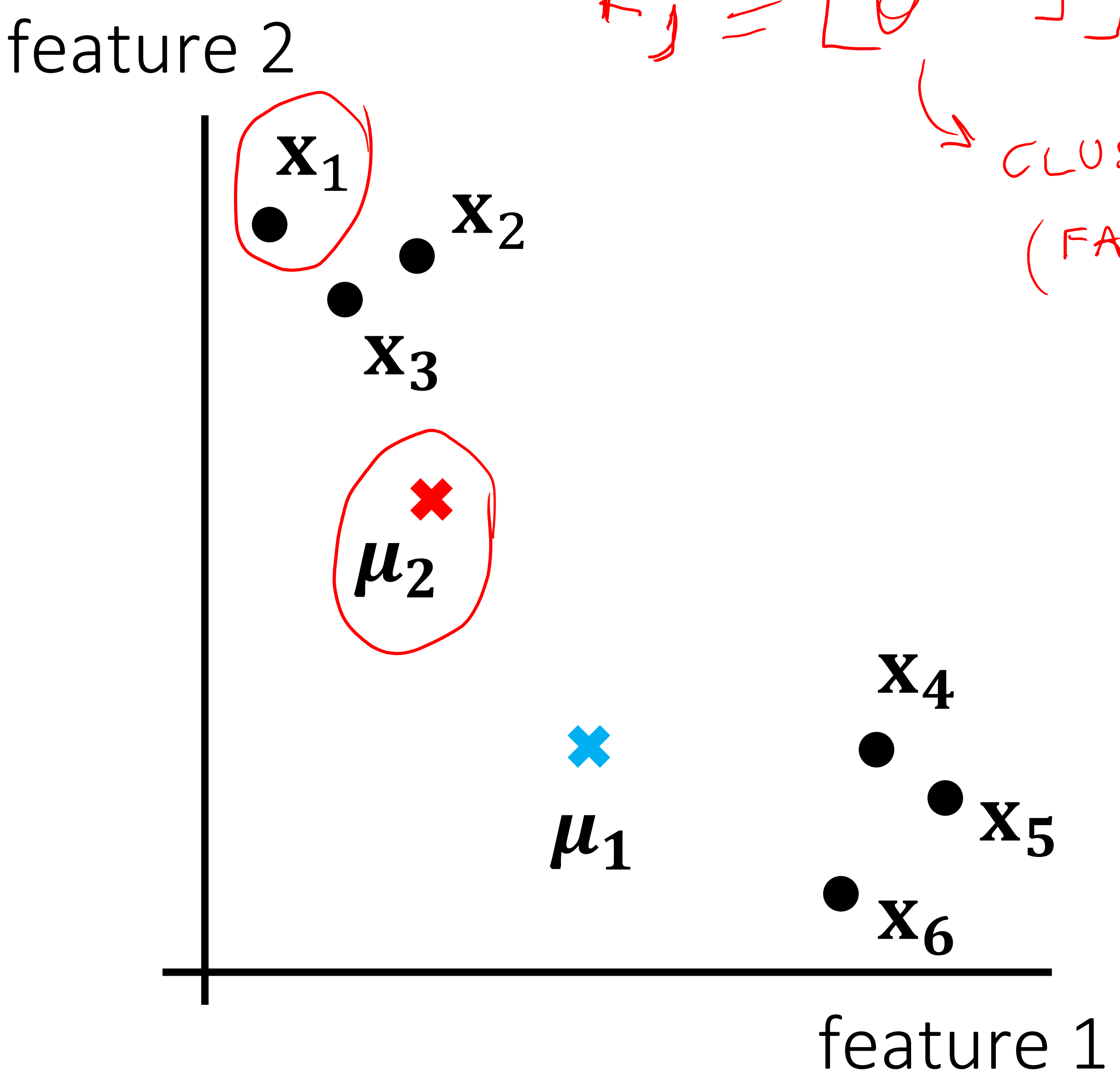
$$\text{Cluster centers: } \mathbf{M} = \begin{bmatrix} \boldsymbol{\mu}_1^T \\ \boldsymbol{\mu}_2^T \end{bmatrix} = \begin{bmatrix} 4.5 & 2.5 \\ 2.5 & 5.0 \end{bmatrix}_{K \times D = 2 \times 2}$$

K-means algorithm: Step 1

CLUSTER 2 (TRUE)

$r_1 = [0 \quad 1]$

CLUSTER 1 (FALSE)

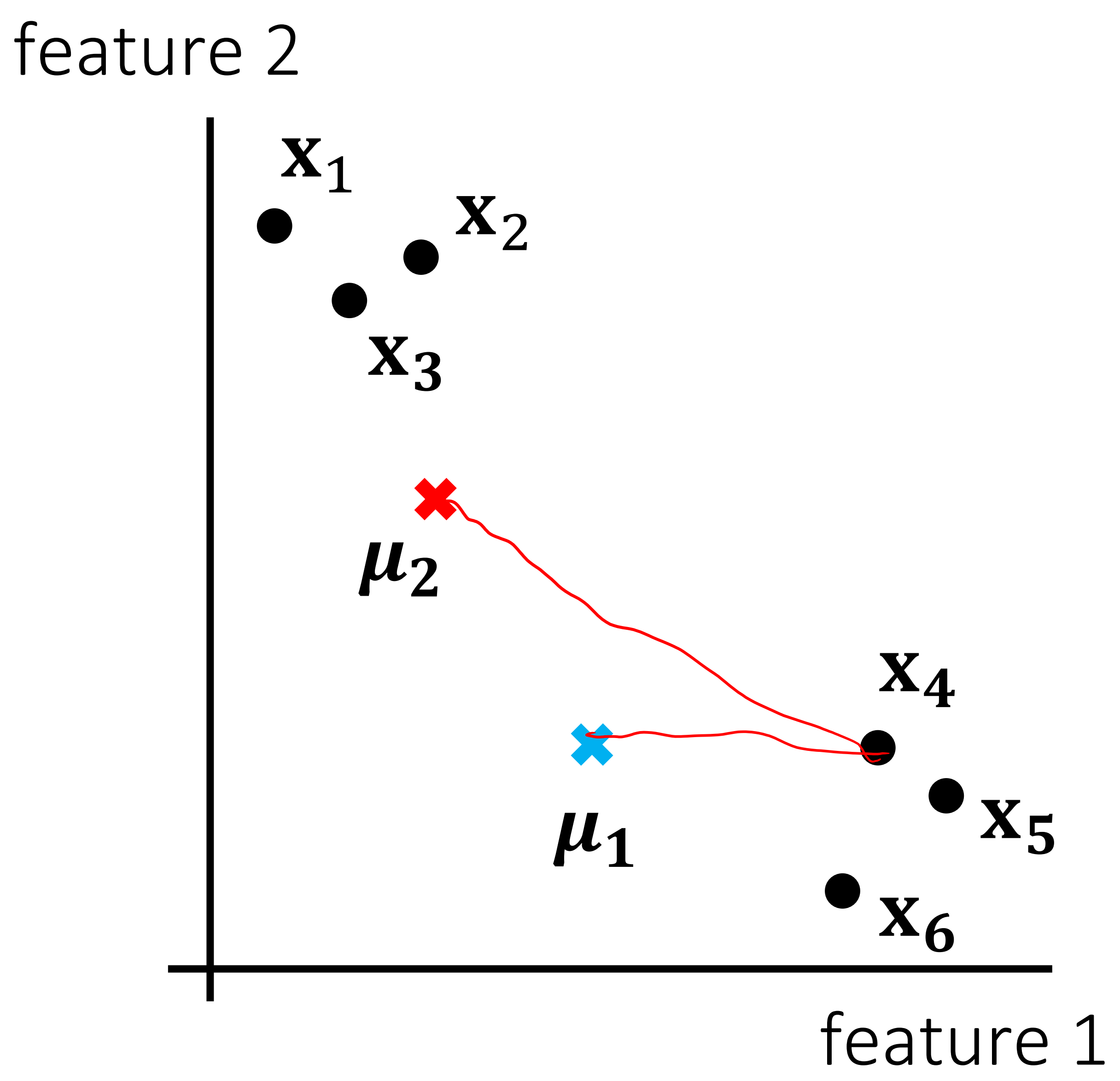


Keeping $\boldsymbol{\mu}_k$ and computing the squared distances between $\mathbf{x}_n$ and $\boldsymbol{\mu}_k$, we can optimize the objective simply by assigning $\mathbf{x}_n$ to the nearest cluster center

$$r_{nk} = \begin{cases} 1 & \text{if } k = \arg\min_j \|\mathbf{x}_n - \boldsymbol{\mu}_k\|_2^2 \\ 0 & \text{otherwise} \end{cases}$$

1-OF-K ENCODING

K-means algorithm: Step 1



$$\underbrace{\|\mathbf{x}_n - \boldsymbol{\mu}_k\|_2^2 = \sum_{i=1}^{D} (x_{i_n} - \mu_{i_k})^2}$$

DATA

CLUSTER CENTER

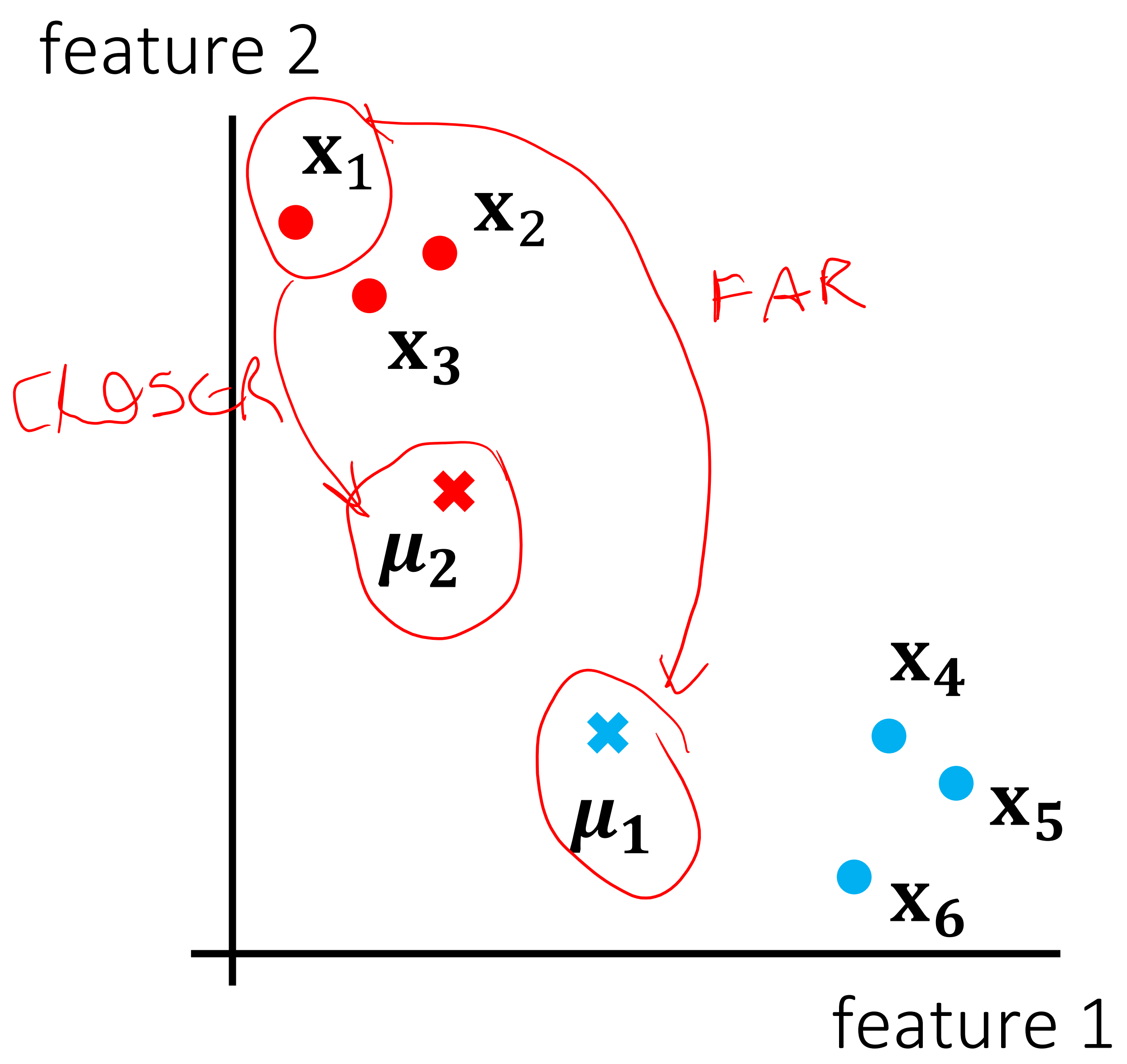
$$\mathbf{X} = \begin{bmatrix} 1.0 & 8.0 \\ 2.5 & 7.5 \\ 2.0 & 7.0 \\ 8.5 & 2.5 \\ 9.0 & 2.0 \\ 8.0 & 1.0 \end{bmatrix} \begin{matrix} \\ \\ x_3 \\ \\ \\ \\ \end{matrix}, \quad \mathbf{M} = \begin{bmatrix} 4.5 & 2.5 \\ 2.5 & 5.0 \end{bmatrix} \begin{matrix} \\ \mu_2 \end{matrix}$$

SQUARED DISTANCE (x_1, μ_2)

Squared pairwise distances: $\mathbf{D} = \begin{bmatrix} 42.5 & 11.25 \\ 29.0 & 6.25 \\ 26.5 & 4.25 \\ 16.0 & 42.25 \\ 20.5 & 51.25 \\ 14.5 & 46.25 \end{bmatrix}_{N \times K}$ $d(x_3, \mu_2)$

$$(2-2.5)^2 + (7-5)^2 = 4.25$$

K-means algorithm: Step 1



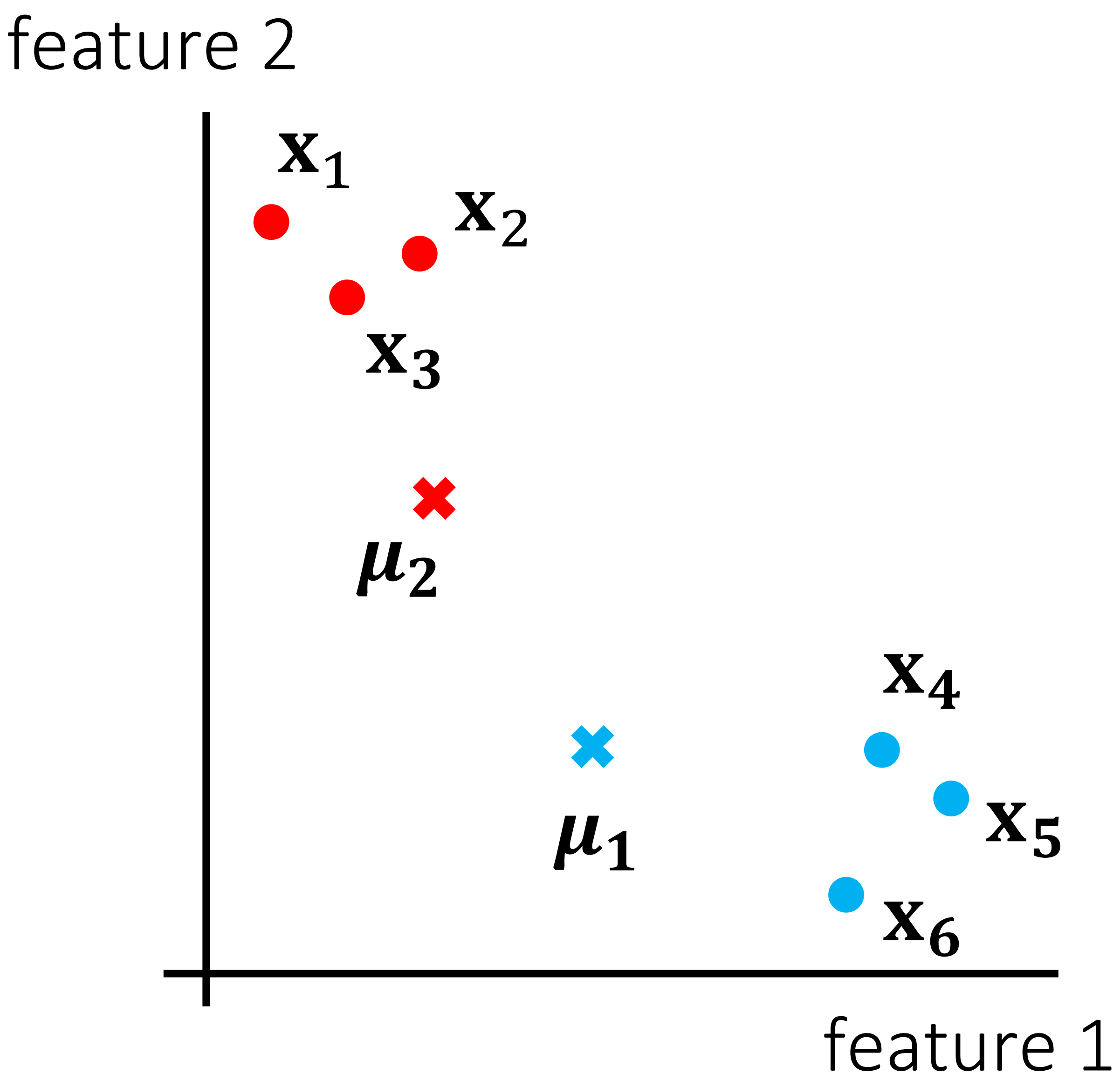
MATRIX D

$$r_{nk} = \begin{cases} 1 & \text{if } k = \arg\min_j \|\mathbf{x}_n - \boldsymbol{\mu}_k\|_2^2 \\ 0 & \text{otherwise} \end{cases}$$

CLUSTER ASSIGNMENT X1

$$\begin{array}{cc} & \begin{array}{cc} \mu_1 & \mu_2 \end{array} \\ \mathbf{D} = \begin{array}{c} x_1 \rightarrow \\ \\ \\ \\ \\ \\ \end{array} & \begin{bmatrix} 42.5 & 11.25 \\ 29.0 & 6.25 \\ 26.5 & 4.25 \\ 16.0 & 42.25 \\ 20.5 & 51.25 \\ 14.5 & 46.25 \end{bmatrix}_{N\times K} \end{array} \rightarrow \mathbf{R} = \begin{bmatrix} 0 & 1 \\ 0 & 1 \\ 0 & 1 \\ 1 & 0 \\ 1 & 0 \\ 1 & 0 \end{bmatrix}_{N\times K}$$

K-means algorithm: Step 2



Keeping r_{nk} fixed we can optimize the objective with respect to $\boldsymbol{\mu}_k$ by setting the derivative wrt to $\boldsymbol{\mu}_k$ to zero and obtain

$$\boldsymbol{\mu}_k = \frac{\sum_n r_{nk}\mathbf{x}_n}{\sum_n r_{nk}}$$

UPDATE CLUSTER CENTERS

K-means algorithm: Step 2

DATASET

$$\mathbf{X} = \begin{bmatrix} 1.0 & 8.0 \\ 2.5 & 7.5 \\ 2.0 & 7.0 \\ 8.5 & 2.5 \\ 9.0 & 2.0 \\ 8.0 & 1.0 \end{bmatrix}_{N \times D} \qquad \mathbf{R} = \begin{bmatrix} 0 & 1 \\ 0 & 1 \\ 0 & 1 \\ 1 & 0 \\ 1 & 0 \\ 1 & 0 \end{bmatrix}_{N \times K}$$

r_{11}

CLUSTER ASSIGNMENT

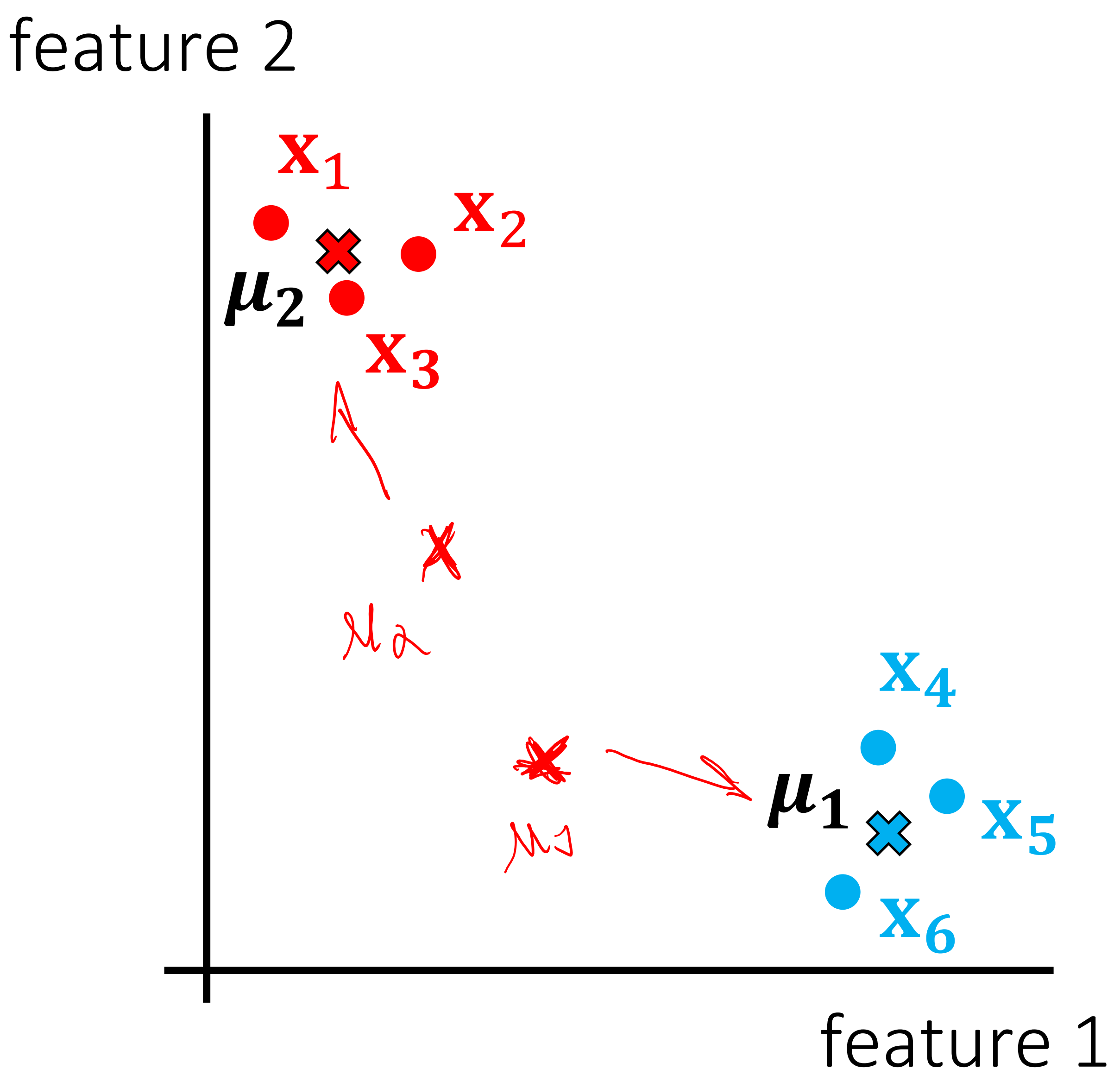
NEW μ_1

$$\boldsymbol{\mu}_1 = \frac{\sum_n r_{nk}\mathbf{x}_n}{\sum_n r_{nk}} = \frac{\cancel{(0 \times \mathbf{x}_1)} + \cancel{(0 \times \mathbf{x}_2)} + \cancel{(0 \times \mathbf{x}_3)} + (1 \times \mathbf{x}_4) + (1 \times \mathbf{x}_5) + (1 \times \mathbf{x}_6)}{0 + 0 + 0 + \underline{1 + 1 + 1}} = \begin{bmatrix} 8.5 \\ 1.83 \end{bmatrix}$$

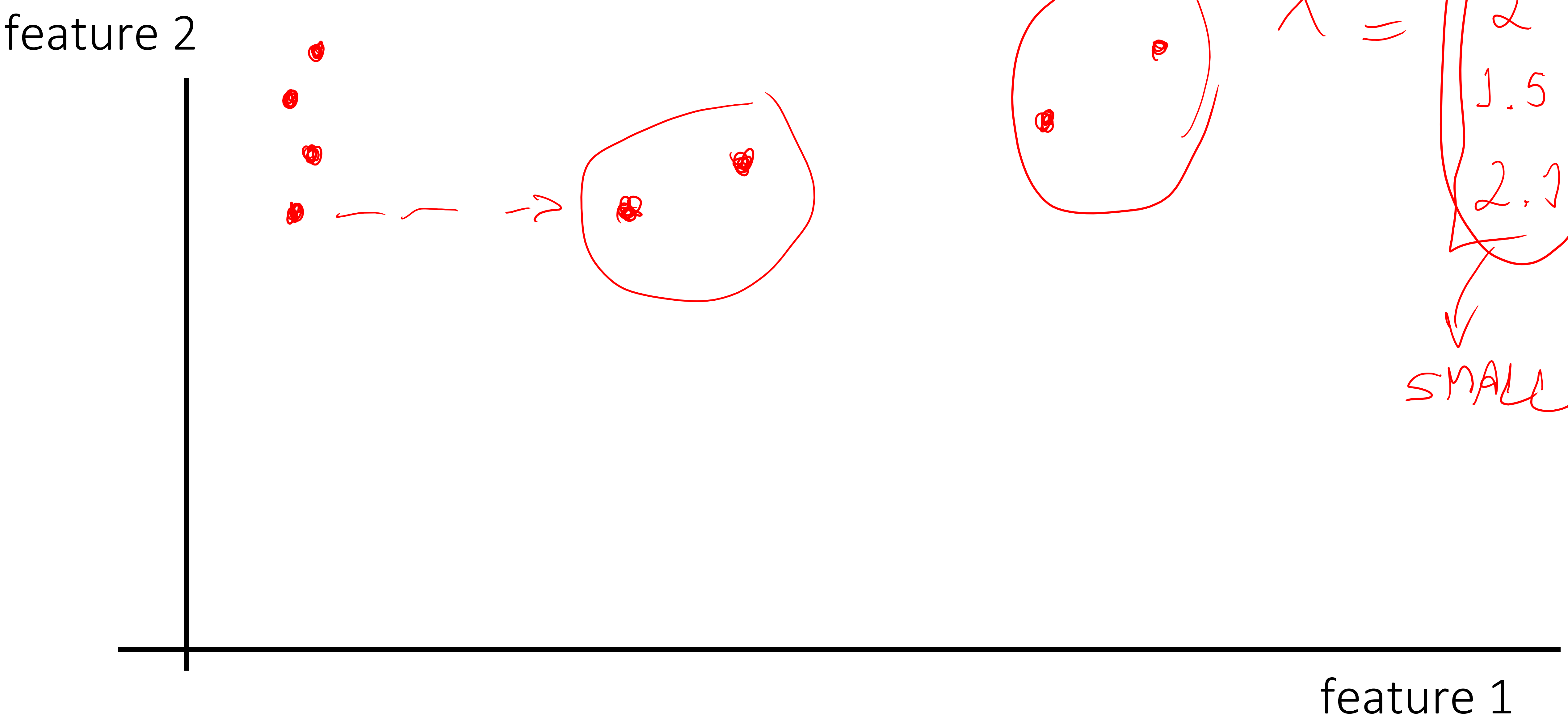
$$\boldsymbol{\mu}_2 = \frac{\sum_n r_{nk}\mathbf{x}_n}{\sum_n r_{nk}} = \frac{(1 \times \mathbf{x}_1) + (1 \times \mathbf{x}_2) + (1 \times \mathbf{x}_3) + (0 \times \mathbf{x}_4) + (0 \times \mathbf{x}_5) + (0 \times \mathbf{x}_6)}{1 + 1 + 1 + 0 + 0 + 0} = \begin{bmatrix} 1.83 \\ 7.5 \end{bmatrix}$$

μ_2

K-means algorithm: Step 2



Practical aspects of K-means: data range

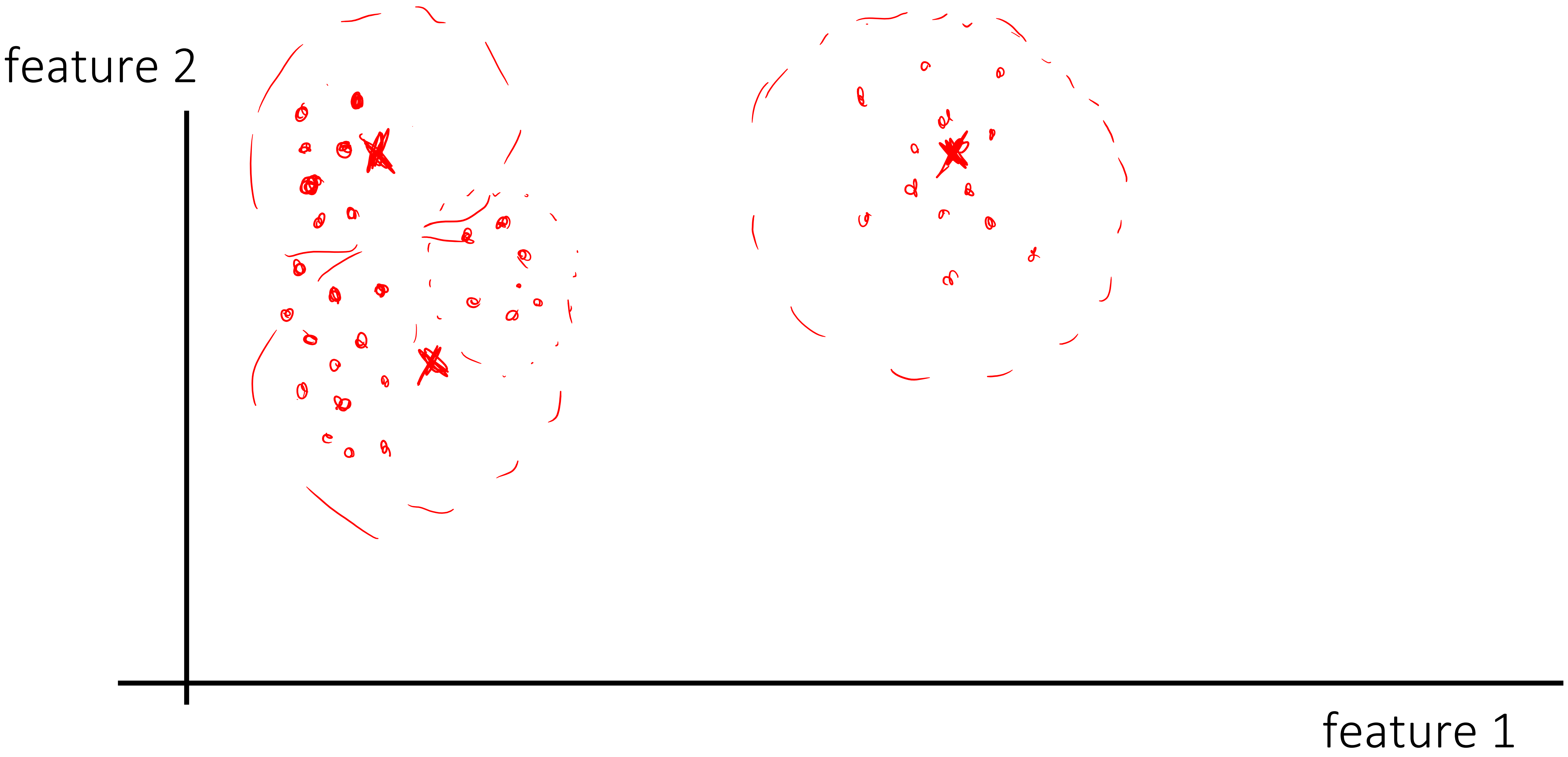


$$X = \begin{bmatrix} 1 & 6 \times 10^3 \\ 2 & 5 \times 10^3 \\ 1.5 & 4 \times 10^3 \\ 2.1 & 4.5 \times 10^3 \end{bmatrix}$$

SMALL

REALLY LARGE

Practical aspects of K-means: distribution



Practical aspects of K-means: selecting K