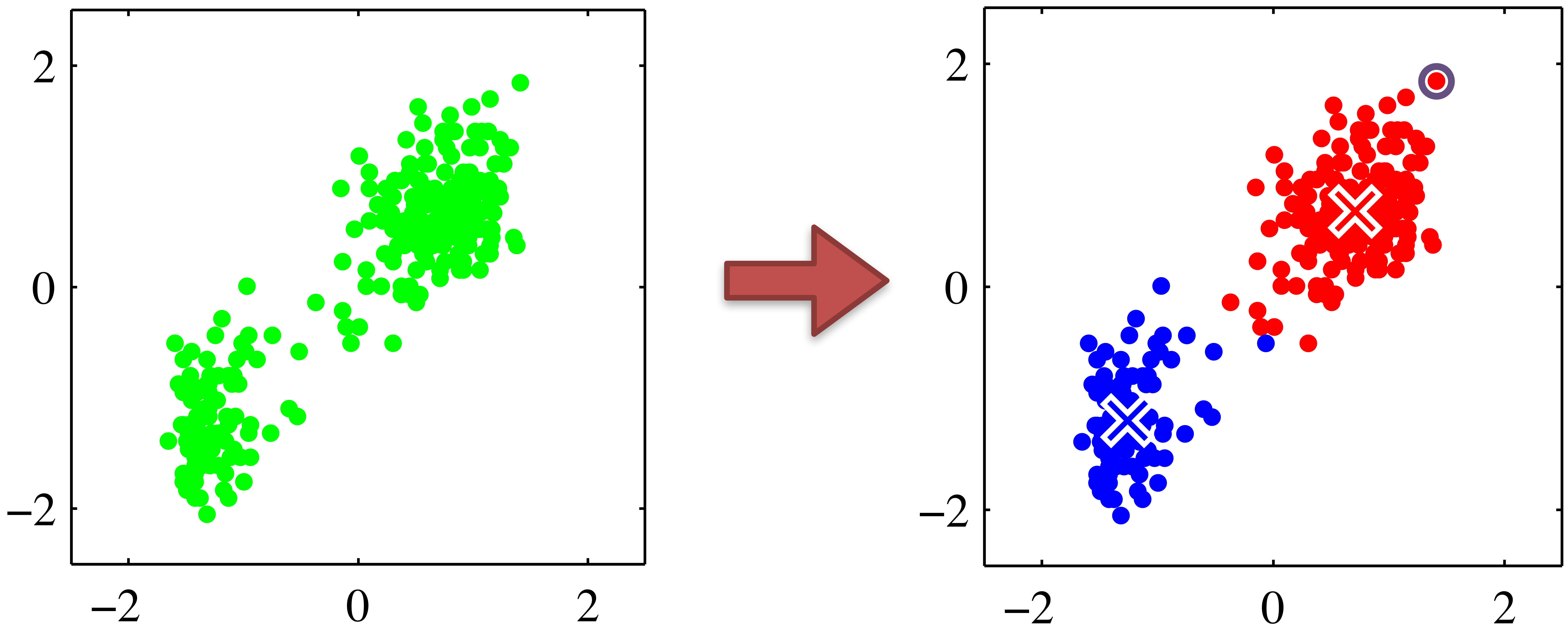
Advanced machine learning

MTH793P 2024

Omer Bobrowski, QMUL

K-MEANS CLUSTERING

K-means clustering



From Bishop. Pattern Recognition & Machine Learning

K-means clustering

Clusters are groups of points whose intra-point distances are small compared to the distances outside the cluster.

Assume we have an unlabelled dataset $\{x_i\}_{i=1}^{s}$ with $x_i \in \mathbb{R}^n$ for all $i \in \{1,\ldots,s\}$

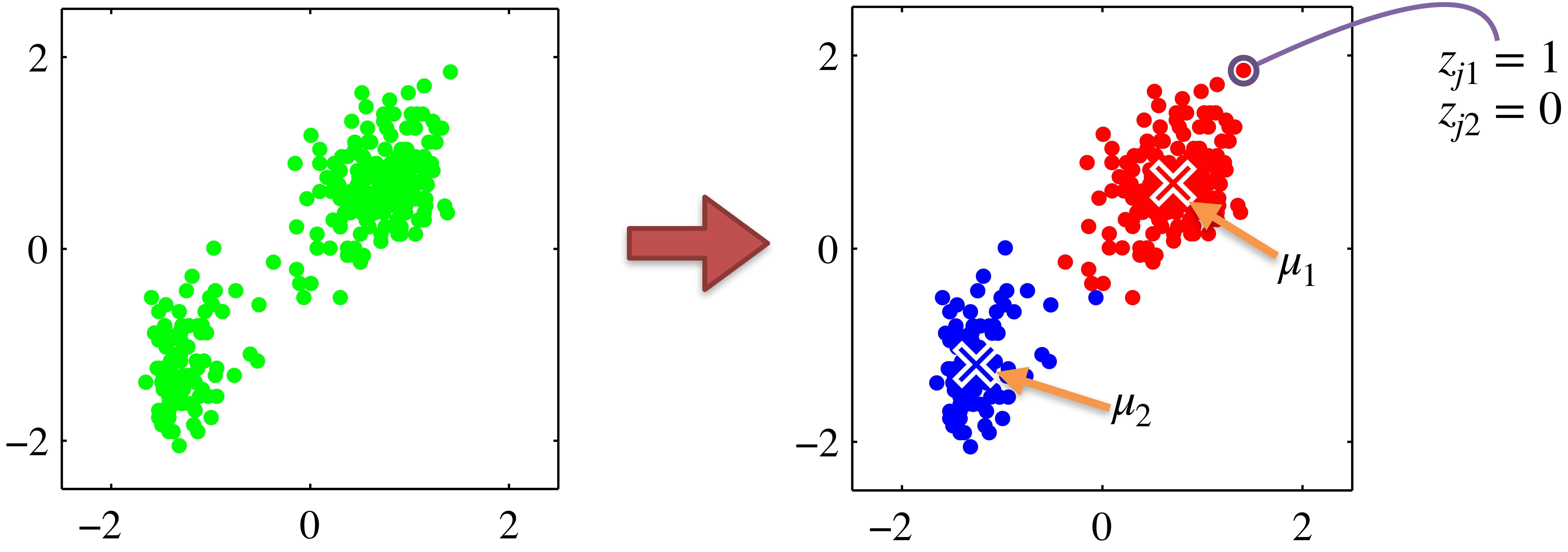
Goal: partition data into K clusters

Introduce so-called prototype vectors $\{\mu_k\}_{k=1}^{K}$ with $\mu_k \in \mathbb{R}^n$ for all $k \in \{1,\ldots,K\}$ that represent the centres of the clusters

Aim is to find these prototype vectors as well as cluster assignments $z_{ik} \in \{0,1\}$ for each data point and each cluster

K-means clustering

Example: $K = 2$



From Bishop. Pattern Recognition & Machine Learning

Old Faithful Dataset

```
Old Faithful Geyser Data

Description: (From R manual):

     Waiting time between eruptions and the duration of the eruption
     for the Old Faithful geyser in Yellowstone National Park, Wyoming,
     USA.

     A data frame with 272 observations on 2 variables.

eruptions  numeric  Eruption time in mins
waiting    numeric  Waiting time to next eruption

References:

     Hardle, W. (1991) Smoothing Techniques with Implementation in S.
     New York: Springer.

     Azzalini, A. and Bowman, A. W. (1990). A look at some data on the
     Old Faithful geyser. Applied Statistics 39, 357-365
```

	eruptions	waiting
1	3.600	79
2	1.800	54
3	3.333	74
4	2.283	62
5	4.533	85
6	2.883	55
7	4.700	88
8	3.600	85
9	1.950	51
10	4.350	85
11	1.833	54
12	3.917	84
13	4.200	78
14	1.750	47
15	4.700	83
16	2.167	52
17	1.750	62
18	4.800	84
19	1.600	52
20	4.250	79

K-means clustering

Optimisation problem:

Cost function: $$L(z,\mu) = \sum_{i=1}^{s}\sum_{k=1}^{K} z_{ik}\|x_i - \mu_k\|^2$$

Find z and μ by solving the following constrained minimisation problem:

$$(z,\mu) = \arg\min_{z,\mu} L(z,\mu)$$

subject to $z_{ik} \in \{0,1\}$ and $\sum_{k=1}^{K} z_{ik} = 1$ for all $i \in \{1,\ldots,s\}$

Is this optimisation problem easy? Convex?

K-means clustering

Algorithm: coordinate descent / alternating minimisation

Given the function $L(z, \mu) = \sum_{i=1}^{s} \sum_{k=1}^{K} z_{ik} \|x_i - \mu_k\|^2$, iteratively compute

1) Fix μ^l : $z^{l+1} = \arg\min_z L(z, \mu^l)$ subject to $z_{ik} \in \{0,1\}, \sum_{k=1}^{K} z_{ik} = 1$

2) Fix z^{l+1} : $\mu^{l+1} = \arg\min_\mu L(z^{l+1}, \mu)$

K-means clustering

Algorithm: coordinate descent / alternating minimisation

Given the function $L(z, \mu) = \sum_{i=1}^{s} \sum_{k=1}^{K} z_{ik} \|x_i - \mu_k\|^2$, iteratively compute

1) $$z_{ik}^{l+1} = \begin{cases} 1 & \text{if } k = \arg\min_{j \in \{1,\ldots,K\}} \left\| x_i - \mu_j^l \right\|^2 \\ 0 & \text{otherwise} \end{cases}$$

assign each point to the nearest center

2) $$\mu^{l+1} = \frac{\sum_{i=1}^{s} z_{ik}^{l+1} x_i}{\sum_{i=1}^{s} z_{ik}^{l+1}}$$

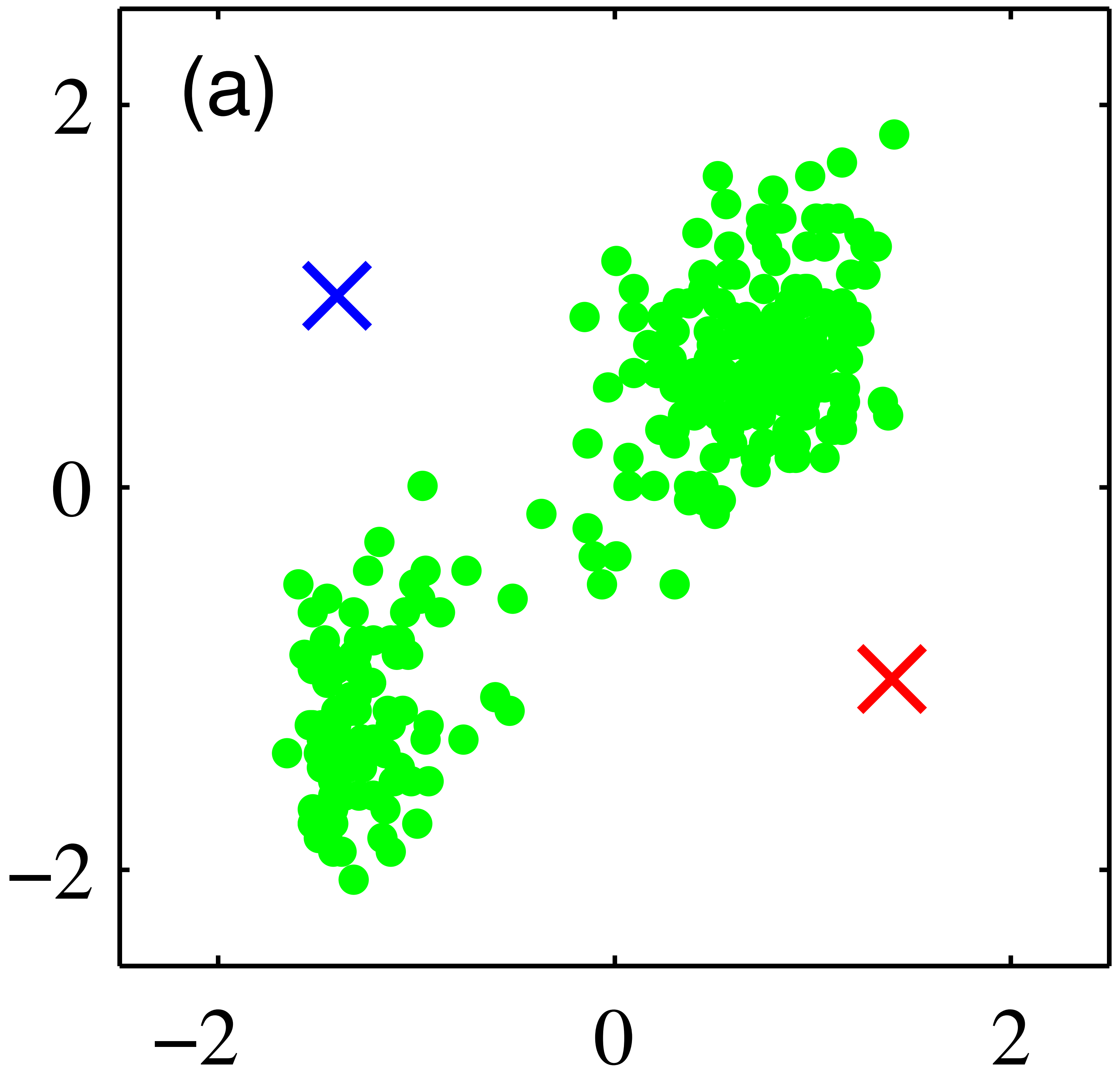
new centers are the average over all points in the cluster

K-means clustering

Example:

1) $z_{ik}^{l+1} = \begin{cases} 1 & \text{if } k = \arg\min_{j\in\{1,\ldots,K\}} \left\| x_i - \mu_j^l \right\|^2 \\ 0 & \text{otherwise} \end{cases}$

2) $\mu^{l+1} = \dfrac{\sum_{i=1}^{s} z_{ik}^{l+1} x_i}{\sum_{i=1}^{s} z_{ik}^{l+1}}$



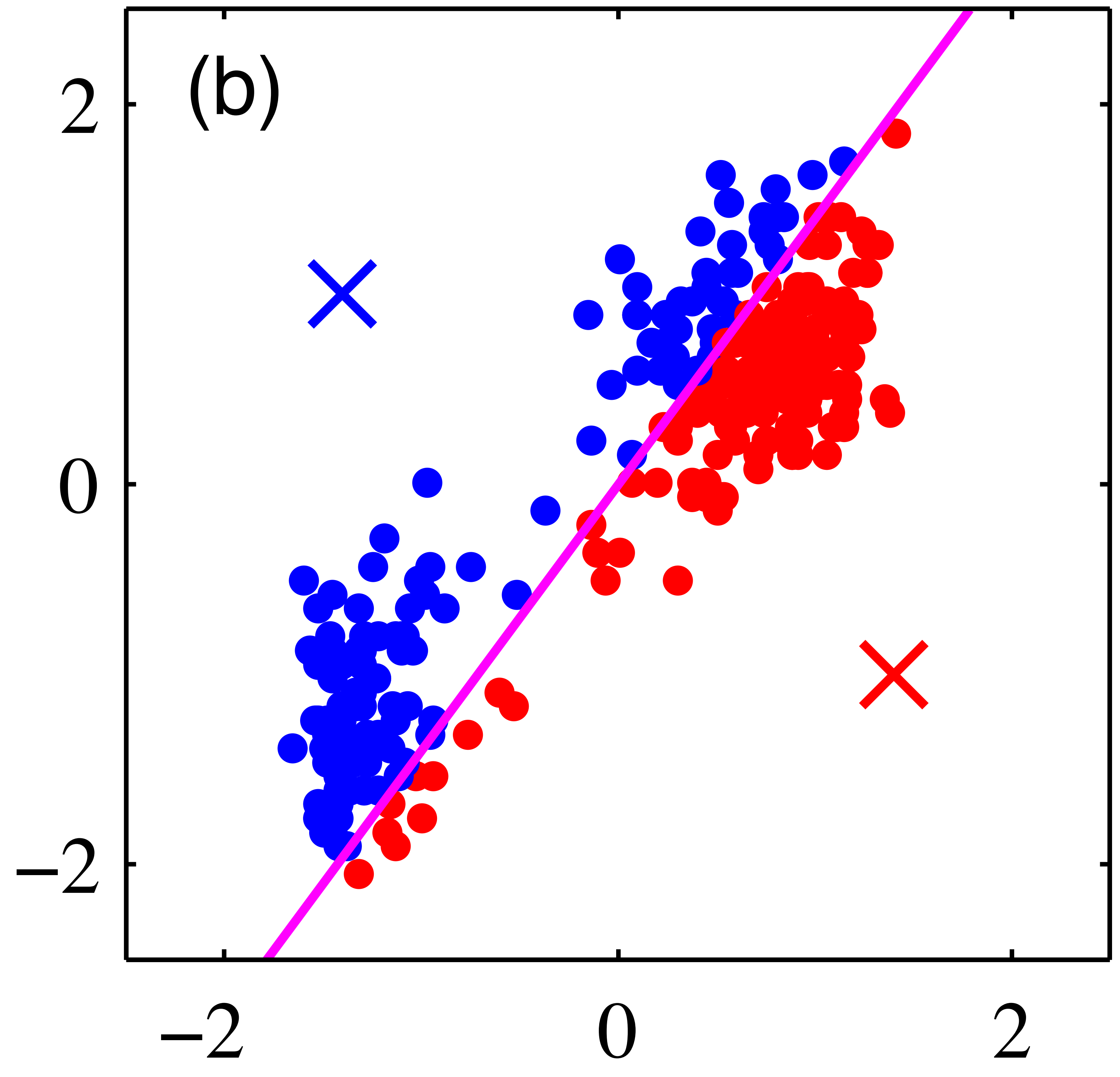
From Bishop. Pattern Recognition & Machine Learning

K-means clustering

Example:

1) $z_{ik}^{l+1} = \begin{cases} 1 & \text{if } k = \arg\min_{j\in\{1,\ldots,K\}} \left\| x_i - \mu_j^l \right\|^2 \\ 0 & \text{otherwise} \end{cases}$

2) $\mu^{l+1} = \frac{\sum_{i=1}^{s} z_{ik}^{l+1} x_i}{\sum_{i=1}^{s} z_{ik}^{l+1}}$



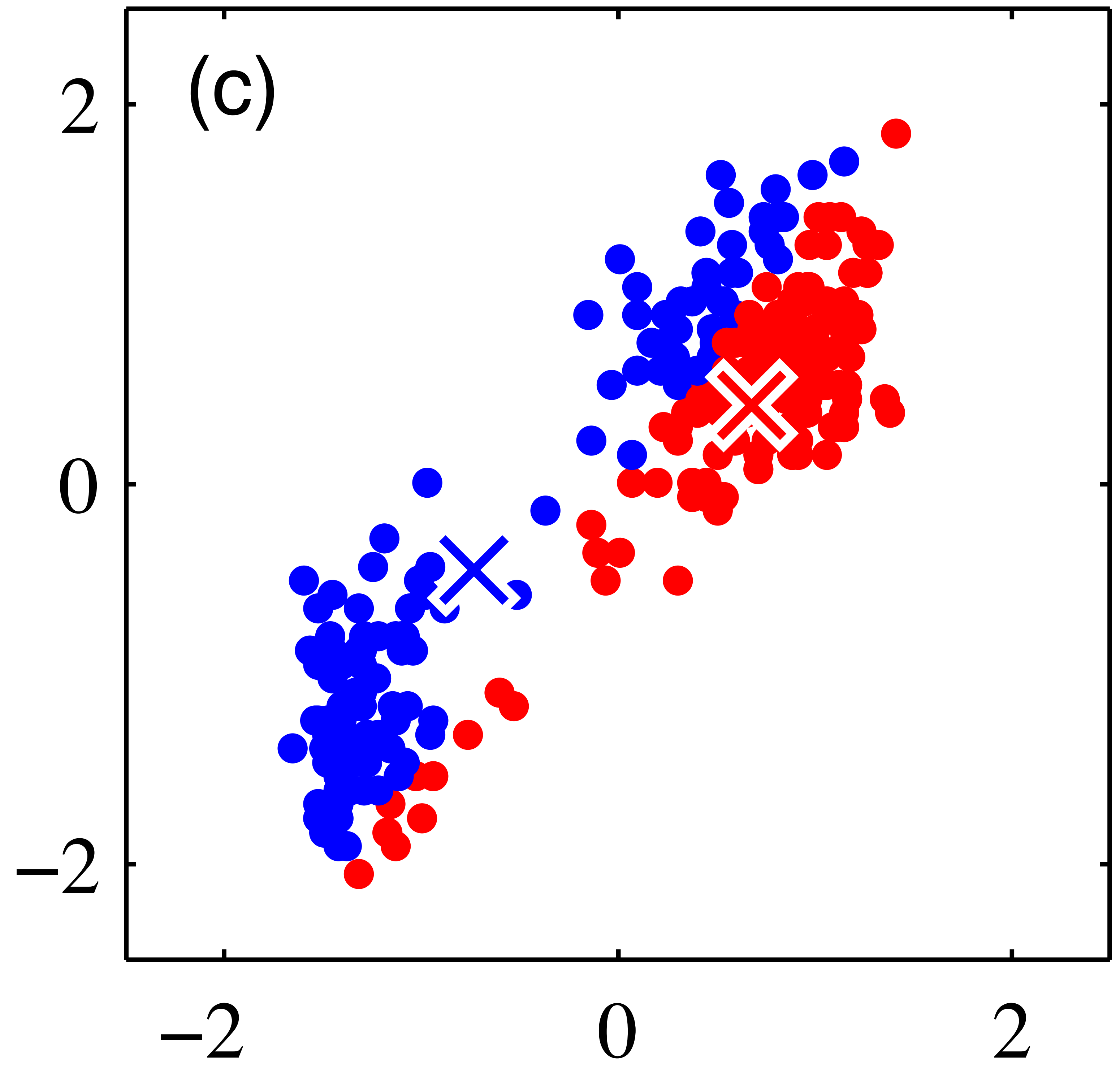
From Bishop. Pattern Recognition & Machine Learning

K-means clustering

Example:

1) $z_{ik}^{l+1} = \begin{cases} 1 & \text{if } k = \arg\min_{j\in\{1,\ldots,K\}} \left\| x_i - \mu_j^l \right\|^2 \\ 0 & \text{otherwise} \end{cases}$

2) $\mu^{l+1} = \dfrac{\sum_{i=1}^{s} z_{ik}^{l+1} x_i}{\sum_{i=1}^{s} z_{ik}^{l+1}}$



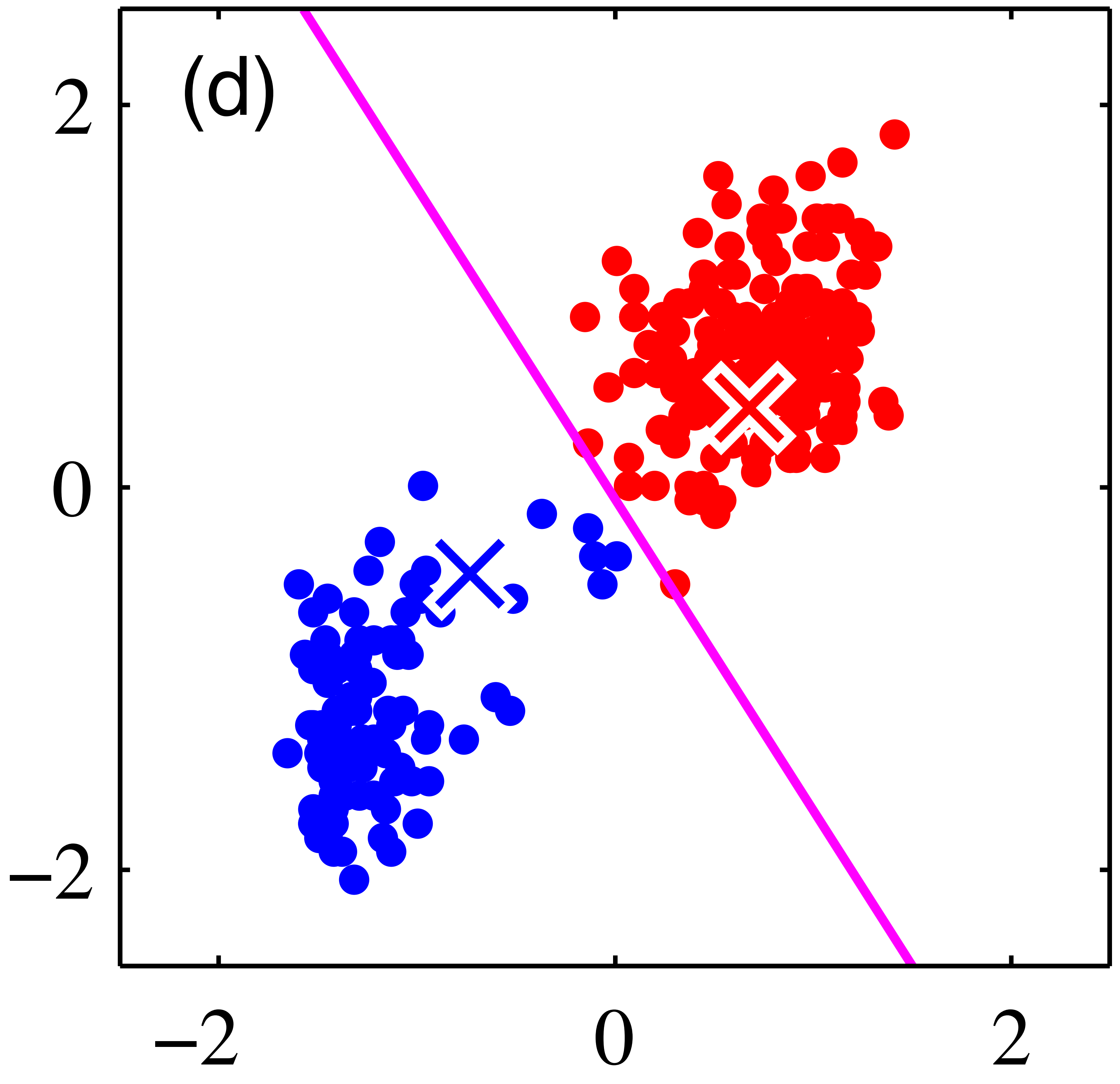
From Bishop. Pattern Recognition & Machine Learning

K-means clustering

Example:

1) $z_{ik}^{l+1} = \begin{cases} 1 & \text{if } k = \arg\min_{j \in \{1,\ldots,K\}} \left\| x_i - \mu_j^l \right\|^2 \\ 0 & \text{otherwise} \end{cases}$

2) $\mu^{l+1} = \dfrac{\sum_{i=1}^{s} z_{ik}^{l+1} x_i}{\sum_{i=1}^{s} z_{ik}^{l+1}}$



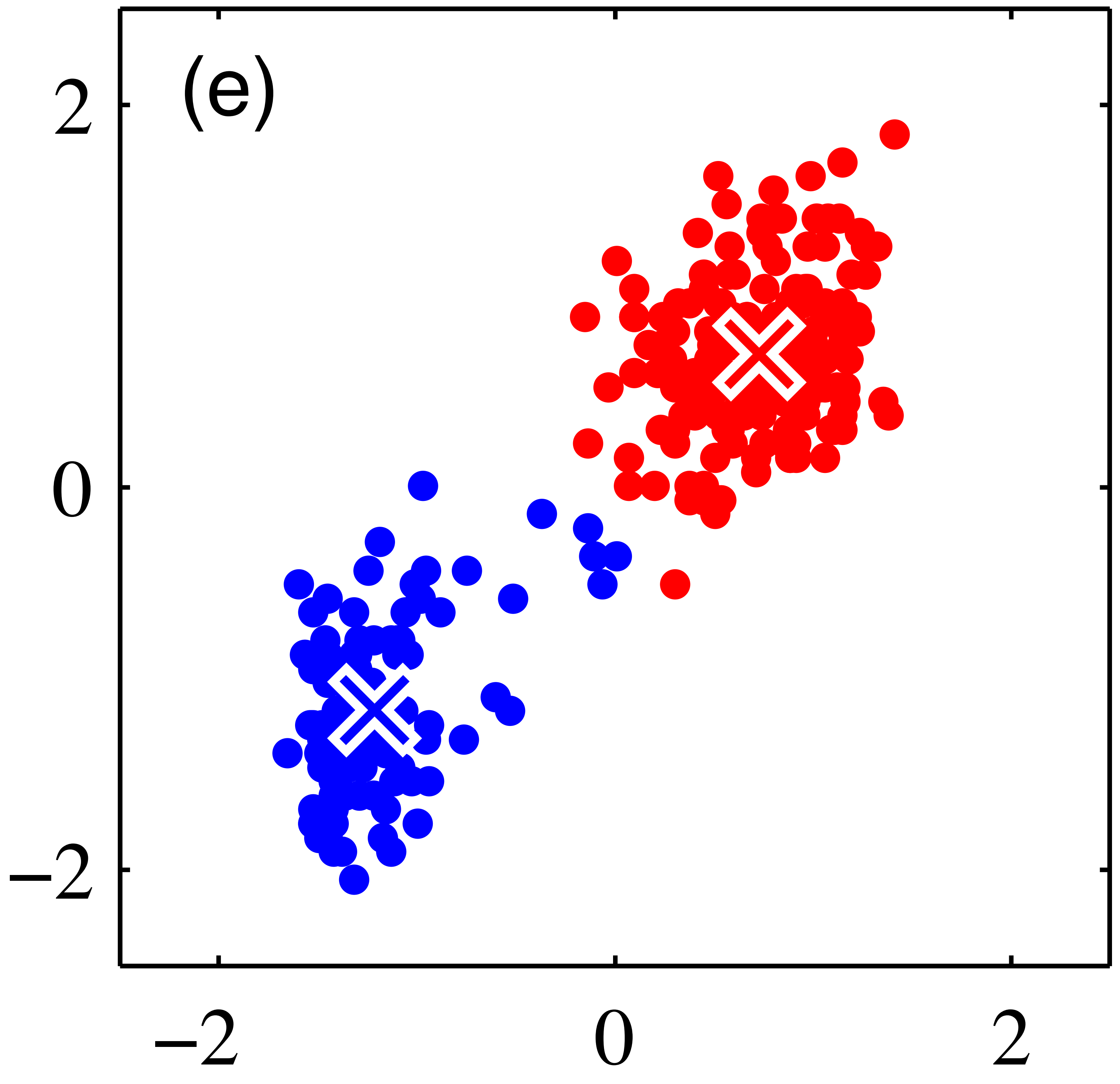
From Bishop. Pattern Recognition & Machine Learning

K-means clustering

Example:

1) $z_{ik}^{l+1} = \begin{cases} 1 & \text{if } k = \arg\min_{j \in \{1,\ldots,K\}} \left\| x_i - \mu_j^l \right\|^2 \\ 0 & \text{otherwise} \end{cases}$

2) $\mu^{l+1} = \frac{\sum_{i=1}^{s} z_{ik}^{l+1} x_i}{\sum_{i=1}^{s} z_{ik}^{l+1}}$



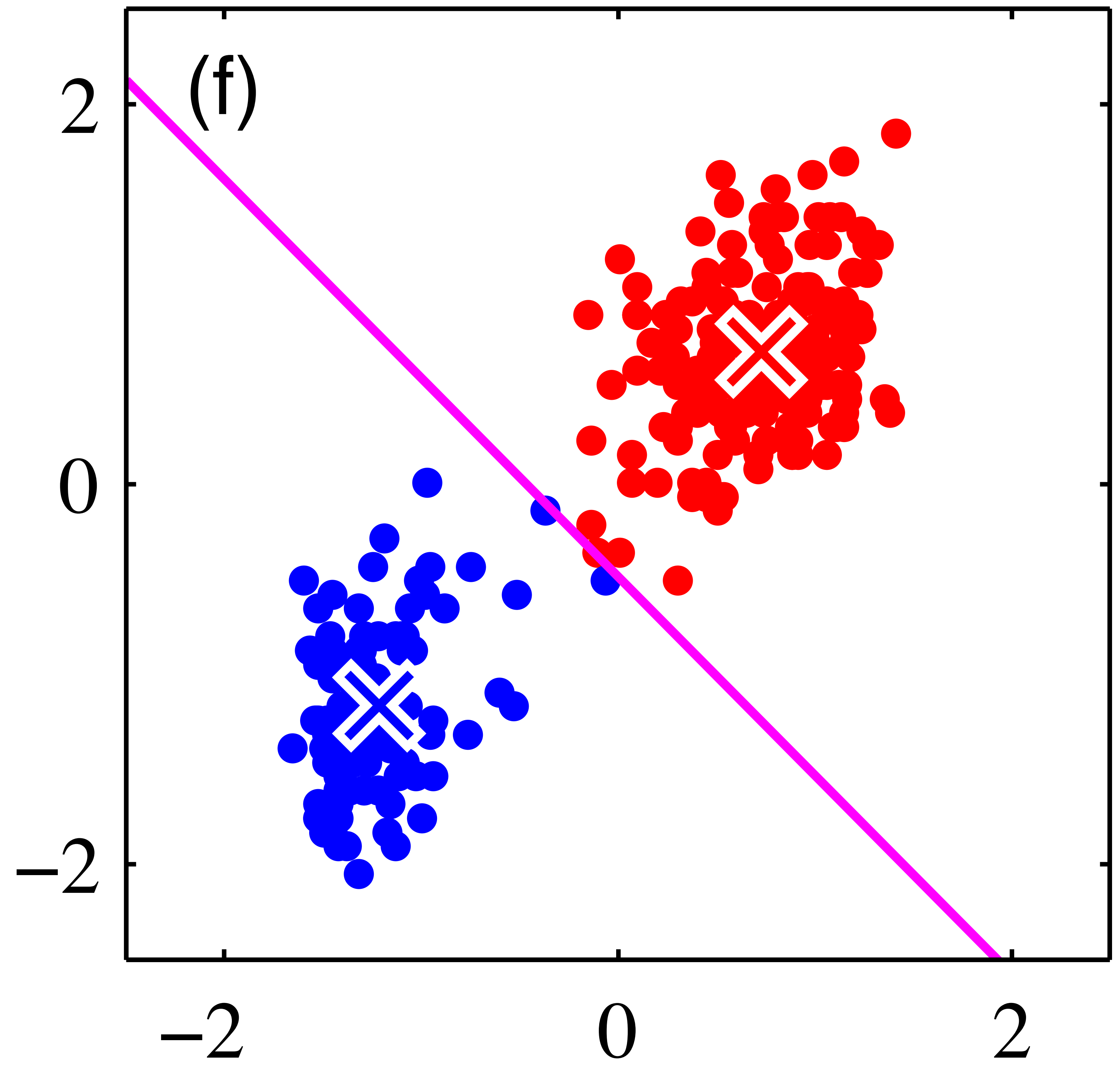
From Bishop. Pattern Recognition & Machine Learning

K-means clustering

Example:

1) $$z_{ik}^{l+1} = \begin{cases} 1 & \text{if } k = \arg\min_{j \in \{1,\ldots,K\}} \left\| x_i - \mu_j^l \right\|^2 \\ 0 & \text{otherwise} \end{cases}$$

2) $$\mu^{l+1} = \frac{\sum_{i=1}^{s} z_{ik}^{l+1} x_i}{\sum_{i=1}^{s} z_{ik}^{l+1}}$$



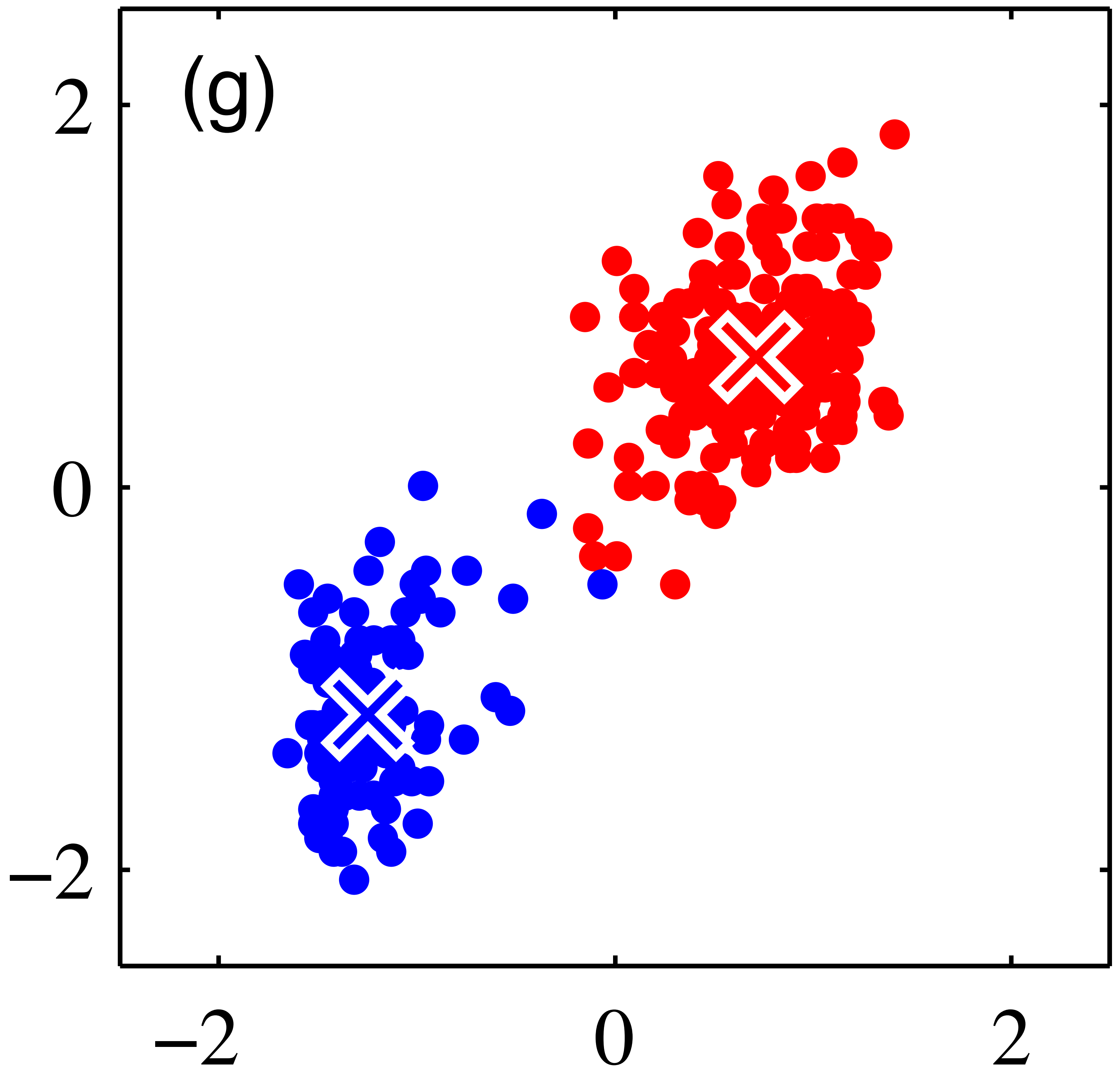
From Bishop. Pattern Recognition & Machine Learning

K-means clustering

Example:

1) $z_{ik}^{l+1} = \begin{cases} 1 & \text{if } k = \arg\min_{j\in\{1,\ldots,K\}} \left\| x_i - \mu_j^l \right\|^2 \\ 0 & \text{otherwise} \end{cases}$

2) $\mu^{l+1} = \dfrac{\sum_{i=1}^{s} z_{ik}^{l+1} x_i}{\sum_{i=1}^{s} z_{ik}^{l+1}}$



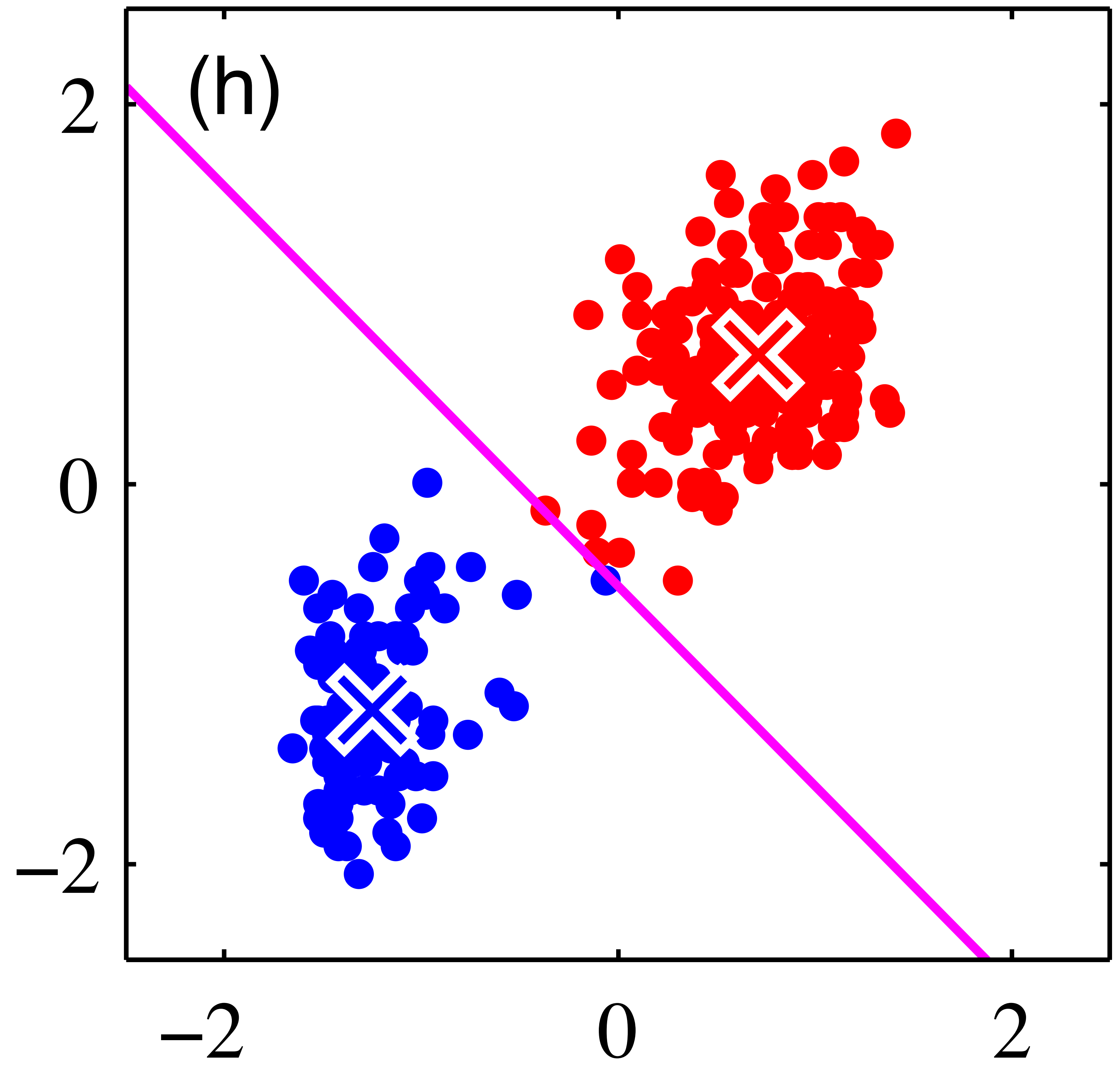
From Bishop. Pattern Recognition & Machine Learning

K-means clustering

Example:

1) $z_{ik}^{l+1} = \begin{cases} 1 & \text{if } k = \arg\min_{j \in \{1,\dots,K\}} \left\| x_i - \mu_j^l \right\|^2 \\ 0 & \text{otherwise} \end{cases}$

2) $\mu^{l+1} = \frac{\sum_{i=1}^{s} z_{ik}^{l+1} x_i}{\sum_{i=1}^{s} z_{ik}^{l+1}}$



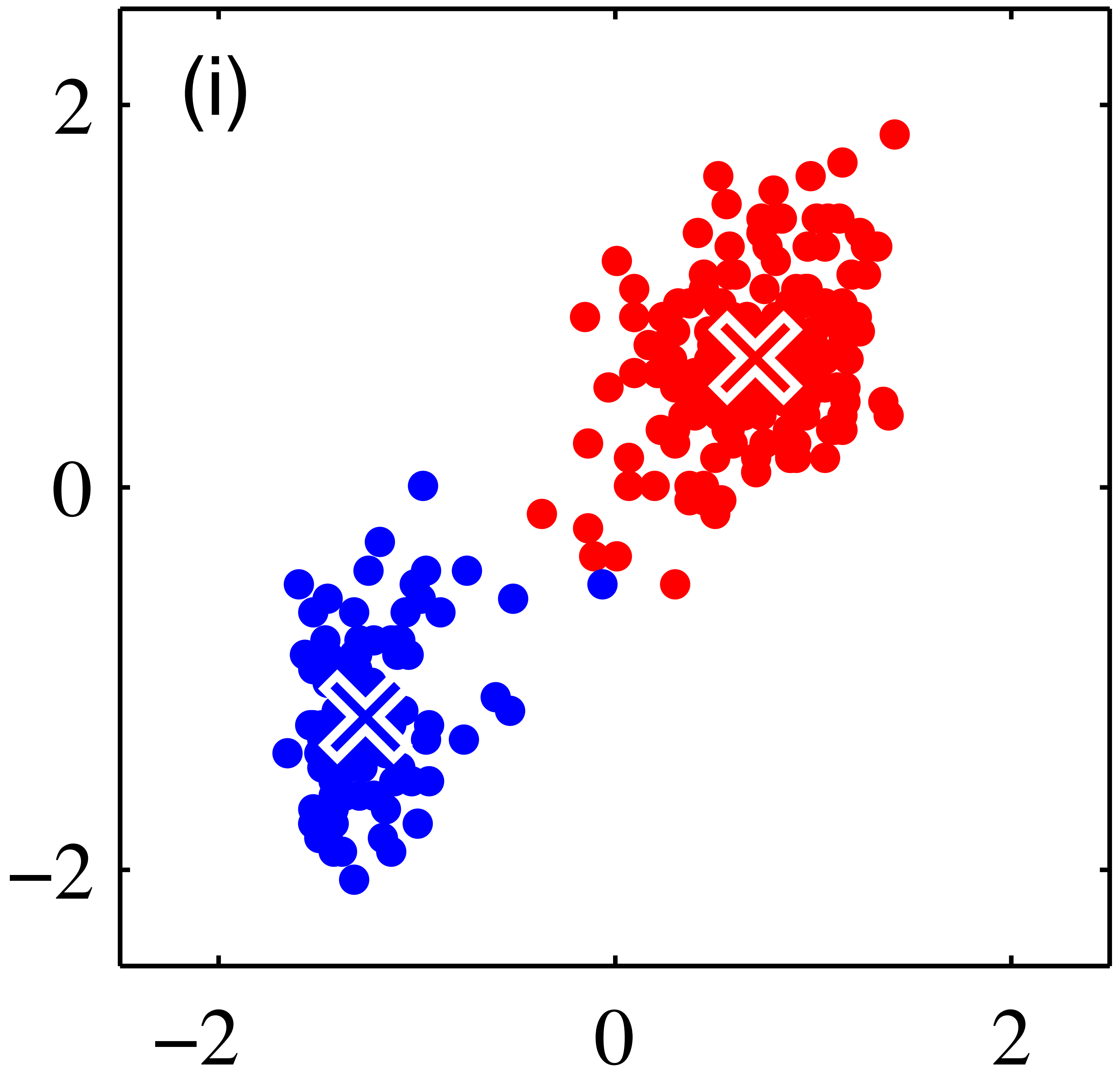
From Bishop. Pattern Recognition & Machine Learning

K-means clustering

Example:

1) $z_{ik}^{l+1} = \begin{cases} 1 & \text{if } k = \arg\min_{j\in\{1,\dots,K\}} \left\| x_i - \mu_j^l \right\|^2 \\ 0 & \text{otherwise} \end{cases}$

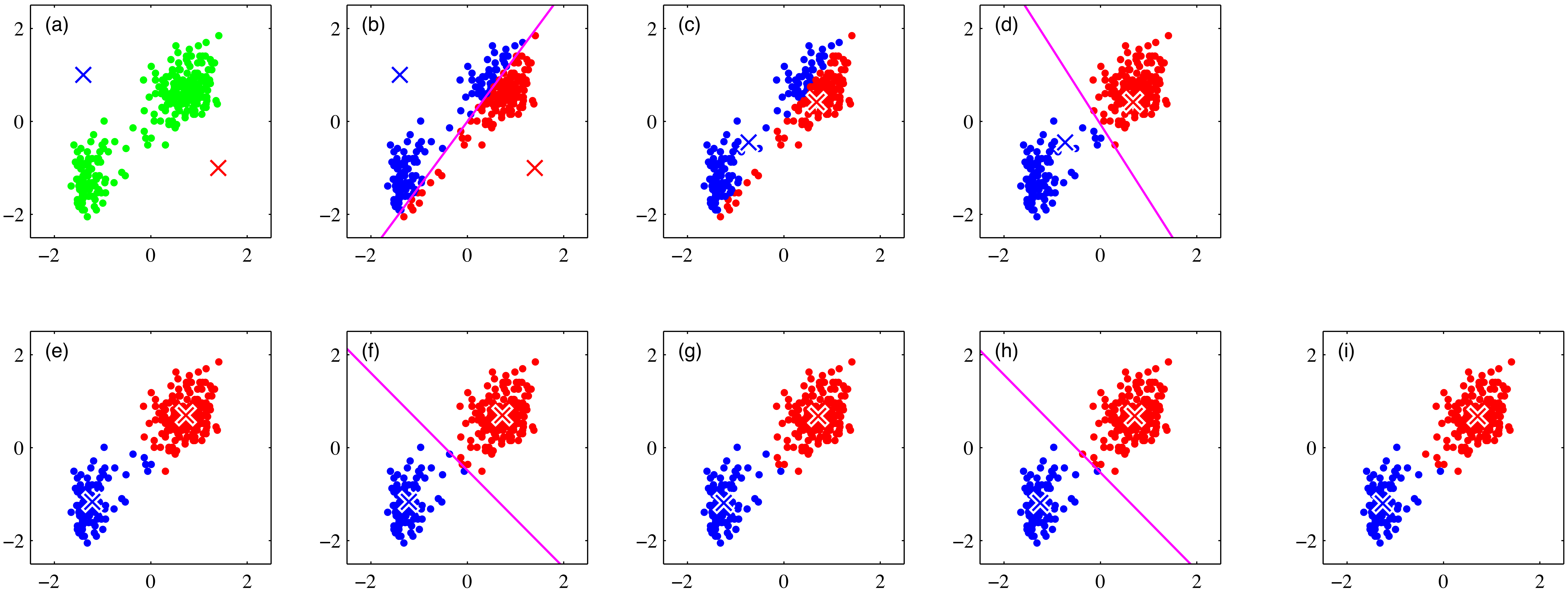
2) $\mu^{l+1} = \dfrac{\sum_{i=1}^{s} z_{ik}^{l+1} x_i}{\sum_{i=1}^{s} z_{ik}^{l+1}}$



From Bishop. Pattern Recognition & Machine Learning

K-means clustering

Example:



From Bishop. Pattern Recognition & Machine Learning

K-means Algorithm

Algorithm 4 k-means clustering.

Specify: the number of clusters k.

Initialise: $\mu^0 \in \mathbb{R}^{n\times k}$

Iterate:

1: **for** $l = 0, \ldots, N-1$ **do**

2: $$z_{ij}^{l+1} = \begin{cases} 1 & j = \arg\min_{r\in\{1,\ldots,k\}} \|x_i - \mu_r^l\|^2 \\ 0 & \text{otherwise} \end{cases}, \quad \text{for all} \quad i \in \{1,\ldots,s\}$$

3: $$\mu_j^{l+1} = \frac{\sum_{i=1}^s z_{ij}^{l+1} x_i}{\sum_{i=1}^s z_{ij}}, \quad \text{for all} \quad j \in \{1,\ldots,k\}$$

4: **end for**

return z^N, μ^N.

When should the algorithm stop?

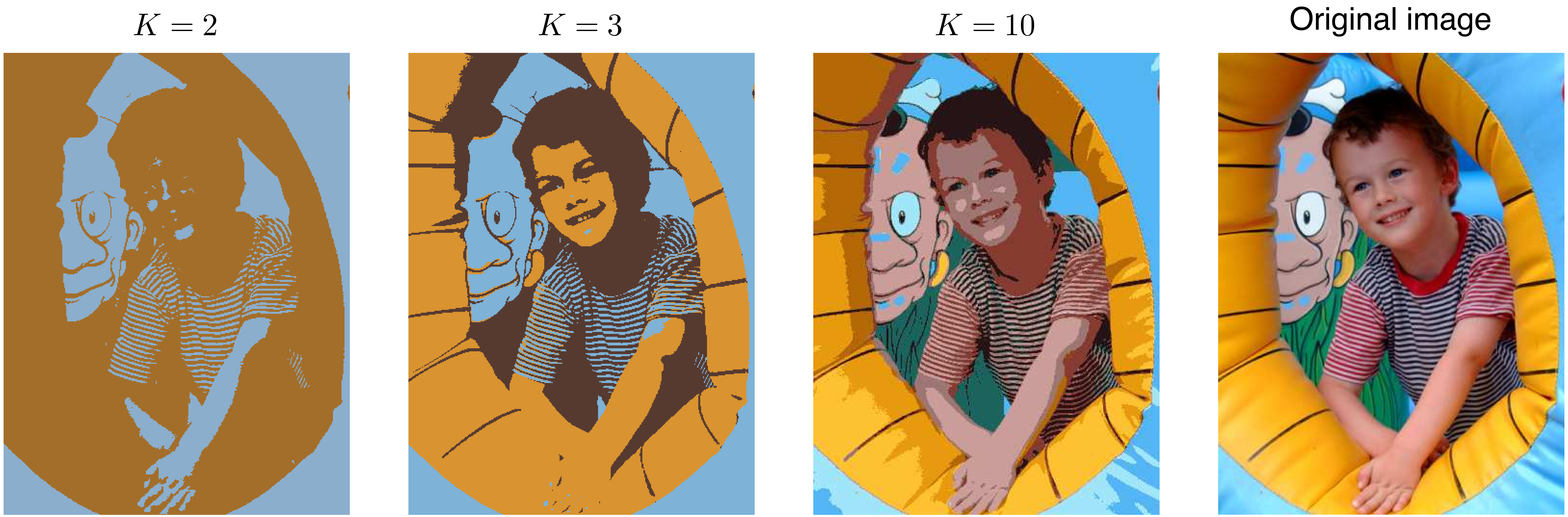
- Recall - cost function: $L(z, \mu) = \sum_{i=1}^{s} \sum_{k=1}^{K} z_{ik} \| x_i - \mu_k \|^2$,
- **Claim:** $L(z^{l+1}, \mu^{l+1}) \leq L(z^l, \mu^l)$
- **Stopping conditions:**
  - Assignments (z_{ik}) do not change
  - Change in the cost function is very small

K-means Properties

- Cost decreases at every step
- Algorithm always stops
- Solution is **not guaranteed** to be optimal, but is often close enough
- Different initializations → different results
- K has to be chosen in advance.
    - small k → larger error, large compression
    - large k → smaller error, poor compression, overfitting

K-means clustering

Example: image compression / quantisation



From Bishop. Pattern Recognition & Machine Learning

K-means clustering

Example: image compression / quantisation

$K = 2$

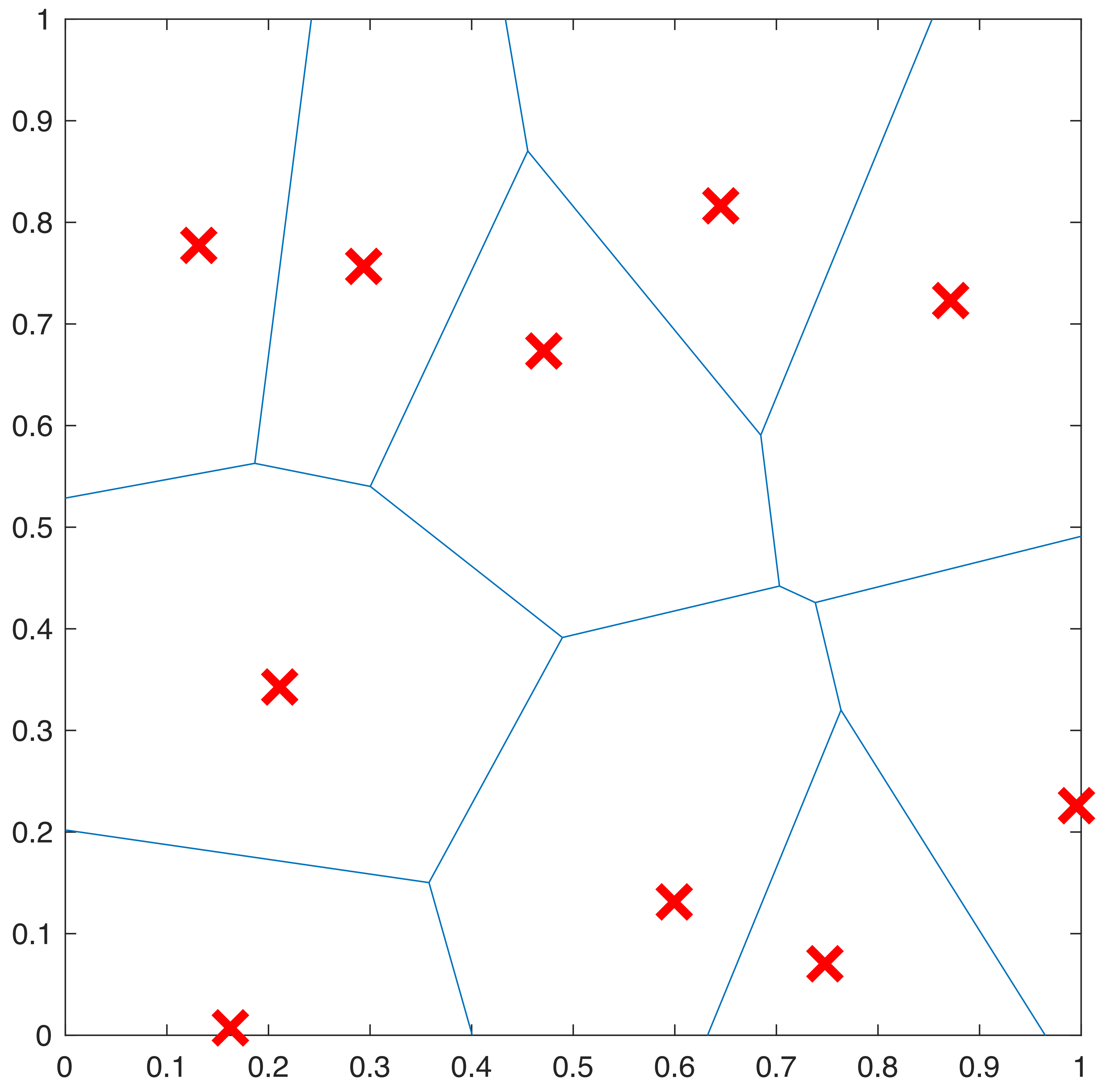
$K = 3$

$K = 10$

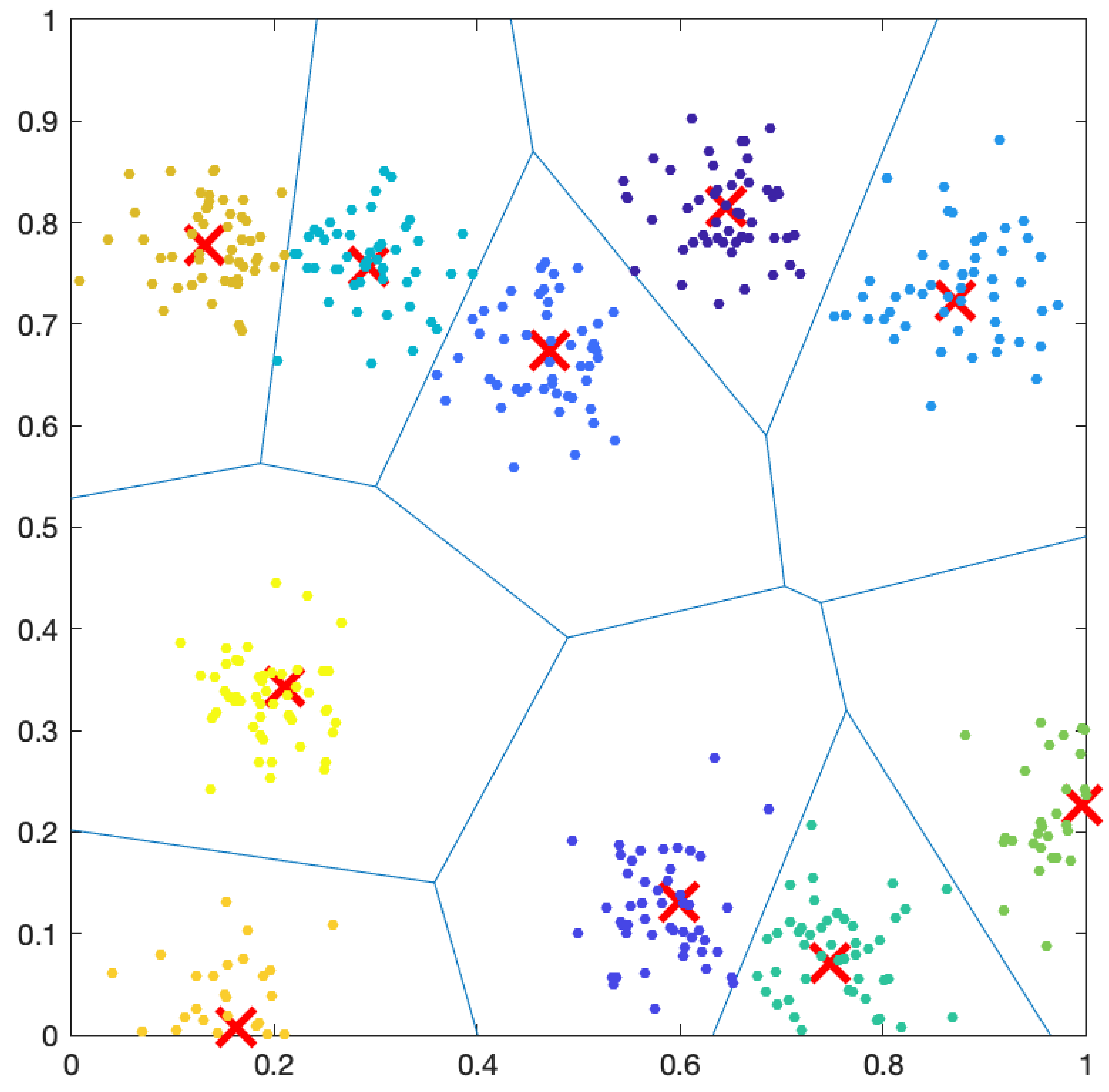
Original image

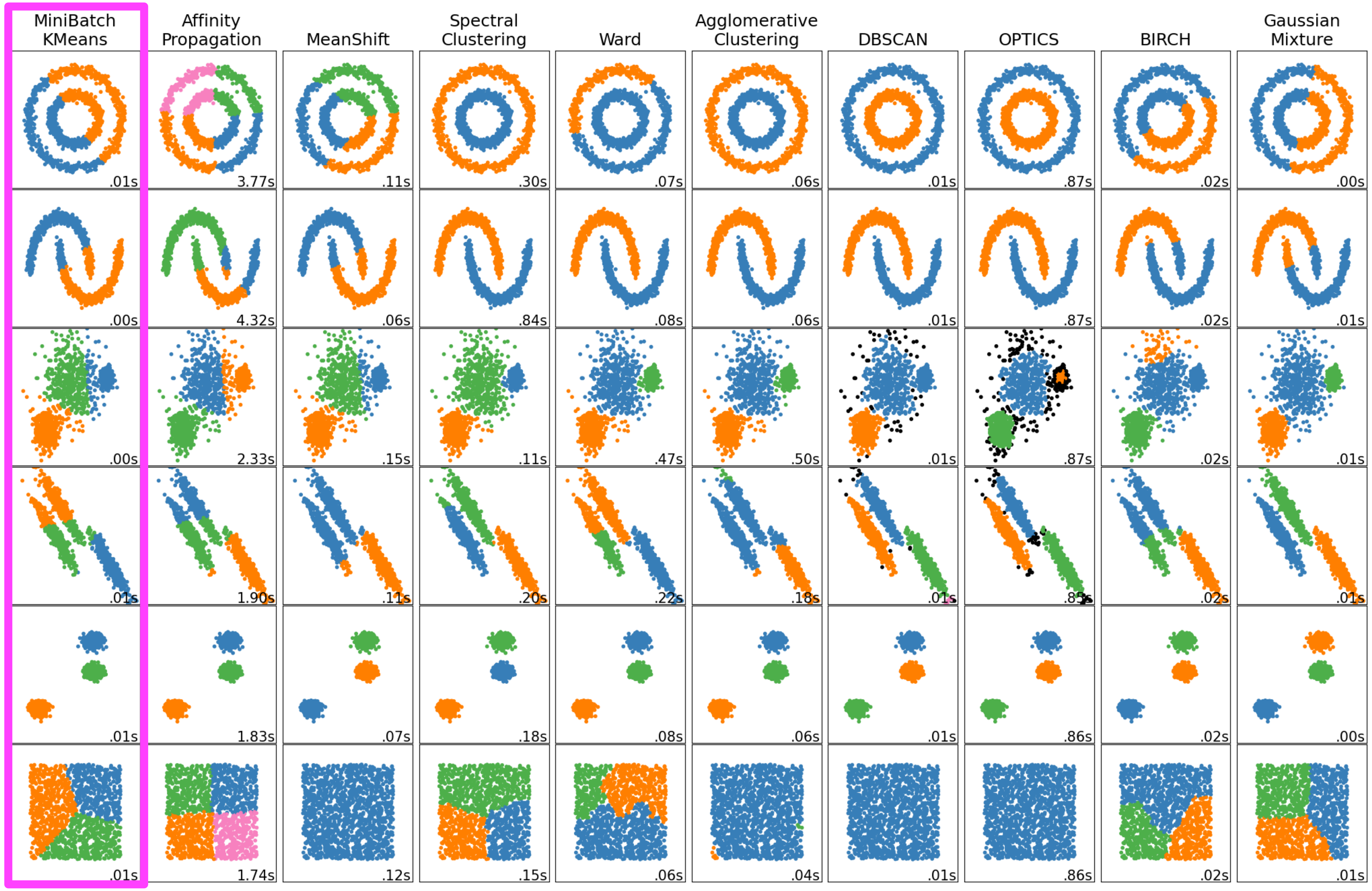
From Bishop. Pattern Recognition & Machine Learning

Voronoi Cells

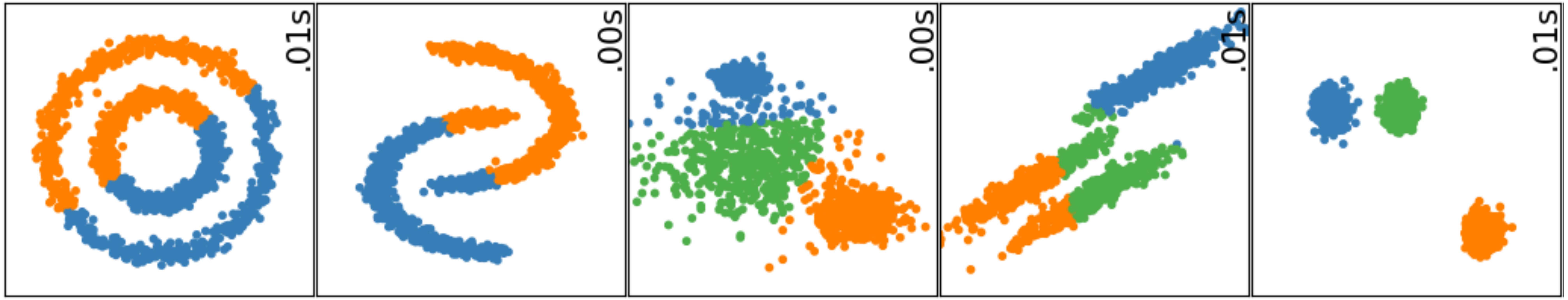


Voronoi Cells





Taken from scikit-learn python package documentation



Taken from scikit-learn python package documentation