ED5340 - Data Science: Theory and Practise

L23 - K-Means

Ramanathan Muthuganapathy (https://ed.iitm.ac.in/~raman)

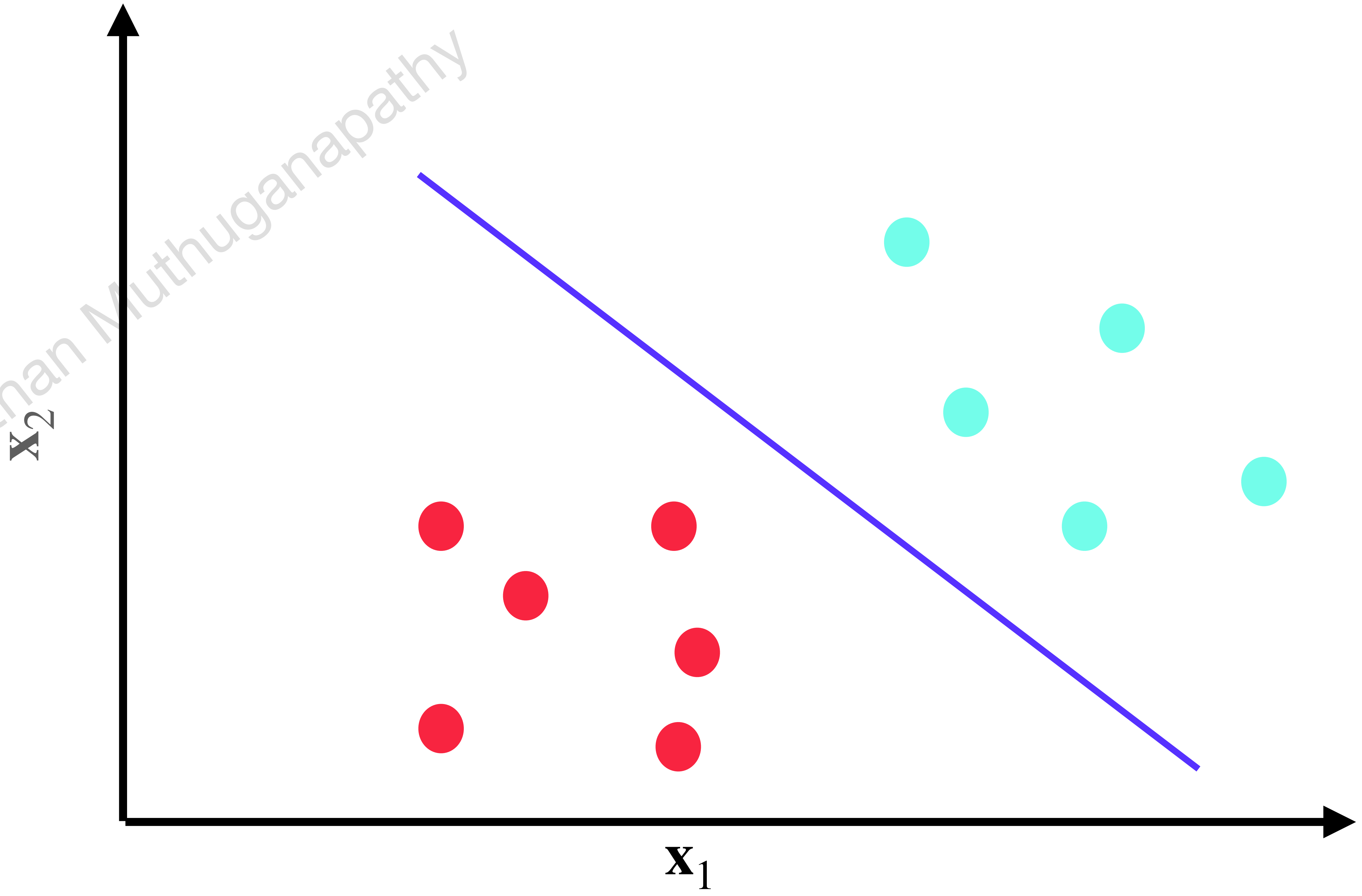
Course web page: https://ed.iitm.ac.in/~raman/datascience.html
Moodle page: Available at https://courses.iitm.ac.in/

Unsupervised

- Unsuperivsed - no labelling available
- Popular clustering technique
- Social Networks Analysis, Market analysis, etc.

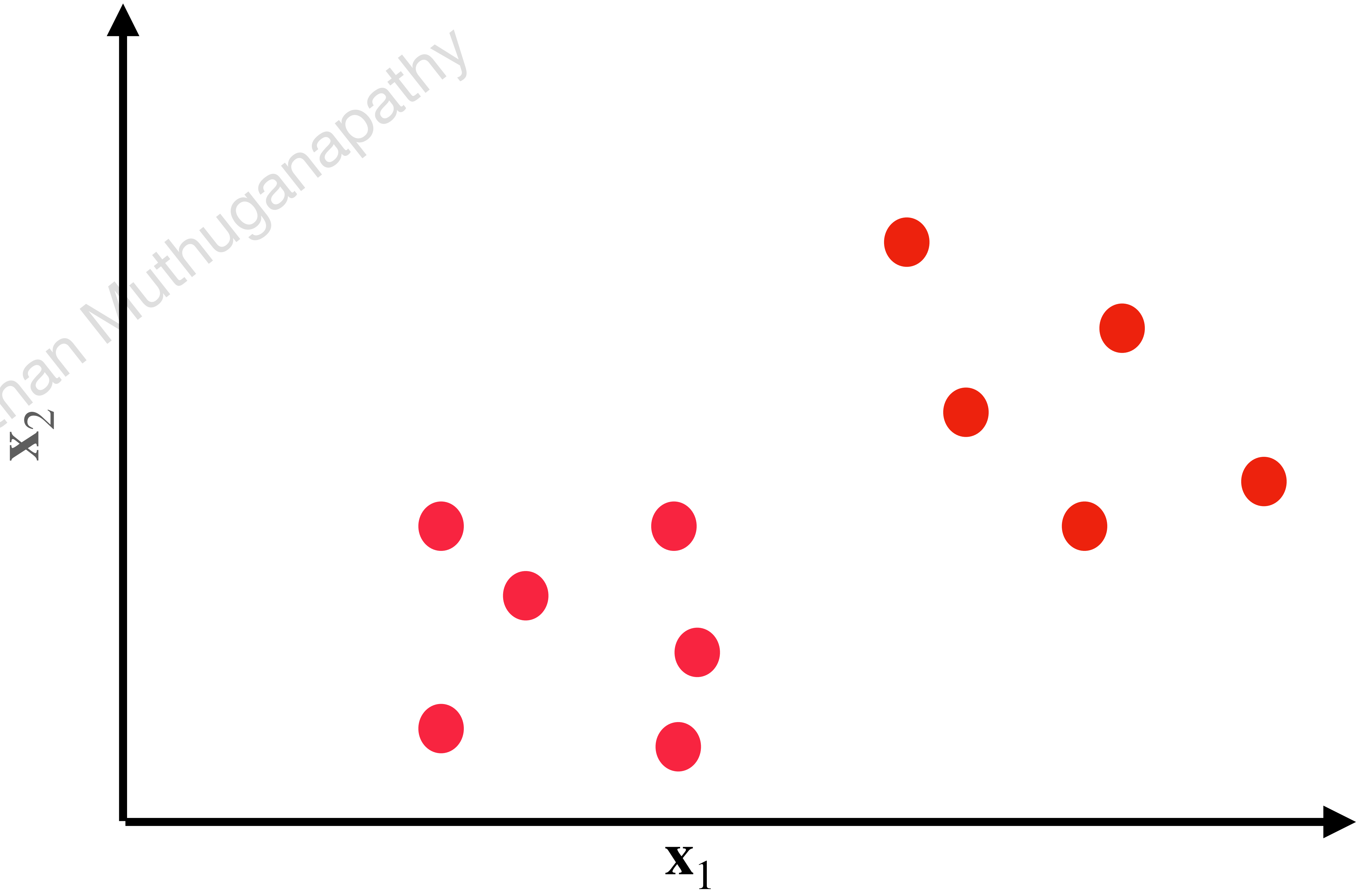
Supervised

- Labelled data
- For classification - logistic regression



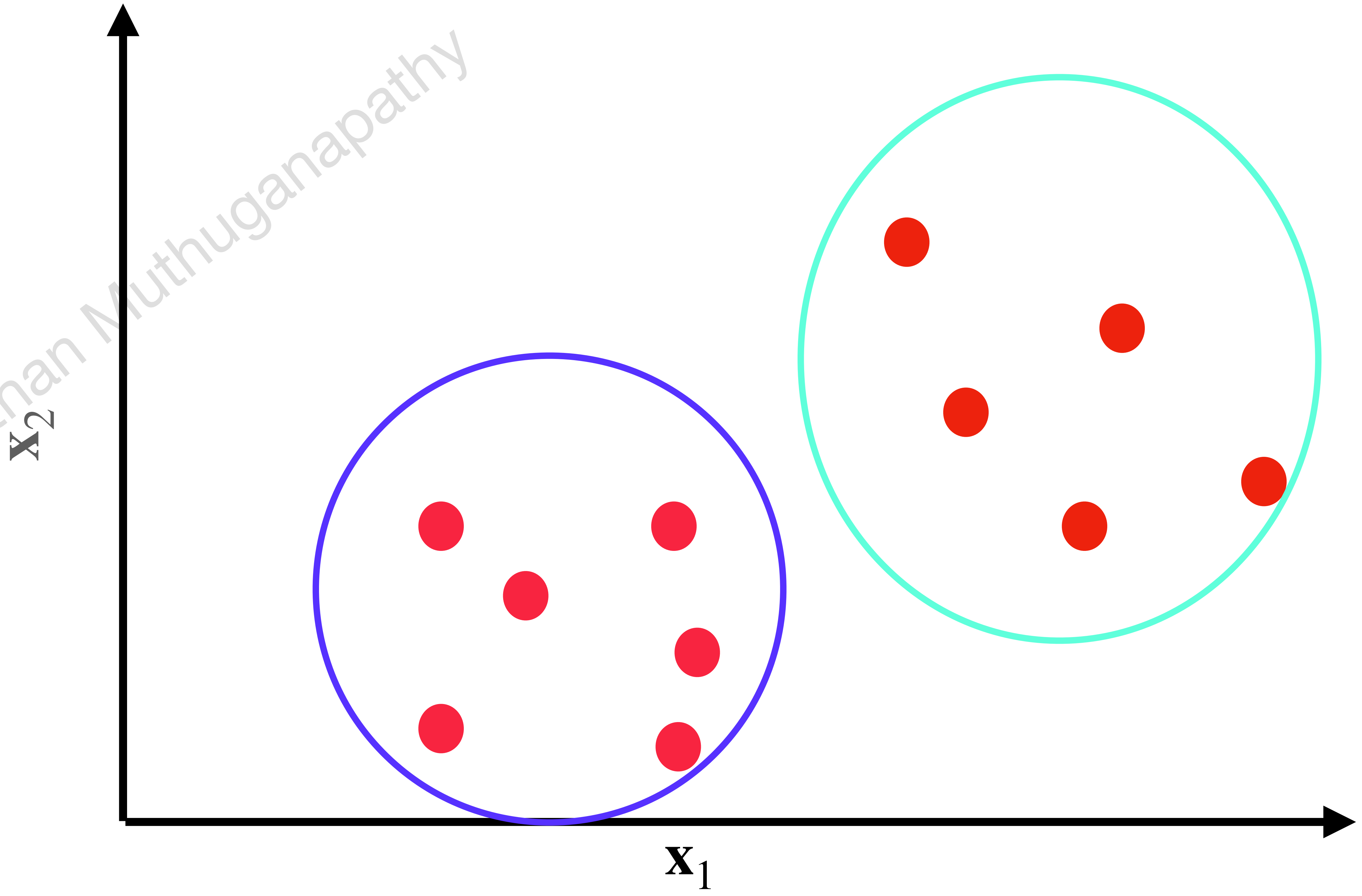
Unsupervised

- No labelling available
- Need to group / cluster them



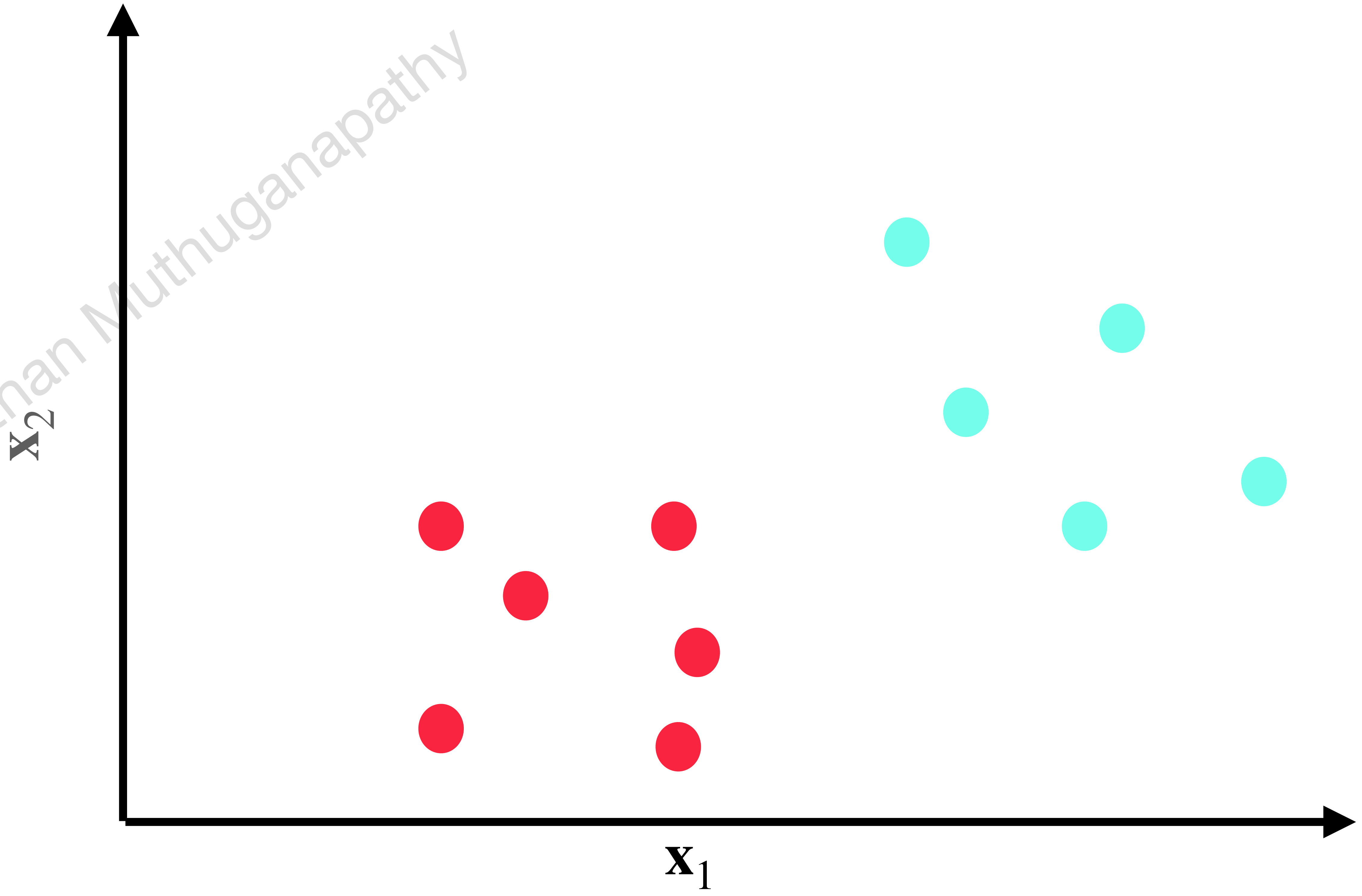
Unsupervised

- Visually two classes



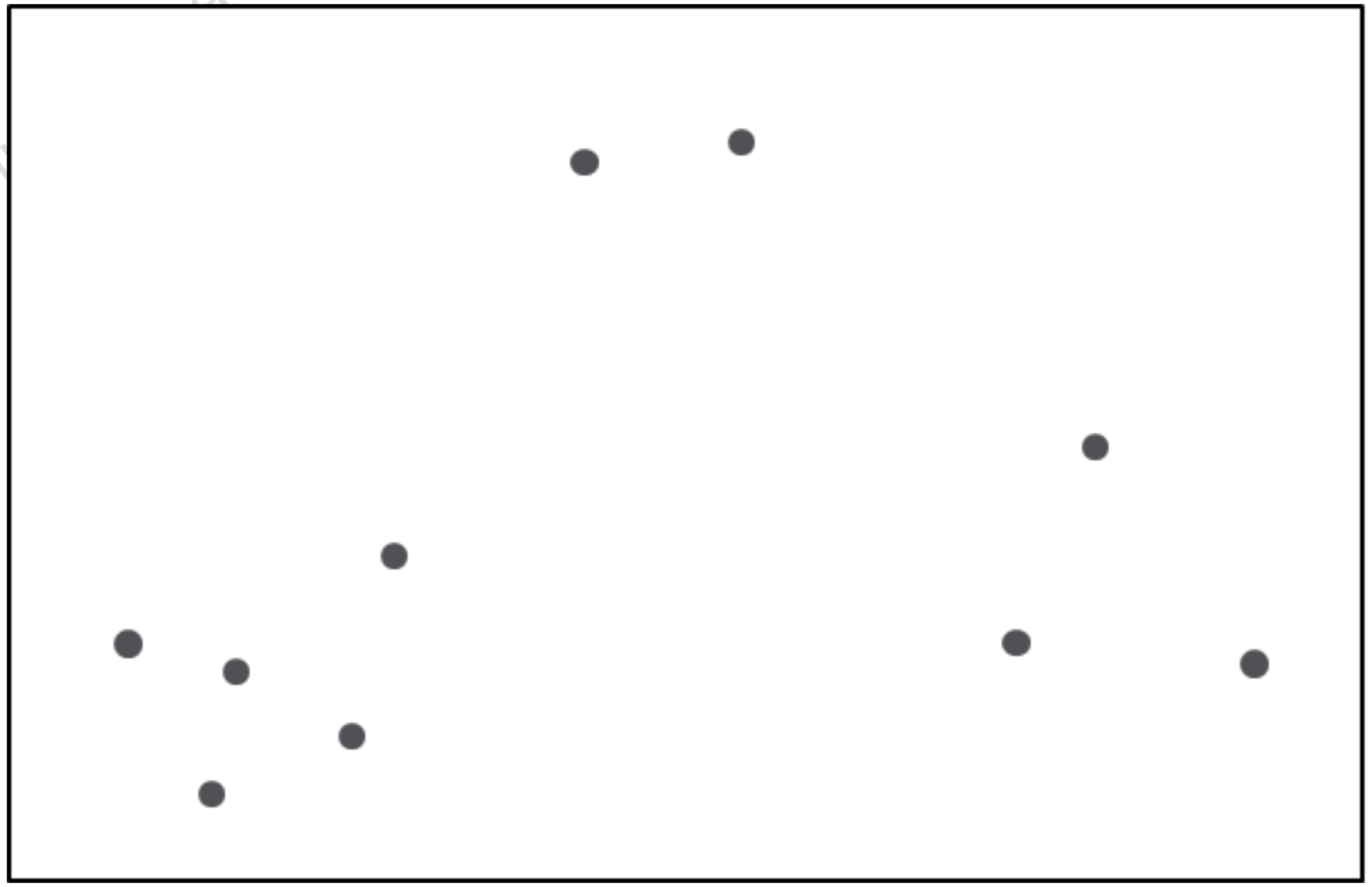
Output

- Output likely be two classes



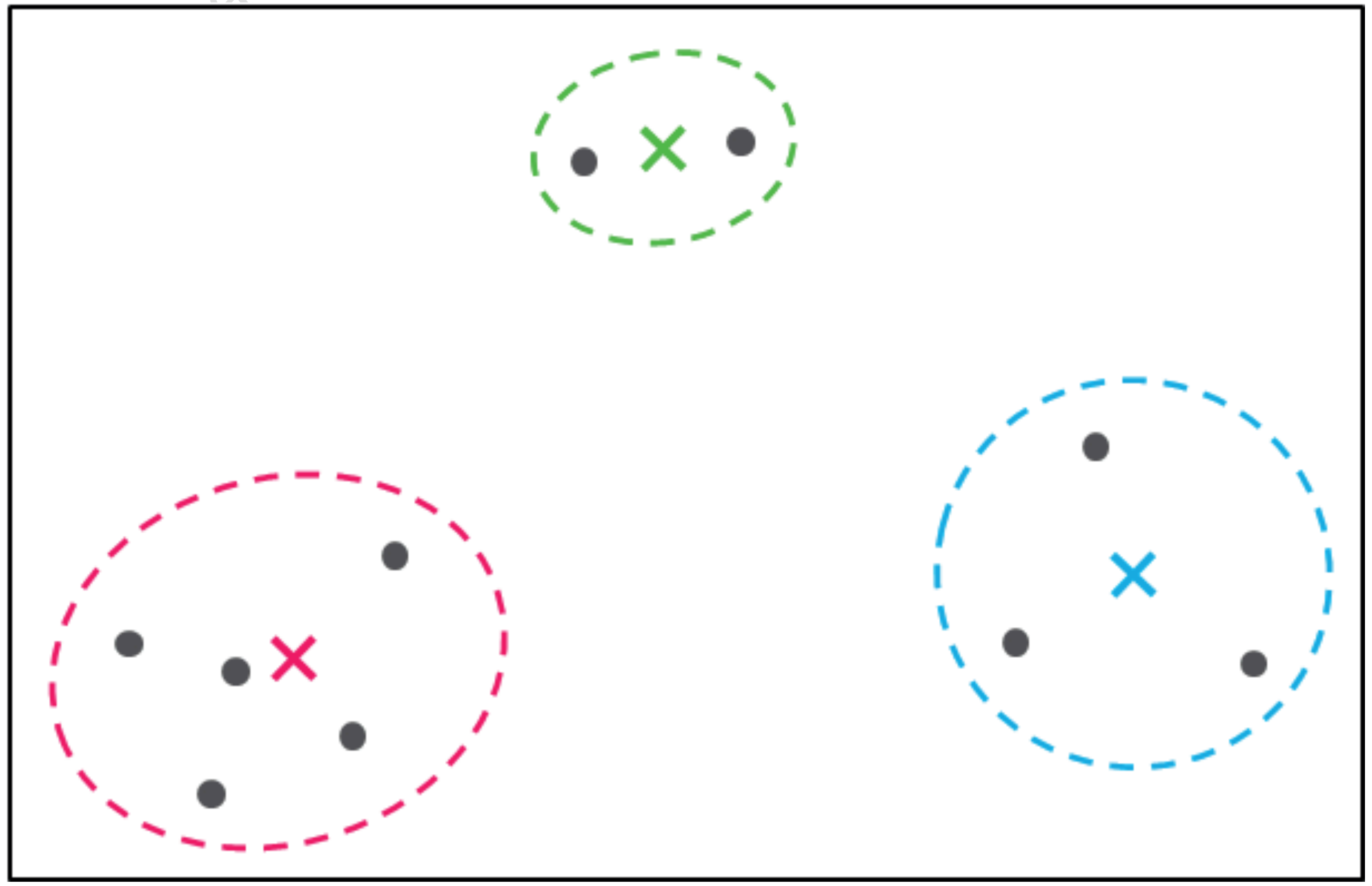
Look at this data

- How many classes (groups)?



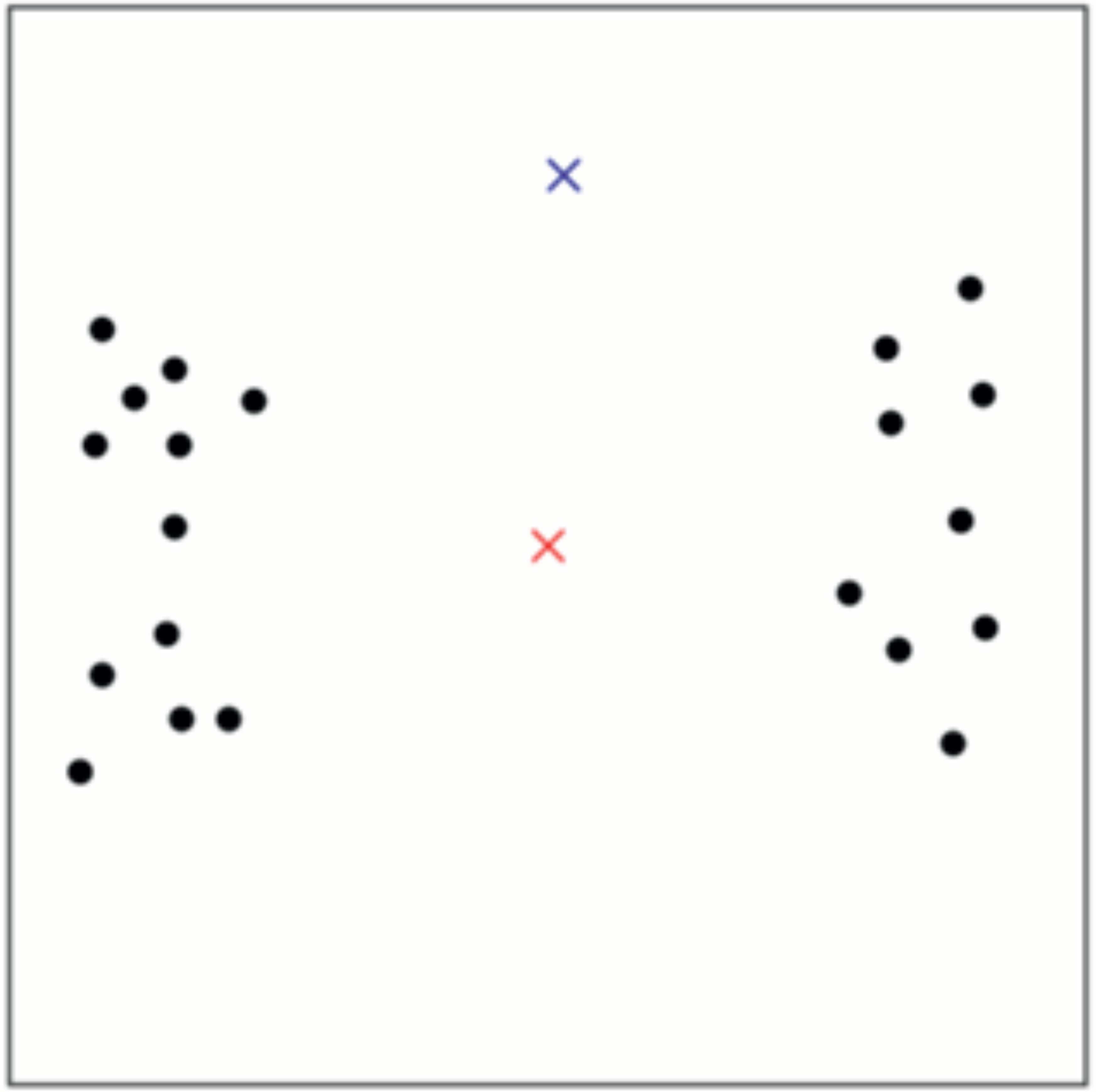
Groupings

- Possibly three
- Centroids are also shown



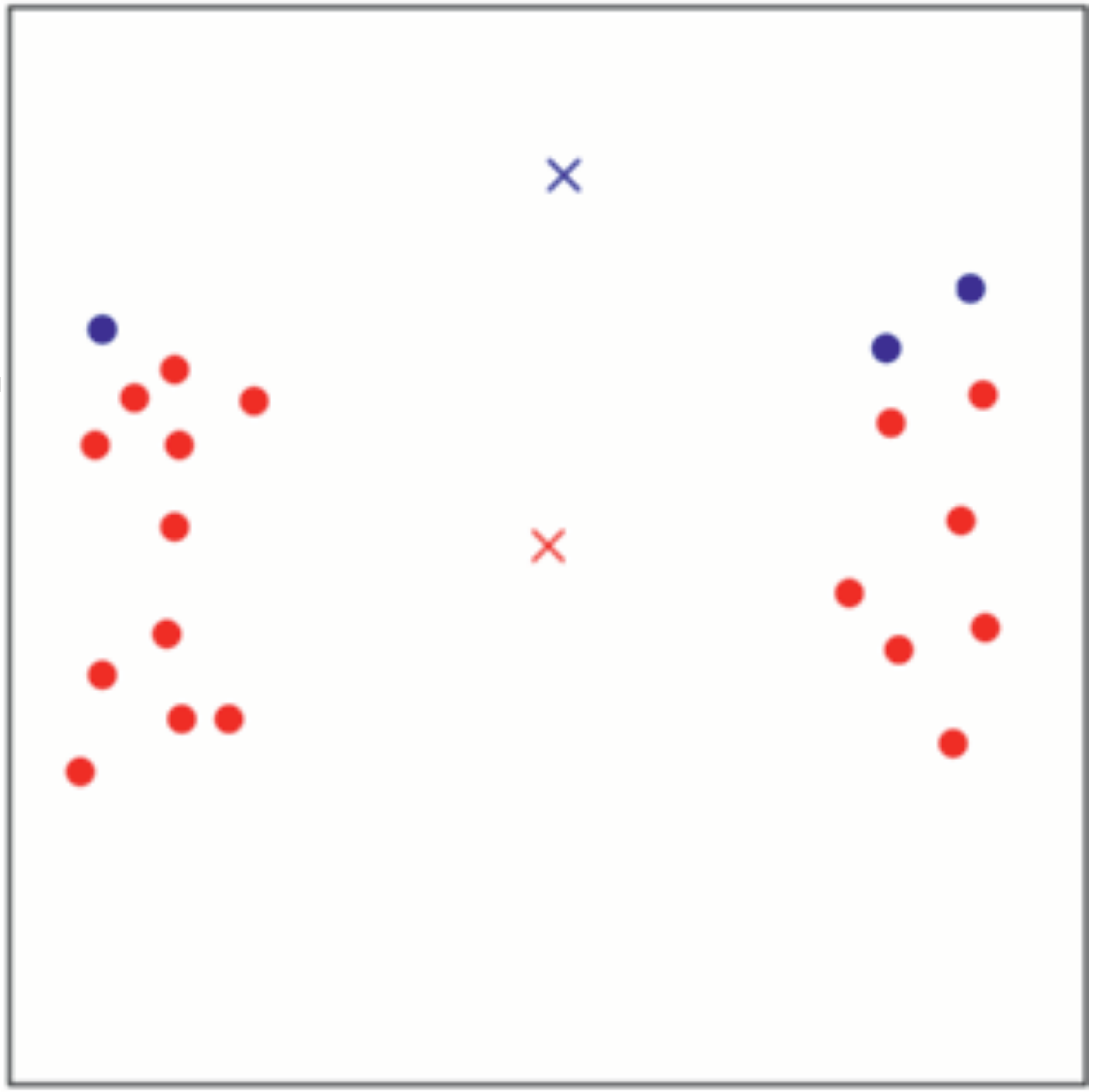
K-Means algorithm

- Input K (number of classes)
- In this case, K = 2
- Initialise TWO centroid locations (red and blue) (μ_1, μ_2)



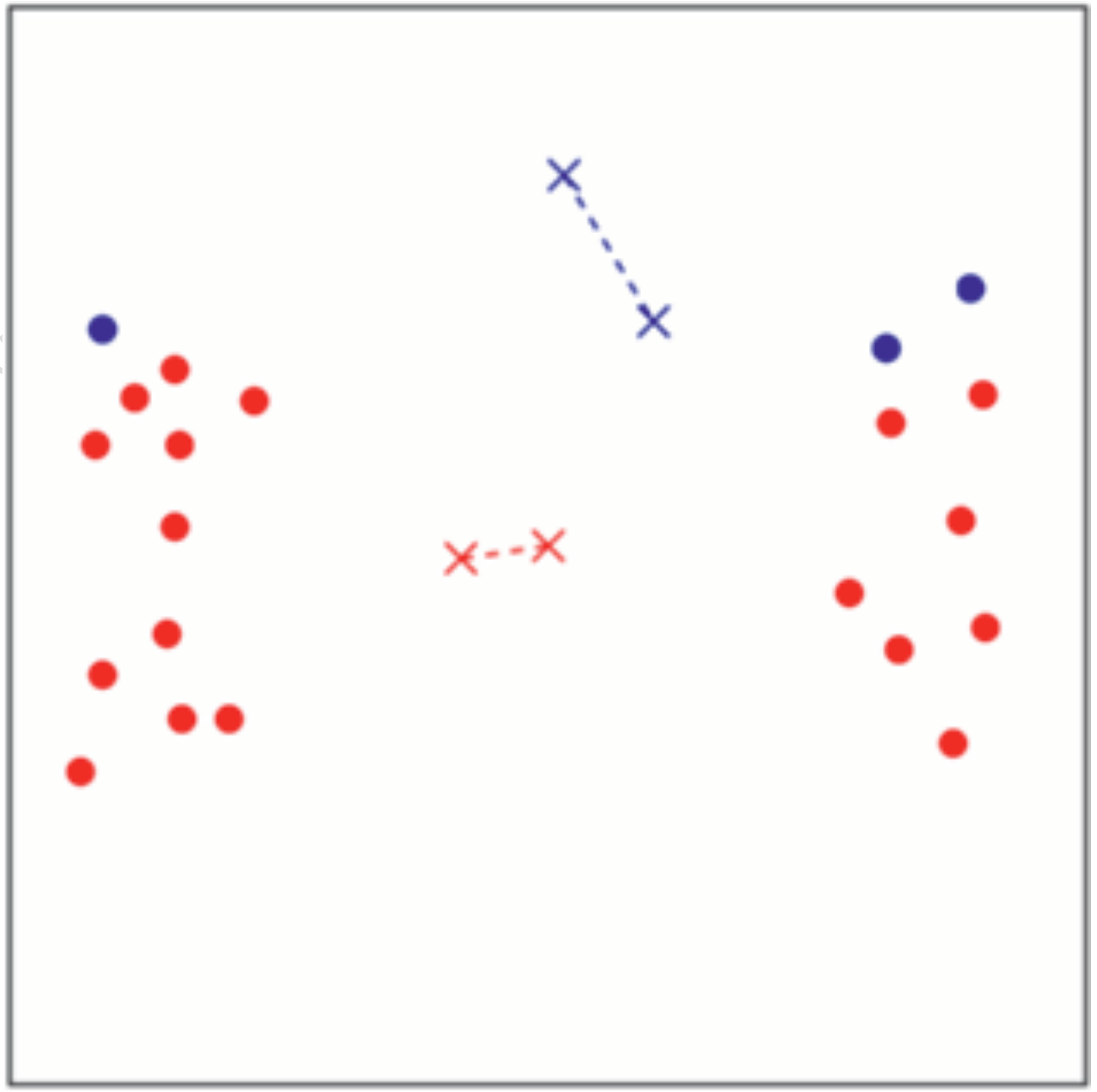
K-Means algorithm

- For each data point, find the closest centroid.
- Assign cluster index i.e., $c^{(i)}$ = index (1 or 2), i = 1 to m



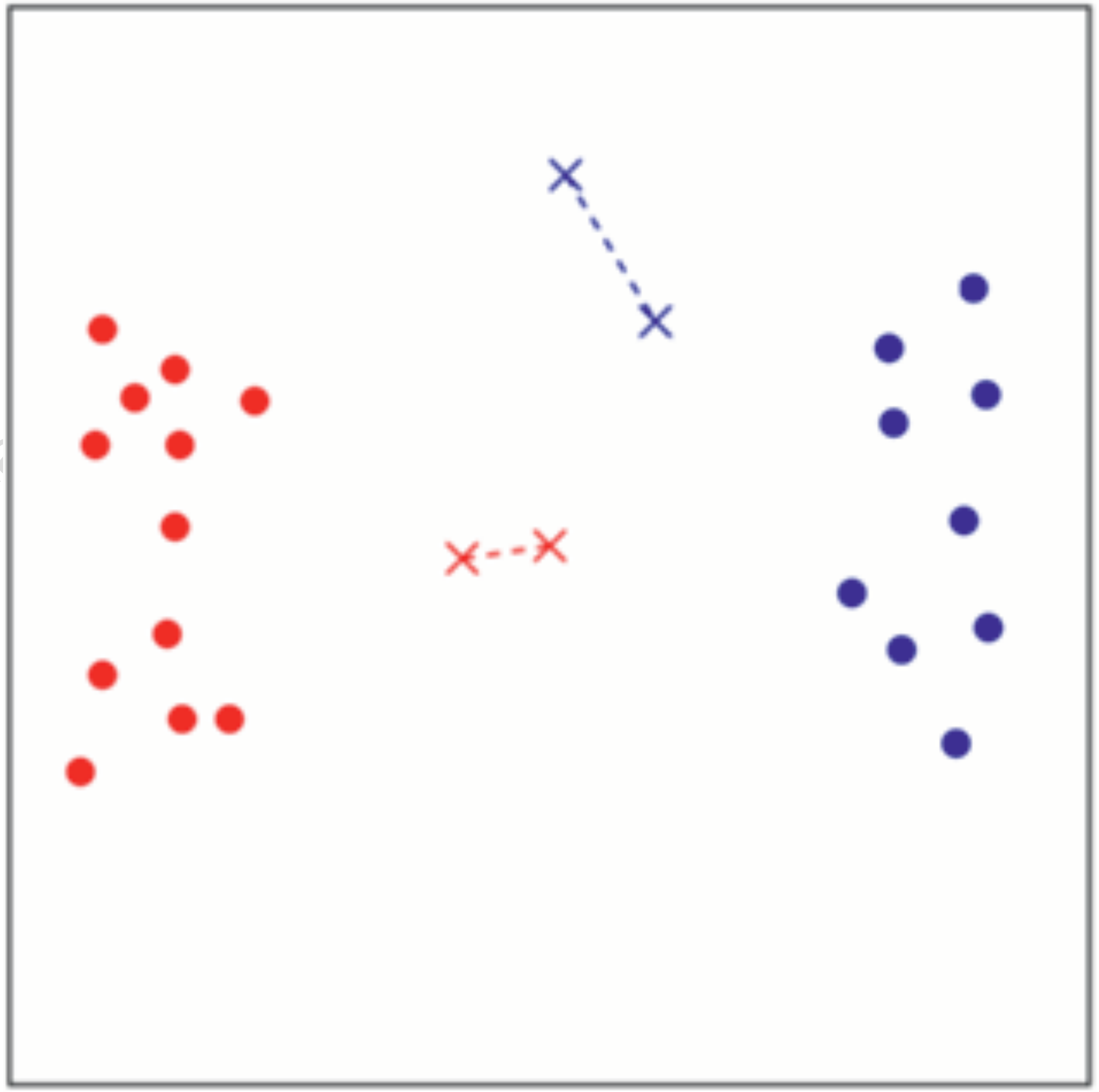
K-Means algorithm

- Update the centroid (μ_1, μ_2).



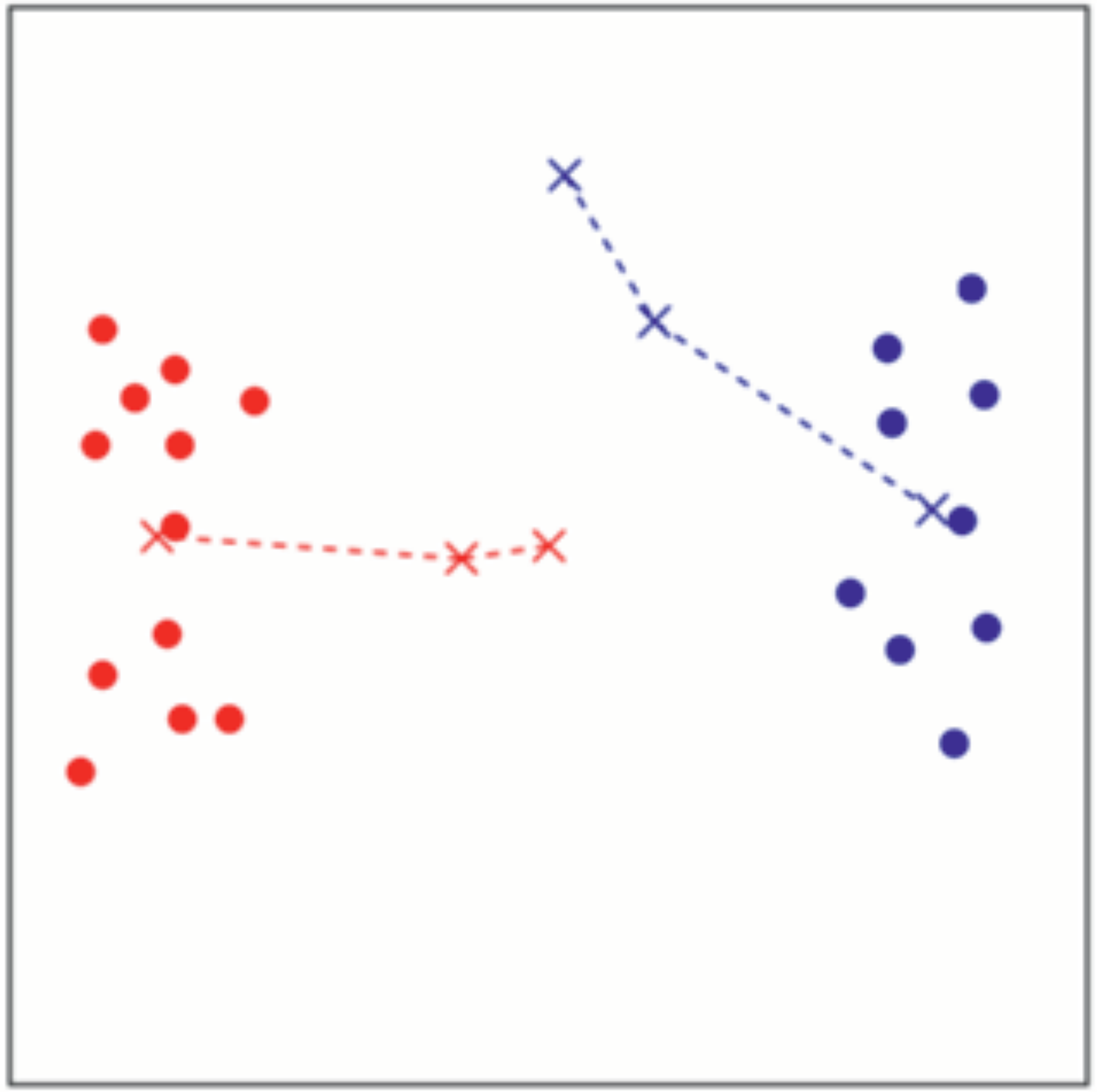
K-Means algorithm

- Update cluster index i.e., $c^{(i)}$ = index (1 or 2), i = 1 to m



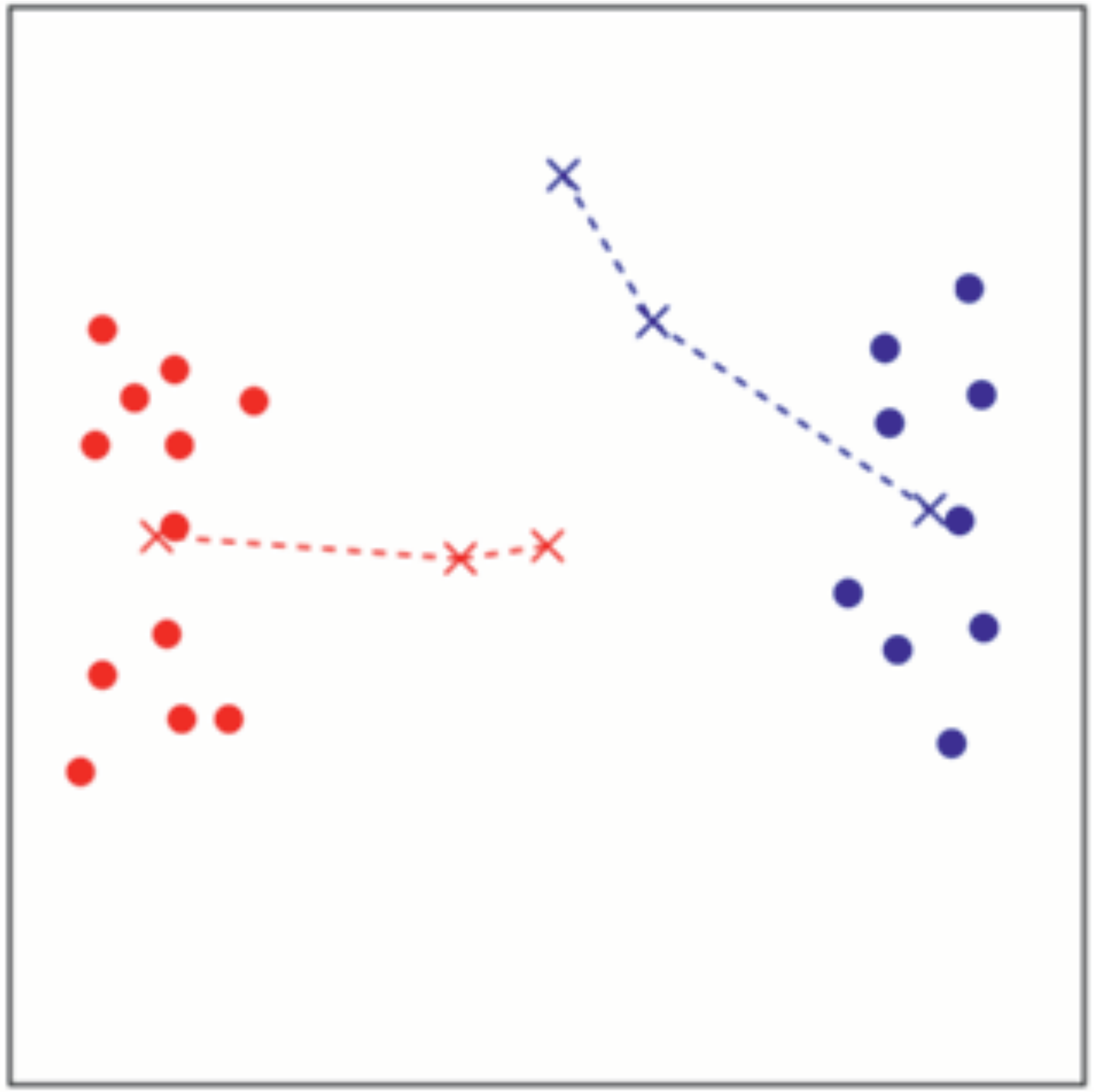
K-Means algorithm

- Update the centroid (μ_1, μ_2).



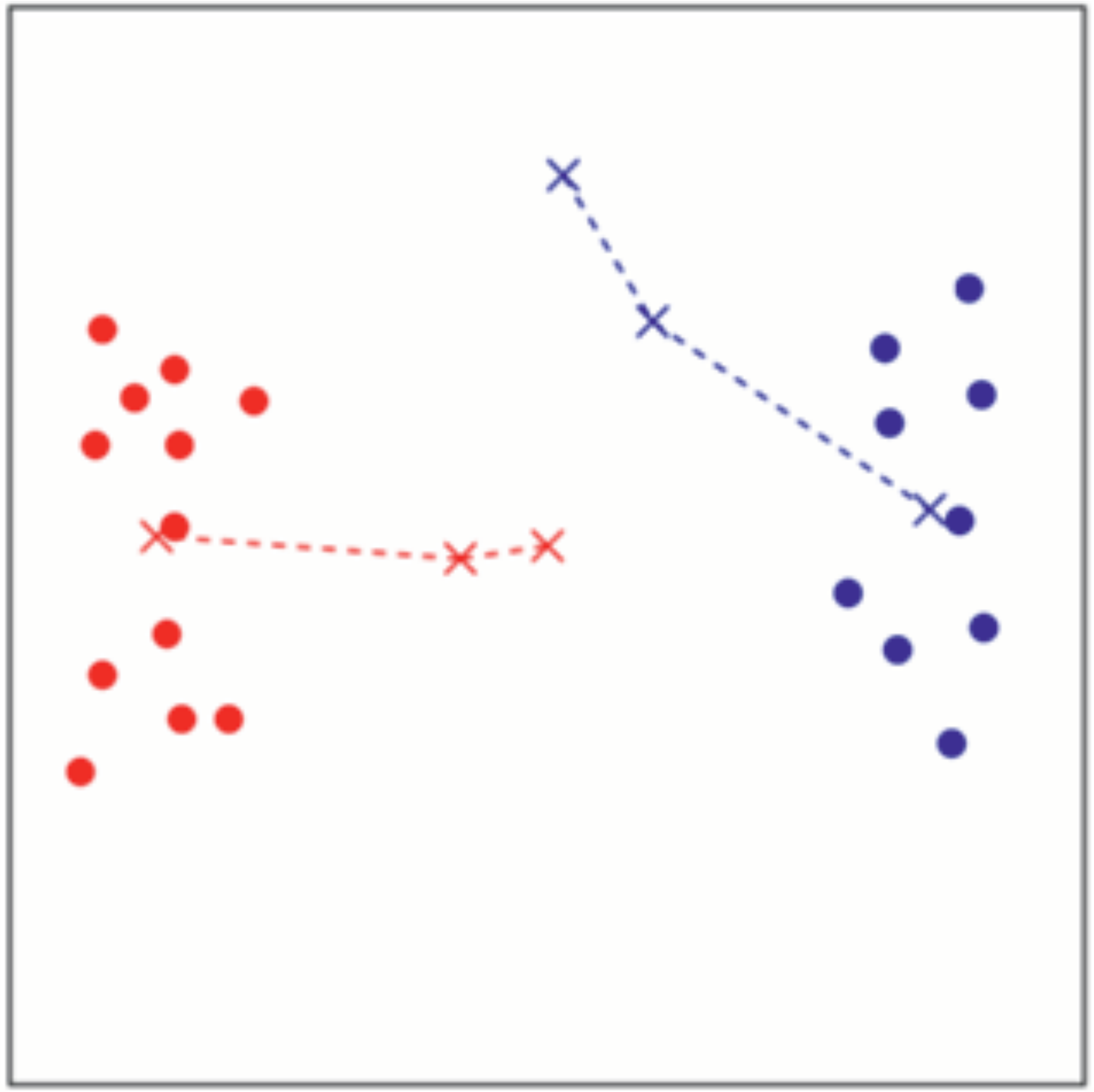
K-Means algorithm

- Update cluster index i.e., $c^{(i)}$ = index (1 or 2), i = 1 to m



K-Means algorithm

- Update the centroid (μ_1, μ_2).
- Algorithm stops as no change in update of the centroids.



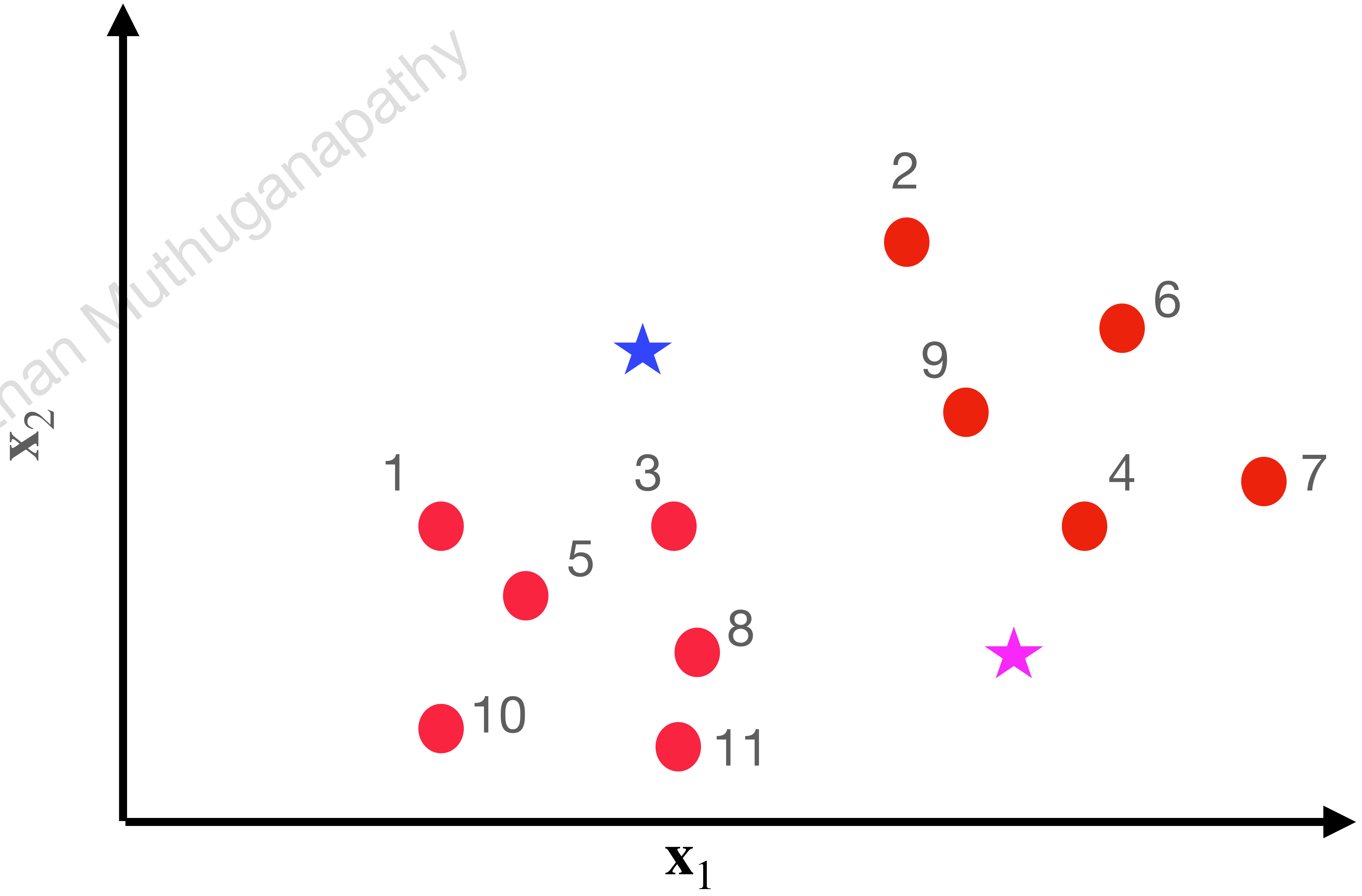
Overall algorithm

Algorithm

```
Randomly initialise $K$ cluster centroids, m-samples $(x^{(1)}, x^{(2)}, .....x^{(m)})$

Repeat {

    for $i = 1\ to\ m$

        $c^{(i)}$ = index (from 1 to $K$ of centroid closest to $x^{(i)}$ )

    for $k = 1\ to\ K$

        $\mu_k =$ average of points assigned to cluster $k$

}
```

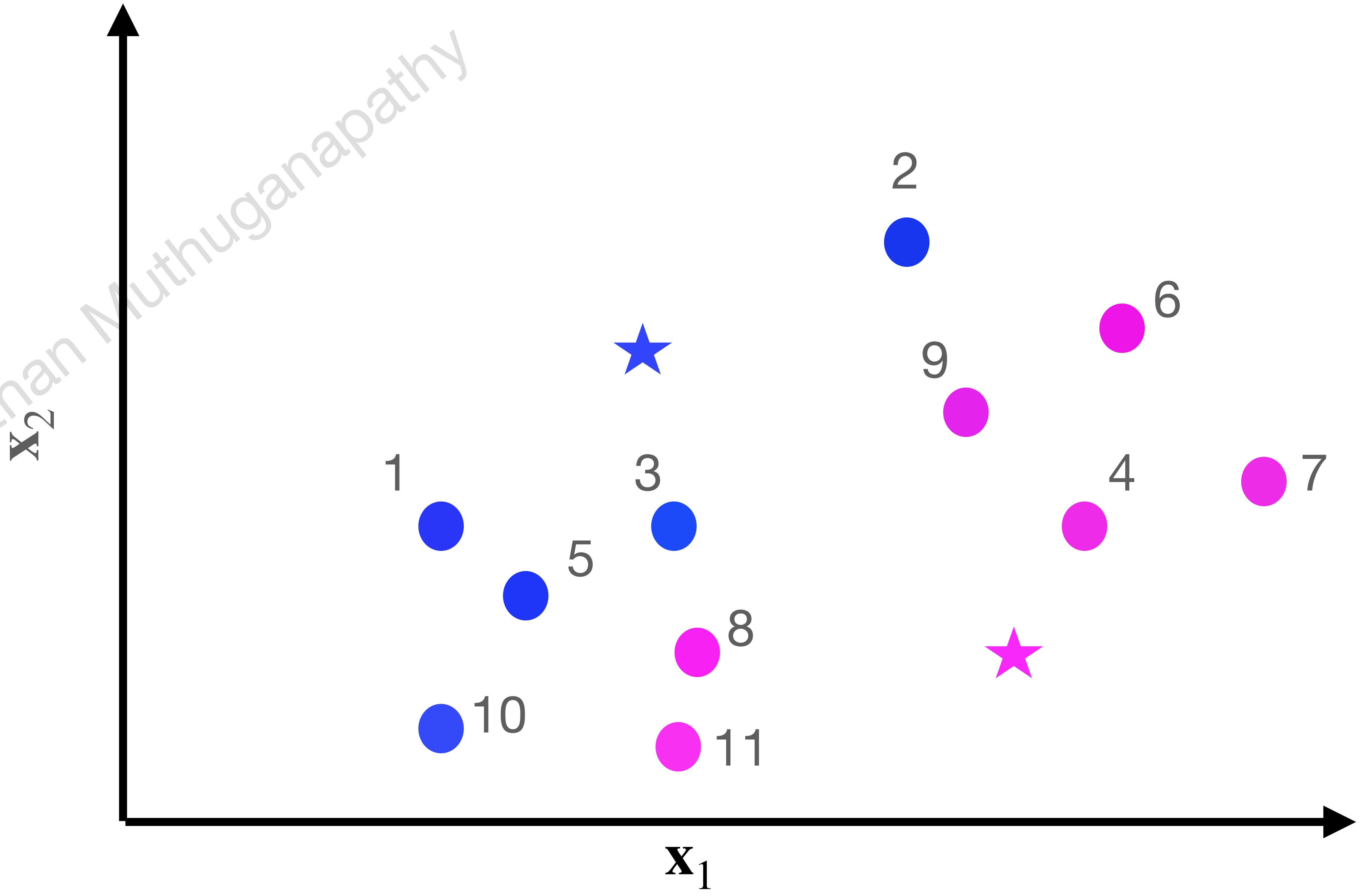
Unsupervised

- No labelling available
- Need to group / cluster them
- Random initialisation of centroids.



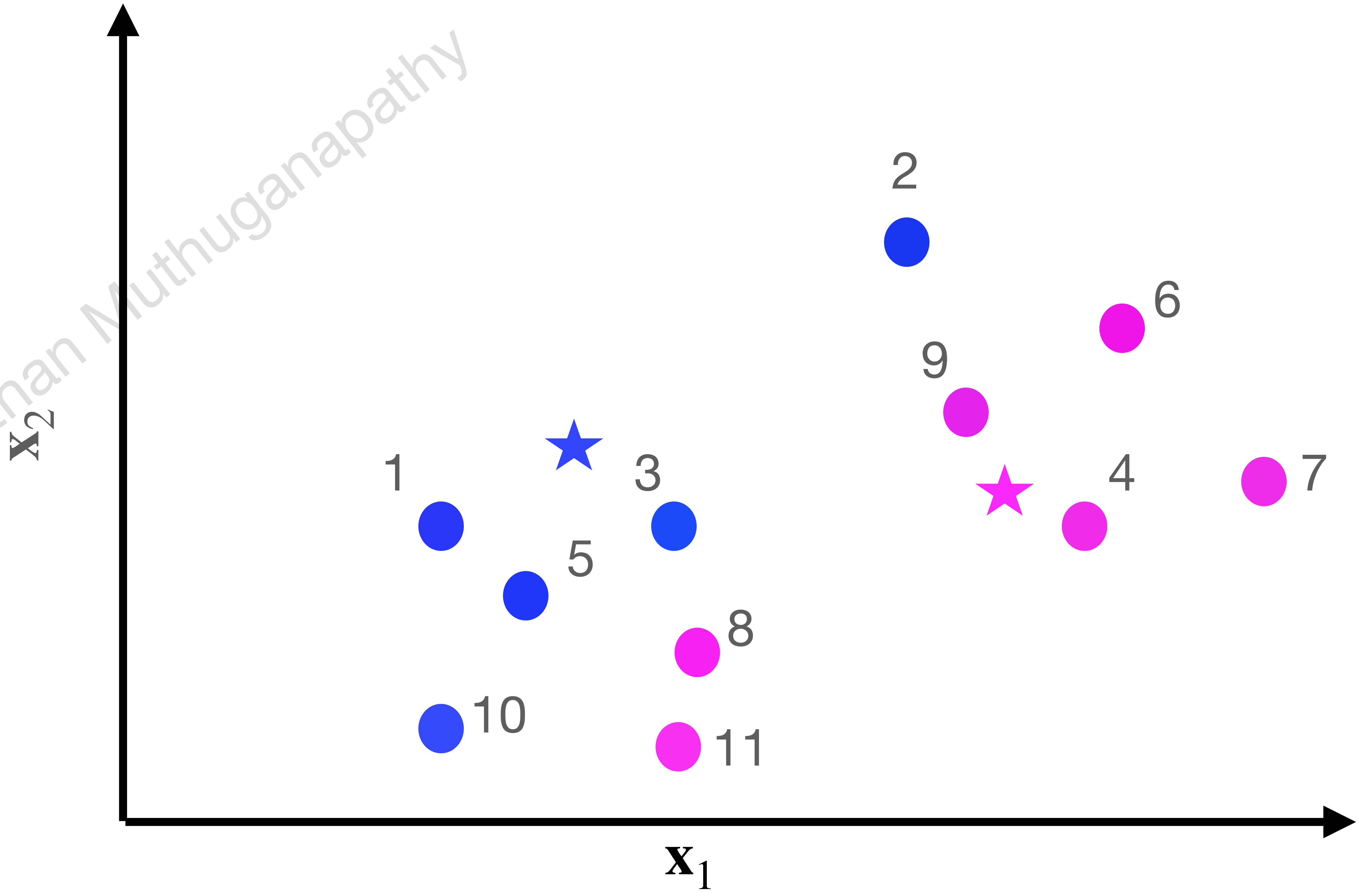
Unsupervised

- $c^{(1)}$ = 1, $c^{(2)}$ = 1, $c^{(3)}$ = 1, $c^{(4)}$ = 2, $c^{(5)}$ = 1, $c^{(6)}$ = 2, $c^{(7)}$ = 2, $c^{(8)}$ = 2, $c^{(9)}$ = 2, $c^{(10)}$ = 1, $c^{(11)}$ = 2
- $\mu_1 = \frac{1}{5}\left(x^{(1)} + x^{(2)} + x^{(3)} + x^{(5)} + x^{(10)}\right)$



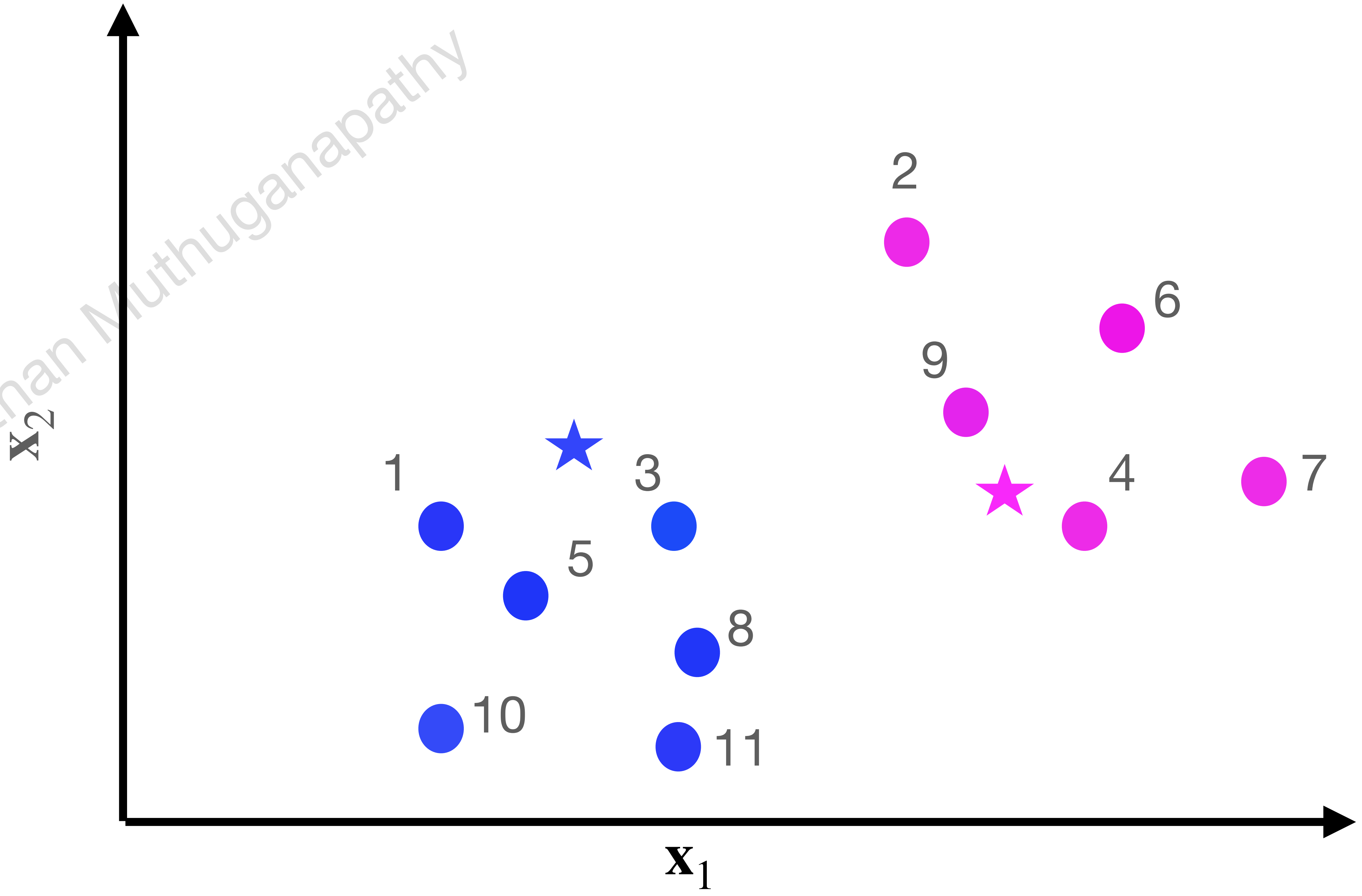
Unsupervised

- Updated Centroids



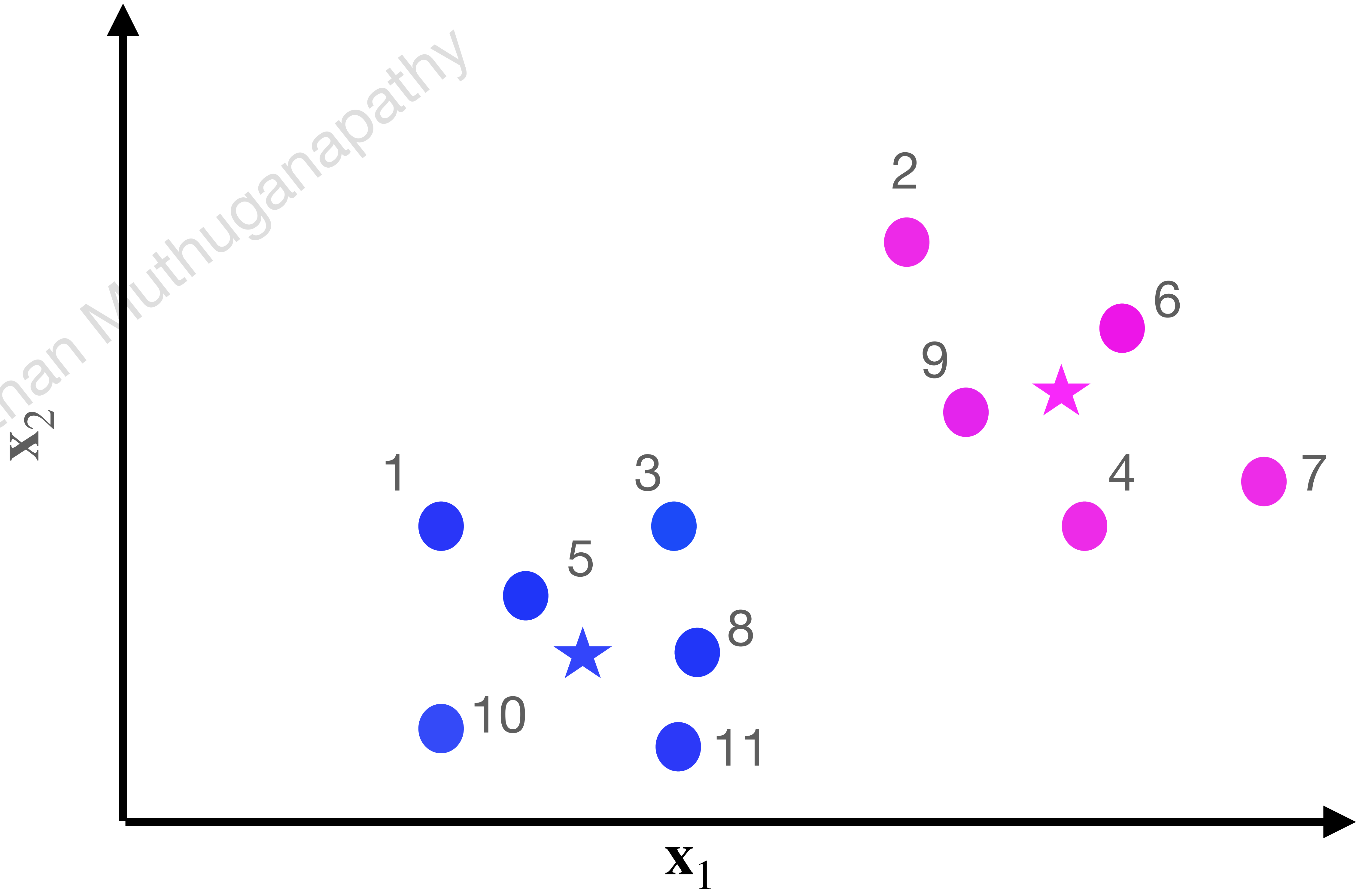
Unsupervised

- $c^{(1)}$ = 1, $c^{(2)}$ = 2, $c^{(3)}$ = 1, $c^{(4)}$ = 2, $c^{(5)}$ = 1, $c^{(6)}$ = 2, $c^{(7)}$ = 2, $c^{(8)}$ = 1, $c^{(9)}$ = 2, $c^{(10)}$ = 1, $c^{(11)}$ = 1
- $\mu_1 = \frac{1}{5}\left(x^{(1)} + x^{(3)} + x^{(5)} + x^{(8)} + x^{(10)} + x^{(11)}\right)$



Unsupervised

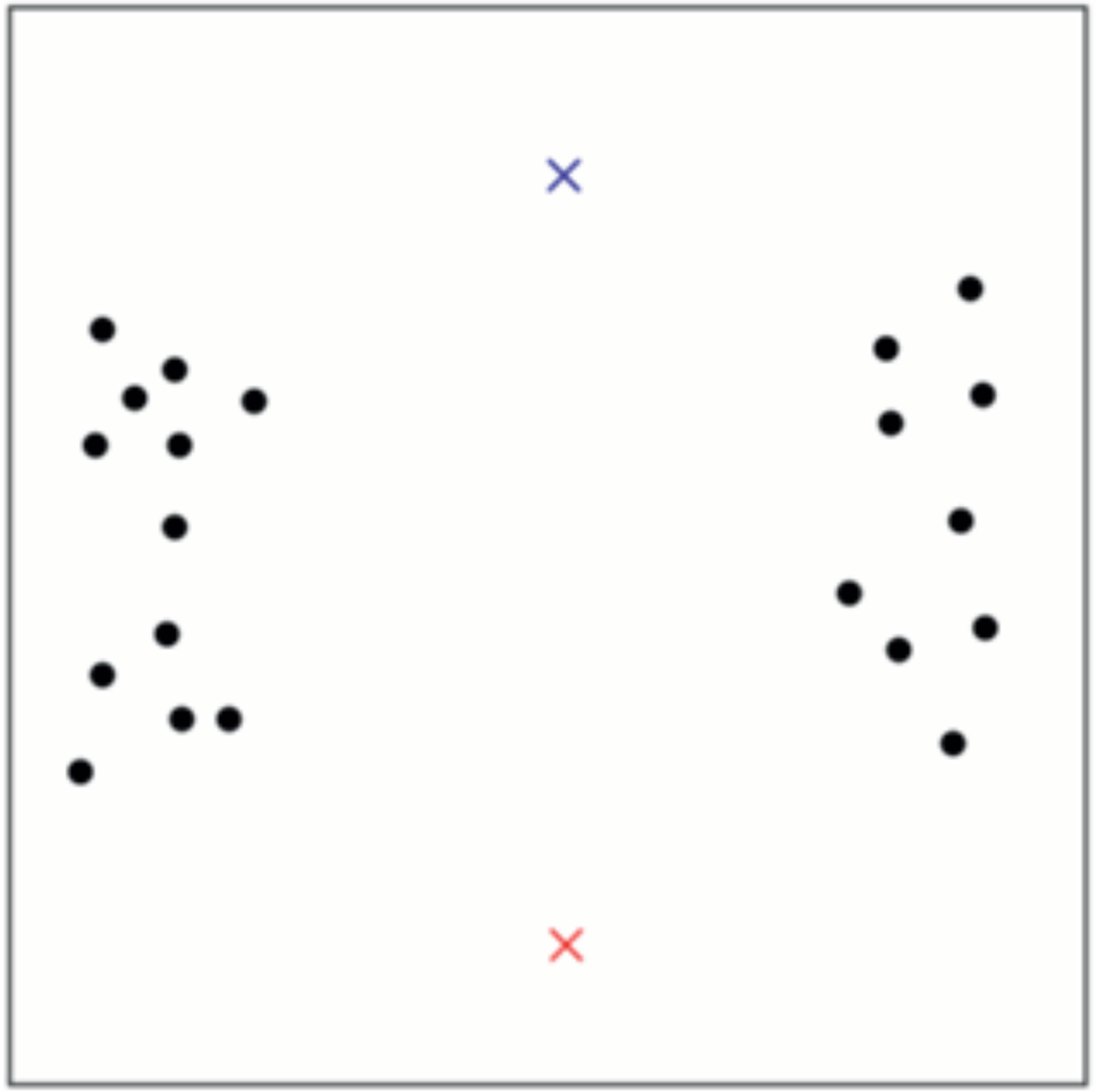
- Updated centroids.
- No further change and hence algorithm stops.



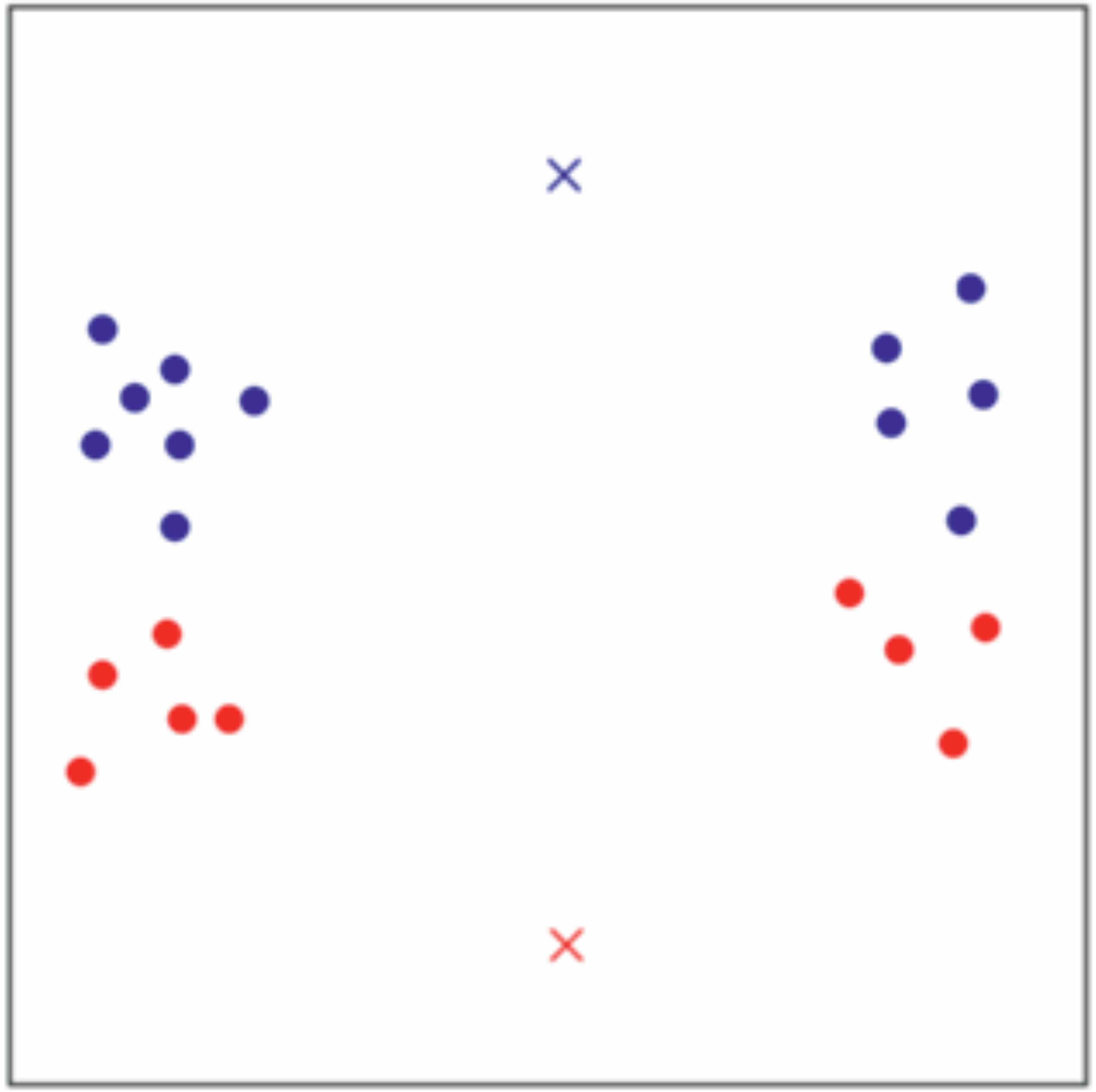
Issues

- Wrong initialisation
- How to choose K?

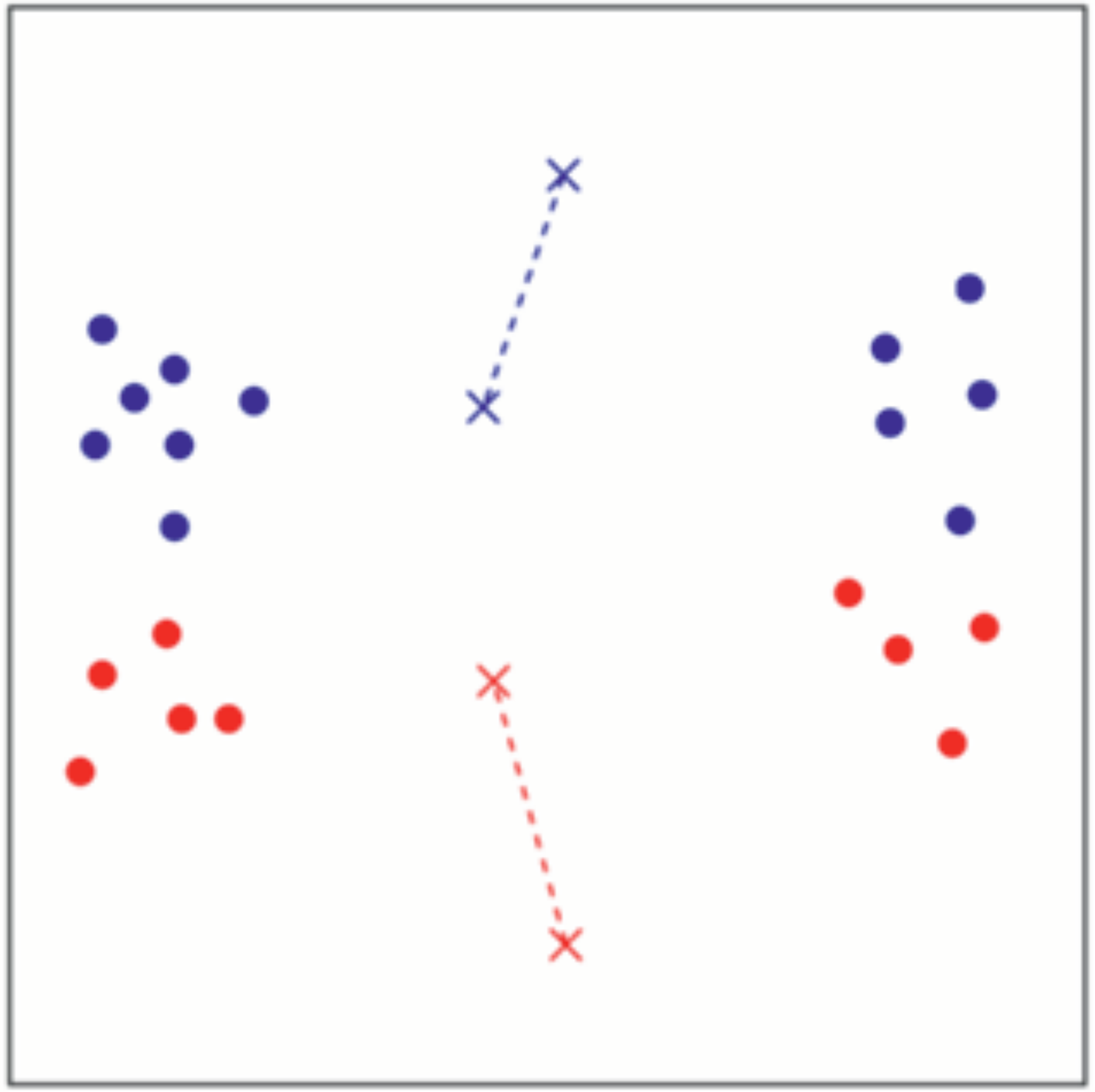
Wrong initialisation



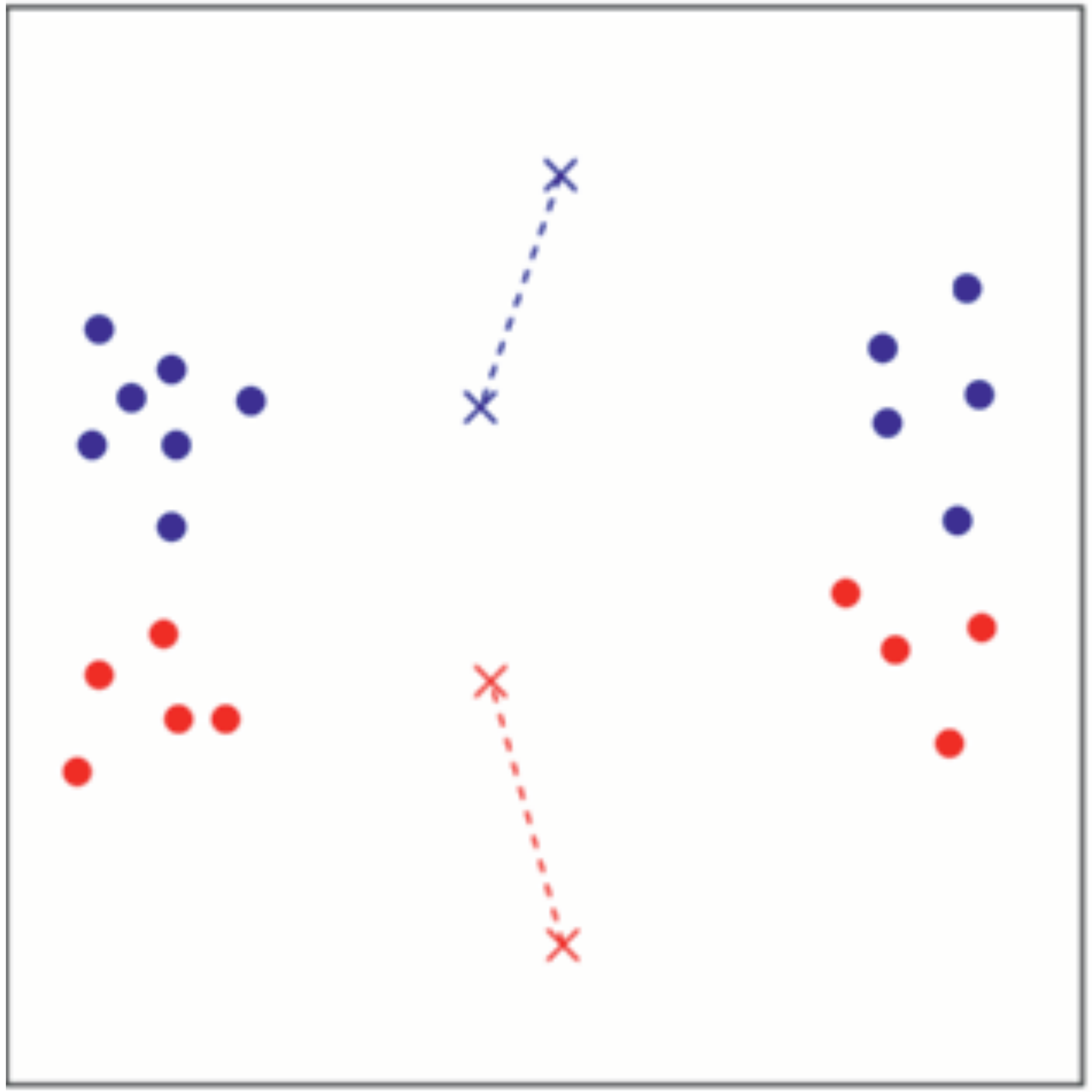
Wrong initialisation



Wrong initialisation



Wrong initialisation



Cost function

- $J = \frac{1}{m} \sum_{i=1}^{m} | | x^{(i)} - \mu_{c^{(i)}} | |$

$\mu_{c^{(i)}} - Centroid\, of\, x^{(i)}$

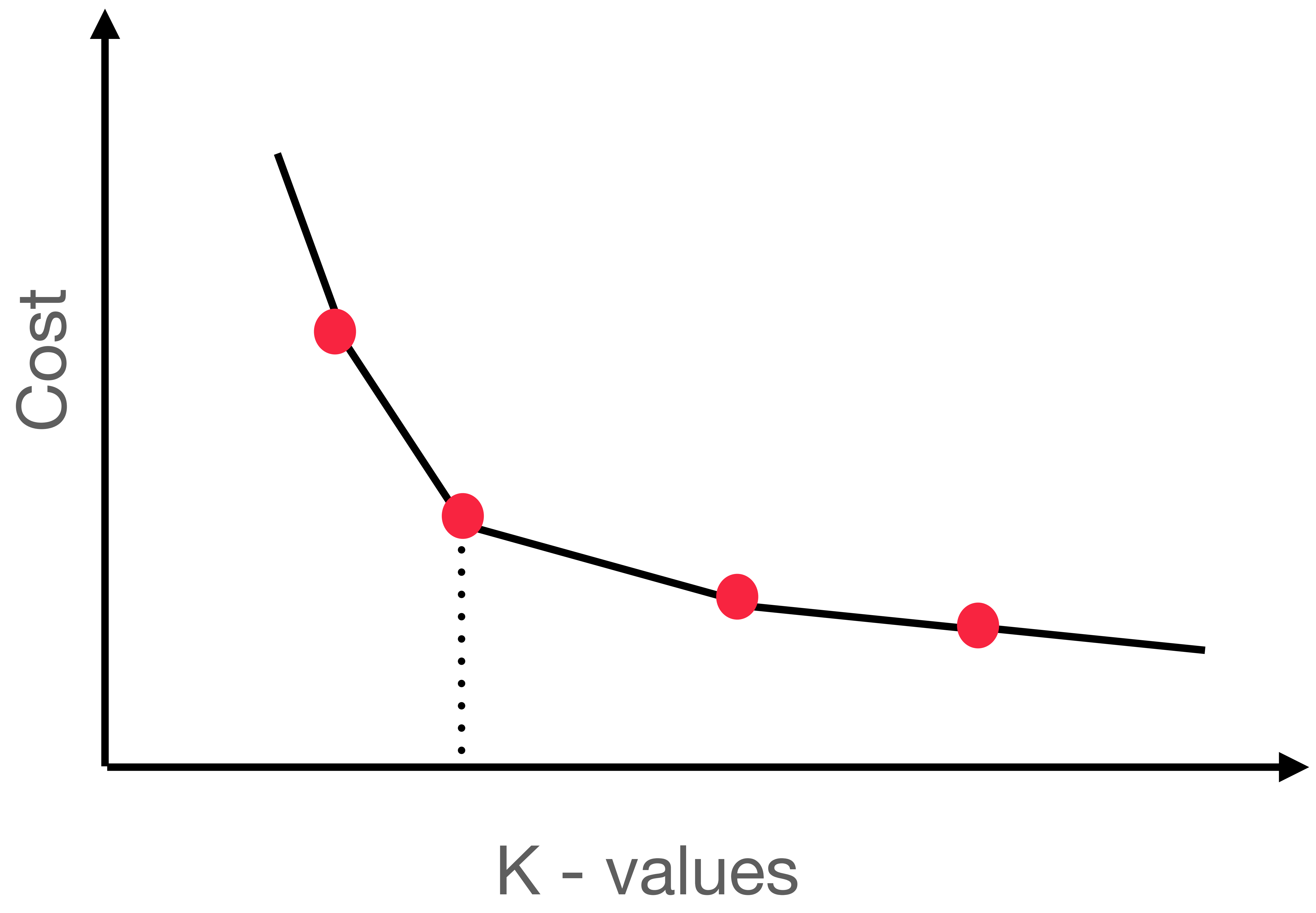
How to choose centroids

- Randomly pick K training examples
- Set them to $(\mu_1, \mu_2 \ldots\ldots, \mu_k)$ as centroids

How to find K?

- for i = 1 to 100
    - Randomly initialise K-means
    - Run K-means
    - Compute cost function
    - Pick the centroids with min J
- Try k = 2 to 10

Elbow method



Further reading

- Agglomerative clustering
- Dendrogram
- DBSCAN, HDBSCAN