Machine Learning: ITCS 4156

Clustering: k-Means and k-Medoids

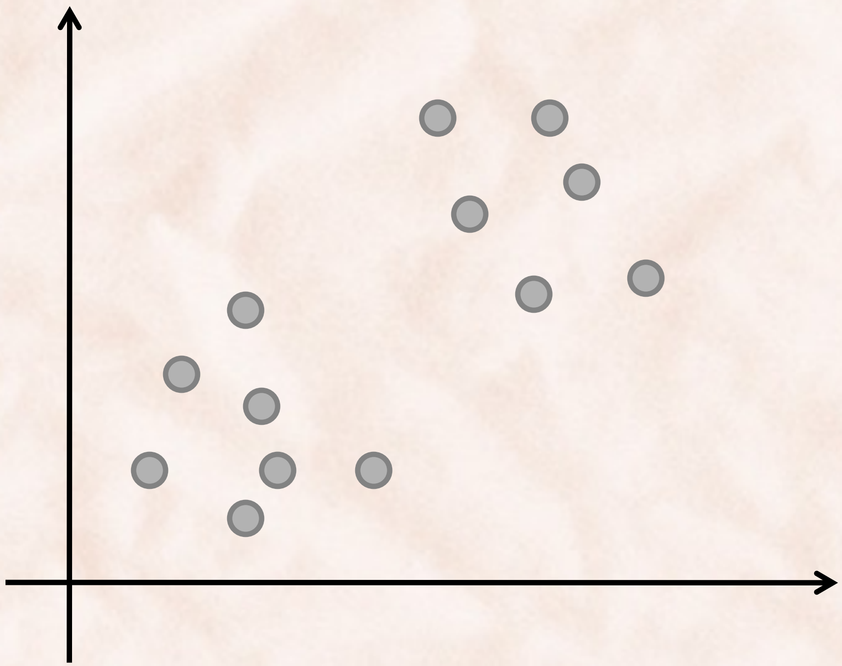
Razvan C. Bunescu

Department of Computer Science @ CCI

rbunescu@uncc.edu

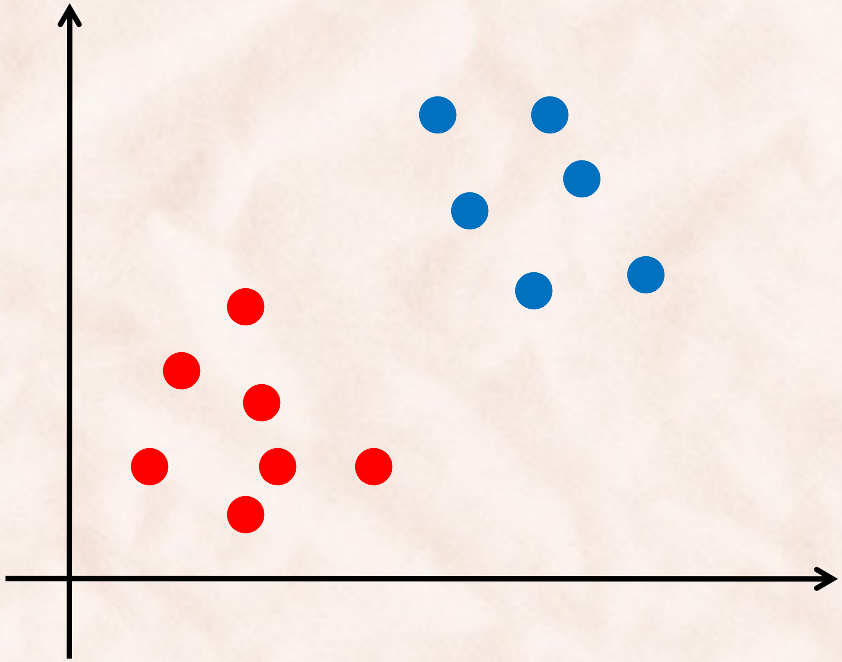
Unsupervised Learning: Clustering

- Partition unlabeled examples into disjoint clusters such that:
  - Examples in the same cluster are very similar.
  - Examples in different clusters are very different.



Unsupervised Learning: Clustering

- Partition unlabeled examples into disjoint clusters such that:
  - Examples in the same cluster are very similar.
  - Examples in different clusters are very different.



Divisive Clustering with *k*-Means

- The goal is to produce k clusters C ={C_1, C_2, …, C_k} such that instances are close to the cluster centroids:
  - The cluster centroid $\mathbf{m}_i$ is the mean of all instances in the cluster C_i.

- Optimization problem:

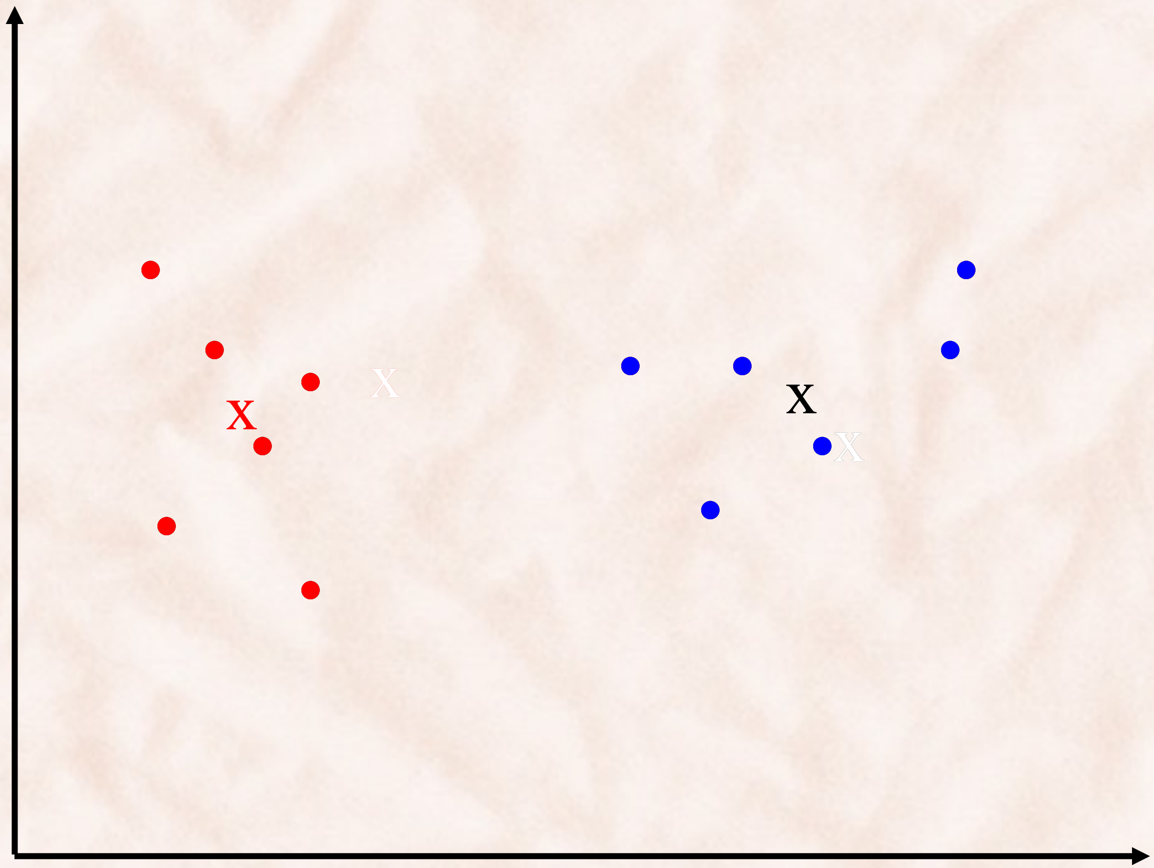
$$\hat{C} = \arg\min_{C} J(C)$$

$$J(C) = \sum_{i=1}^{k} \sum_{\mathbf{x} \in C_i} \| \mathbf{x} - \mathbf{m}_i \|^2$$

The k-Means Algorithm

1. start with some seed centroids $\mathbf{m}_1^{(0)}, \mathbf{m}_2^{(0)}, \ldots, \mathbf{m}_k^{(0)}$
2. **set** $t \leftarrow 0$.
3. **while** not converged:
4. **for** each $\mathbf{x}$:
5. **set** $\mathbf{m}^{(t)}(\mathbf{x}) \leftarrow \arg\min_{\mathbf{m}_i^{(t)}} \left\| \mathbf{x} - \mathbf{m}_i^{(t)} \right\|$ ← [E] step
6. **set** $C_i^{(t+1)} \leftarrow \left\{ \mathbf{x} \mid \mathbf{m}^{(t)}(\mathbf{x}) = \mathbf{m}_i^{(t)} \right\}$
7. **set** $\mathbf{m}_i^{(t+1)} \leftarrow \frac{1}{\left| C_i^{(t+1)} \right|} \sum_{\mathbf{x} \in C_i^{(t+1)}} \mathbf{x}$ ← [M] step
8. **set** $t \leftarrow t + 1$

The k-Means Algorithm ($k = 2$)



Pick seeds

Reassign clusters

Compute centroids

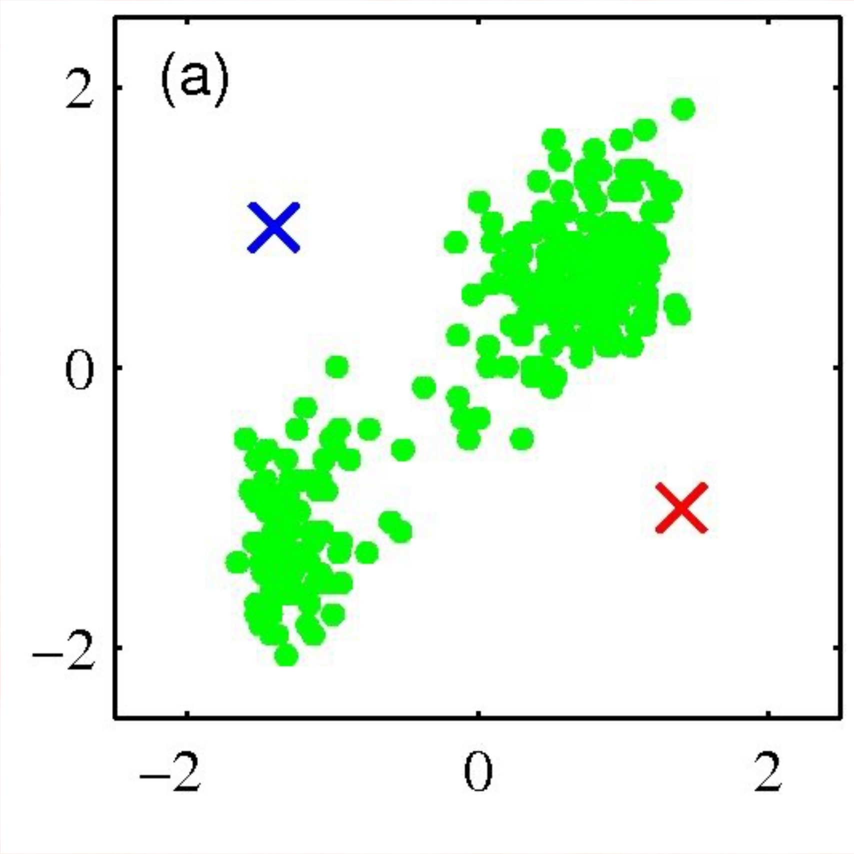
Reasssign clusters

Compute centroids

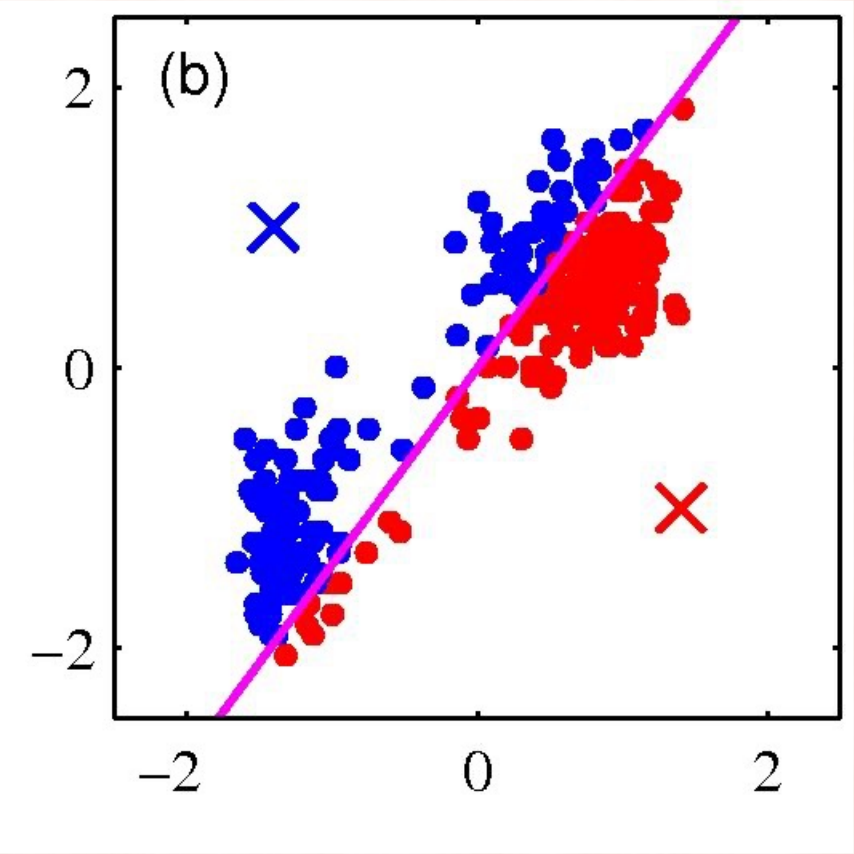
Reassign clusters

Converged!

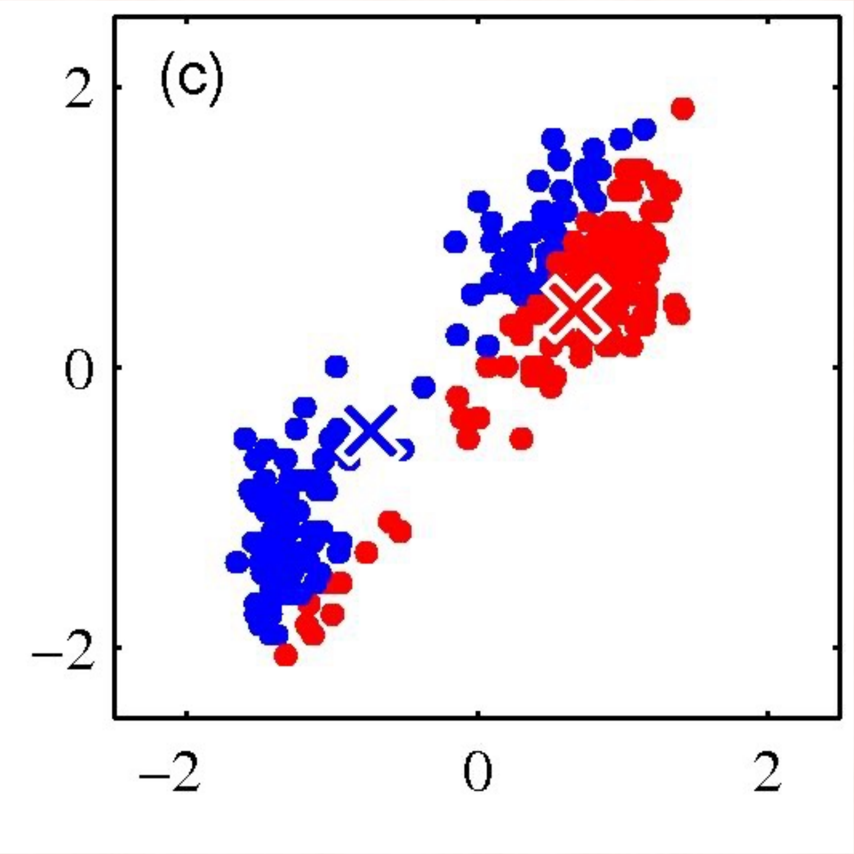
The k-Means Algorithm ($k = 2$)



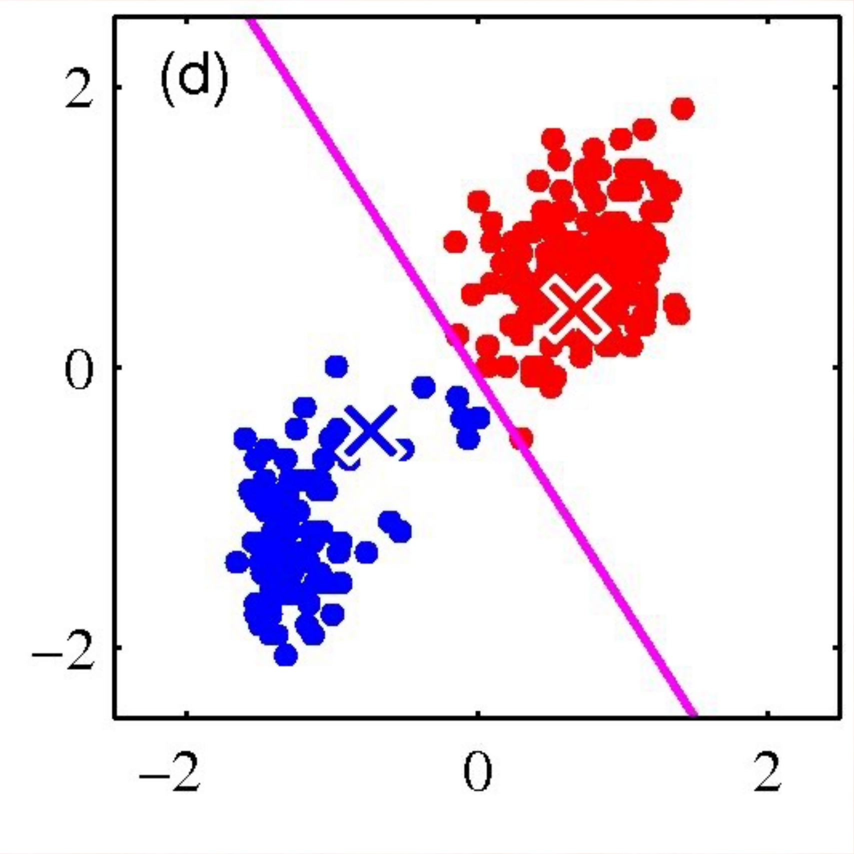
The k-Means Algorithm ($k = 2$)



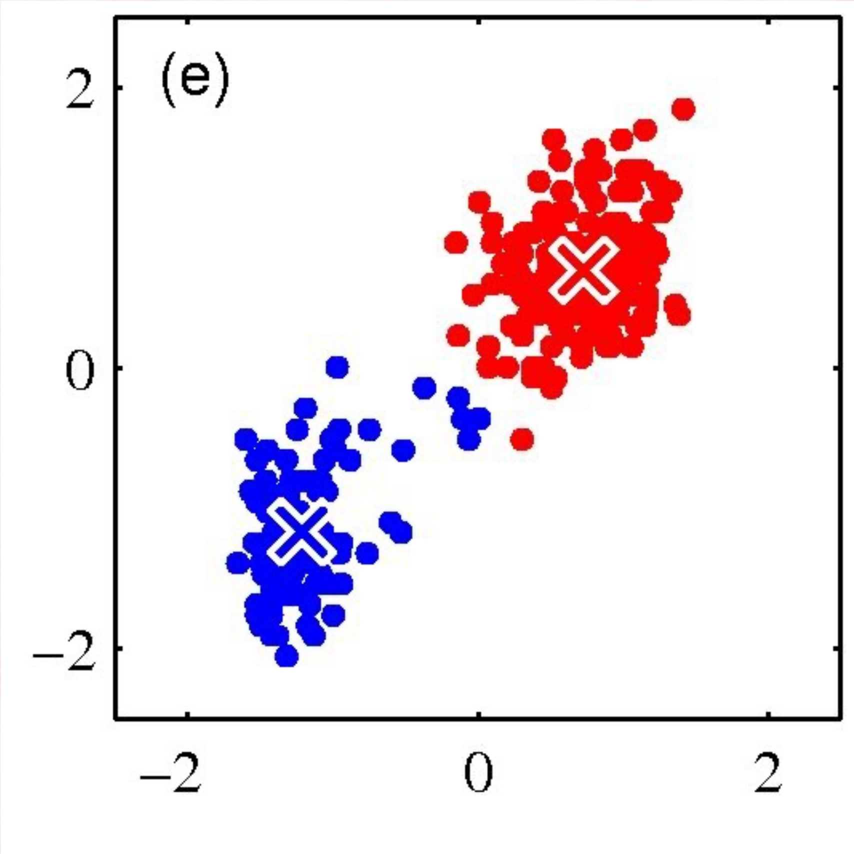
The k-Means Algorithm ($k = 2$)



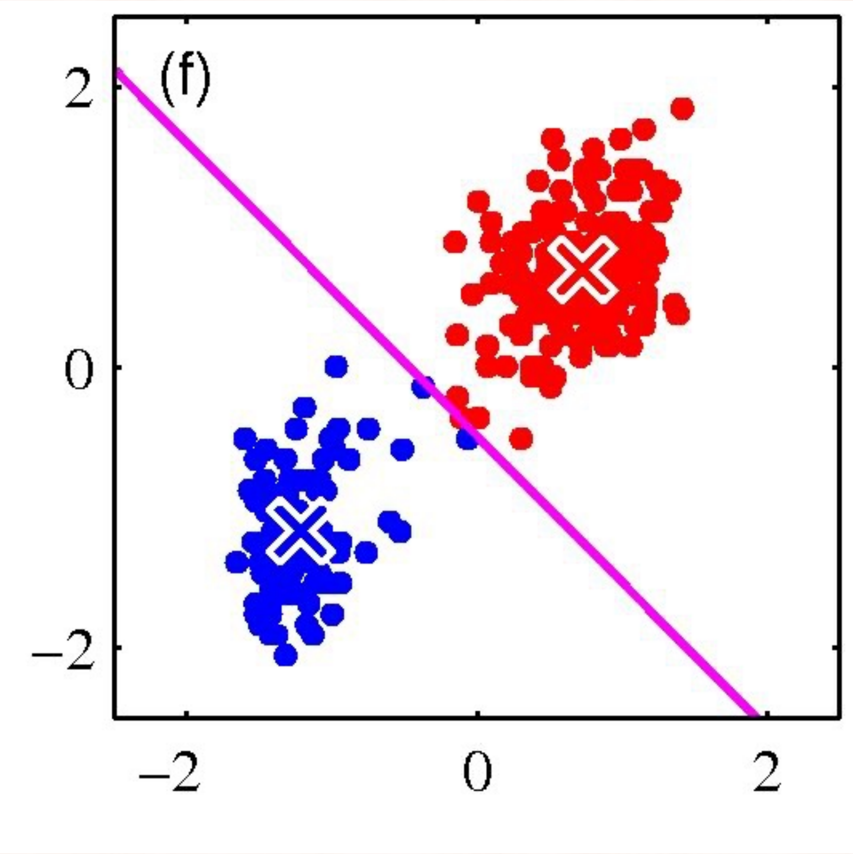
The k-Means Algorithm ($k = 2$)



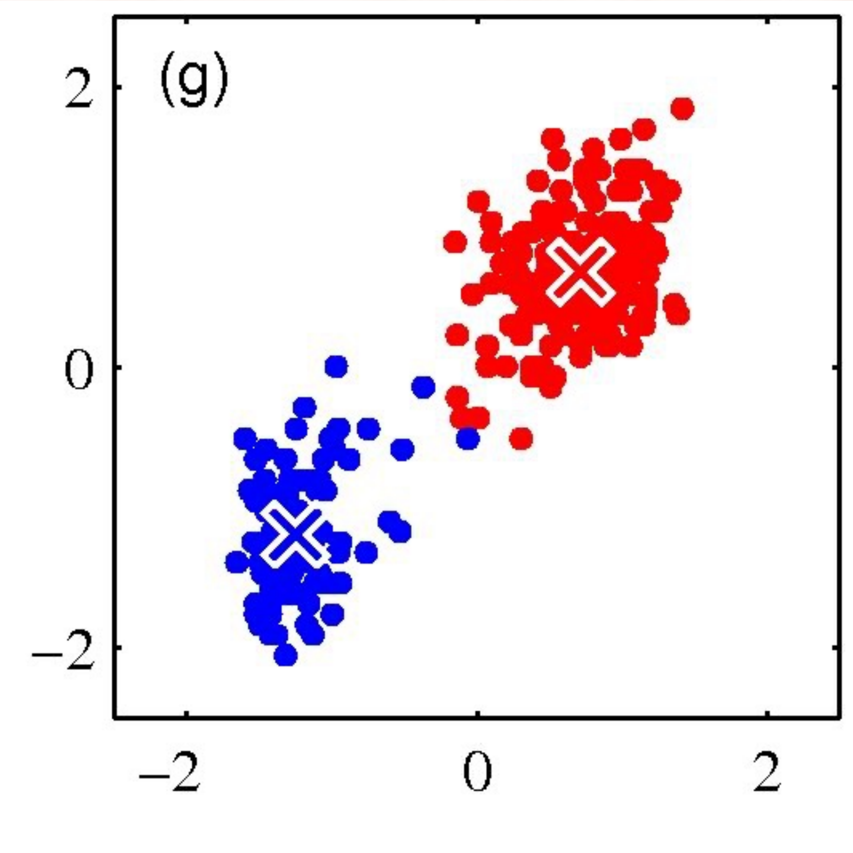
The k-Means Algorithm ($k = 2$)



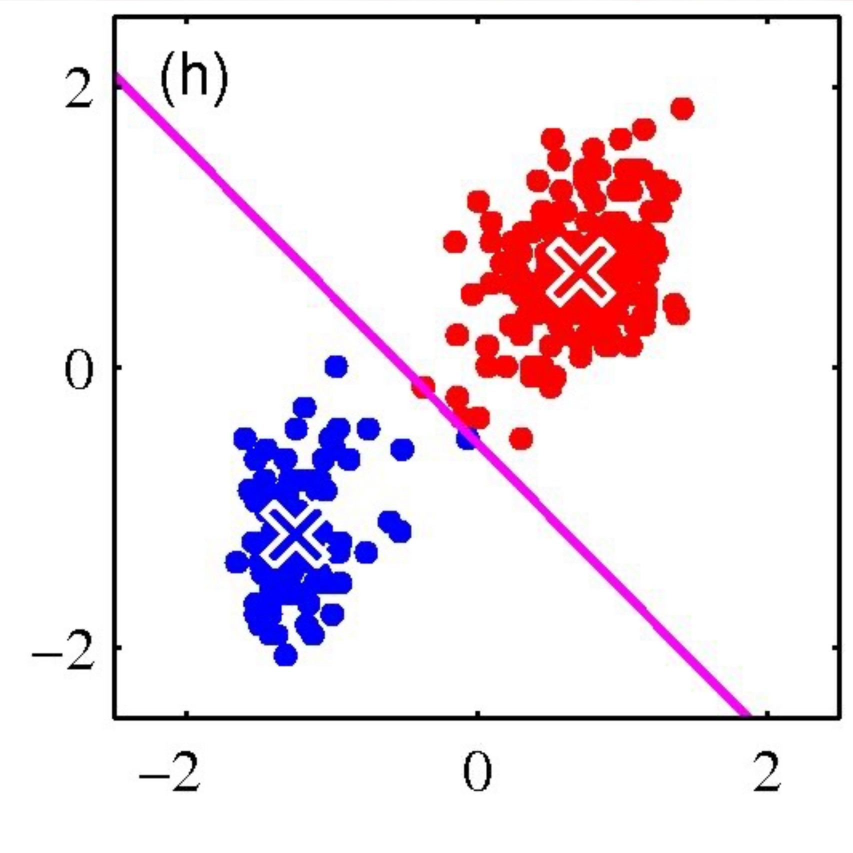
The k-Means Algorithm ($k = 2$)



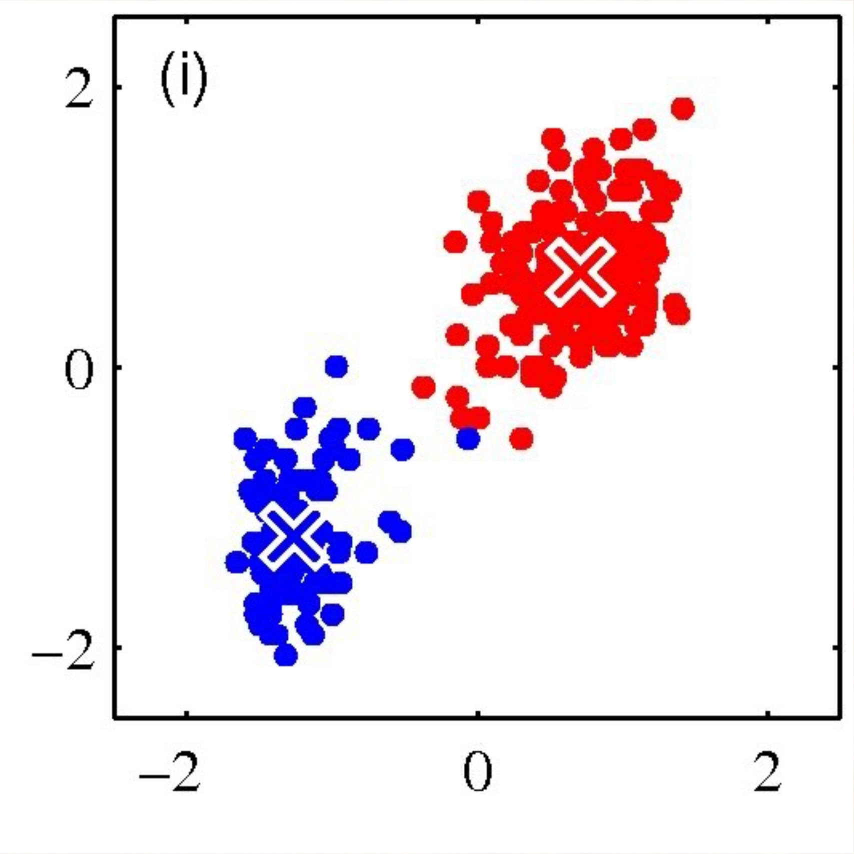
The k-Means Algorithm ($k = 2$)



The k-Means Algorithm ($k = 2$)



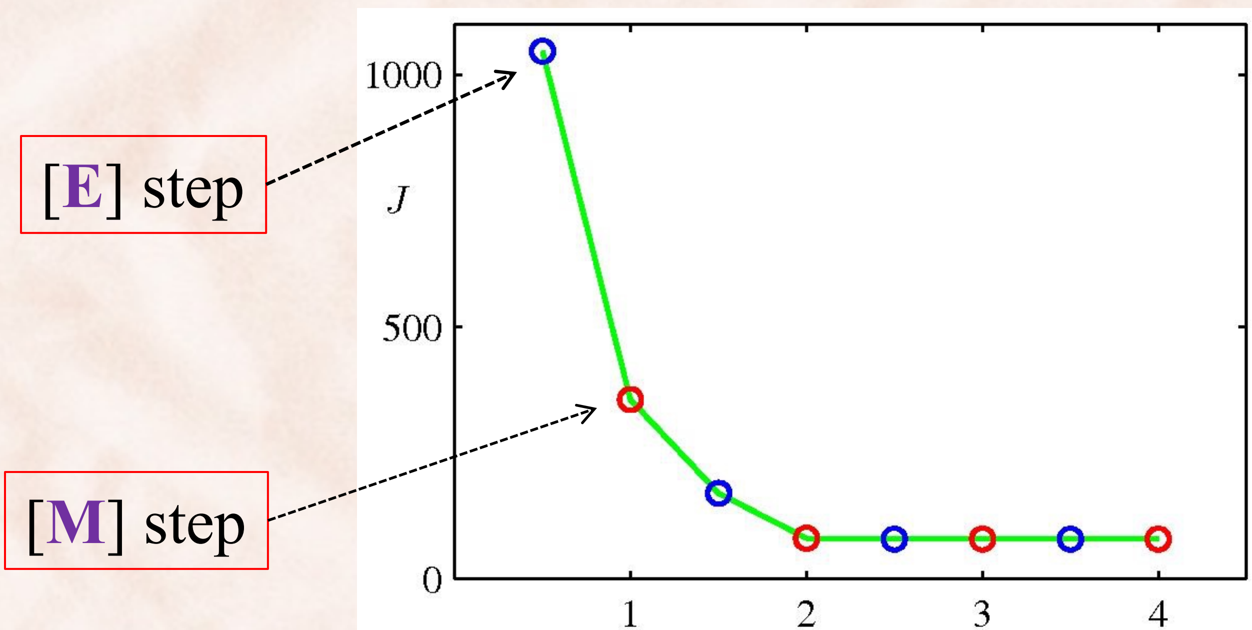
The k-Means Algorithm ($k = 2$)



The *k*-Means Algorithm

- The objective function monotonically decreases at every iteration:

$$J^{(t)} \geq J^{(t+1)}$$



The *k*-Means Algorithm

- Optimization problem is NP-hard:
  - Results depend on seed selection.
  - Improve performance by providing *must-link* and/or *cannot-link* constraints ⇒ semi-supervised clustering.

- Time complexity for each iteration is O(*knm*):
  - number of clusters is *k*.
  - feature vectors have dimensionality *m*.
  - total number of instances is *n*.

The k-Means Algorithm

1. start with some seed centroids $\mathbf{m}_1^{(0)}, \mathbf{m}_2^{(0)}, \ldots, \mathbf{m}_k^{(0)}$
2. **set** $t \leftarrow 0$.
3. **while** not converged:
4. **for** each $\mathbf{x}$:
5. **set** $\mathbf{m}^{(t)}(\mathbf{x}) \leftarrow \arg\min_{\mathbf{m}_i^{(t)}} \left\| \mathbf{x} - \mathbf{m}_i^{(t)} \right\|$ ← [**E**] step
6. **set** $C_i^{(t+1)} \leftarrow \left\{ \mathbf{x} \mid \mathbf{m}^{(t)}(\mathbf{x}) = \mathbf{m}_i^{(t)} \right\}$
7. **set** $\mathbf{m}_i^{(t+1)} \leftarrow \frac{1}{\left|C_i^{(t+1)}\right|} \sum_{\mathbf{x} \in C_i^{(t+1)}} \mathbf{x}$ ← [**M**] step
8. **set** $t \leftarrow t + 1$

The k-Medoids Algorithm

1. start with some random seed centroids $\mathbf{m}_1^{(0)}, \mathbf{m}_2^{(0)}, \ldots, \mathbf{m}_k^{(0)}$
2. **set** $t \leftarrow 0$.
3. **while** not converged:
4. **for** each $\mathbf{x}$:
5. **set** $\mathbf{m}^{(t)}(\mathbf{x}) \leftarrow \arg\min_{\mathbf{m}_i^{(t)}} d\left(\mathbf{x} - \mathbf{m}_i^{(t)}\right)$ ←---------- [E] step
6. **set** $C_i^{(t+1)} \leftarrow \left\{ \mathbf{x} \mid \mathbf{m}^{(t)}(\mathbf{x}) = \mathbf{m}_i^{(t)} \right\}$
7. **set** $\mathbf{m}_i^{(t+1)} \leftarrow \arg\min_{\mathbf{x} \in C_i^{(t+1)}} \sum_{\mathbf{y} \in C_i^{(t+1)}} d(\mathbf{x}, \mathbf{y})$ ←---------- [M] step
8. **set** $t \leftarrow t + 1$