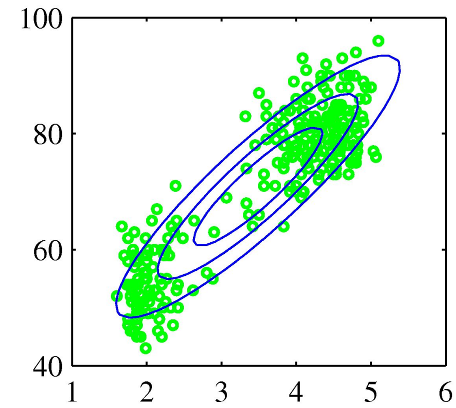
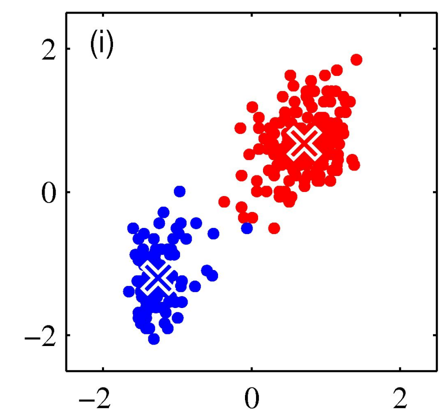
K-Means Clustering

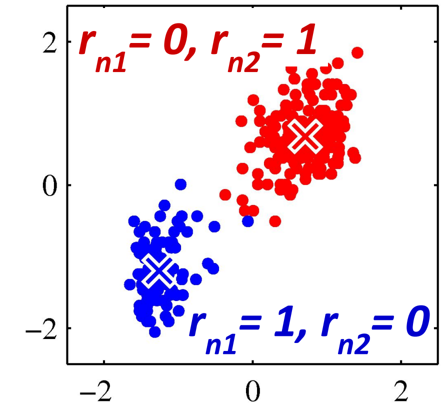
The K-Means Algorithm

- Given ***unlabeled*** data $\{\mathbf{x}^{(n)}\}$ $(n = 1, \ldots, N)$,
- and believing it belongs in K clusters (say $K = 2$ here),
- How do we find the clusters?
  - What would be the objective function?





The K-Means Algorithm



- Use indicator variables $r_{nk} \in \{0,1\}$:
  - $r_{nk} = 1$ if $\mathbf{x}^{(n)}$ is in cluster k
  - and $r_{nk} = 0$ for all $j \neq k$
- Find cluster centers μ_k and assignments r_{nk} to minimize the distortion measure J:
  - Sum of squared distance of points from the center of its own cluster (*Intra-cluster variation*):

$$J = \sum_{k=1}^{K} \sum_{n=1}^{N} r_{nk} \|\mathbf{x}^{(n)} - \mu_k\|^2 \qquad \mu_k = \frac{1}{N_k} \textstyle\sum_{n:\, \mathbf{x}^{(n)} \in \text{cluster } k} \mathbf{x}^{(n)} = \frac{\sum_{n=1}^{N} r_{nk} \mathbf{x}^{(n)}}{\sum_{n=1}^{N} r_{nk}}$$

The K-Means Algorithm

- Initialize the cluster centers (centroids)
- Repeat the following update until convergence:

1. $r := \arg\min_r J(r, \mu)$
2. $\mu := \arg\min_\mu J(r, \mu)$

where $$J = \sum_{k=1}^{K} \sum_{n=1}^{N} r_{nk} \|\mathbf{x}^{(n)} - \mu_k\|^2$$

$$\mu_k = \frac{1}{N_k} \sum\nolimits_{n:\, \mathbf{x}^{(n)} \in \text{cluster } k} \mathbf{x}^{(n)} = \frac{\sum_{n=1}^{N} r_{nk} \mathbf{x}^{(n)}}{\sum_{n=1}^{N} r_{nk}}$$

The K-Means Algorithm

Note: E, M stands for:

- E: Expectation
- M: Maximization

(We will revisit EM later.)

- Initialize the cluster centers.
- Repeat until convergence:
  - **Cluster assignment ("E-Step"):** assign each point to closest center.

$$r_{nk} = \begin{cases} 1 & \text{if } k = \arg\min_j \|\mathbf{x}^{(n)} - \mu_j\|^2 \\ 0 & \text{otherwise} \end{cases}$$

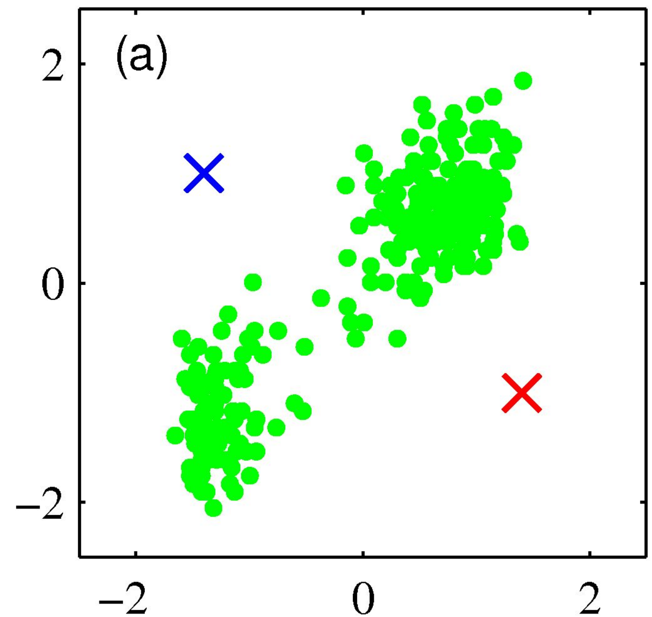
  - **Parameter update ("M-Step"):** update the centers

$$\mu_k = \frac{\sum_n r_{nk}\mathbf{x}^{(n)}}{\sum_n r_{nk}}$$

Q. Verify this

K-Means Clustering

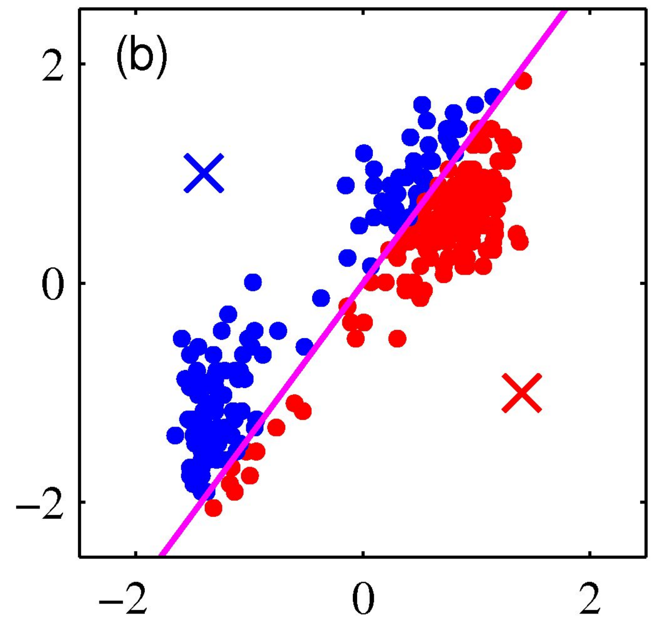
- Select K. Pick random centroids.
  - Here K=2.



K-Means Clustering

Cluster assignment Step ("E-Step")

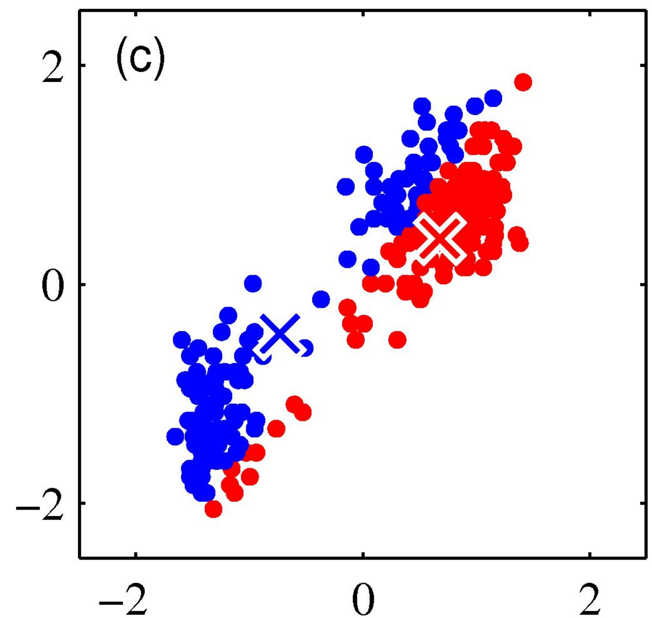
- Assign each point to the nearest center.



K-Means Clustering

Update parameters (centroids) ("M-Step")

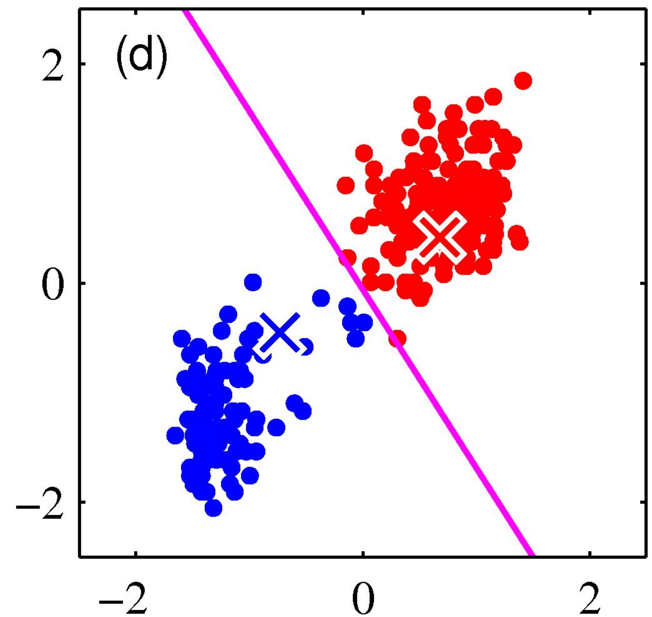
- Compute new centers for each cluster.



K-Means Clustering

Cluster assignment Step (“E-Step”) again

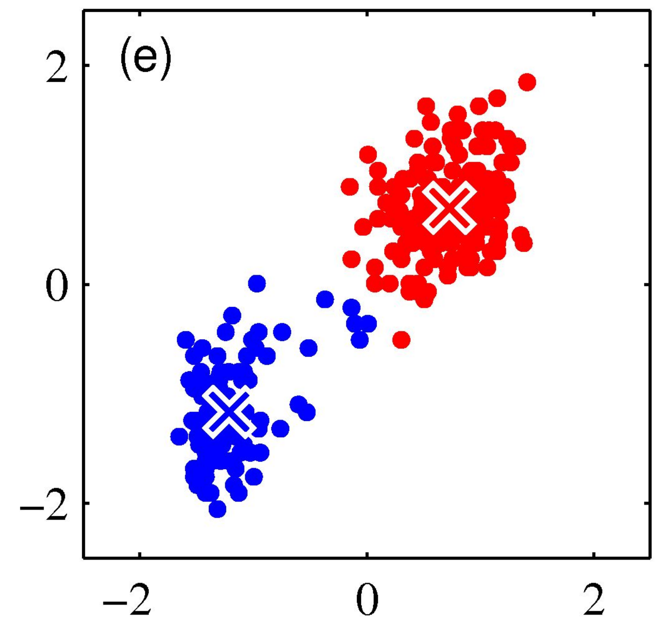
- Re-assign points to the now-nearest center.



K-Means Clustering

Update parameters (centroids) (“M-Step”) again

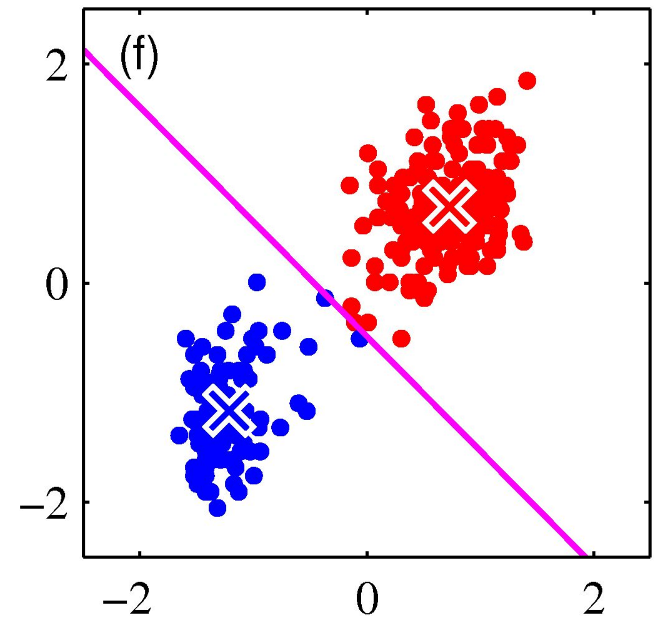
- Compute centers for the new clusters.



K-Means Clustering

Another Cluster assignment Step (“E-Step”)

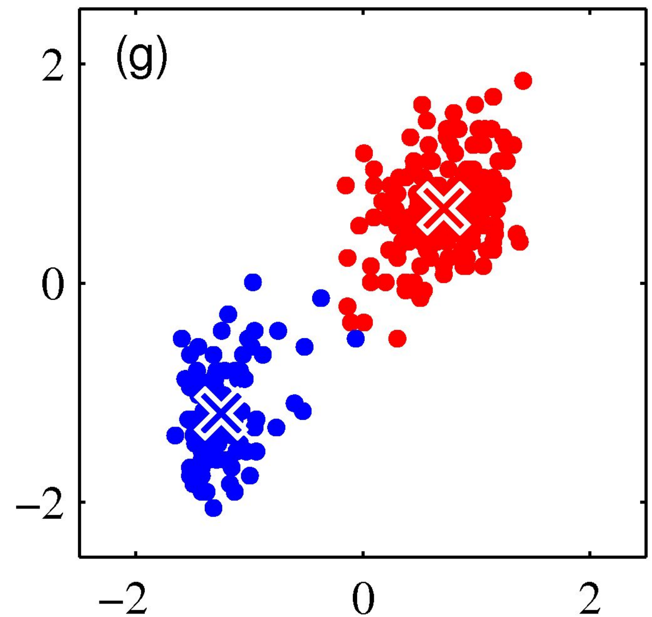
- Reassign the points to centers.



K-Means Clustering

Update parameters (centroids) (“M-Step”) again

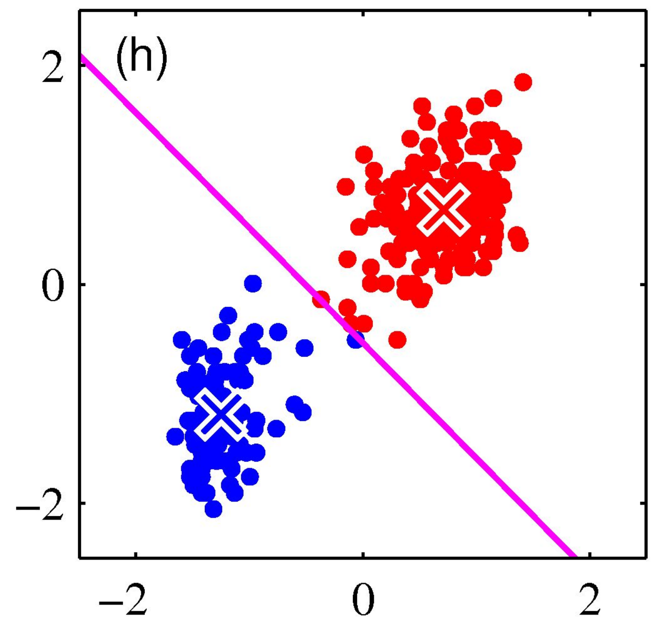
- New centers.



K-Means Clustering

Another Cluster assignment Step ("E-Step")

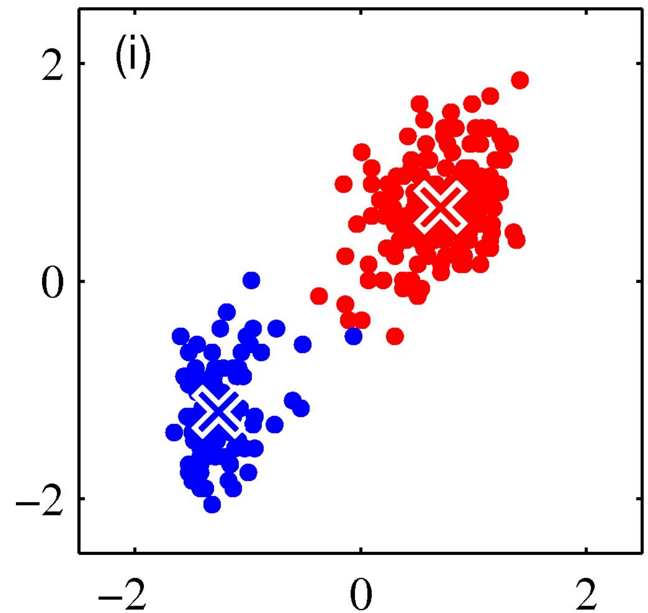
- New cluster assignments.



K-Means Clustering

Update parameters (centroids) (“M-Step”) again

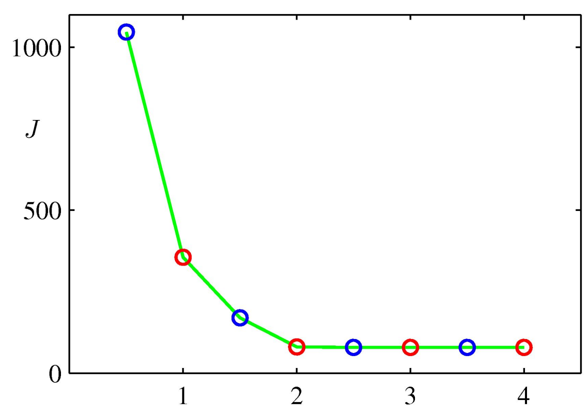
- The cluster centers have stopped changing.



Convergence

- The objective function of K-means decreases monotonically as the K-means procedure reduces J in both E-step and M-step.

- Convergence is relatively quick, in steps.
  – blue circles after E-step: assign each point to a cluster
  – red circles after M-step: recompute the cluster centers
  – However, all those distance computations are expensive.



Convergence

- No guarantee that we found the globally optimal solution. The quality of local optimum depends on the initial values.
- The following clustering is a stable local optima