CSci5521: Machine Learning Fundamentals

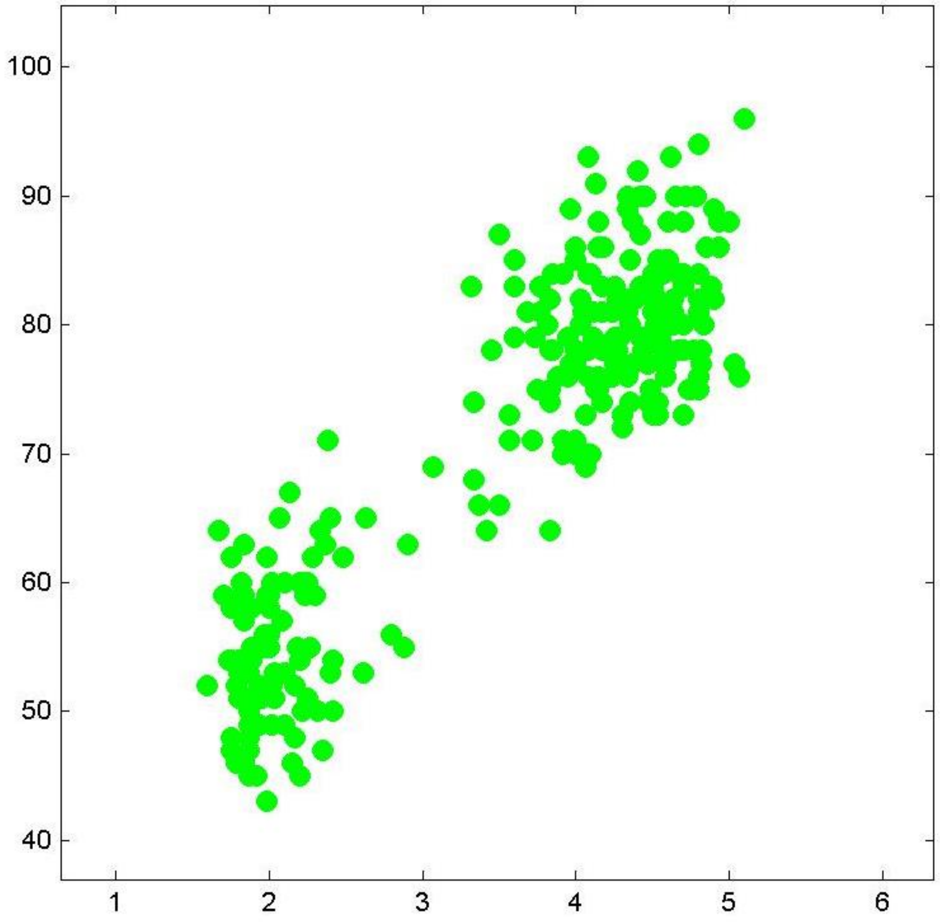
- Mixture Models and EM

Catherine Qi Zhao
Computer Science and Engineering
University of Minnesota

Old Faithful

Old Faithful Data Set

Time between eruptions (minutes)



Duration of eruption (minutes)

K-means Algorithm

- Goal: represent a data set in terms of K clusters each of which is summarized by a prototype μ_k
- Initialize prototypes, then iterate between two phases:
  - E-step: assign each data point to the nearest prototype
  - M-step: update prototypes to be the cluster means

K-Means Clustering: Distortion Measure

IFF nth data point is assigned to cluster k.

vector w/ d-dimensional

- Dataset $\{\mathbf{x}_1, \ldots, \mathbf{x}_N\}$
- Partition in K clusters

mean @ cluster k

- Cluster prototype: μ_k — $k = 1, 2, \ldots, K$ — # μs: K — # rs: $N \cdot k$
- Binary indicator variable, 1-of-K Coding scheme
  $r_{nk} \in \{0, 1\}$ — r_n k (↑ data point, ↑ cluster)
  $r_{nk} = 1$, and $r_{nj} = 0$ for $j \neq k$. o.w.
  Hard assignment. → basically only one data pt. could be assigned to one (1)
- Distortion measure — objective func.

for x_1, $(n=1)$

$r_{11} = 0$
$r_{12} = 0$
$\vdots = 1$
$r_{1K} = 0$

$$\Sigma \; r_{11}\|x_1 - \mu_1\|^2 + r_{12}\|x_1 - \mu_2\|^2 + \|x_1 - \mu_k\|^2 + r_{1K}\|x_1 - \mu_K\|^2$$

loop through all n's.

$$J = \sum_{n=1}^{N} \boxed{\sum_{k=1}^{K} r_{nk} \|\mathbf{x}_n - \mu_k\|^2} \qquad (9.1)$$

$\in \{0, 1\}$

$\|x_1 - \mu_k\|^2$

k. reduce to a single n @ k

L2 Norm

K-Means Clustering: Expectation Maximization

- Find values for $\{r_{nk}\}$ and $\{\mu_k\}$ to minimize:

naturally goes away in M step

$$J = \sum_{n=1}^{N} \sum_{k=1}^{K} r_{nk} \|\mathbf{x}_n - \mu_k\|^2 \qquad (9.1)$$

only 1 can be 1

for n = 1

$\|x_1 - \mu_1\|^2$

$\|x_1 - \mu_2\|^2$

$\vdots$

$\|x_1 - \mu_k\|^2$

basically we want a low J value

- Iterative procedure:
  1. Minimize J w.r.t. r_{nk}, keep μ_k fixed (Expectation)

$$r_{nk} = \begin{cases} 1 & \text{if } k = \arg\min_j \|\mathbf{x}_n - \mu_j\|^2 \\ 0 & \text{otherwise} \end{cases} \qquad (9.2)$$

  2. Minimize J w.r.t. μ_k, keep r_{nk} fixed (Maximization)

pick the one with min $\|x_1 - \mu_k\|^2$

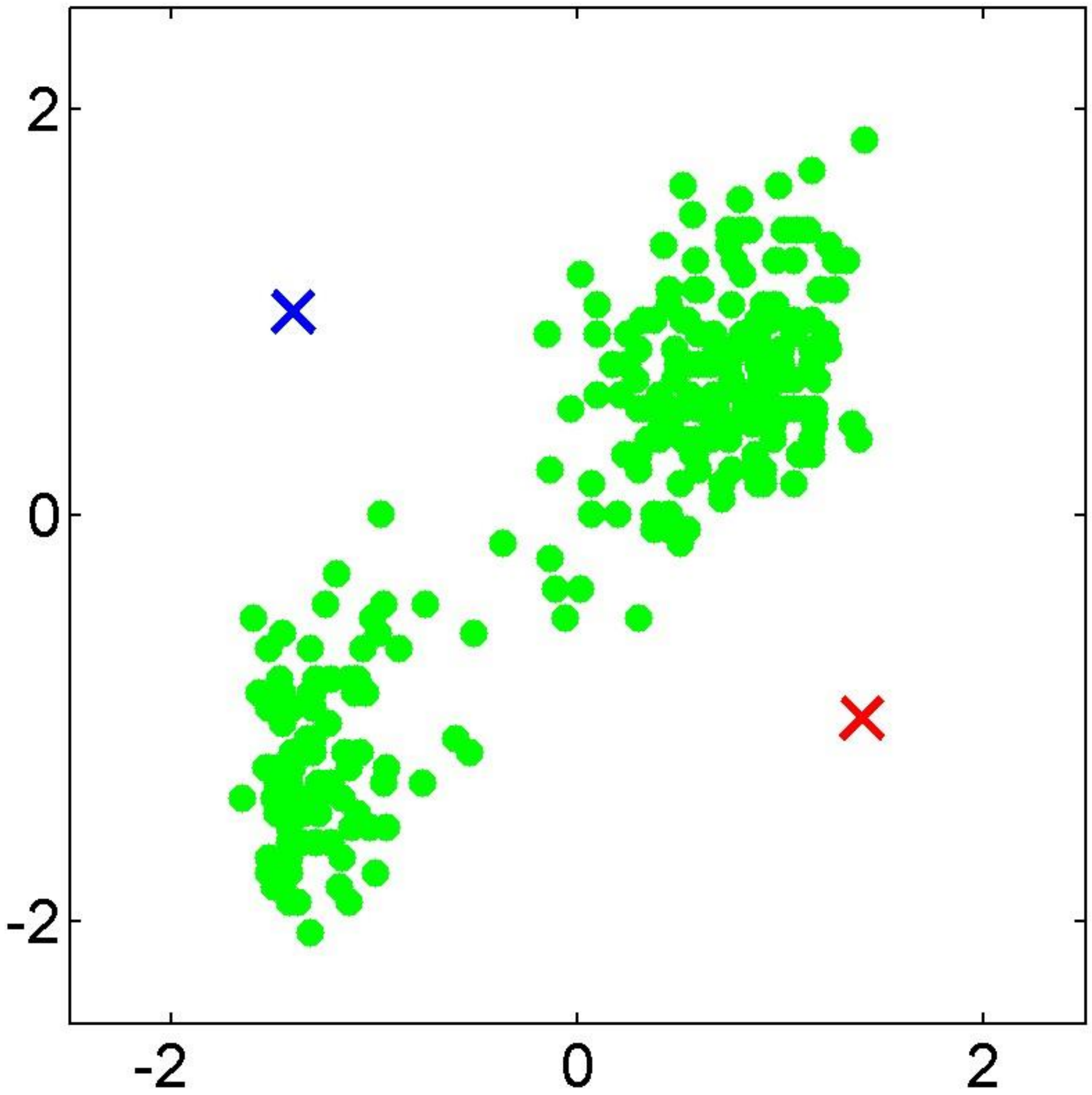
distance

$\frac{\partial J}{\partial \mu}$

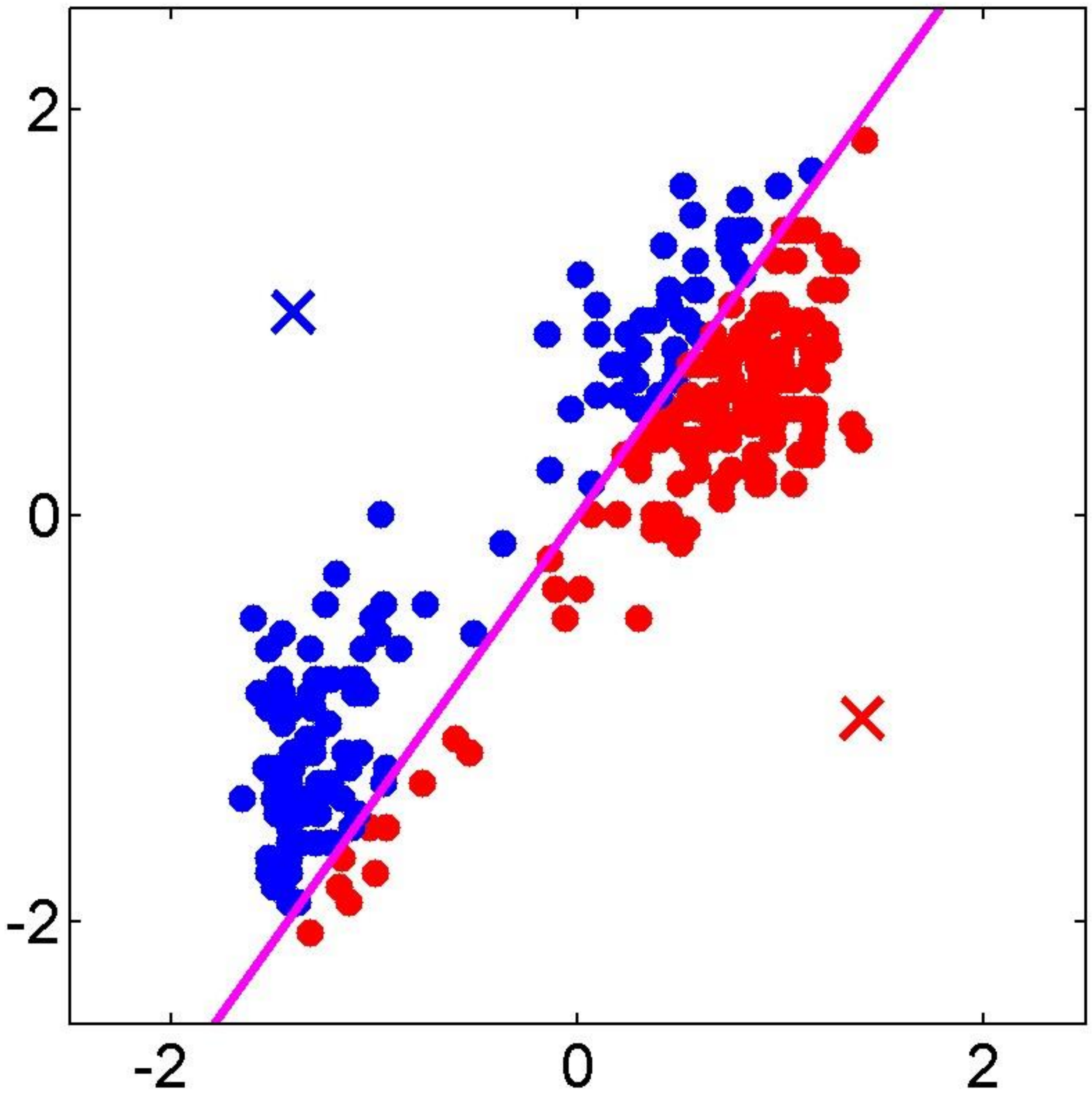
$$2 \sum_{n=1}^{N} r_{nk} (\mathbf{x}_n - \mu_k) = 0 \qquad (9.3)$$

$$\mu_k = \frac{\sum_n r_{nk} \mathbf{x}_n}{\sum_n r_{nk}} \qquad (9.4)$$

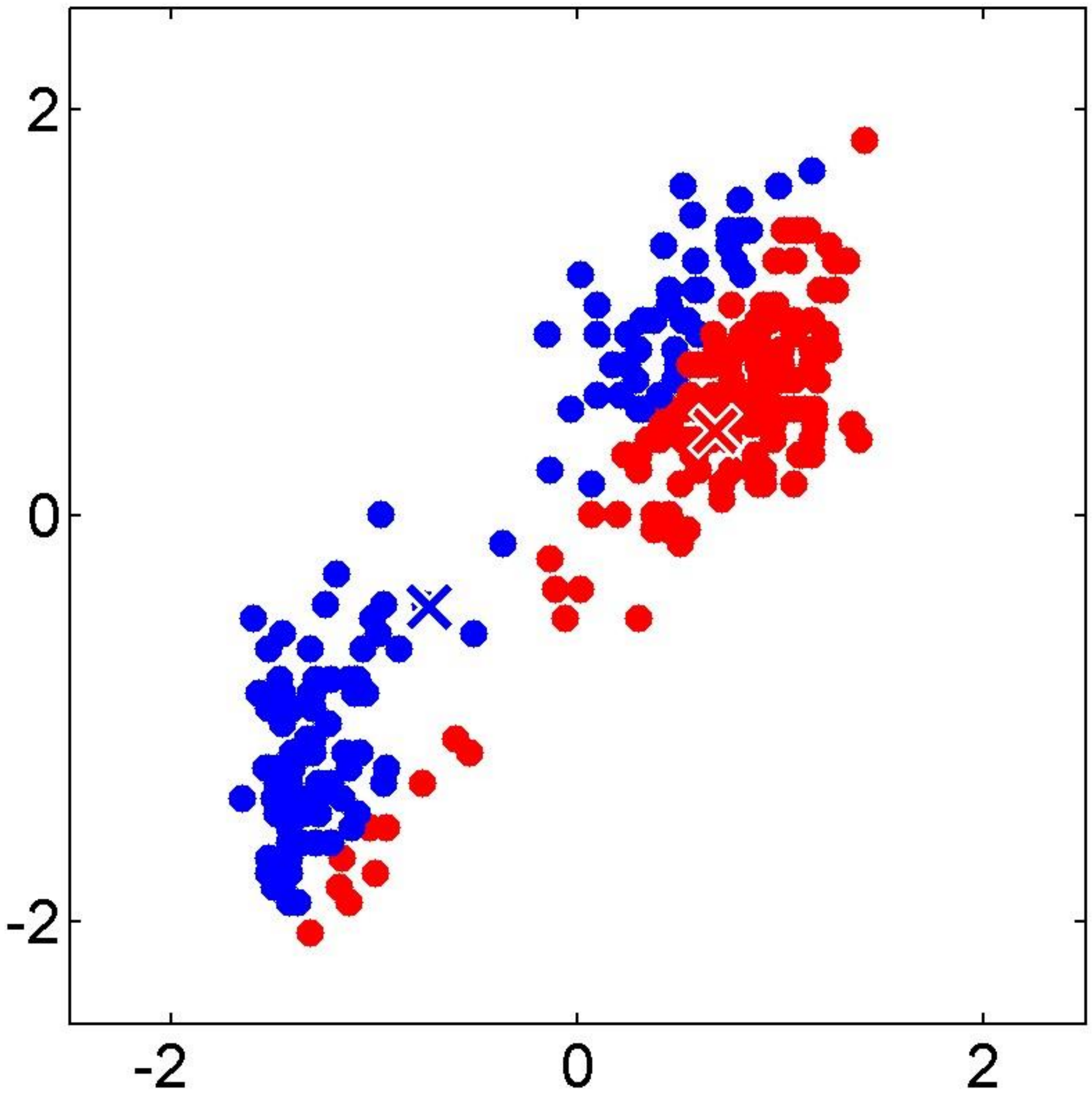
sample mean can be interpreted as a label

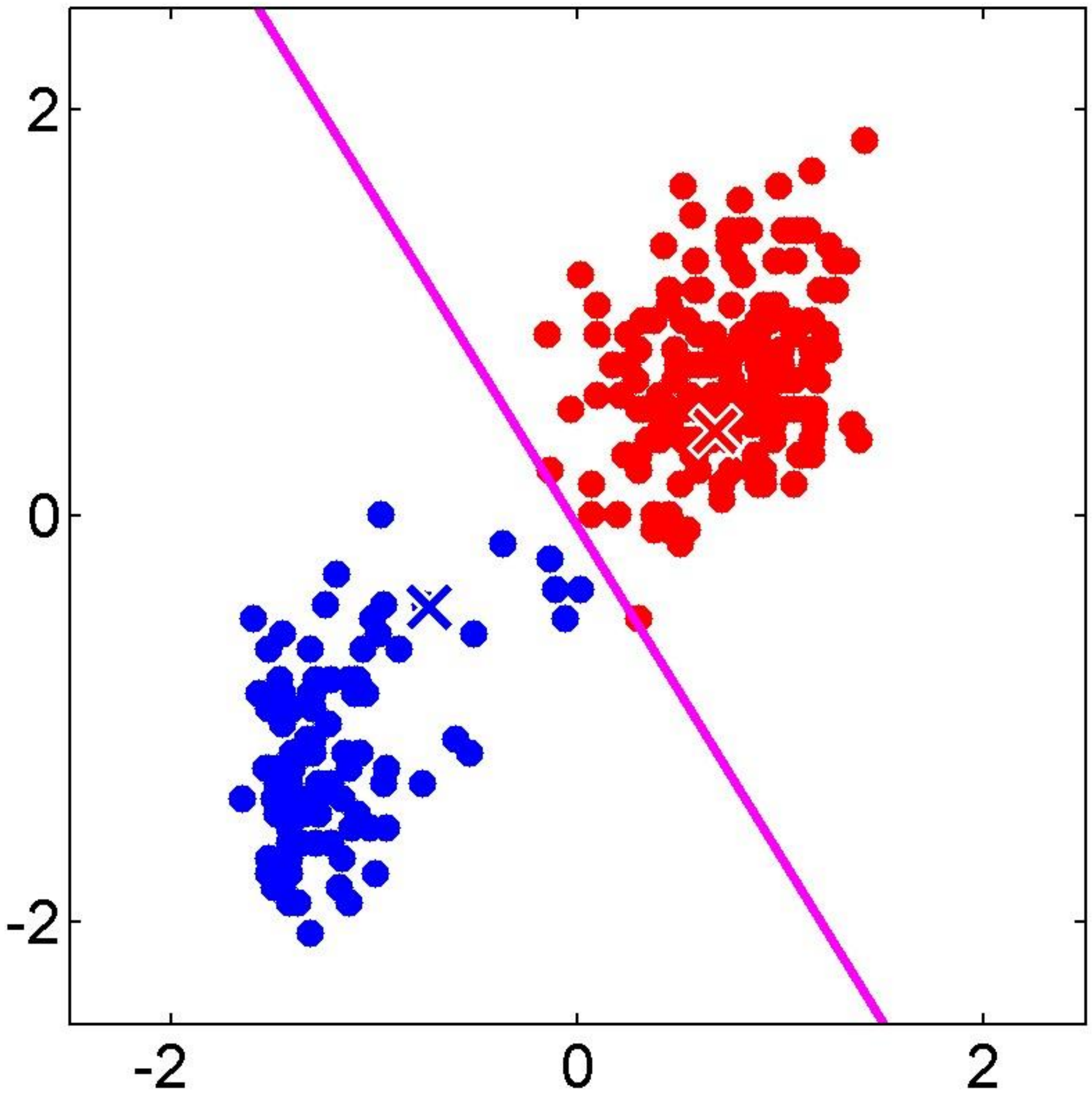

2
0
-2
-2
0
2

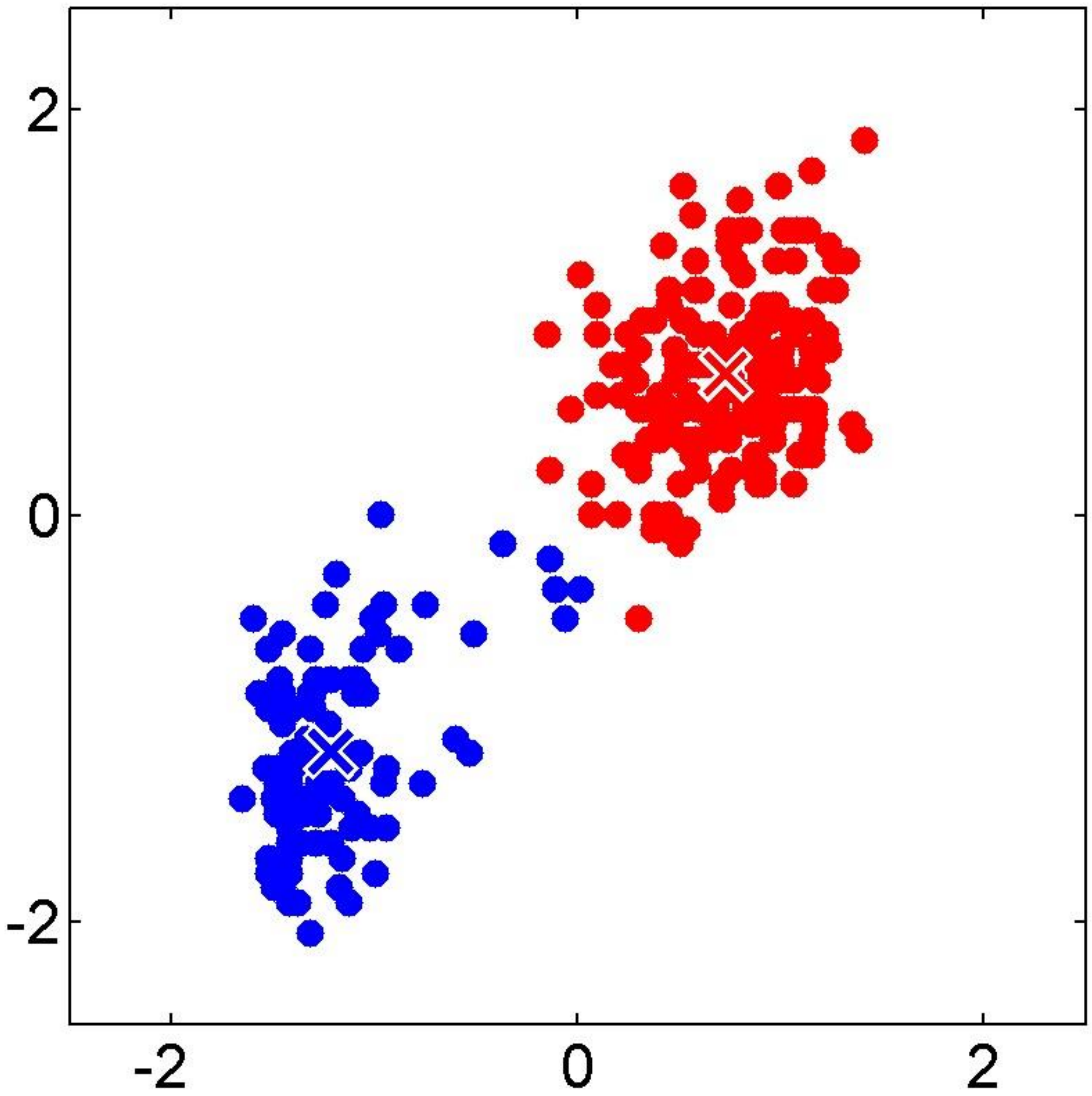
after E step

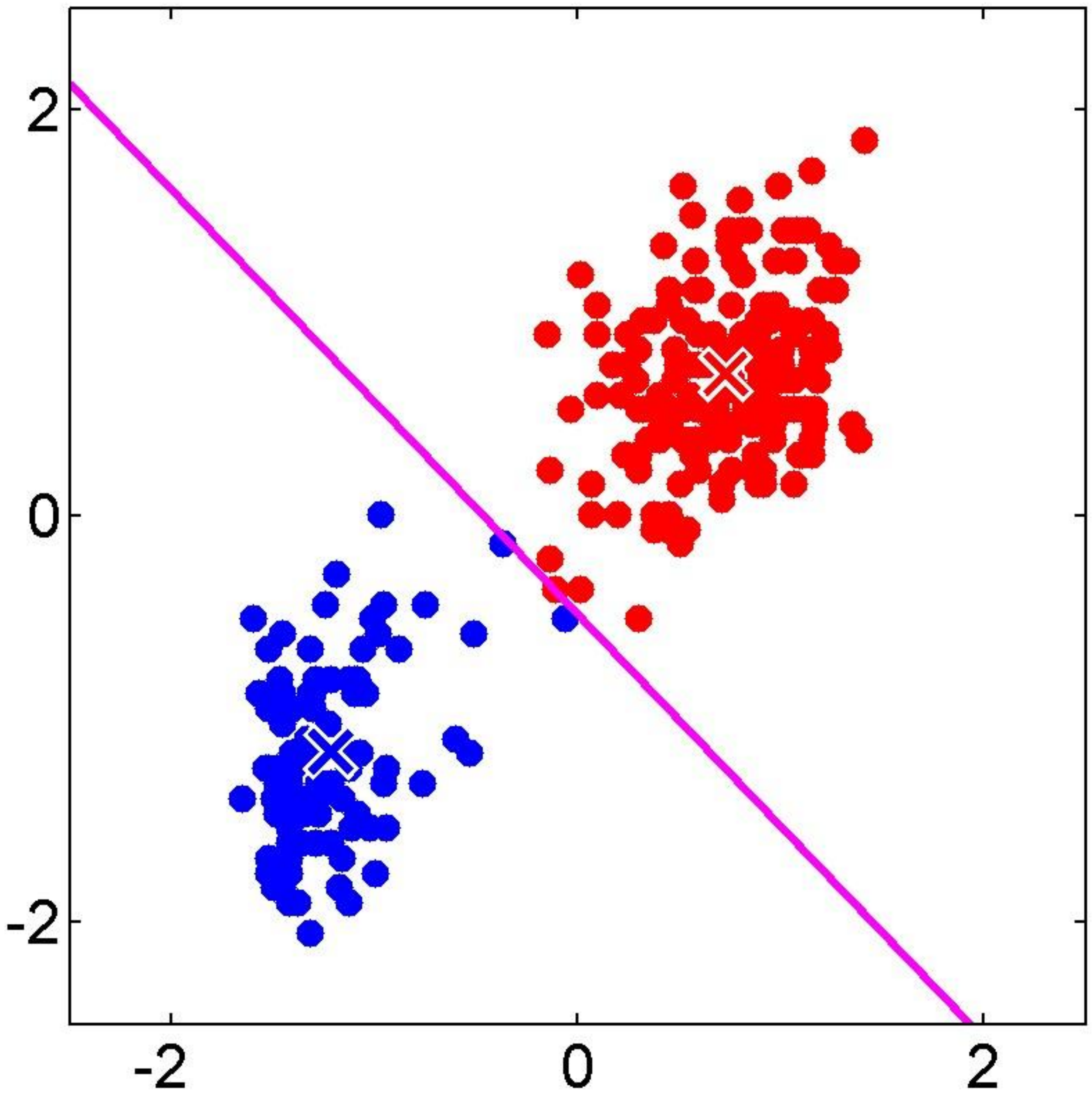


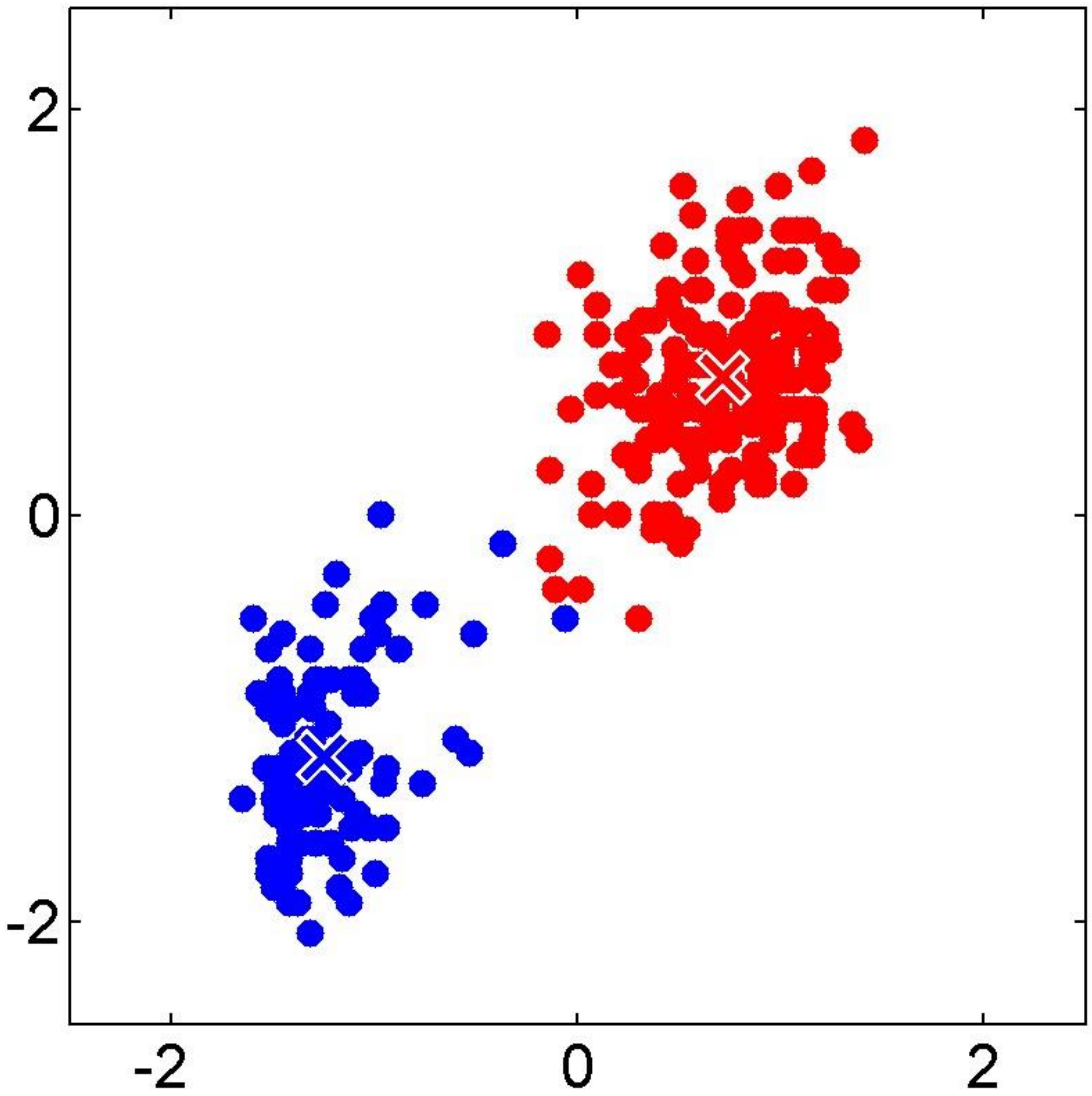
after M

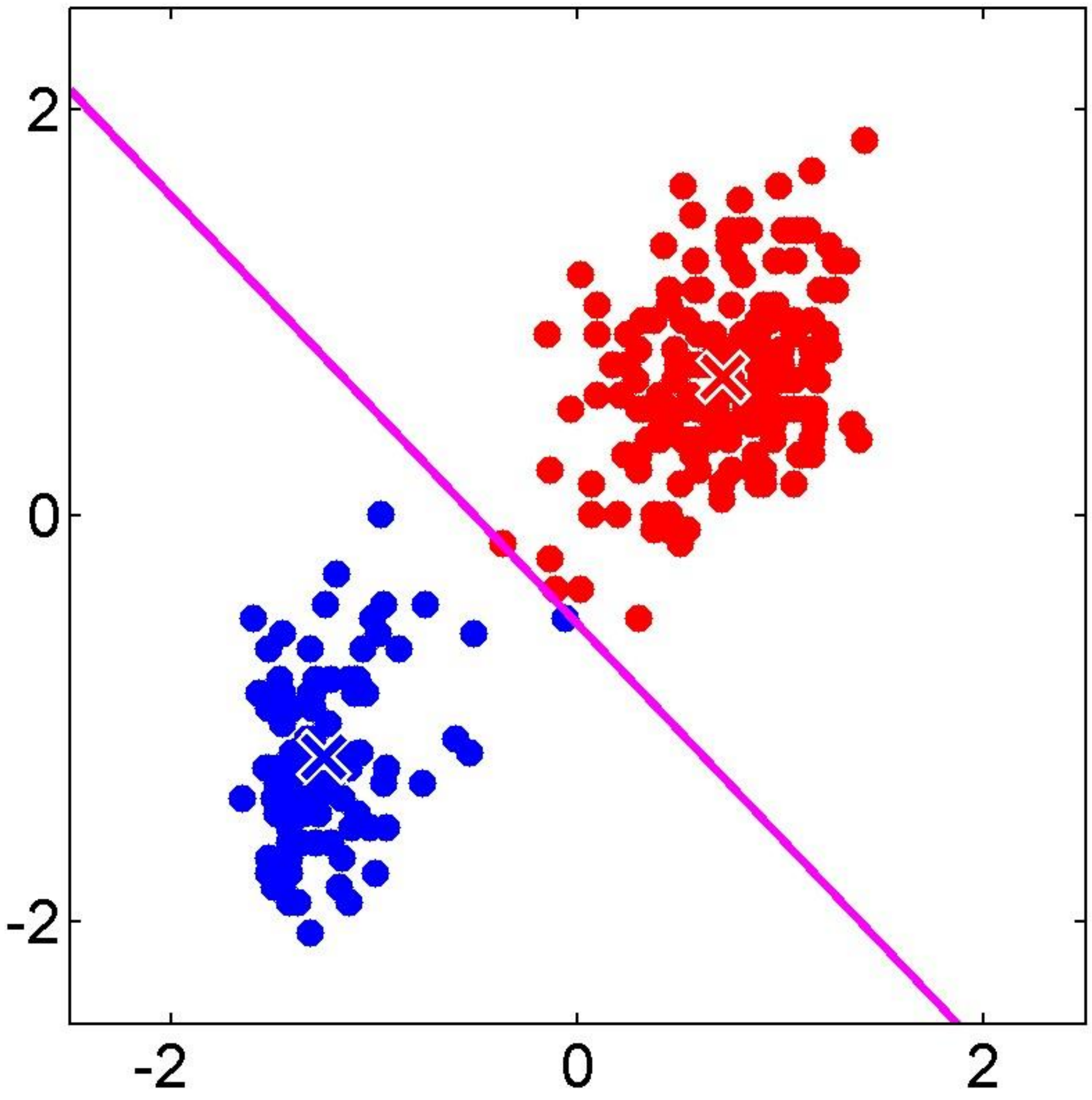



2
0
-2
-2
0
2

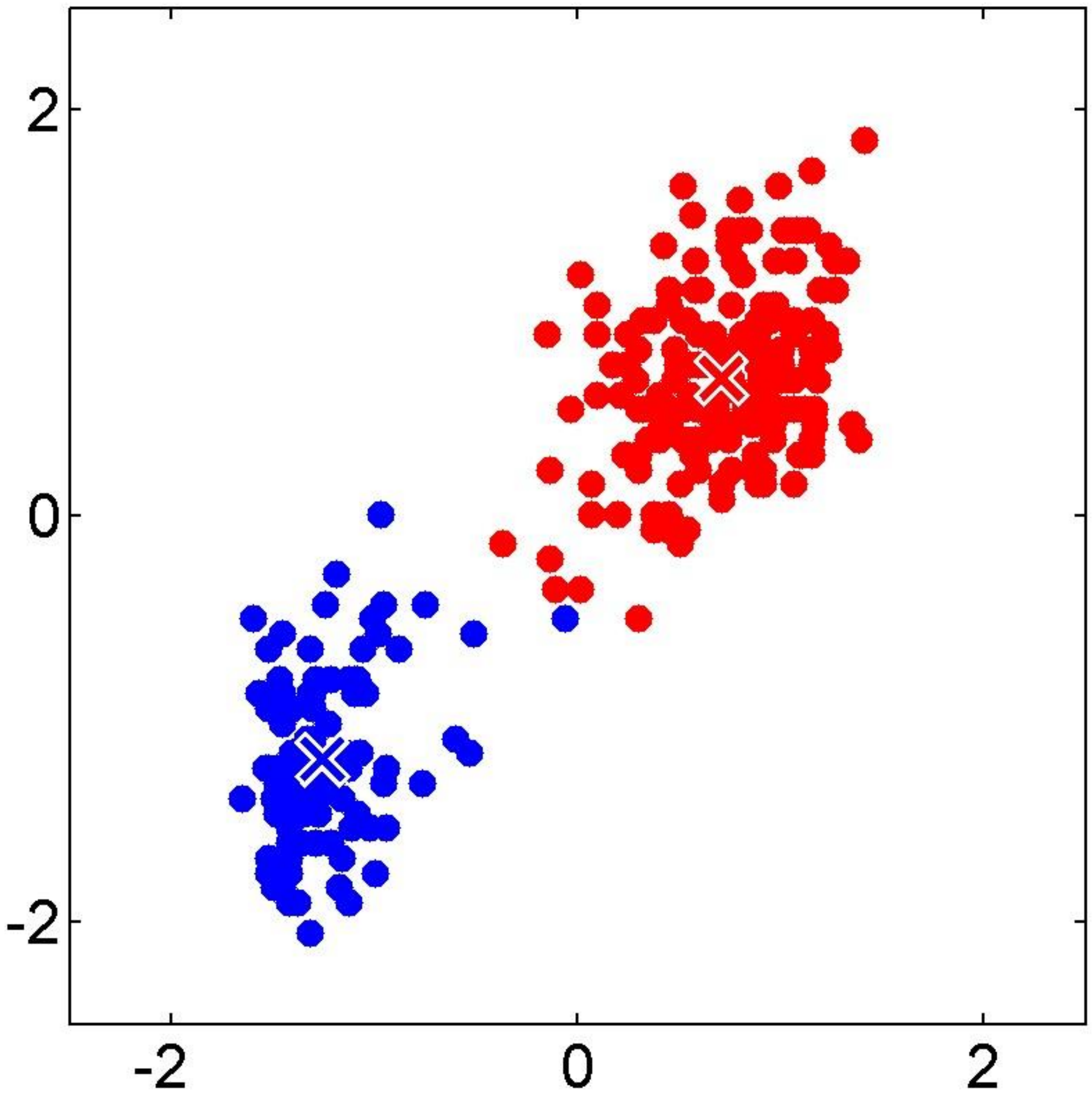

2
0
-2
-2
0
2


2
0
-2
-2
0
2


2
0
-2
-2
0
2


2
0
-2
-2
0
2

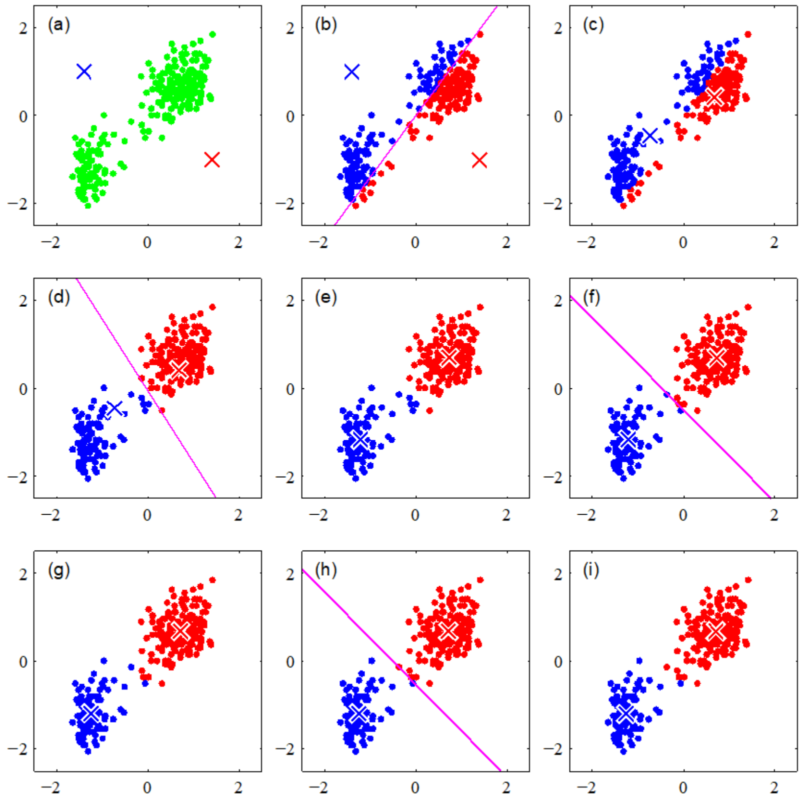
M

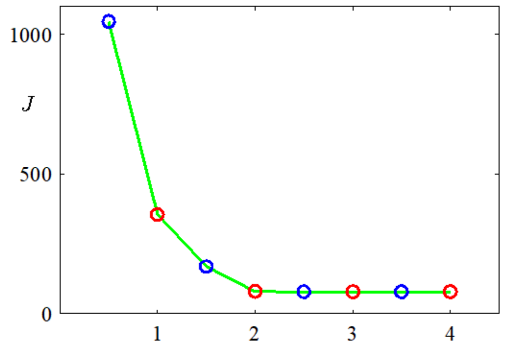


in other words, when do we stop?

K

K-Means Clustering: Example





- Each E or M step reduces the value of the objective function J
- Convergence to a global or local maximum

K-Means Clustering: Concluding remarks

1. Direct implementation of K-Means can be slow
2. Online version:

$$\mu_k^{\text{new}} = \mu_k^{\text{old}} + \eta_n(\mathbf{x}_n - \mu_k^{\text{old}}) \tag{9.5}$$

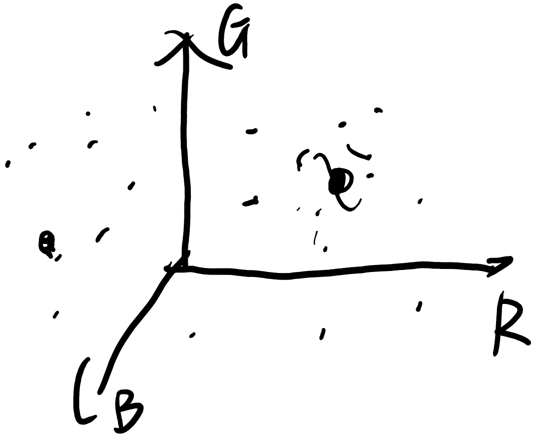
x_n

3. K-mediods, general distortion measure

$$\tilde{J} = \sum_{n=1}^{N} \sum_{k=1}^{K} r_{nk} \mathcal{V}(\mathbf{x}_n, \mu_k) \tag{9.6}$$

where $\mathcal{V}(\cdot, \cdot)$ is any kind of dissimilarity measure

4. Image segmentation and compression example: RGB



4.2 % | 8.3 % | 16.7 % | 100 %

600×800