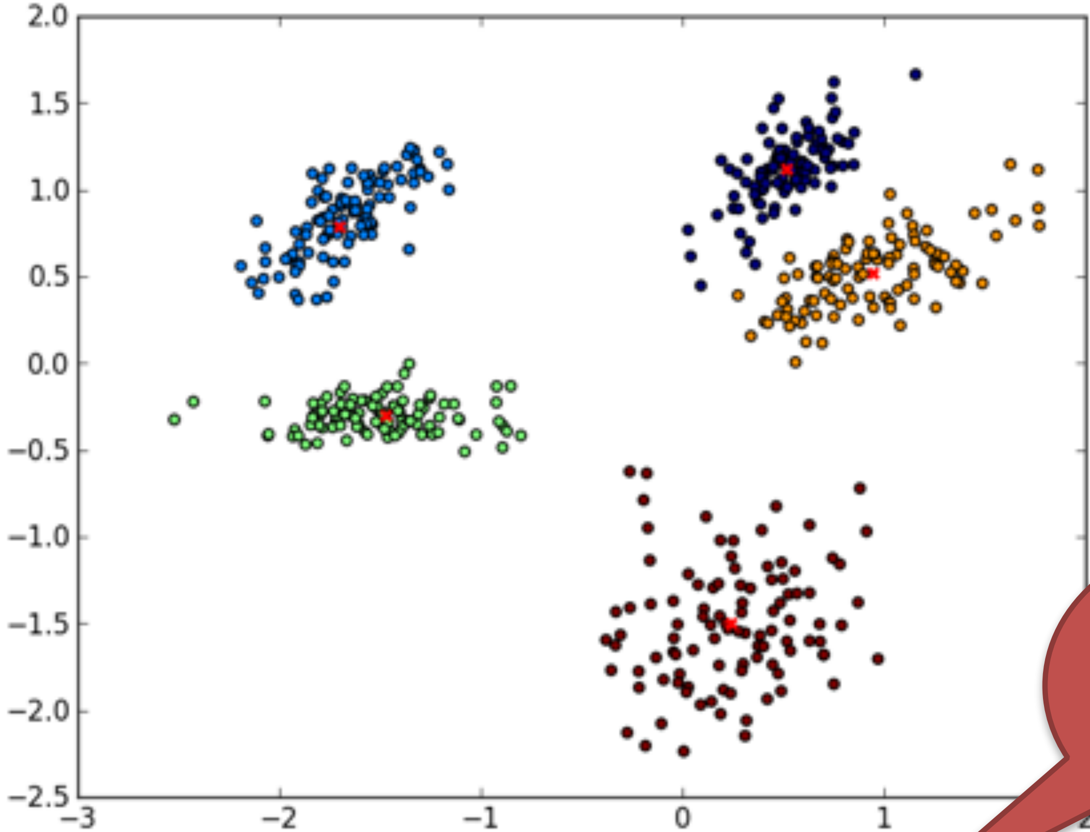
Machine Learning
Lab Course: Lecture 2

Clustering:

K-Means and Gaussian Mixture Models

Felix Brockherde <felix.brockherde@tu-berlin.de>

Clustering



What is the correct number?

Divide data points into a fixed number of disjoint sets.

K-Means Algorithm

Algorithm 6 K-means clustering

Input: data points $x_1, \ldots, x_n \in \mathbb{R}^d$, number of clusters k, maximum number of iterations m.

Output: cluster centres $\mu_1, \ldots, \mu_k \in \mathbb{R}^d$, assignment vector $r \in$

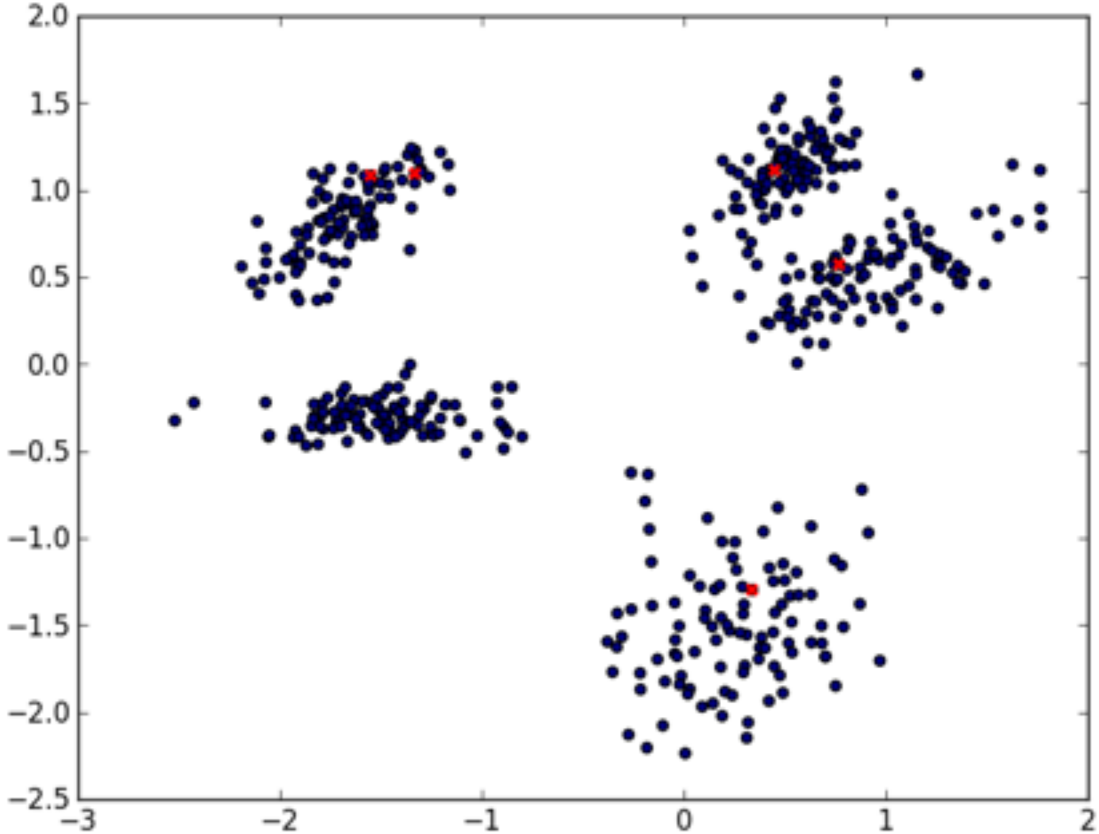
Choose random data points as initial cluster centres $\mu_1 \leftarrow x$ … where $i_j \neq i_l$ for all $j \neq l$.
$r \leftarrow \mathbf{0}_n$
$r' \leftarrow \mathbf{0}_n$
$i \leftarrow 0$
while $i < m$ **do**
  for $j \leftarrow 1$ to n **do**
    Find nearest cluster centre $r'_j \leftarrow \text{argmin}_{1 \leq l \leq k} \|x_j - \mu_l\|_2$
  end for
  for $j \leftarrow 1$ to k **do**
    Compute new cluster centre $\mu_j \leftarrow \frac{1}{|\{l : r'_l = j\}|} \sum_{l : r'_l = j} x_l$
  end for
  if $r = r'$ **then**
    break
  end if
  $r \leftarrow r'$
  $i \leftarrow i + 1$
end while

No of clusters as input

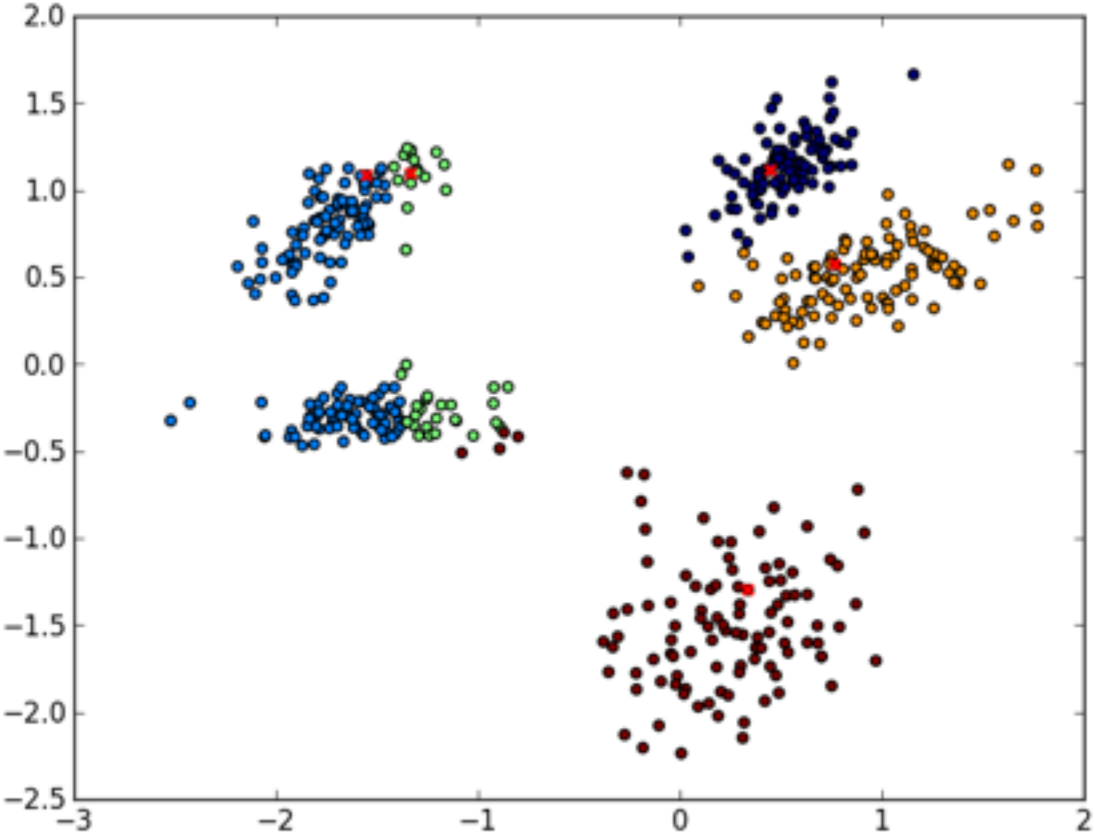
Step 1: Reassign data points to clusters

Step 2: Update cluster centers

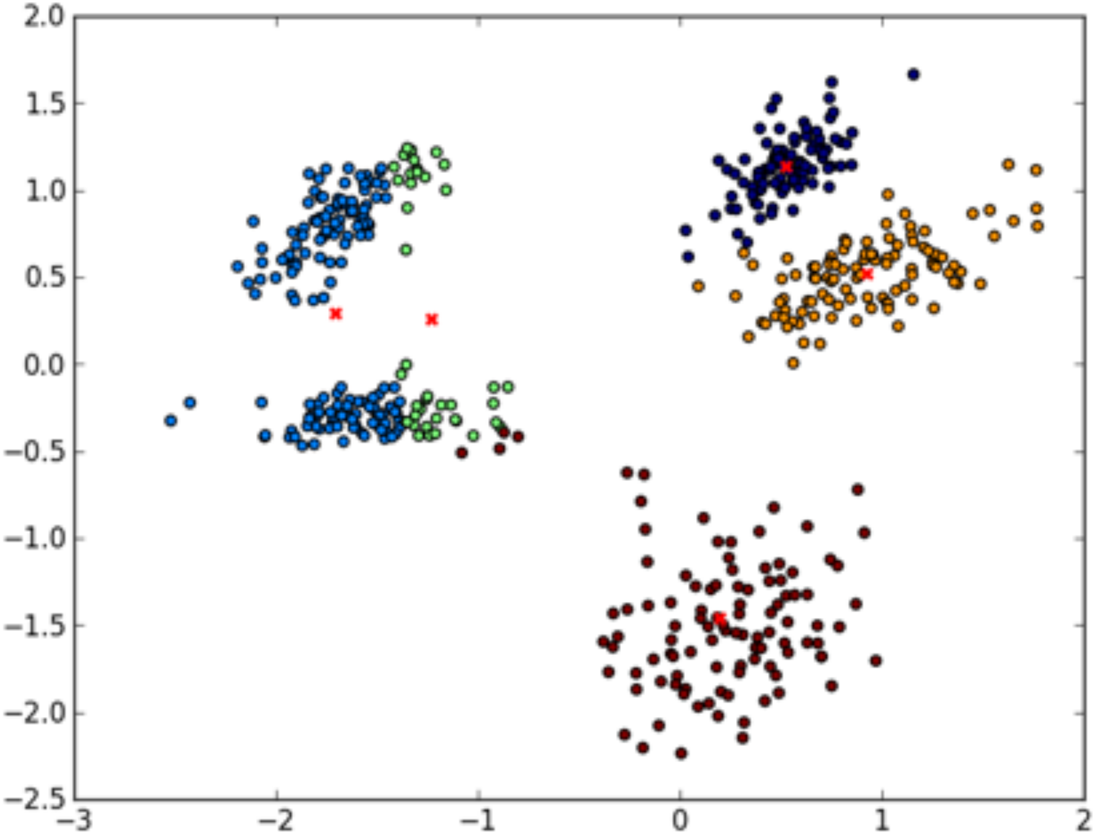
K-Means: Example



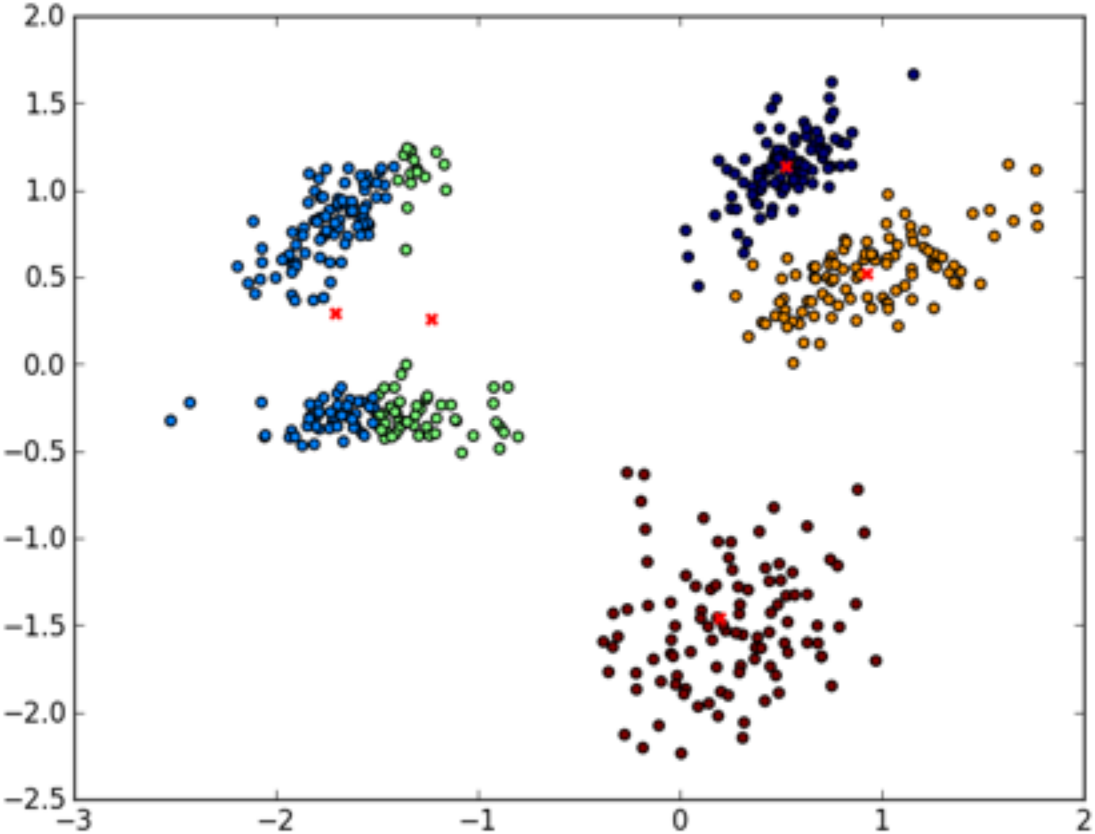
K-Means: Example



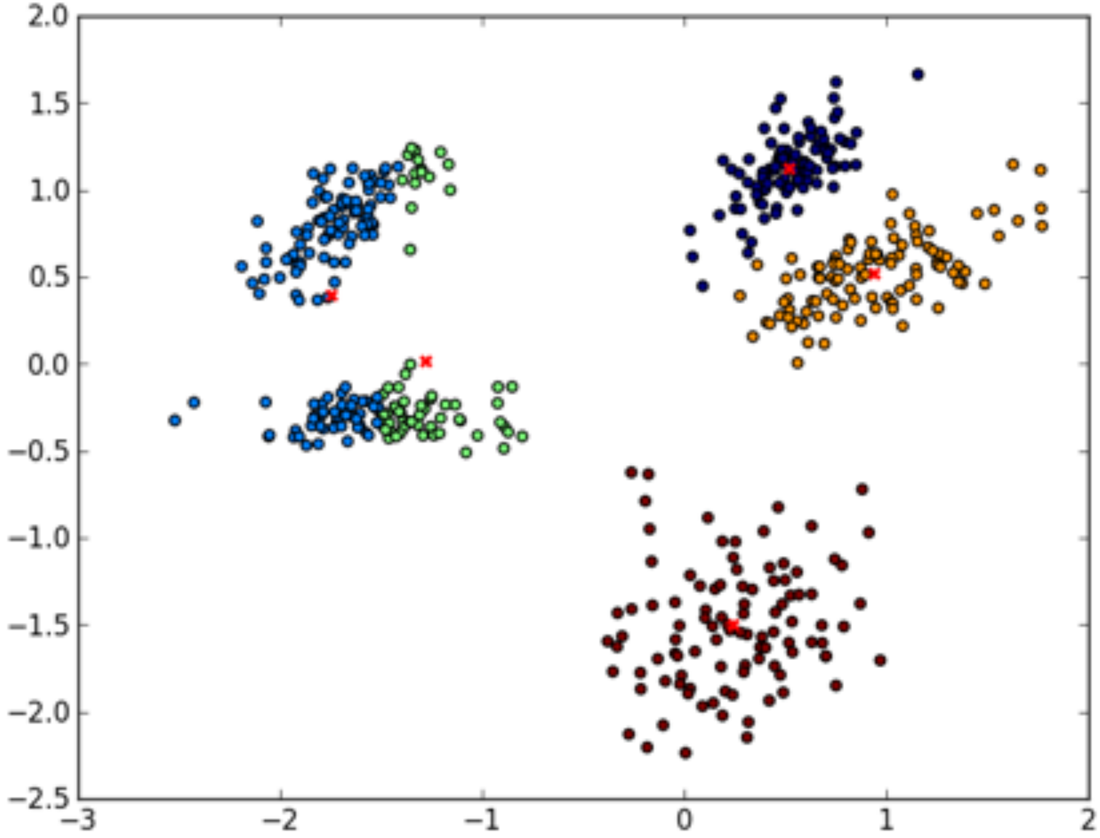
K-Means: Example



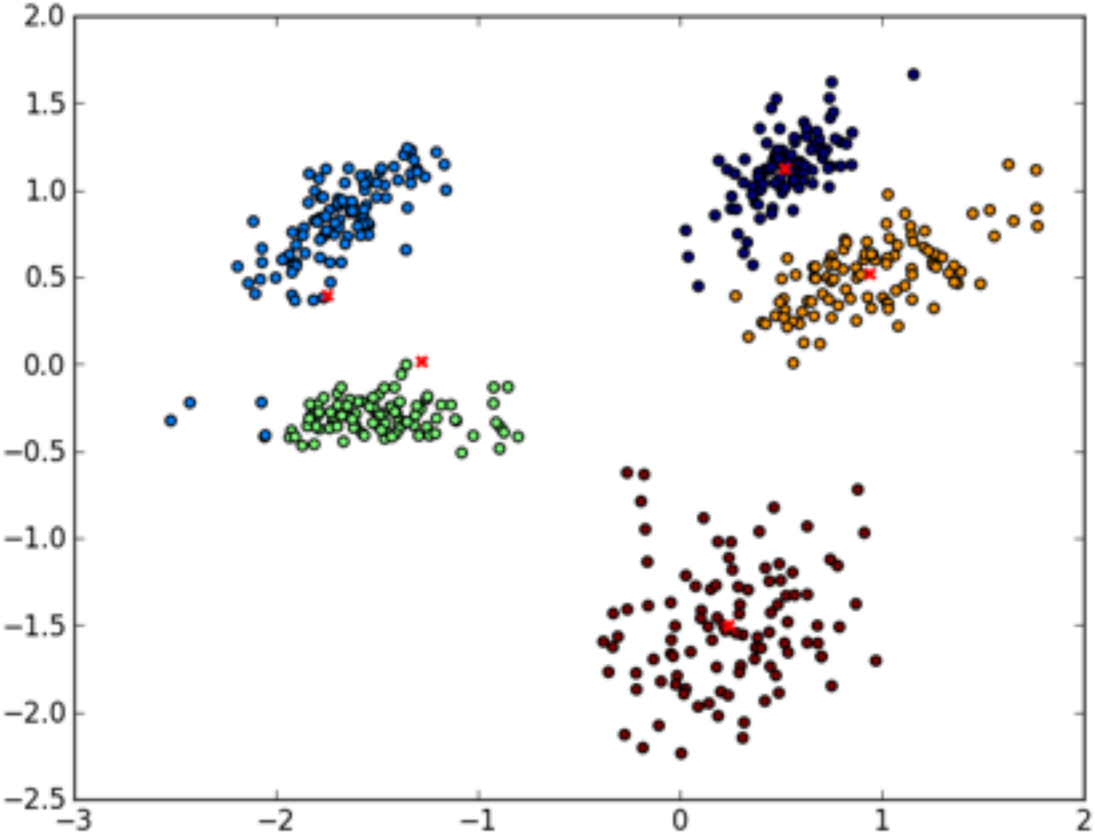
K-Means: Example



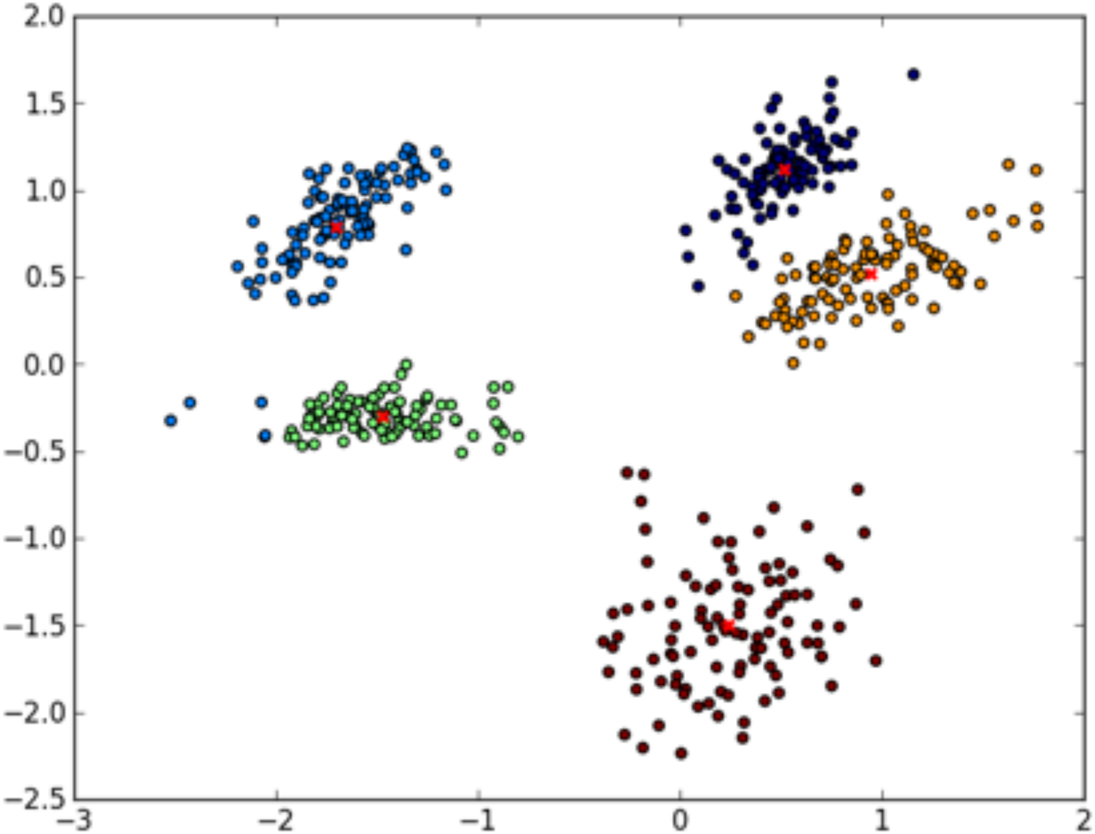
K-Means: Example



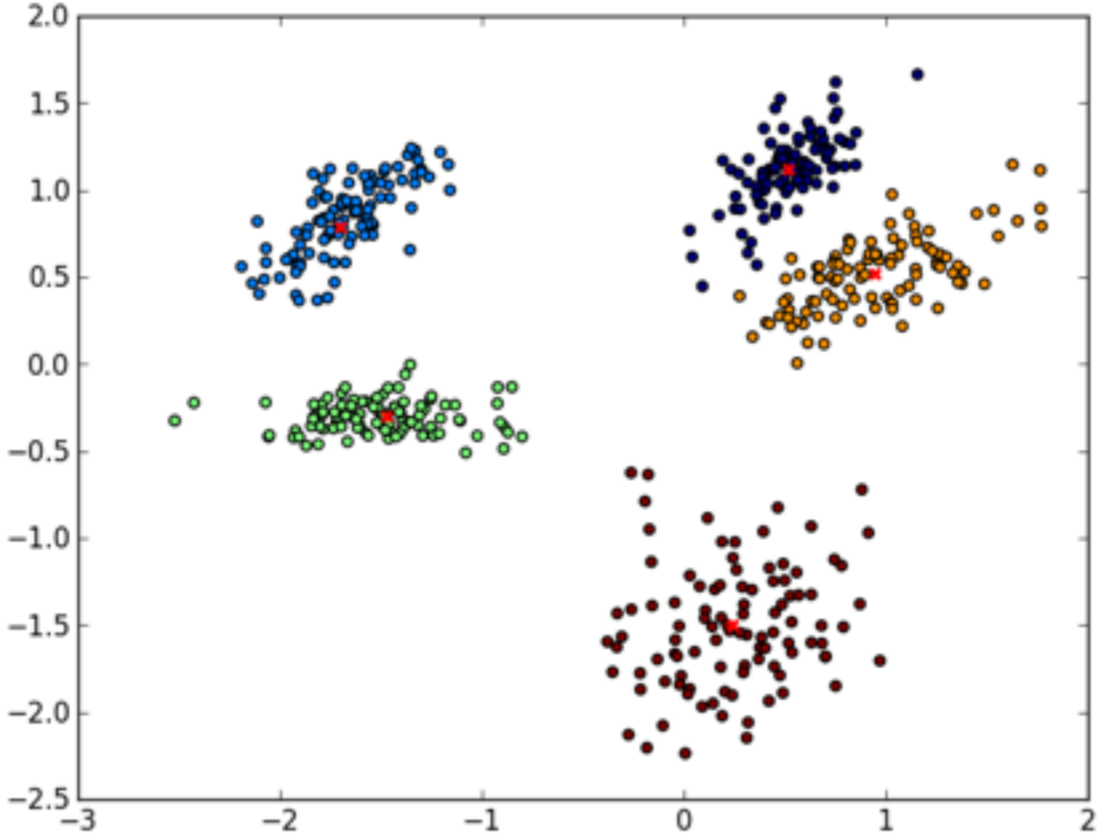
K-Means: Example



K-Means: Example



K-Means: Example



Hierarchical Clustering

Input: Data points, assignment to clusters, clustering cost function

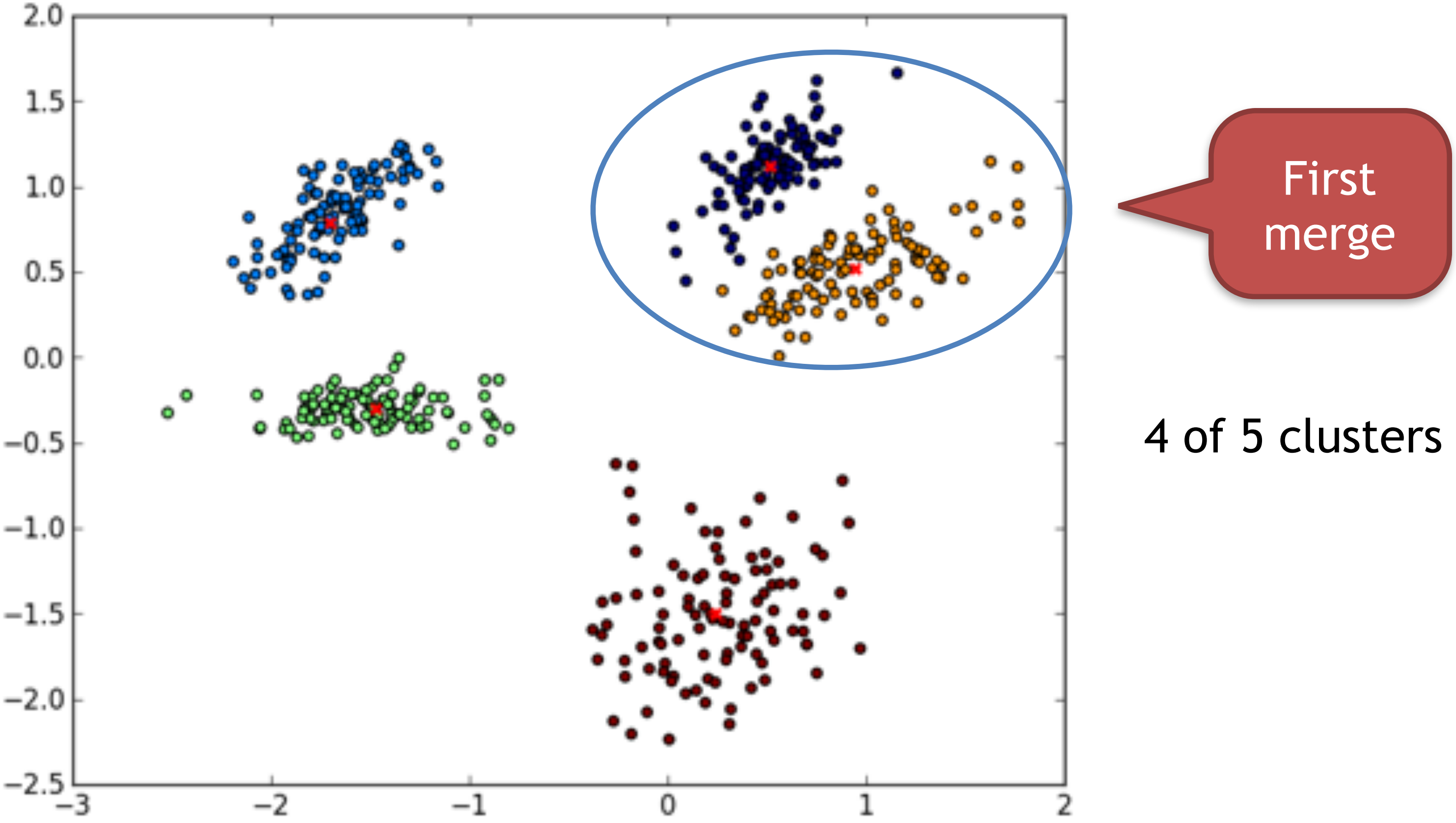
for *i=1* to *(no of clusters)-2* **do**

- find two clusters c_1, c_2 so that if we merge the clusters c_1, c_2 the cost function is minimal for all possible mergers
- merge c_1, c_2

Cost function: E.g. k-means criterion. Sum of distances to cluster center for each point.

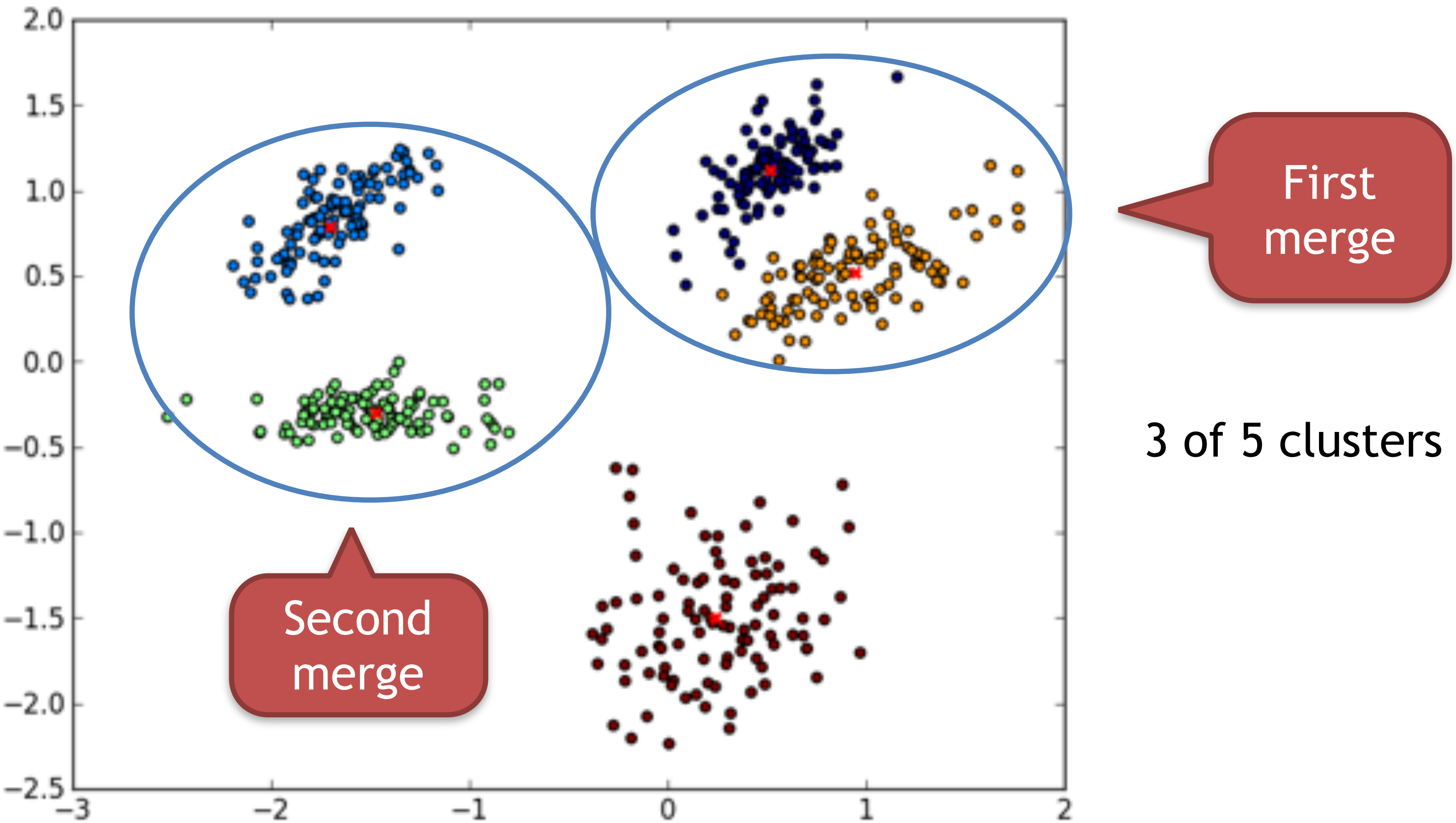
$$l(\{x_1, \ldots, x_n\}, r) = \sum_{i=1}^{n} \|x_i - \mu_{r_i}\|$$

Hierarchical Clustering: Example



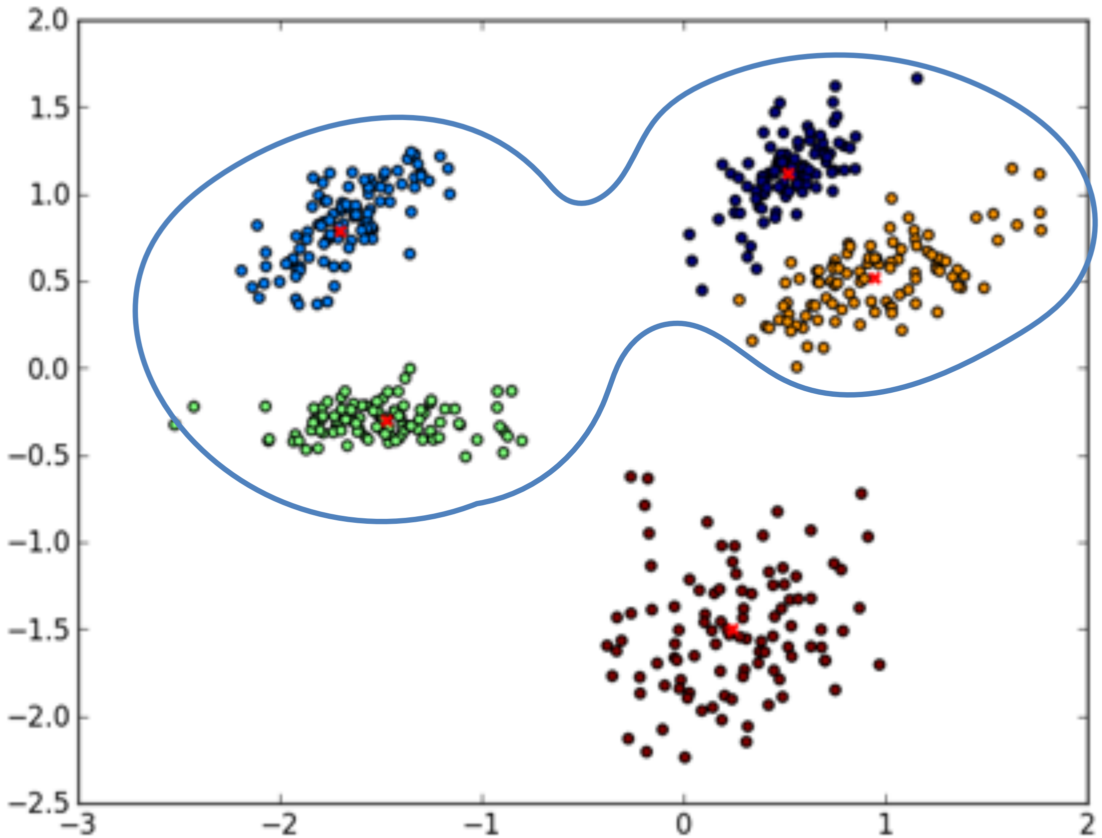
4 of 5 clusters remaining

Hierarchical Clustering: Example



3 of 5 clusters remaining

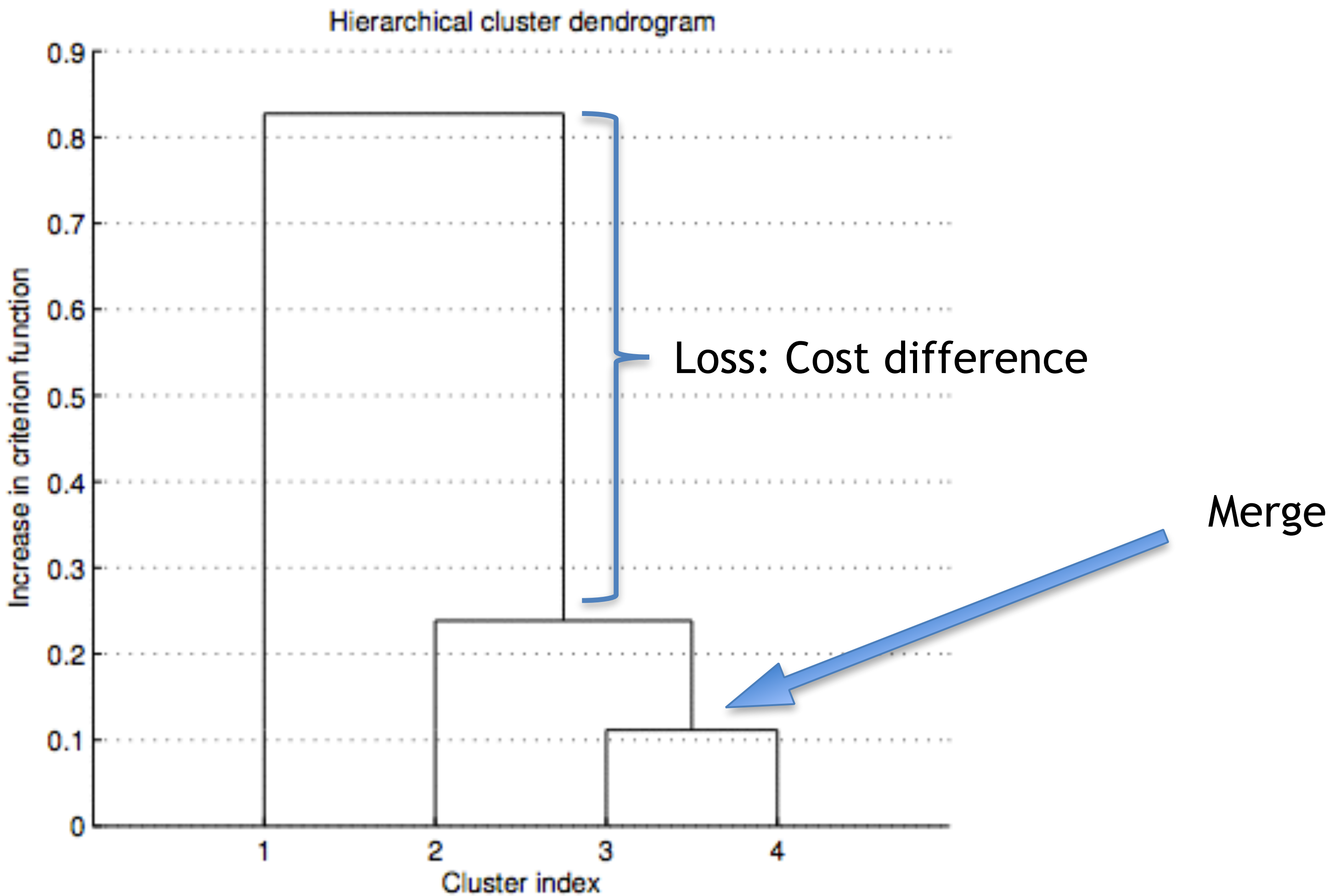
Hierarchical Clustering: Example



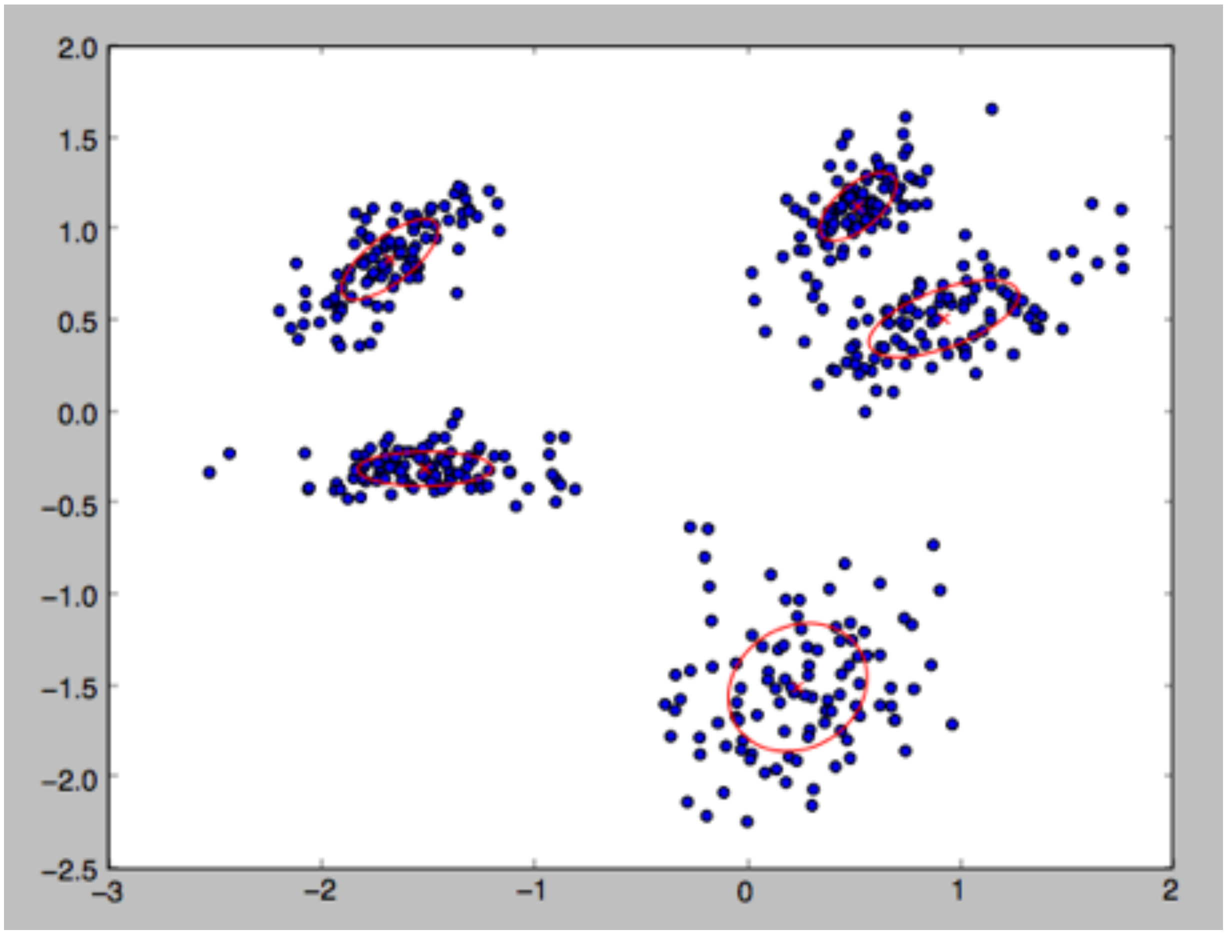
Third merge

2 of 5 clusters remaining

Visualization: Dendrogram plot



Mixture of Gaussians



View clusters as <u>Mixtures of Gaussians</u>

Consider not only cluster centers, but also Covariance matrices

Gaussian probability density function:

$$g(x,\mu,\Sigma) = \frac{1}{\sqrt{(2\pi)^k|\Sigma|}} exp\left(-\frac{1}{2}(x-\mu)^\intercal\Sigma^{-1}(x-\mu)\right)$$

Probability density function for a mixture of gaussians:

$$p(x) = \sum_{k=1}^{K} \pi_k g(x,\mu_k,\Sigma_k) \qquad \sum_{k=1}^{K} \pi_k = 1,\ \pi_k \geq 0$$

where K is the number of gaussians, μ_k and Σ_k are mean and covariance matrix of the gaussians, and π_k are class priors and describe how probable a draw from that cluster is.

We have these variables:

- X_n: observed data points
- γ_n: latent (hidden) variable that describes to which cluster X_n belongs
- μ_k, Σ_k, π_k: unknown parameters (mean, covariance, and class priors of gaussians)

We would like to maximize the likelihood

$$L(\mu, \Sigma, \pi; X, \gamma) = p(X, \gamma | \mu, \Sigma, \pi) = \sum_{n=1}^{N} \sum_{k=1}^{K} \delta_{\gamma_n = k} g(X_i | \mu_k, \Sigma_k, \pi_k)$$

However, this is intractable. Differentiating either variable will lead to an equation that depends on the other variables in a non-linear way.

The E-M algorithm calculates the cluster assignments γ as an intermediate step and iterates over:

- **E-Step:** Calculate the expected value of the log-likelihood (the cluster assignments) given the current parameters μ, Σ, π

$$Q(\mu, \Sigma, \pi | \mu^{(t)}, \Sigma^{(t)}, \pi^{(t)}) = E_{\gamma | X, \mu^{(t)}, \Sigma^{(t)}, \pi^{(t)}} [\log L(\mu, \Sigma, \pi; X, \gamma)]$$

- **M-Step:** Find the parameters μ, Σ, π that maximize Q:

$$\mu^{(t+1)}, \Sigma^{(t+1)}, \pi^{(t+1)} = \arg\max_{\mu, \Sigma, \pi} Q(\mu, \Sigma, \pi | \mu^{(t)}, \Sigma^{(t)}, \pi^{(t)})$$

This indeed maximizes the likelihood. (Proof not trivial.)

EM for Mixture of Gaussians

Algorithm

$\hat{\pi}_k \leftarrow 1/K$ Prior distribution of cluster assignments

$\hat{\mu}_k \leftarrow$ random points out of $X_1, \ldots, X_n$

$\hat{\Sigma}_k \leftarrow \mathbf{I}_d$

Step 1 (E-Step)

for $k \leftarrow 1$ to K **do**

  for $n \leftarrow 1$ to N **do**

    Set $\gamma_{nk} \leftarrow \dfrac{\hat{\pi}_k g(X_n; \hat{\mu}_k, \hat{\Sigma}_k)}{\sum_{k'=1}^{K} \hat{\pi}_{k'} g(X_n; \hat{\mu}_{k'}, \hat{\Sigma}_{k'})}$

  end for

end for

Compute likelihood that point n belongs to cluster k given the cluster centers and covariance matrices

g is the Gaussian probability density function

Step 2 (M-Step)

for $k \leftarrow 1$ to K **do**

  $N_k \leftarrow \sum_{n=1}^{N} \gamma_{nk}$

  $\hat{\pi}_k \leftarrow N_k/N$

  $\hat{\mu}_k \leftarrow \frac{1}{N_k} \sum_{n=1}^{N} \gamma_{nk} X_n$

  $\hat{\Sigma}_k \leftarrow \frac{1}{N_k} \sum_{n=1}^{N} \gamma_{nk} (X_n - \hat{\mu}_k)(X_n - \hat{\mu}_k)^\top$

end for

Computer new cluster centers + covariance matrices + priors