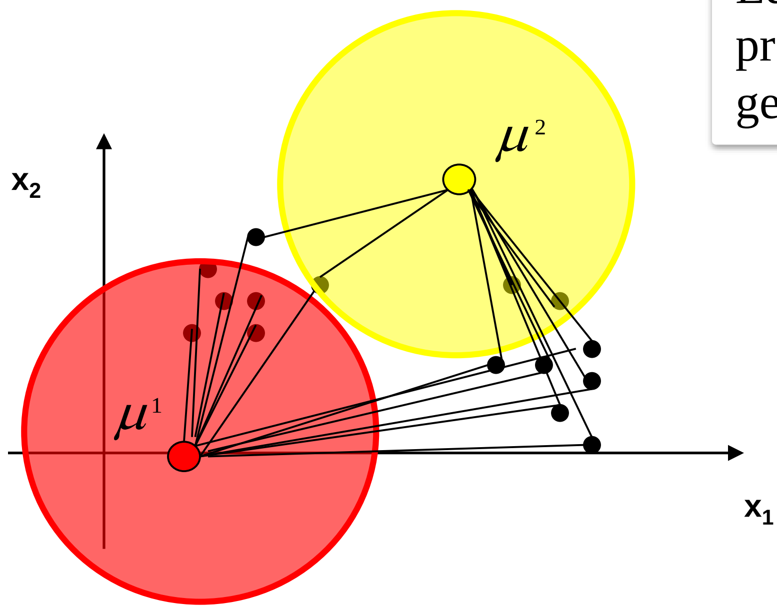
APPLIED MACHINE LEARNING

From soft K-means Clustering
to
Density Modeling of Data Clusters with Mixture of Gaussian PdF

Soft K-means Clustering (probabilistic interpretation)



Equivalent to computing the relative probability that the data point has been generated by the k-th cluster (d:norm-2).

$$r_i^k = \frac{e^{\left(-\beta \cdot d\left(\mu^k, x^i\right)\right)}}{\sum_{k'} e^{\left(-\beta \cdot d\left(\mu^{k'}, x^i\right)\right)}} \in [0,1], \quad \beta \in \mathbb{R}_+$$

The likelihood of the k-th model is:

$$L\left(\mu^k; X\right) \sim \prod_{i=1}^{M} e^{\left(-\beta \cdot d\left(\mu^k, x^i\right)\right)}$$

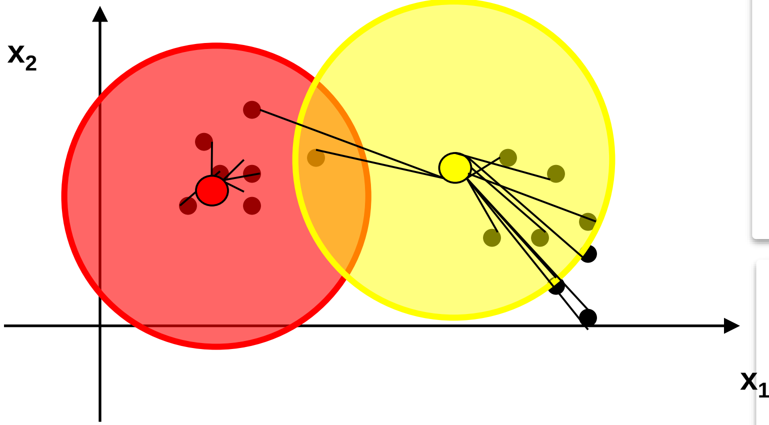
Assignment Step (~E-step)

Assume density under cluster *k* follows an (**un-normalized**) Gaussian pdf with σ=1/β variance, centered on the cluster's centroid. Un-normalized Gaussian pdf are called Radial Basis Function – RBF.

We can derive the likelihood that each cluster has generated the dataset.

E-step: expectation step ~ expectation of the likelihood

K-means Clustering (probabilistic interpretation)



The new centroid is closer to the datapoints after the update step,

$$\boxed{\mu^k = \frac{\sum_i r_i^k x^i}{\sum_i r_i^k}} \qquad \boxed{\Rightarrow d(x^i, \mu^k) \searrow}$$

→ The likelihood of the k-th model increases

$$L(\mu^k; X) = \prod_{i=1}^{M} e^{\left(-\beta \cdot d\left(\mu^k, x^i\right)\right)} \nearrow$$

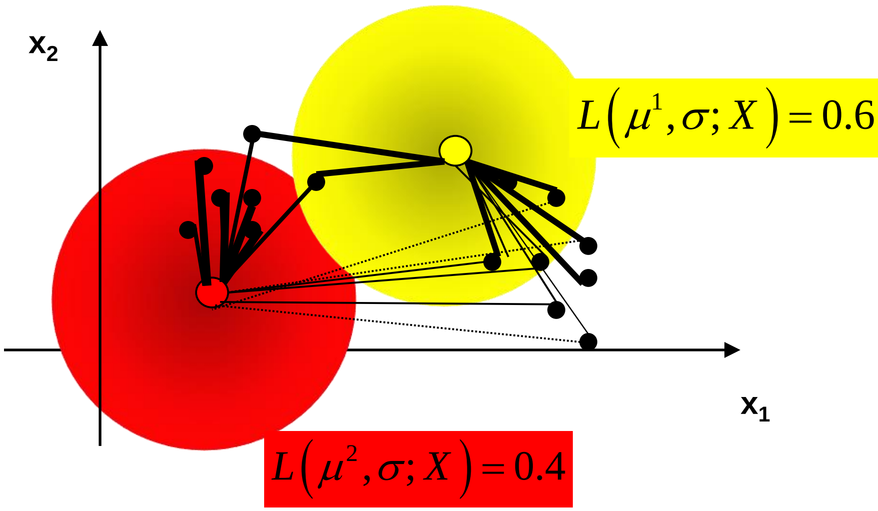
Update Step (~M-step)**:**
Update the position of the centroids.

Modify the parameters to maximize likelihood of the pdf.

Soft-K-means (probabilistic interpretation)

Soft K-means is similar to fitting the data distribution with a *mixture of isotropic (spherical) unnormalized Gaussian pdf-s and same variance (the stiffness).*

Assignment-update ~ E-M on the parameters of each Gaussian to optimize the likelihood that the Gaussians represent the distribution of the datapoints.



Should we weight equivalently the likelihoods of each cluster?

From Soft-K-means to GMM

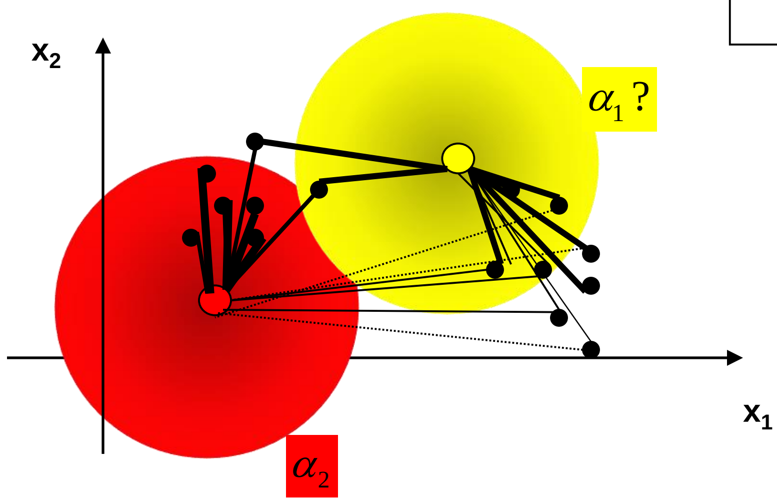
The responsibility factor gives a measure of the likelihood that cluster k generated the dataset.

r_i^k: responsibility of cluster k for point x^i

$$r_i^k = \frac{\boxed{\alpha_k} p\left(x^i;\mu^k,\sigma\right)}{\sum_{k'} \boxed{\alpha_{k'}} p\left(x^i;\mu^{k'},\sigma\right)}, \qquad \alpha_k \in [0,1]$$

$p\left(x^i;\mu^k,\sigma\right) \in [0,1]$: Gauss pdf evaluated at x^i

Normalized over clusters: $\sum_k r_i^k = 1$



Relative importance of each of the K clusters.

It should give a measure of the likelihood that the Gaussian k (or cluster k) generated the whole dataset.

$$\boxed{\alpha_k = \frac{\sum_i r_i^k}{\sum_k \sum_i r_i^k}}$$

One step towards Gaussian Mixture Model with Spherical Gaussians

r_i^k: responsibility of cluster k for point x^i

$$r_i^k = \frac{\boxed{\alpha_k} p\left(x^i; \boxed{\mu^k}, \boxed{\sigma^k}\right)}{\sum_{k'} \alpha_{k'} p\left(x^i; \mu^{k'}, \sigma^k\right)}, \qquad \alpha_k \in [0,1]$$

$p\left(x^i; \mu^k, \sigma\right) \in [0,1]$: Gauss pdf evaluated at x^i

Normalized over clusters: $\sum_k r_i^k = 1$

Update Step (M-Step)

$$\mu^k = \frac{\sum_i r_i^k x^i}{\sum_i r_i^k}$$

$$\left(\sigma^k\right)^2 = \frac{\sum_i r_i^k \left\| x^i - \mu^k \right\|^2}{N \cdot \sum_i r_i^k}$$

$$\alpha_k = \frac{\sum_i r_i^k}{\sum_k \sum_i r_i^k}$$

This fits a mixture of *spherical* Gaussians. The variance of each Gauss pdf fits the spread of the data around its mean.

From spherical to diagonal Gaussian pdf-s.

r_i^k: responsibility of cluster k for point x^i

$$r_i^k = \frac{\alpha_k p\left(x^i;\mu^k,\sigma_j^k\right)}{\sum_{k'} \alpha_{k'} p\left(x^i;\mu^{k'},\sigma_j^{k'}\right)}$$

$j = 1,\ldots N$: dimension of dataset

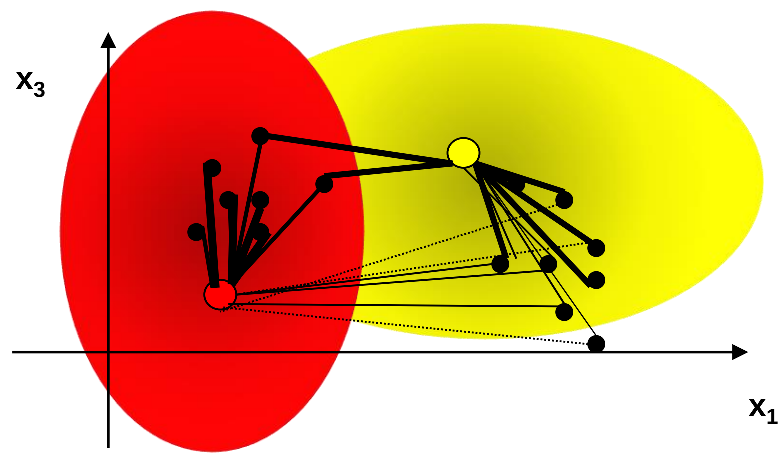
$p\left(x^i;\mu^k,\sigma^k\right) \in [0,1]$: Gauss pdf evaluated at x^i

Normalized over clusters: $\sum_k r_i^k = 1$

Update Step (M-Step)

$$\left(\sigma_j^k\right)^2 = \frac{\sum_i r_i^k \left(x_j^i - \mu_j^k\right)^2}{\sum_i r_i^k}$$

One covariance element per dimension, aligned with the axes of the original frame of reference.



Clustering with Mixture of Gaussians

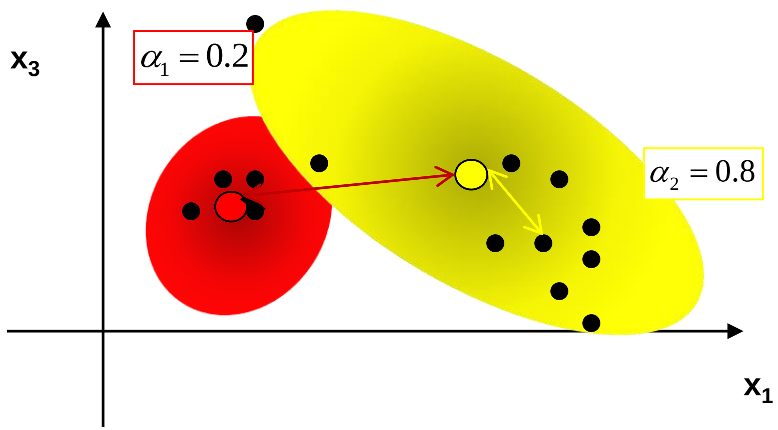
Likelihood of the mixture of Gaussians: $L\left(\Theta=\left\{\alpha_k,\mu^k,\Sigma^k\right\}_{k=1}^{K};x\right)=\sum_{k=1}^{K}\boxed{\alpha_k}\cdot p\left(x;\mu^k,\Sigma^k\right)$

with $\boxed{p\left(X;\mu^k,\Sigma^k\right)\sim\prod_{i=1}^{M}e^{-\left(x^i-\mu_k\right)^T\left(\Sigma^k\right)^{-1}\left(x^i-\mu_k\right)}}$ (omit normalization factor)

μ^k,Σ^k : mean and covariance matrix of Gaussian k

The mixing Coefficients are normalized. $\sum_{k=1}^{K}\alpha_k=1$

$$\alpha_k\sim\frac{1}{M}\sum_{i=1}^{M}\frac{p\left(x^i;\mu^k,\Sigma^k\right)}{\sum_{k'}p\left(x^i;\mu^{k'},\Sigma^{k'}\right)}$$



E-M Steps for GMM

Initialization:

The priors $\alpha_1,..,\alpha_k$ can be uniform for starters.

The means $\mu^1,..,\mu^K$ can be initialized with K-means.

Calculate the initial value of the likelihood

$$p\left(X \mid \Theta^{(t)}\right)=\prod_{i=1}^{M}\sum_{k=1}^{K}\alpha_{k(t)}\,p\left(x^i;\mu^{k(t)},\Sigma^{k(t)}\right)$$

Expectation Step (E-step):

Evaluate responsibilities of each cluster *k* over each sample x^i *using current parameters*

$$r_i^k=\frac{\alpha_k\,p\left(x^i;\mu^k,\Sigma^k\right)}{\sum_{k'}\alpha_{k'}\,p\left(x^i;\mu^{k'},\Sigma^{k'}\right)}$$

Maximization (Update step) Step (M-step):

Recompute the means, covariances matrices and prior probabilities so as to maximize the log – likelihood of the current estimate: $\log\left(L\left(\Theta^{(t)} \mid X\right)\right)$

$$\boxed{\mu_k^{(t+1)} = \frac{1}{M_k}\sum_{i=1}^{M} r_k^i x_i}$$

$$\boxed{\Sigma_k^{(t+1)} = \frac{1}{M_k}\sum_{i=1}^{M} r_k^i (x_i - \mu_k^{(t+1)})(x_i - \mu_k^{(t+1)})^T}$$

$$\boxed{\alpha_k^{(t+1)} = \frac{M_k}{M}}$$

where: $M_k = \sum_{i=1}^{M} r_k^i$

The E and M steps alternate until the log-likelihood reaches a plateau.