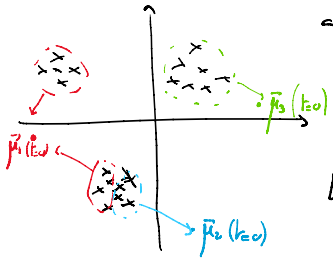
II Gaussian Mixture model

A] First approach : k-means → intuitive & simple approach of clustering

Based on the following idea : . a cluster of points is a set of points nearby each-other
→ we put them together according to a distance measure

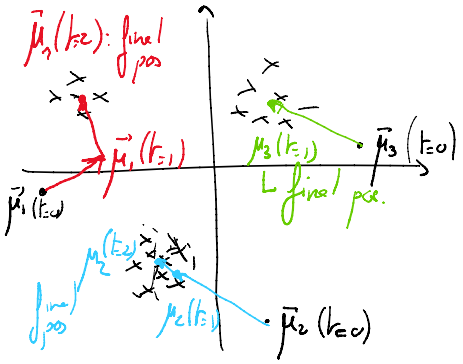


Let's $\{\vec{x}_i\} \in \mathbb{R}^{m \times d}$: $\begin{cases} m \text{ is the nb of pts} \\ d \text{ is the dim} \end{cases}$

Let's define C centers (means) : $\{\vec{\mu}_a\} \in \mathbb{R}^{C \times d}$

Algo : i) assign each pt to the closest center

$$r_a^i = \begin{cases} 1 & \text{if the center } a \text{ is the closest of the data } i \\ 0 & \text{otherwise} \end{cases}$$



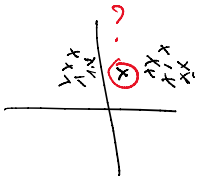
ii) move the centers to the center of mass of the datapts that have been assigned :

$$\vec{\mu}_a^{(t+1)} = \frac{\sum_{i=1}^{m} \vec{x}_i \cdot r_a^i}{\sum_{i=1}^{m} r_a^i}$$

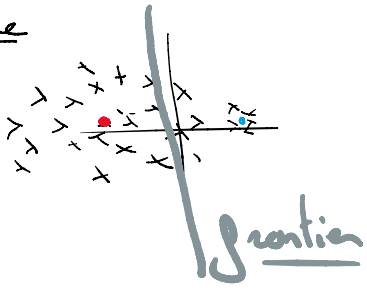
→ very intuitive

Some drawbacks :

. hard assignement



. no scale



. no geometry

. no quality indicator

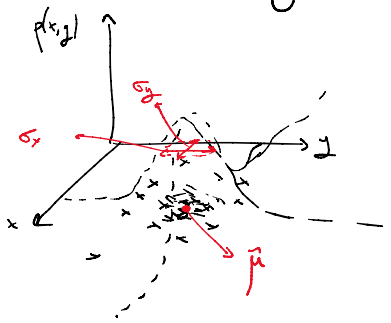
. hard to choose initial cdt.

. _____ the number C of clusters

B] A bayesian approach: GMM

we need a probabilistic model of our clusters

A simple choice: describe a cluster by a multi-dim
a gaussian distribution



$$p(\vec{x}|\theta) = \sum_{h=1}^{C} \rho_h \prod_{i=1}^{d} \left[\frac{1}{\sqrt{2\pi\sigma_{hd}^2}} \exp\left(-\frac{(x_d - \mu_{hd})^2}{2\sigma_{hd}^2} \right) \right]$$

θ :
- ρ_h : density of cluster h ($\sim \frac{\#\text{ of pts in clus } h}{\text{Total \# of pts}}$)
- $\vec{\mu}_h$: position of the center (means)
- $\vec{\sigma}_h$: variance in all direction of clus. h.

⚠ How the variances is parametrized:

- here we have a variance for each direction along the axis

Other cases: . having the same variance in all directions

. having a full covariance matrix

⚠ huge # of parameters: $\frac{d(d-1)}{2} \cdot k \sim O(d^2 k)$

Bayes Thm: $$p(A|B) = \frac{p(B|A) \cdot p(A)}{p(B)}$$

posterior dist. | likelihood | a prior

$$p(\theta \,|\, \{\vec{x}\}) = \frac{p(\{\vec{x}\}|\theta) \cdot p(\theta)}{p(\{\vec{x}\})} \quad \left(= \int d\theta \, p(\{\vec{x}\}|\theta) \, p(\theta)\right)$$

probe of my param θ given the data.

evidence → we ignore it

Goal: to maximize $p(\theta|\{\vec{x}\})$: $\theta^* = \arg\max p(\theta|\{\vec{x}\})$

when using a uniform prior for $p(\theta)$

$$\rightarrow \boxed{\theta^* = \arg\max p(\{\vec{x}\}|\theta)}$$

- ——— the number C of clusters

Lets do it in 1D : impossible to obtain directly the global max !

→ gradient ascent

log-likelihood $\mathcal{L} = \log p(\{x\}|\theta)$

$$\frac{\partial \mathcal{L}}{\partial \mu_h} = \sum_{i=1}^{m} \left[\frac{\rho_h \frac{(x_i - \mu_h)}{\sigma_h^2} \cdot e^{-\frac{(x_i-\mu_h)^2}{2\sigma_h^2}}}{\sum_{h'=1} \rho_{h'} \frac{1}{\sqrt{2\pi\sigma_{h'}^2}} \cdot \exp\left[-\frac{(x_i-\mu_{h'})^2}{2\sigma_{h'}^2} \right]} \right] = 0$$

no explicit form for the gradient

To find the max : using [Expectation, Maximization] ~ similar to the way k-means work

Update eqs with EM

1) We assign (in probas) each data to a cluster

the responsability ← $r_a^i = \dfrac{\frac{\rho_a}{\sqrt{2\pi\sigma_a^2}} \exp\left(-\frac{(x_i-\mu_a)^2}{2\sigma_a^2}\right)}{\sum_h \frac{\rho_h}{\sqrt{2\pi\sigma_h^2}} \exp\left[-\frac{(x_i-\mu_h)^2}{2\sigma_h^2}\right]}$: Expectation

2) update of the parameters :

$\Sigma : R \times X$

$\mathbb{R}^{c\times m}\ \mathbb{R}^{m\times d}$

$$\begin{cases} \mu_a = \dfrac{\sum_{i=1}^{m} r_a^i x_i}{\sum_{i=1}^{m} r_a^i} & (R_a := \sum_i r_a^i) \\ \sigma_a = \dfrac{\sum_{i=1}^{m} r_a^i (x_i-\mu_a)^2}{R_a} \\ \rho_a = \dfrac{R_a}{\sum_a R_a} \end{cases}$$

Maximization step