CLUSTERING

classifi: $\mathcal{X} = \{(x_i, y_i)\}_{i=1}^{N}$

x_i → data points, y_i → class labels

→ $y_i \in \{0, 1\}$ or $\{-1, +1\}$

→ $y_i \in \{1, 2, \ldots, K\}$

clustering: $\mathcal{X} = \{x_i\}_{i=1}^{N}$ ⇐ NO CLASS LABELS!

PARAMETRIC CLASSIFICATION

- We assumed that each class follows a certain density. class conditional density

$$p(x \mid y=c) \rightarrow \text{class conditional density}$$

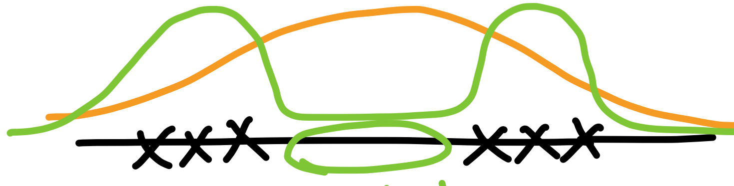
- We estimated the parameters

$$\underbrace{p(x \mid y=1)}_{\hat{\mu}_1,\ \hat{\Sigma}_1},\ \underbrace{\Pr(y=1)}_{\hat{P}r(y=1)},\ \ldots\ldots,\ \underbrace{p(x \mid y=K)}_{\hat{\mu}_K,\ \hat{\Sigma}_K},\ \underbrace{\Pr(y=K)}_{\hat{P}r(y=K)}$$

$\Pr(y=c \mid x) = ?$ posterior probability

Mixture Densities

K different clusters (unknown)



C_k = cluster #k

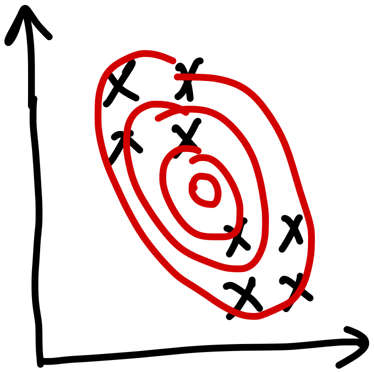
$$p(x) = \sum_{k=1}^{K} \underbrace{p(x \mid C_k)}_{\text{component densities}} \underbrace{Pr(C_k)}_{\text{mixture proportions}}$$

$N(0,1)$

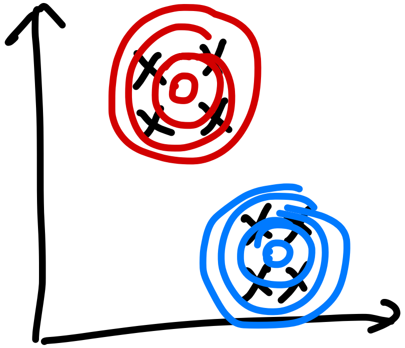
$N(4,2)$

$$\boxed{0.6\, N(0,1) + 0.4\, N(4,2)}$$

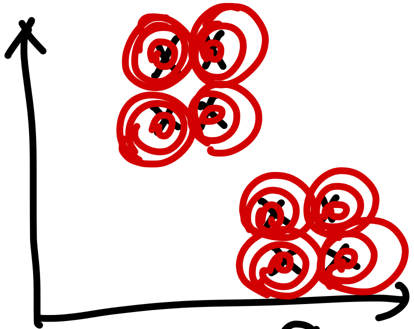
$$\begin{cases} p(x) \geq 0 \quad \forall x \\ \int_{-\infty}^{+\infty} p(x)\,dx = 1 \end{cases}$$



k = 1



k = 2



k = 8

$$\int_{-\infty}^{+\infty} [0.6\, p_1(x) + 0.4\, p_2(x)]\, dx$$

$$0.6 \underbrace{\int_{-\infty}^{+\infty} p_1(x)\,dx}_{1} + 0.4 \underbrace{\int_{-\infty}^{+\infty} p_2(x)\,dx}_{1}$$

K = # of components (groups) (clusters)

$$\Phi = \{\hat{Pr}(C_k), \hat{\mu}_k, \hat{\Sigma}_k\}_{k=1}^{K}$$

cluster membership

$$y_{ik} = \begin{cases} 1 & \text{if } x_i \text{ belongs to component/cluster/group } k \\ 0 & \text{otherwise} \end{cases}$$

$\hookrightarrow \hat{y}_{ik}$

WE DO NOT KNOW y_{ik} VALUES!!!

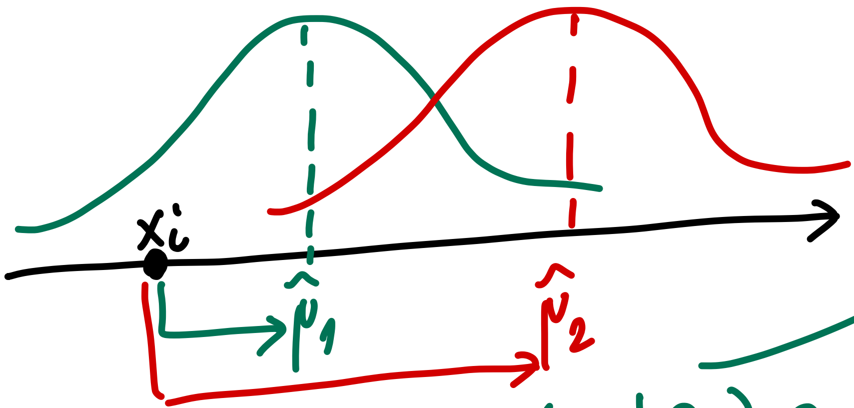
Iterative Algorithm:

STEP 1: Estimate the cluster memberships ($\hat{y}_{ik}$)

STEP 2: Estimate the parameters.

$$\hat{Pr}(C_k) = \frac{\sum_{i=1}^{N} \hat{y}_{ik}}{N} \qquad \hat{\mu}_k = \frac{\sum_{i=1}^{N} \hat{y}_{ik} \cdot x_i}{\sum_{i=1}^{N} \hat{y}_{ik}} \qquad \hat{\Sigma}_k = \frac{\sum_{i=1}^{N} \hat{y}_{ik} (x_i - \hat{\mu}_k)(x_i - \hat{\mu}_k)^T}{\sum_{i=1}^{N} \hat{y}_{ik}}$$

K-MEANS CLUSTERING



$$\cancel{\frac{1}{\sqrt{2\pi\hat{\sigma}_1^2}}} \cdot \exp\left[-\frac{(x_i - \hat{\mu}_1)^2}{\cancel{2\hat{\sigma}_1^2}}\right]$$

$$\cancel{\frac{1}{\sqrt{2\pi\hat{\sigma}_2^2}}} \cdot \exp\left[-\frac{(x_i - \hat{\mu}_2)^2}{\cancel{2\hat{\sigma}_2^2}}\right]$$

$$Pr(C_1 \mid x_i) = \frac{p(x_i \mid C_1)\, \cancel{Pr(C_1)}}{\cancel{p(x_i)}}$$

equal frequencies ← Assuming $Pr(C_1) = Pr(C_2)$ & $\sigma_1^2 = \sigma_2^2$

$$Pr(C_2 \mid x_i) = \frac{p(x_i \mid C_2)\, \cancel{Pr(C_2)}}{\cancel{p(x_i)}}$$

Compare $\|x_i - \hat{\mu}_1\|_2$ & $\|x_i - \hat{\mu}_2\|_2$

if $\|x_i - \hat{\mu}_1\|_2 < \|x_i - \hat{\mu}_2\|_2 \Rightarrow C_1$

if $\|x - \hat{\mu}_1\|_2 > \|x_i - \hat{\mu}_2\|_2 \Rightarrow C_2$

$$\text{Error} = \sum_{i=1}^{N} \sum_{k=1}^{K} b_{ik} \|x_i - \hat{\mu}_k\|_2^2 \Rightarrow \text{total squared distance to nearest centroids.}$$

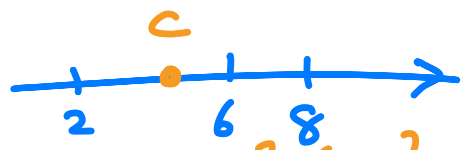
$$\boxed{b_{ik}} = \begin{cases} 1 & \text{if } \|x_i - \hat{\mu}_k\|_2 = \min_{c=1}^{K} \|x_i - \hat{\mu}_c\|_2 \\ 0 & \text{otherwise} \end{cases}$$

hard clustering.

minimize $\sum_{i=1}^{N} \sum_{k=1}^{K} b_{ik} \|x_i - \hat{\mu}_k\|_2^2$ → lowercase

with respect to: $\hat{\mu}_1, \hat{\mu}_2, \ldots, \hat{\mu}_K$, $\{b_{ik}\}_{i=1, k=1}^{N, K}$

uppercase



$(c-2)^2 + (c-6)^2 + (c-8)^2$

$c^2 - 4c + 4$
$c^2 - 12c + 36$
$c^2 - 16c + 64$

$3c^2 - 32c$

$\Downarrow$

$6c - 32 = 0$

$c = \frac{16}{3}$

- Initialize $\hat{\mu}_1, \hat{\mu}_2, \ldots, \hat{\mu}_K$ randomly

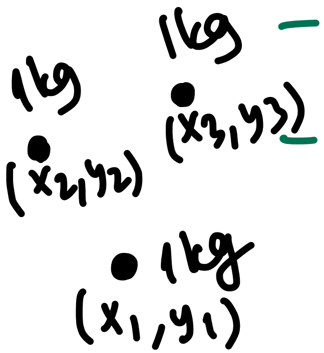
- Repeat

[for all x_i:

$$b_{ik} = \begin{cases} 1 & \text{if } \|x_i - \hat{\mu}_k\|_2 = \min_{c=1}^{K} \|x_i - \hat{\mu}_c\|_2 \\ 0 & \end{cases}$$

[for all $\hat{\mu}_k$:

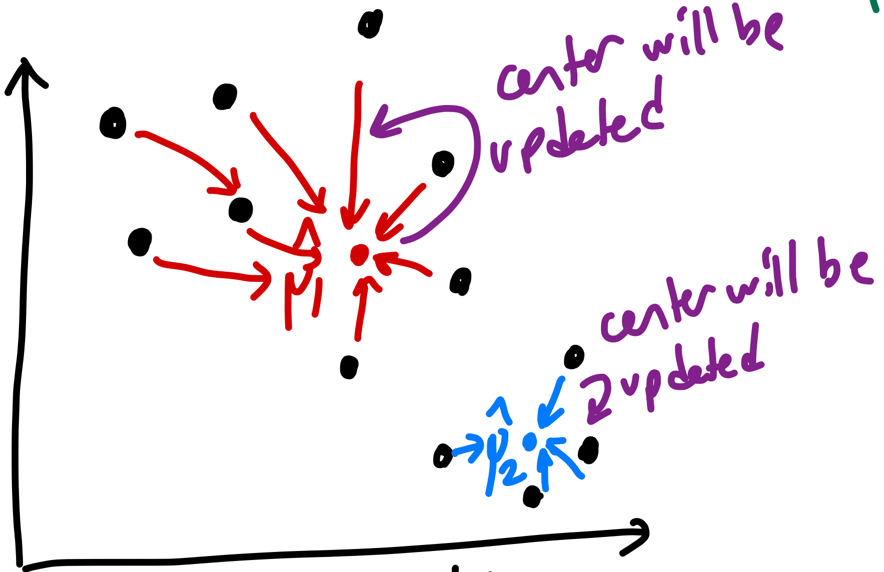
$$\hat{\mu}_k = \frac{\sum_{i=1}^{N} b_{ik} \cdot x_i}{\sum_{i=1}^{N} b_{ik}}$$]



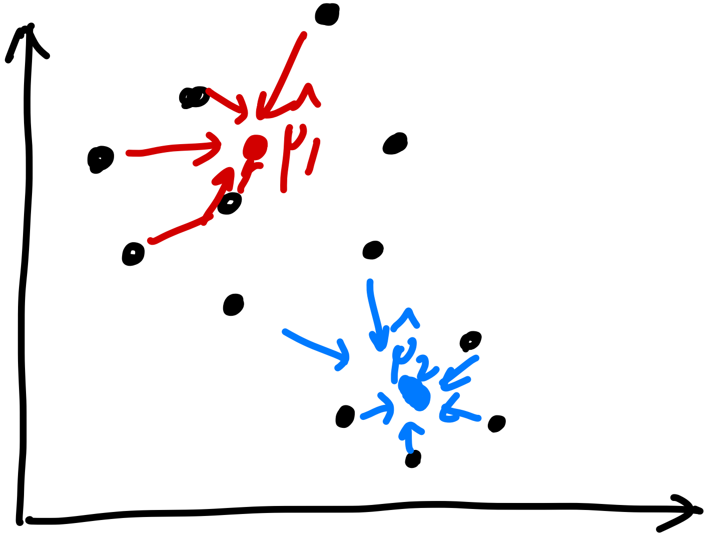
center of mass

$$\left(\frac{x_1 + x_2 + x_3}{3}, \frac{y_1 + y_2 + y_3}{3}\right)$$

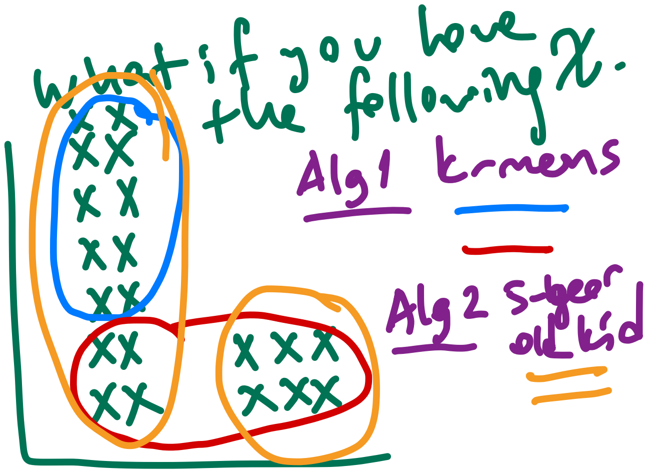
- Until convergence
  if [[all b_{ik}'s stay the same]] or
  [[all $\hat{\mu}_k$'s stay the same]]

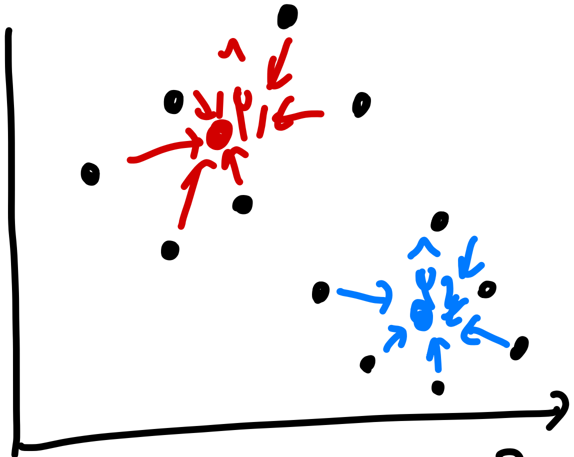


Iteration #1



Iteration #2





Iteration #3

b_{ik}'s stayed the same

STOP!

Expectation-Maximization Algorithm (EM Algorithm)

$\mathcal{X} = \{x_i\}_{i=1}^{N}$ likelihood $\Rightarrow \mathcal{L}(\Phi \mid \mathcal{X}) = \prod_{i=1}^{N} p(x_i \mid \Phi)$

log likelihood $\Rightarrow \log \mathcal{L}(\Phi \mid \mathcal{X}) = \sum_{i=1}^{N} \log \left[\sum_{k=1}^{K} p(x_i \mid C_k) \Pr(C_k) \right]$

two sets of random variables

Z = cluster memberships (hidden variables)

Φ = parameters $[\hat{\mu}_1, \hat{\mu}_2, \ldots, \hat{\mu}_k, \hat{\Sigma}_1, \hat{\Sigma}_2, \ldots, \hat{\Sigma}_k]$

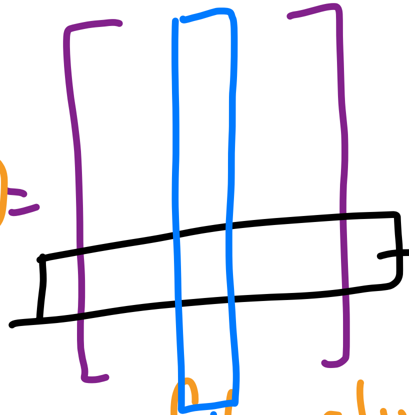
E-Step: $E\left[\mathcal{L}(\Phi \mid \mathcal{X}, Z) \mid \mathcal{X}, \Phi^{(t)}\right]$

M-Step: $\Phi^{(t+1)} = \arg\max_{\Phi} E\left[\mathcal{L}(\Phi \mid \mathcal{X}, Z) \mid \mathcal{X}, \Phi^{(t)}\right]$

E-Step Updates:

$$h_{ik} = E\left[z_{ik} \mid \mathcal{X}, \Phi^{(t)}\right] = \frac{p(x_i \mid C_k, \Phi^{(t)})\, Pr(C_k)}{\sum_{c=1}^{K} p(x_i \mid C_c, \Phi^{(t)})\, Pr(C_c)}$$

$h_{ik} \geq 0$ $\quad \sum_{k=1}^{K} h_{ik} = 1 \quad \forall i$

$H =$ 

→ summation should be 1.

→ soft clustering.

→ column sum would give us relative weight of cluster k.

M-Step Updates:

$$\hat{Pr}^{(t+1)}(C_k) = \frac{\sum_{i=1}^{N} h_{ik}}{N}$$

$$\hat{\mu}_k^{(t+1)} = \frac{\sum_{i=1}^{N} h_{ik} \cdot x_i}{\sum_{i=1}^{N} h_{ik}}$$

$$\hat{\Sigma}_k^{(t+1)} = \frac{\sum_{i=1}^{N} h_{ik} \left(x_i - \hat{\mu}_k^{(t+1)}\right)\left(x_i - \hat{\mu}_k^{(t+1)}\right)^T}{\sum_{i=1}^{N} h_{ik}}$$