Clustering

Classification $\left[\mathcal{X} = \{(x_i, y_i)\}_{i=1}^{N} \right.$

x_i → data points; y_i → class labels

Binary classification

$y_i \in \{0, 1\}$ or $y_i \in \{-1, +1\}$

Multiclass classification

$y_i \in \{1, 2, \ldots, K\}$

Clustering $\left[\mathcal{X} = \{x_i\}_{i=1}^{N} \right.$ NO CLASS LABELS!

PARAMETRIC CLASSIFICATION

- We assumed that each class follows a certain density $p(x \mid y=c)$
- We estimated the parameters

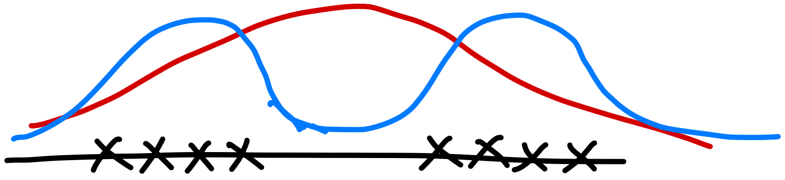
$$p(x \mid y=1) \quad P(y=1) \quad \cdots\cdots \quad p(x \mid y=K) \quad P(y=K)$$

$$\downarrow \qquad\qquad \downarrow \qquad\qquad\qquad \downarrow \qquad\qquad \downarrow$$

$$\hat{\mu}_1, \hat{\Sigma}_1 \quad \hat{P}(y=1) \quad \cdots \quad \hat{\mu}_K, \hat{\Sigma}_K \quad \hat{P}(y=K)$$

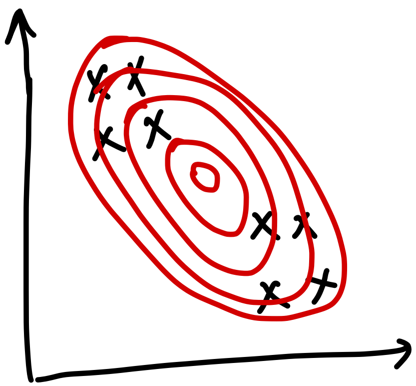
$P(y=c \mid x) = ?$

Mixture Densities K different clusters (unknown)



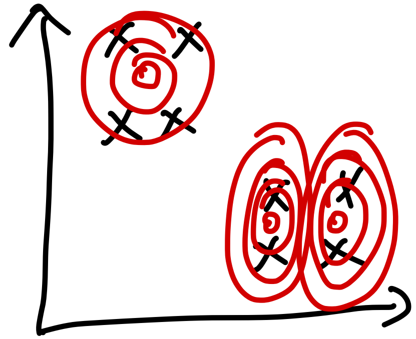
C_k = cluster # k

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



K=1

K=2



K=3

K = # of components (clusters) (groups)

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

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

↳ cluster/component/group membership

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

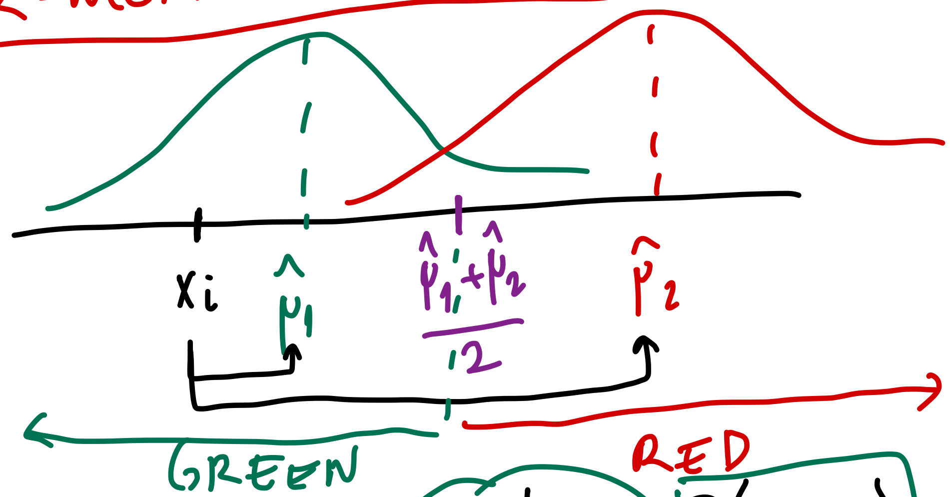
Iterative Algorithm:

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

STEP 2 : Estimate the parameters.

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

K-MEANS CLUSTERING



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

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

$$\boxed{\hat{\sigma}_1^2 = \hat{\sigma}_2^2} \Leftarrow$$

$$P(y=1 \mid x) = \frac{p(x \mid y=1)\, \boxed{P(y=1)}}{p(x)} \approx$$

$$\hat{\sigma}_1^2 = \hat{\sigma}_2^2$$

$$p(y=2 \mid x) = \frac{p(x \mid y=2)\, \boxed{P(y=2)}}{p(x)}$$

$$\|x_i - \hat{\mu}_1\|_2 \quad \underbrace{\|x_i - \hat{\mu}_2\|_2}_{} \quad \cdots \quad \|x_i - \hat{\mu}_k\|_2$$

assume that 2nd distance is minimum.

$$\hat{y}_{i1} = 0 \quad \hat{y}_{i2} = 1 \quad \hat{y}_{i3} = 0 \quad \cdots \quad \hat{y}_{ik} = 0$$

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

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

$$B = \underbrace{\begin{bmatrix} \quad \end{bmatrix}}_{K} \Big\} N$$

each row contains exactly one 1.

MIP

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

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

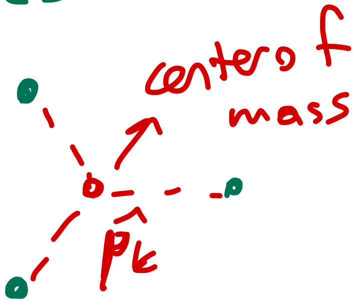
- Initialize $\hat{\mu}_1, \hat{\mu}_2, \ldots, \hat{\mu}_K$ randomly
- Repeat

E-STEP a → for all x_i:

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

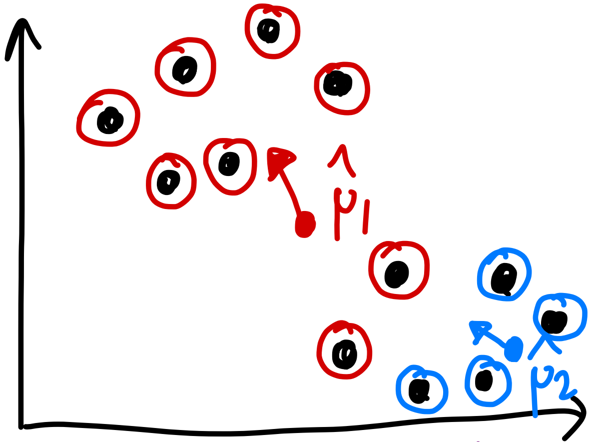
M-STEP b → for all $\hat{\mu}_k$:

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

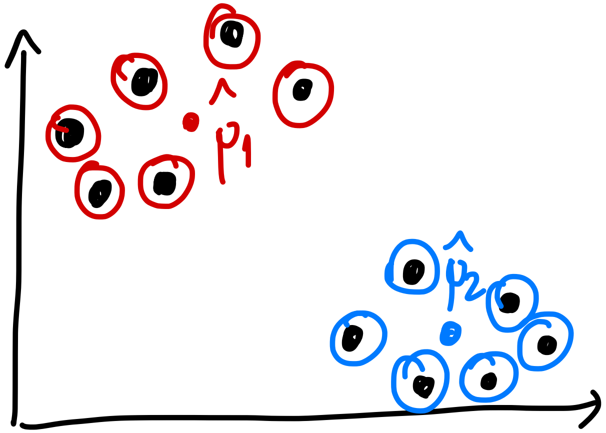


- Until convergence [all b_{ik}'s stay the same] or [all $\hat{\mu}_k$'s stay the same]

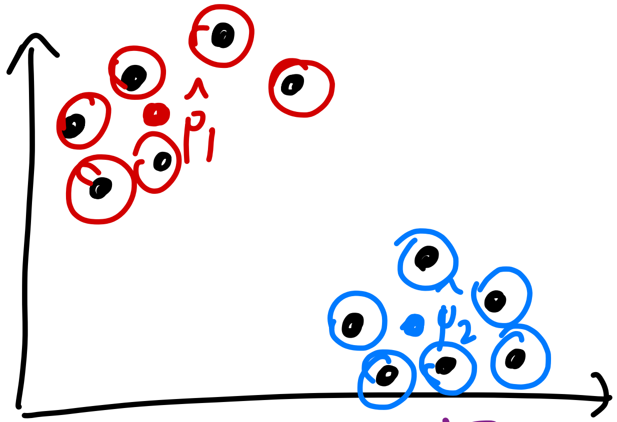
K=2



Iteration #1

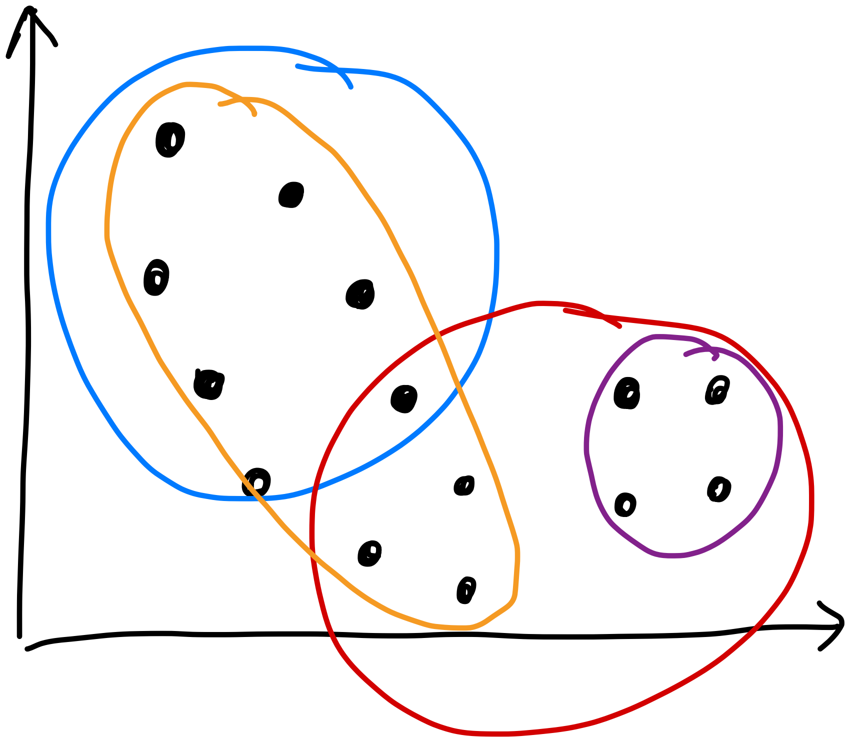


Iteration #2



Iteration #3

bits stay the same ⇒ STOP



assumed shared covariance

assumed different covariances

Expectation - Maximization Algorithm

$X = \{x_i\}_{i=1}^{N}$ log likelihood $\Rightarrow L(\Phi \mid \mathcal{X}) = \log \underbrace{\left[\prod_{i=1}^{N} p(x_i \mid \Phi)\right]}$

$$\log L(\Phi \mid \mathcal{X}) = \sum_{i=1}^{N} \log \left[\underbrace{\sum_{k=1}^{K} p(x_i \mid C_k) P(C_k)}_{\text{mixture densities}}\right]$$

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

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

E-STEP:

$$h_{ik} = E\left[z_{ik} \mid \mathcal{X}, \Phi^{(t)}\right] = \frac{p(x_i \mid C_k, \Phi^{(t)}) \cdot P(C_k)}{\sum_{c=1}^{K} \underbrace{p(x_i \mid C_c, \Phi^{(t)})}_{\text{multivariate Gaussians}} P(C_c)}$$

$$H = \overbrace{\begin{bmatrix} \quad & \quad \\ & \\ & \end{bmatrix}}^{K} \Bigg\} N$$

each row sums up to 1.

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

M-STEP:

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

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

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