Discussion 06

2022/4/21

https://www.youtube.com/watch?v=REypj2sy_5U&list=PLBv09BD7ez_4e9LtmK626Evn1ion6ynrt

Hard clustering: clusters do not overlap

Soft clustering: clusters my overlap

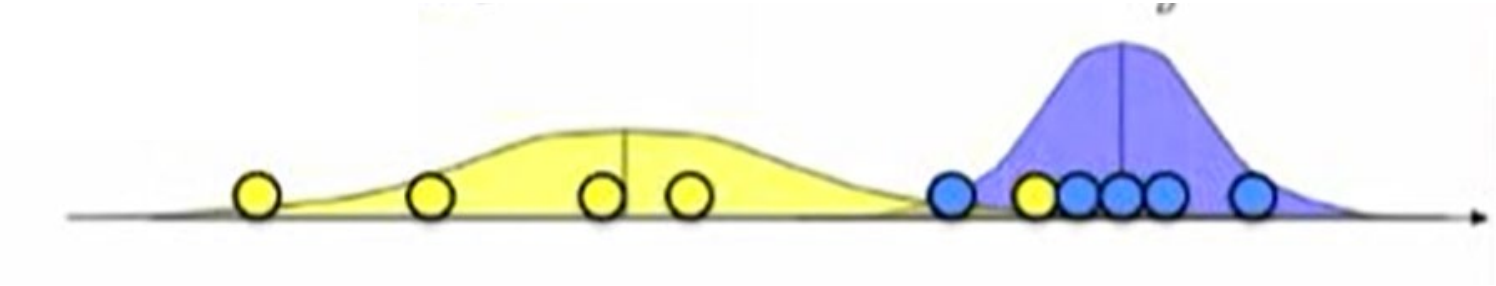
Case 1

- Observations $x_1, x_2, \ldots\ldots, x_n$

a : yellow
b : blue

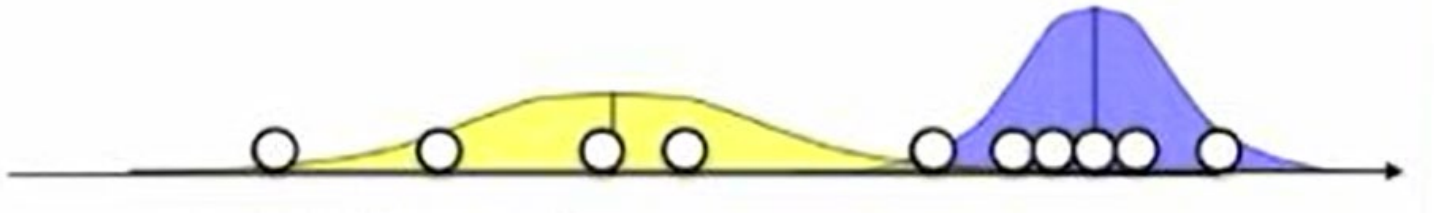
- From 2 Gaussians with unknown $(\mu_a, \sigma_a^2)\ and\ (\mu_b, \sigma_b^2)$

- $\mu_a = \frac{x_1 + x_2 + \ldots + x_{na}}{n_a}$
- $\sigma_a^2 = \frac{(x_1 - \mu_a)^2 + (x_2 - \mu_a)^2 + \ldots + (x_{na} - \mu_a)^2}{n_a}$

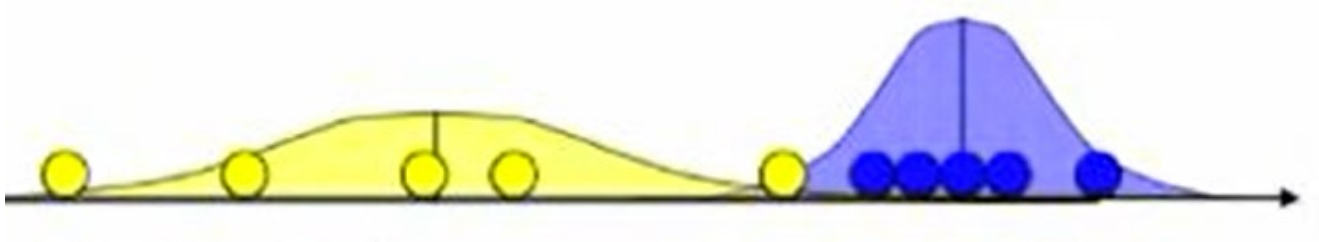


Case 2

- What if do not know the source ?

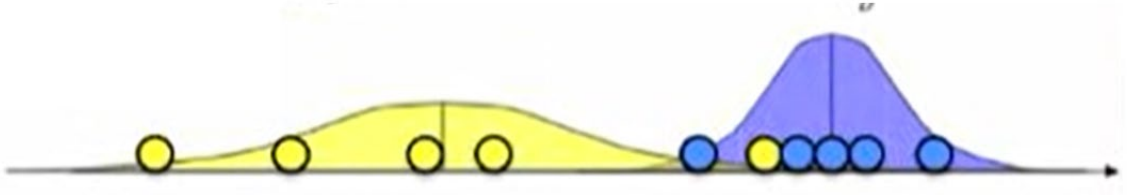


- If we know the parameters of the 2 Gaussians $\mu_a, \sigma_a^2 \ \ and \ \mu_b, \sigma_b^2$
- Guess whether a point is more likely to be a or b
- $P(x_i \mid a) = \frac{1}{\sqrt{2\pi\sigma_a^2}} e^{-\frac{(x_i - \mu_a)^2}{2\sigma_a^2}}$
- $P(a \mid x_i) = \frac{P(x_i \mid a)\, P(a)}{P(x_i \mid a)\, P(a) + P(x_i \mid b)\, P(b)}$, $P(b \mid x_i) = 1 - P(a \mid x_i)$

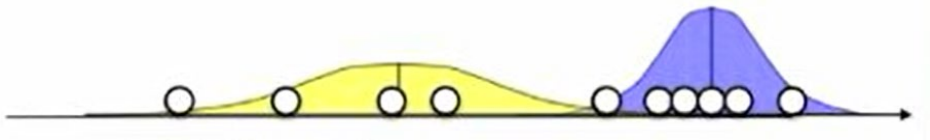


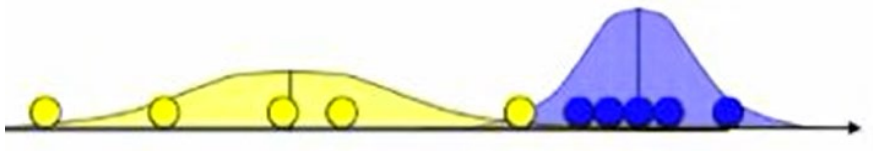
Chicken and egg problem

- Case 1: Need known sources to estimate $(\mu_a, \sigma_a^2\ )\ and\ (\mu_b, \sigma_b^2)$



- Case 2: Need $(\mu_a, \sigma_a^2\ )\ and\ (\mu_b, \sigma_b^2)$ to guess source of points

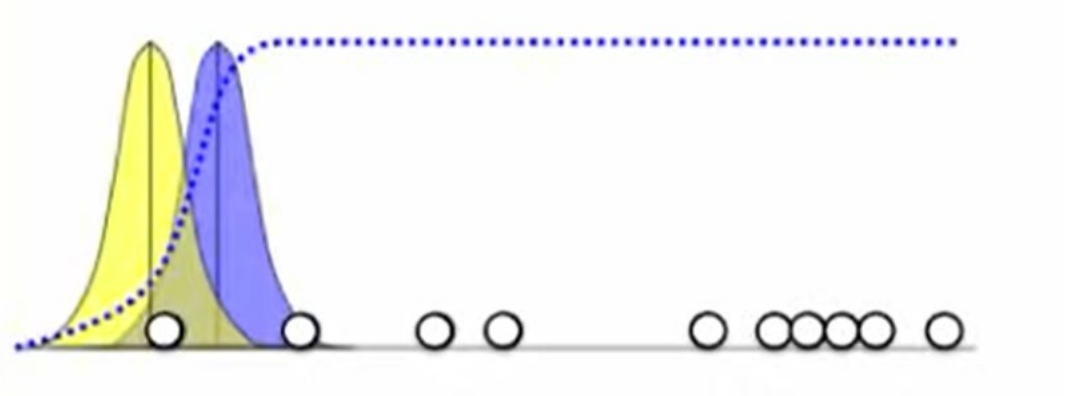




EM

- 1）Randomly choose (μ_a, σ_a^2) and (μ_b, σ_b^2)
- 2）For each point x_i, calculate $P(a \mid x_i)$ and $P(b \mid x_i)$ **E-step**
- 3）Adjust (μ_a, σ_a^2) and (μ_b, σ_b^2) **M-step**
- 4）Repeat 2）and 3）until convergence

1）Randomly choose (μ_a, σ_a^2) and (μ_b, σ_b^2)



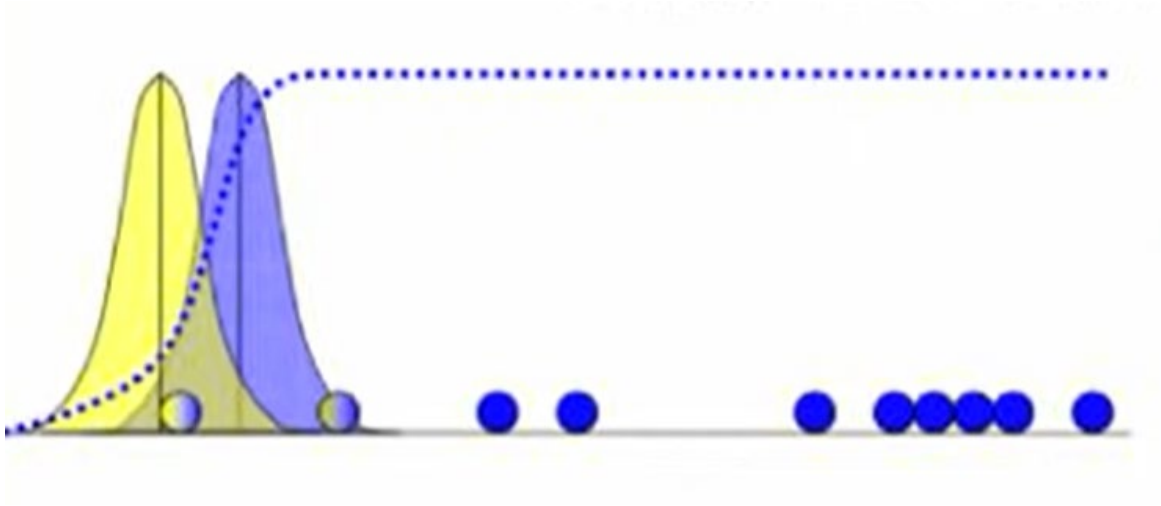
2) For each point x_i, calculate $P(a \mid x_i)$ and $P(b \mid x_i)$

- For each point x_i, calculate $P(a \mid x_i)$ and $P(b \mid x_i)$

$$P(x_i \mid a) = \frac{1}{\sqrt{2\pi\sigma_a^2}} e^{-\frac{(x_i - \mu_a)^2}{2\sigma_a^2}}$$

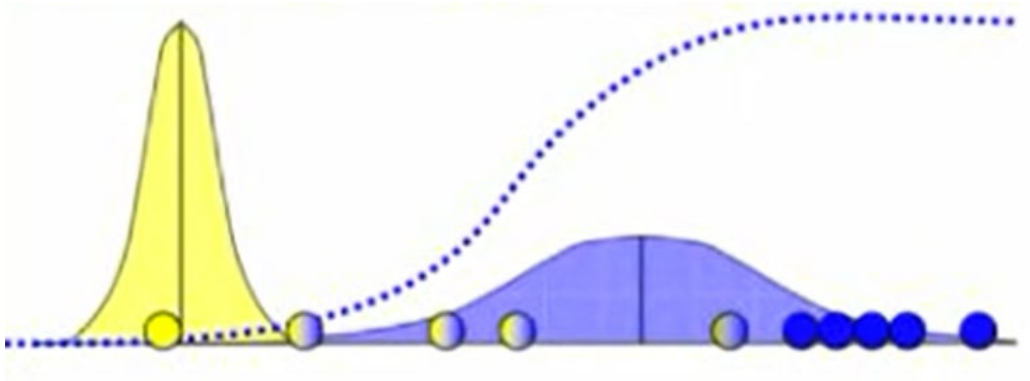
$$a_i = P(a \mid x_i) = \frac{P(x_i \mid a)\, P(a)}{P(x_i \mid a)\, P(a) + P(x_i \mid b)\, P(b)}$$

$$b_i = 1 - a_i$$



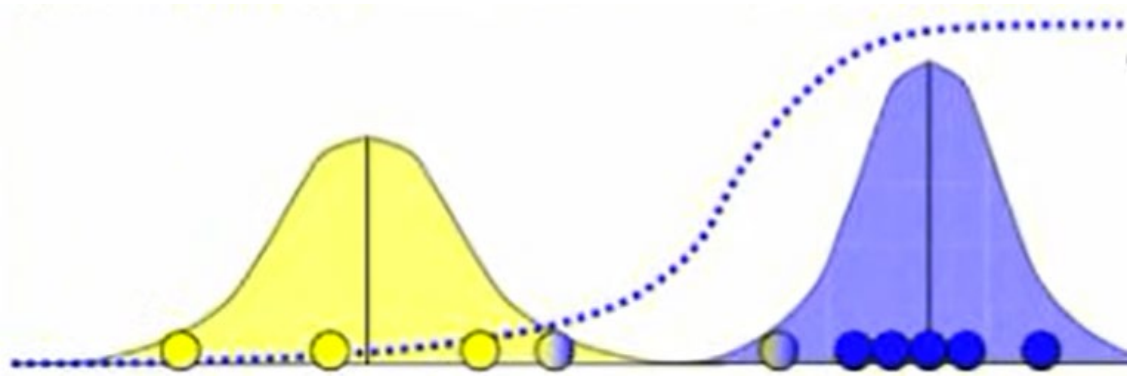
3) Adjust $(\mu_a, \sigma_a^2)\ and\ (\mu_b, \sigma_b^2)$

- Adjust $(\mu_a, \sigma_a^2)\ and\ (\mu_b, \sigma_b^2)$
- $\mu_a = \frac{a_1x_1+a_2x_2+\ \ldots+a_nx_n}{a_1+a_2+\cdots+a_n}$
- $\sigma_a^2 = \frac{a_1(x_1-\mu_a)^2+a_2(x_2-\mu_a)^2+\ \ldots+a_n(x_n-\mu_a)^2}{a_1+a_2+\cdots+a_n}$

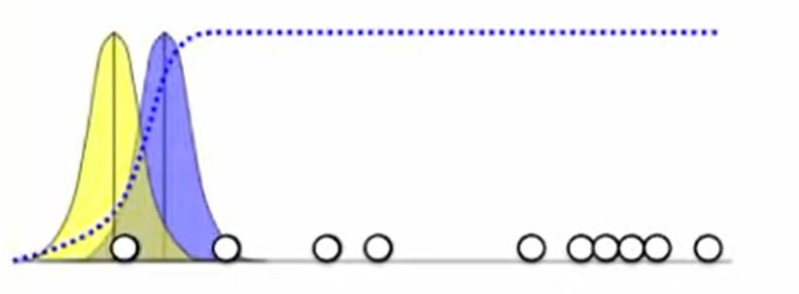


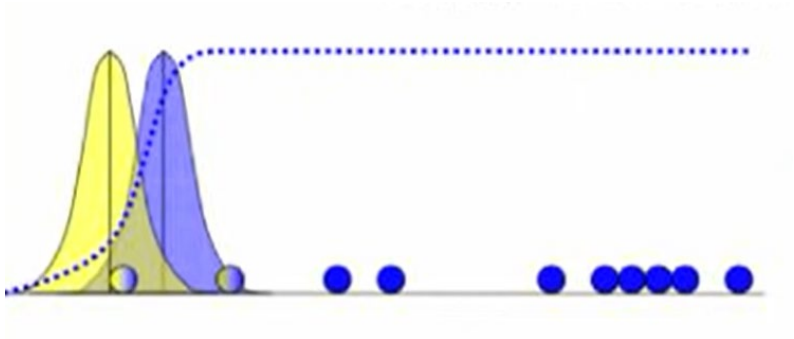
4）Repeat 2）and 3）until convergence

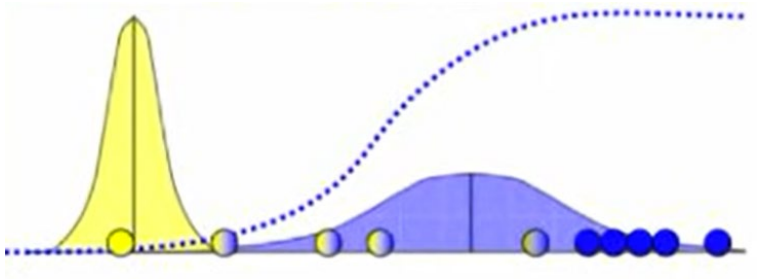
- Repeat until convergence

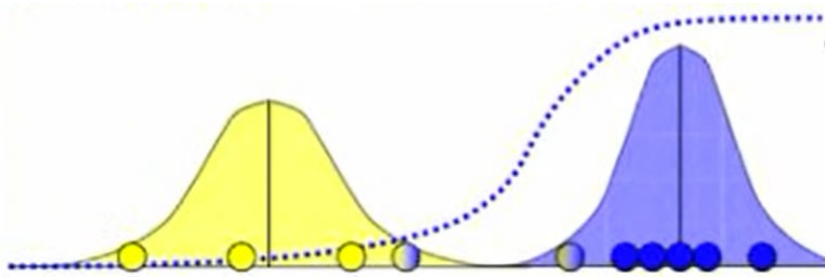


Process









THANKS