Advanced Machine Learning

Lecture 6: Mixture models fitting

Nora Ouzir : nora.ouzir@centralesupelec.fr
Lucca Guardiola : lucca.guardiola@centralesupelec.fr

Oct. - Nov. 2020

CentraleSupélec

Content

Mixture Models Fitting

- Data-to-knowledge
  - Statistical model fitting $\rightarrow$ model learning
  - Feature extraction: behavior, shapes...
  - Data characterisation $\rightarrow$ Complex modelling
- Complex estimation problems, e.g. many parameters, non parametric estimation...
- Clustering / Classification: Modes $\simeq$ clusters / classes
- Dealing with missing (latent) data: unknown labels can be generalized to unobserved data...

How to fit a mixture model to data? Inference/ Learning

Today’s Lecture

1. The Gaussian Mixture Model
    1. Two component case
    2. Generalization

2. EM algorithm

Today's course

1. The Gaussian Mixture Model
    1. Two component case
    2. Generalization

2. EM algorithm

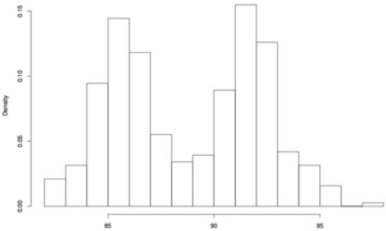
Gaussian Mixture Model

Example

Sizes of small animals coming from two different regions

Length	82	83	84	85	86	87	88	89
Observations	5	3	12	36	55	45	21	13

Length	90	91	92	93	94	95	96	98
Observations	15	34	59	48	16	12	6	1



Corresponding histogram

Whiteboard

Today's course

1. The Gaussian Mixture Model
    1. Two component case
    2. Generalization

2. EM algorithm

Gaussian Mixture Model: two component case

In our previous example...

There seems to be two separate underlying regimes, so we model X as a mixture of two normal distributions:

$$
\begin{aligned}
Y_1 &\sim \mathcal{N}(\mu_1, \sigma_1^2) \\
Y_2 &\sim \mathcal{N}(\mu_2, \sigma_2^2) \\
X &= Z\,Y_1 + (1 - Z)\,Y_2
\end{aligned}
$$

where $Z \sim \mathcal{B}(1, p)$

- $P(Z = 1) = p$ and $P(Z = 0) = 1 - p$.
- The data *follows the first distribution / belongs to the first cluster* with a probability p.

$\rightarrow$ Generative representation: generate $Z \in \{0, 1\}$ with probability p, and then depending on the outcome, deliver either Y_1 or Y_2.

Gaussian Mixture Model : two components

- Generative model $P(z,x) = P(z)P(x|z)$
- The pdf over x is defined by marginalizing (summing out z)

$$f_X(x) = \sum_{k=1}^{2} P(Z=k)P(x|Z=k)$$

Denote $\phi_\theta(x)$ the Gaussian PDF with parameters $\theta = (\mu, \sigma^2)$:

$\rightarrow$ PDF for X:

$$f_X(x) = p\,\phi_{\theta_1}(x) + (1-p)\,\phi_{\theta_2}(x)$$

$\rightarrow$ log-likelihood for n observations $(X_1, \ldots, X_n)$

$$\ell(\theta; \mathbf{x}) = \sum_{i=1}^{n} \log\left(p\,\phi_{\theta_1}(x_i) + (1-p)\,\phi_{\theta_2}(x_i)\right)$$

How to estimate the unknown parameters p, θ_1, θ_2? MLE...

Gaussian Mixture Model: MLE

Maximizing $\ell(\theta; \mathbf{x})$ is difficult...

- $\theta = (p, \theta_1, \theta_2)$, 5 unknown parameters in the simplest case...
- The sum inside the log couples all the parameters of all the component Gaussian distributions of the mixture

Idea: consider unobserved latent variables $(Z_1, \ldots, Z_n)$ where Z_i is the latent class of X_i $\rightarrow$ Computing MLEs becomes trivial...

$$\ell(\theta; \mathbf{x}, \mathbf{z}) = \sum_{i=1}^{n} \left(z_i \log(\phi_{\theta_1}(x_i)) + (1 - z_i) \log(\phi_{\theta_2}(x_i))\right) + \sum_{i=1}^{n} \left(z_i \log(p) + (1 - z_i) \log(1 - p)\right)$$

where $\mathbf{x} = (x_1, \ldots, x_n)$ and $\mathbf{z} = (z_1, \ldots, z_n)$.

Gaussian Mixture Model: MLE

Maximizing $\ell(\theta; \mathbf{x})$ is difficult...

- $\theta = (p, \theta_1, \theta_2)$, 5 unknown parameters in the simplest case...
- The sum inside the log couples all the parameters of all the component Gaussian distributions of the mixture $\rightarrow$ **Unseparable!**

Idea: consider unobserved latent variables $(Z_1, \ldots, Z_n)$ where Z_i is the latent class of X_i $\rightarrow$ Computing MLEs becomes trivial...

$$\ell(\theta; \mathbf{x}, \mathbf{z}) = \sum_{i=1}^{n} \left(z_i \log(\phi_{\theta_1}(x_i)) + (1 - z_i) \log(\phi_{\theta_2}(x_i))\right) + \sum_{i=1}^{n} \left(z_i \log(p) + (1 - z_i) \log(1 - p)\right)$$

where $\mathbf{x} = (x_1, \ldots, x_n)$ and $\mathbf{z} = (z_1, \ldots, z_n)$. $\rightarrow$ **Separable!**

But Z is unknown in practice...

Gaussian Mixture Model : posterior inference

- Let's consider that the parameters are known
- A GMM with known parameters defines a joint distribution over (X_i, Z_i) $\rightarrow$ probabilistic/posterior inference

We infer the posterior over Z using Bayes' rule (*e.g., k=1*):

$$
\begin{aligned}
P(Z_i = 1|x_i) &= \frac{P(Z_i = 1)P(X_i|Z_i = 1)}{P(X_i)} \\
&= \frac{p_1\phi_{\theta_1}(x_i)}{p\phi_{\theta_1}(x_i) + (1-p)\phi_{\theta_2}(x_i)}
\end{aligned}
$$

Responsibility γ_i

The expected value of Z_i conditional to the observed data and known parameters

$$\gamma_i^k(\theta) = E[Z_i|\theta, \mathbf{x}] = P(Z_i = k|\theta, \mathbf{x})$$

Gaussian Mixture Model : EM algorithm

- Chicken and egg problem
- Use an iterative approach : alternately fix the parameters/the latent variables

Algorithm: Expectation-Maximization (EM)

- Random initialization of $\theta^{(0)}$
- Repeat until CV for $t = 0, 1, \ldots$
  (a) **E-Step:** Compute the responsibilities
  $$\hat{\gamma}_i = \frac{\hat{p}\,\phi_{\hat{\theta}_1}(x_i)}{\hat{p}\,\phi_{\hat{\theta}_1}(x_i) + (1-\hat{p})\,\phi_{\hat{\theta}_2}(x_i)}, \text{ for } i = 1, \ldots, n$$
  (b) **M-Step:** Compute the parameters...
  $$\hat{\mu}_1 = \frac{\sum_i \hat{\gamma}_i x_i}{\sum_i \hat{\gamma}_i}, \hat{\sigma}_1^2 = \frac{\sum_i \hat{\gamma}_i (x_i - \hat{\mu}_1)^2}{\sum_i \hat{\gamma}_i}, \ldots \text{ and } \hat{p} = \sum_i \hat{\gamma}_i / n.$$

Today's course

1. The Gaussian Mixture Model
    1. Two component case
    2. Generalization

2. EM algorithm

Mixture Model

Goal: Model the statistical behaviour of several populations, groups or classes...

- different objects x_i in an image containing N pixels
- population of animals: x_i corresponds to the size of the i^{th} animal, classes correspond to age/sex/origin (young, old, female, male)...

- n observations of i.i.d. random variables/vectors $(X_1, \ldots, X_n)$
- K different clusters containing n_k observations with $n = \sum_{k=1}^{K} n_k$
- p_k the probability of belonging to the k^{th} class and f_k the PDF of r.v. in this class.

PDF of a mixture

$$f(x) = \sum_{k=1}^{K} p_k \times f_k(x)$$

Gaussian Mixture Model: GMM

Gaussian Mixture Model

$$f(x) = \sum_{k=1}^{K} p_k \times \frac{1}{\sqrt{2\pi\sigma_k^2}} \exp\left(-\frac{(x-\mu_k)^2}{2\sigma_k^2}\right)$$

with $\sum_{k=1}^{K} p_k = 1$ and $\forall k \in \{1, \ldots, K\}, \mu_k \in \mathbb{R}, \sigma_k \in \mathbb{R}_+^*$.

Challenges

- Many unknown parameters $\theta = (p_k, \mu_k, \sigma_k)_{k=1,\ldots,K}$
- What about K ? Known, unknown ?

But useful for modelling a wide range of distributions!

GMMs: Examples

(a) $\frac{1}{5}\mathcal{N}(0,1) + \frac{1}{5}\mathcal{N}(1/2,(2/3)^2) + \frac{3}{5}\mathcal{N}(13/15,(5/9)^2)$,

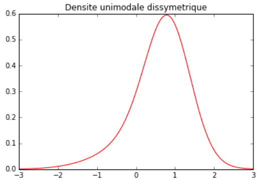
(b) $\sum_{k=0}^{7} \mathcal{N}(3((2/3)^k - 1),(2/3)^{2k})$

(c) $\frac{1}{2}\mathcal{N}(-1,(2/3)^2) + \frac{1}{2}\mathcal{N}(1,(2/3)^2)$

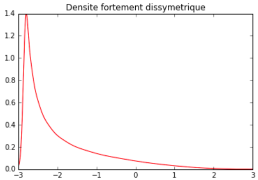
(d) $\frac{3}{4}\mathcal{N}(0,1) + \frac{1}{4}\mathcal{N}(3/2,(1/3)^2)$

(e) $\frac{9}{2}0\mathcal{N}(-6/5,(3/5)^2) + \frac{9}{2}0\mathcal{N}(6/5,(3/5)^2) + \frac{1}{1}0\mathcal{N}(0,(1/4)^2)$

(f) $\frac{1}{2}\mathcal{N}(0,1) + \sum_{k=-2}^{2} \frac{2^{1-k}}{31}\mathcal{N}(k+1/2,(2^{-k}/10)^2)$

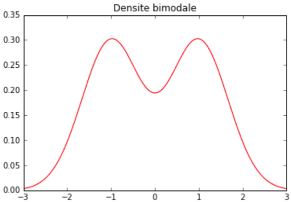


(a) Asymmetric unimodal PDF

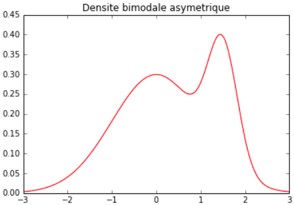


(b) Strongly asymmetric unimodal PDF

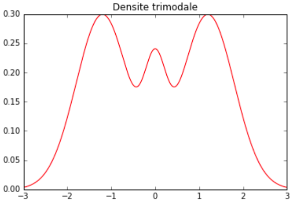
GMMs: Examples



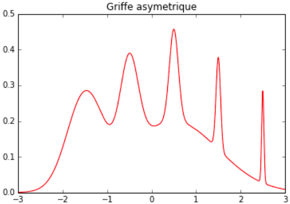
(c) Bimodal PDF



(d) Asymmetric bimodal PDF



(e) Tri-modal PDF



(f) More complex PDF

GMM: simulation

In order to simulate the mixture $f(x) = \sum_{k=1}^{K} p_k \times \frac{1}{\sqrt{2\pi\sigma_k^2}} \exp\left(-\frac{(x-\mu_k)^2}{2\sigma_k^2}\right)$, one needs to introduce a latent variable Z (or missing data) corresponding to the class of the variable X.

Now, the complete data $T = (X, Z)$ is defined by:

- Z follows a discrete distribution $(p_1, \dots, p_K)$ on $\{1, \dots, K\}$ such that $\forall k$, one has (Multinomial distribution)

$$P(Z = k) = p_k, \text{ with } \sum_k p_k = 1$$

- $\forall k \in \{1, \dots, K\}$, conditionally to $\{Z = k\}$, X has a PDF f_k:

$$\mathcal{L}(x|Z = k) = f_k(x)$$

$\rightarrow$ Goal: estimation of $\theta = (p_k, \mu_k, \sigma_k)_{k=1,\dots,K}$

Today’s course

1. The Gaussian Mixture Model
    1. Two component case
    2. Generalization

2. EM algorithm

Reminders: Bayesian probabilities/statistics

For two events (or r. v. ...), one has:

- Conditional probabilities

$$P(A|B) = \frac{P(A \cap B)}{P(B)}$$

- Bayes rule

$$P(B|A) = \frac{P(A|B)\,P(B)}{P(A)}$$

- if $B_1, \ldots, B_n$ is a partition of Ω, i.e. $\bigcup_{i=1}^{n} B_i = \Omega$ and $\forall i \neq j, B_i \cap B_j = \emptyset$, then

$$P(A) = \sum_{i=1}^{n} P(A \cap B_i)$$

EM algorithm

Let us start by considering Z known

- we observe $(x_i, z_i)_{i=1,\ldots,n}$ instead of (only) $(x_i)_{i=1,\ldots,n}$
- this is the maximum-likelihood step $\rightarrow$ again trivial!

ML estimates of θ: K classes

Let the observations be $(x_i, z_i)_{i=1,\ldots,n}$, then $\forall k \in \{1, \ldots, K\}$, one has

$$\hat{p}_k = \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}_{z_i=k} \tag{1}$$

$$\hat{\mu}_k = \frac{1}{n\,\hat{p}_k}\sum_{i|z_i=k} x_i \tag{2}$$

$$\hat{\sigma}_k^2 = \frac{1}{n\,\hat{p}_k}\sum_{i|z_i=k} (x_i - \hat{\mu}_k)^2 \tag{3}$$

EM algorithm

However one only observes $(x_1, \ldots, x_n)$ and again...

Maximizing $\ell(x_1, ..., x_n; \theta)$ is difficult

$$\ell_{obs}(x_1, \ldots, x_n; \theta) = \sum_{i=1}^{n} \log \left(\sum_{k=1}^{K} p_k \times f_k(x_i) \right)$$

where $\theta = (p_k, \mu_k, \sigma_k)_{k=1,\ldots,K}$

BUT

EM algorithm

However one only observes $(x_1, \dots, x_n)$ and again...

Maximizing $\ell(x_1, ..., x_n; \theta)$ is difficult

$$\ell_{obs}(x_1, \dots, x_n; \theta) = \sum_{i=1}^{n} \log \left(\sum_{k=1}^{K} p_k \times f_k(x_i) \right)$$

where $\theta = (p_k, \mu_k, \sigma_k)_{k=1,\dots,K}$

BUT one can make assumptions on the unobserved $(Z_1, \dots, Z_n)$

EM algorithm

However one only observes $(x_1, \ldots, x_n)$ and again...

Maximizing $\ell(x_1, ..., x_n; \theta)$ is difficult

$$\ell_{obs}(x_1, \ldots, x_n; \theta) = \sum_{i=1}^{n} \log \left(\sum_{k=1}^{K} p_k \times f_k(x_i) \right)$$

where $\theta = (p_k, \mu_k, \sigma_k)_{k=1,...,K}$

BUT one can make assumptions on the unobserved $(Z_1, \ldots, Z_n)$

For $\theta \in \Theta, x \in \mathbb{R}$ and $k \in \{1, \ldots, K\}$, one has

$$P_\theta (Z = k | X = x) = \frac{p_k \times f_k(x)}{\sum_{l=1}^{K} p_l \times f_l(x)} \tag{4}$$

$\rightarrow$ Intuition: thanks to some θ_{old}, one can assign a z_i to each x_i (4) and thanks to (1-3), one can compute a θ_{new}...

Whiteboard

General EM algorithm : variants

k-means

Hard assignment: Assign a class to each x_i according to

$$z_i = \arg\max_k P_{\theta_{old}}(Z = k | X_i = x_i)$$

SEM

Randomly assign a class to each x_i according to the distribution

$$P_{\theta_{old}}(Z = .|X_i = x_i)$$

More flexible!

N-SEM

Randomly assign N classes to each x_i

EM: Limit of N-SEM when $N \to \infty$ Very flexible and robust!

k-means

$\rightarrow$ One has to make very strong assumptions:

$$p_1 = \ldots = p_K = \frac{1}{K} \text{ and } \sigma_1 = \ldots = \sigma_K$$

$$\forall \theta, \forall x \in \mathbb{R} \quad \arg\max_k P_\theta \left(Z = k | X = x\right) = \arg\min_k |x - \mu_k|$$

k-means

- Randomly initialize $(z_1, \ldots, z_K)$
- Repeat until CV:
  - for $k \in \{1, \ldots, K\}$, $\mu_k = \frac{1}{n} \sum_{i=1}^{n} x_i \, \mathbb{1}_{z_i = k}$
  - for $i \in \{1, \ldots, n\}$, $z_i = \arg\min_k |x - \mu_k|$

Advantages / Drawbacks ...

Stochastic EM

→ General idea: Stochastic version of the k-means algorithm...

SEM

- Randomly initialize $(z_1, \ldots, z_K)$
- Repeat until CV:
  (a) Compute (MLE)

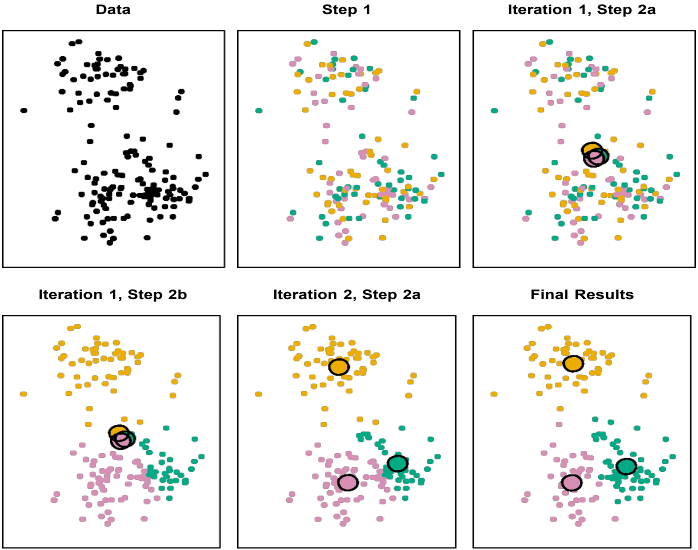
$$\hat{\theta} = \arg\max_{\theta} \ell_{obs}[(x_1, z_1), \ldots, (x_n, z_n); \theta]$$

  (b) for $i \in \{1, \ldots, n\}$, randomly choose z_i according to

$$P_{\hat{\theta}}(Z = .|X_i = x_i)$$

  given by Eq. (4).

Stochastic EM



Stochastic EM - N trials

N-SEM (1)

- Replicate the observations N times: $(x_1, \ldots, x_n) \to \left(x_i^{(j)}\right)_{1 \leq i \leq n, 1 \leq j \leq N}$
- Apply SEM algo to this new dataset.

N-SEM (2)

- Randomly initialize N classes $z_i^1, \ldots, z_i^N \in \{1, \ldots, K\}, \forall i$
- Repeat until CV
  (a) Compute (MLE)
  $$\hat{\theta} = \arg\max_{\theta} \ell_{obs}\left((x_i, z_i^1)_{i=1,\ldots,n} \cup \ldots \cup (x_i, z_i^N)_{i=1,\ldots,n}; \theta\right)$$
  (b) for $i \in \{1, \ldots, n\}$, randomly choose $z_i^1, \ldots, z_i^N$ (independently!) according to
  $$P_{\hat{\theta}}\left(Z = .|X_i = x_i\right)$$
  given by Eq. (4).

EM algorithm

→ General idea: N-SEM when $N \to +\infty$ ...

Given $(x_i)_{1\leq i\leq n}$ and associated classes for N trials $(z_i^k)_{1\leq i\leq n, 1\leq k\leq K}$:

$$\forall\theta, \ell_{obs}\left(\left(x_i, z_i^1\right)_{i=1,\ldots,n} \cup \ldots \cup \left(x_i, z_i^N\right)_{i=1,\ldots,n}; \theta\right) = \sum_{j=1}^N \ell_{obs}\left(\left(x_i, z_i^j\right)_{i=1,\ldots,n}; \theta\right)$$

Theorem [Part I]

Given the observations $(x_i)_{1\leq i\leq n}$ and $\theta_{old} \in \Theta$.

(a) Let $Z_1, ..., Z_n$ independent r.v. such that $Z_i \sim \mathcal{L}_{\theta_{old}}(Z|X = x_i)$. One has $\forall\theta = (p_k, \mu_k, \sigma_k)_{1\leq k\leq K} \in \Theta$,

$$E[\ell\left(\left(x_i, z_i\right)_{i=1,\ldots,n}; \theta\right)] = \sum_{i=1}^n \sum_{k=1}^K P_{\theta_{old}}(Z = k|X = x_i) \log\left(p_k \times f_k(x_i)\right)$$

where $P_{\theta_{old}}(Z = .|X = x_i)$ given by Eq. (4).

EM algorithm

Theorem [Part II] Given the observations $(x_i)_{1\le i\le n}$ and $\theta_{old} \in \Theta$,

(b) One has that $\arg\max_\theta E[\ell\left((x_i, z_i)_{i=1,\dots,n}\,; \theta\right)]$ is given by:

- Class probabilities: $\forall k = 1, \dots, K$,

$$p_k^{argmax} = \frac{1}{n}\sum_{i=1}^{n} P_{\theta_{old}}\left(Z = k | X = x_i\right)$$

- Class means: $\forall k = 1, \dots, K$,

$$\mu_k^{argmax} = \frac{1}{n\, p_k^{argmax}}\sum_{i=1}^{n} P_{\theta_{old}}\left(Z = k | X = x_i\right)\, x_i$$

- Class variances: $\forall k = 1, \dots, K$,

$$(\sigma_k^{argmax})^2 = \frac{1}{n\, p_k^{argmax}}\sum_{i=1}^{n} P_{\theta_{old}}\left(Z = k | X = x_i\right)\, (x_i - \mu_k^{argmax})^2$$

Expectation-Maximization algorithm

$\rightarrow$ So far, our theoretical algorithm looks like...

EM: Theory

- Randomly initialization of θ_0
- Repeat until CV for $t = 0, 1, \ldots$
  (a) **E-Step:** Compute

$$L_t(\theta) = E\left[\ell\left(\left(X_i, Z_i^t\right)_{i=1,\ldots,n}; \theta\right)\right] \; \left(Q(\theta, \theta_t) = E\left(l(\theta; \mathbf{t}) | \mathbf{x}, \theta_t\right)\right)$$

  where $Z_1^t, \ldots, Z_n^t$ are i.i.d. with $Z_i^t \sim \mathcal{L}_{\theta_t}(Z | X = x_i)$

  (b) **M-Step:** Maximize $L_t(\theta)$ to obtain $\theta_{t+1} = \arg\max_\theta L_t(\theta)$

- **E** for *Expectation*
- **M** for *Maximization*

Whiteboard

Whiteboard

A different view - *Maximization-Maximization*

- Consider the function $F(\theta, \mathbf{P}) = E_{\mathbf{P}}[l_0(\theta; \mathbf{t})] - E_{\mathbf{P}}[\log(\mathbf{P}(\mathbf{z}))]$
- $\mathbf{P}$ can be any distribution for the *latent* variables $\mathbf{z}$.
- Note that F evaluated at $\mathbf{P}(\mathbf{z}) = P(\mathbf{z}|\mathbf{x}, \theta)$ is the log-likelihood of the observed data.
- EM algo can be viewed as a joint maximization method for F over θ and $\mathbf{P}(\mathbf{z})$. Maximizer over $\mathbf{P}(\mathbf{z})$ for fixed θ can be shown to be $\mathbf{P}(\mathbf{z}) = P(\mathbf{z}|\mathbf{x}, \theta)$. (dist. computed at the E-step).
- M-step: Maximize $F(\theta, \mathbf{P})$ over θ for fixed $\mathbf{P}(\mathbf{z})$, $\iff$ maximizing $E_{\mathbf{P}}[l_0(\theta; \mathbf{t})|\mathbf{x}, \theta^*]$ (2nd term do not depend on θ).

Since $F(\theta, \mathbf{P})$ and the obs. data log-likelihood agree when $\mathbf{P}(\mathbf{z}) = P(\mathbf{z}|\mathbf{x}, \theta)$, maximization of the former accomplishes maximization of the latter.

EM algorithm: In practice

EM Algorithm

- Randomly initialization of θ_0
- Repeat until CV for $t = 0, 1, \ldots$
  (a) **E-Step:** Compute the matrix $(1 \leq i \leq n, 1 \leq k \leq K)$
  $$[P_{\theta_t}(Z = k | X = x_i)] = \left[\frac{p_k^t \times f_{k,t}(x_i)}{\sum_{l=1}^{K} p_l^t \times f_{l,t}(x_i)} \right]$$
  (b) **M-Step:** Compute θ_{t+1}, for all $k = 1, \ldots, K$,
  $$\hat{p}_k^{t+1} = \frac{1}{n} \sum_{i=1}^{n} P_{\theta_t}(Z = k | X = x_i), \tag{5}$$
  $$\hat{\mu}_k^{t+1} = \frac{1}{n\, \hat{p}_k^{t+1}} \sum_{i=1}^{n} x_i P_{\theta_t}(Z = k | X = x_i) \tag{6}$$
  $$\left(\hat{\sigma}_k^{t+1}\right)^2 = \frac{1}{n\, \hat{p}_k^{t+1}} \sum_{i=1}^{n} P_{\theta_t}(Z = k | X = x_i) \left(x_i - \hat{\mu}_k^{t+1}\right)^2 \tag{7}$$

EM example