The Expectation-Maximization algorithm

Masmoudi Afif

Laboratoire de probabilités et statistique
Université de Sfax
Journées de Statistique Mathématique et Data Science
(JSMDS 2019)

Plan

Plan

Plan

Plan

Introduction

- A finite mixture models is a combination of two or more probability density functions.
- It is a probabilistic model used to represent a subpopulation within an overall population.
- It represents a natural extension of the finite Gaussian mixture of distributions.
- It has been adopted in such diverse areas as ecology, bioinformatics, computer science and biostatistics..

Plan

A Finite Gaussian mixture model

Gaussian mixture distribution

The distribution f is a mixture of K components Gaussian distributions defined by

$$f(x) = f(x; \Theta) = \sum_{k=1}^{K} \pi_k N_k(x; \mu_k, \sigma_k) \tag{1}$$

where,

- π_k is the mixing weights, verify $0 < \pi_k < 1$, and $\sum_{k=1}^{K} \pi_k = 1$.

A Finite Gaussian mixture models

- $\Theta = (\pi_1, \ldots, \pi_K, \mu_1, \ldots, \mu_K, \sigma_k, \ldots, \sigma_K)$ is the parameter vector of the Gaussian mixture models.

Plan

Parameters estimation

Let $(X_1, ..., X_N)$ be N independent random variables with the same Gaussian mixture density $f(.; \Theta)$ and let $(x_1, ..., x_N)$ be N associated observations.

The incomplete likelihood function

The incomplete likelihood function l is given by

$$\begin{aligned} l(x_1, x_2, \ldots, x_N; \Theta) &= \prod_{i=1}^{N} f(x_i; \Theta) \\ &= \prod_{i=1}^{N} \left(\sum_{k=1}^{K} \pi_k \frac{1}{\sigma_k \sqrt{2\pi}} e^{\frac{-1}{2}\left(\frac{x_i - \mu_k}{\sigma_k}\right)^2} \right) \end{aligned}$$

Parameters estimation

The incomplete log -likelihood

the log-likelihood L is given by

$$\begin{aligned} L(x_1, x_2, \ldots, x_N; \Theta) &= \log l(x_1, x_2, \ldots, x_N; \Theta) \\ &= \sum_{i=1}^{N} \log \left(\sum_{k=1}^{K} \pi_k N_k(x_i; \sigma_k, \mu_k) \right) \\ &= \sum_{i=1}^{N} \log \left(\sum_{k=1}^{K} \pi_k \frac{1}{\sigma_k \sqrt{2\pi}} e^{\frac{-1}{2} \left(\frac{x_i - \mu_k}{\sigma_k} \right)^2} \right) \end{aligned}$$

Parameters estimation

For each random variable X_i, we associate a random vector $Z_i = (Z_{i1}, ..., Z_{iK})$ following a multivariate Bernoulli distribution with vector parameters $(\pi_1, ..., \pi_K)$. That is

$$\mathbb{P}(Z_i = z_i) = \prod_{k=1}^{K} \pi_k^{z_{ik}}, \tag{2}$$

where

- $z_i = (z_{i1}, ..., z_{iK}) \in \{0, 1\}^K$, and $\sum_{k=1}^{K} z_{ik} = 1$.

Parameters estimation

The maximum likelihood function from complete data

The maximum likelihood function from complete data l_c is given by

$$
\begin{aligned}
l_c(x_1, .., x_N, z_1, .., z_N; \Theta) &= \prod_{i=1}^{N} \mathbb{P}(Z_i = z_i) \prod_{k=1}^{K} N_k^{z_{ik}}(x_i; \sigma_k, \mu_k) \\
&= \prod_{i=1}^{N} \Big(\prod_{k=1}^{K} \big(\pi_k N_k(x_i; \sigma_k, \mu_k)\big)^{z_{ik}} \Big) \\
&= \prod_{i=1}^{N} \Big(\prod_{k=1}^{K} \Big(\pi_k \frac{1}{\sigma_k \sqrt{2\pi}} e^{\frac{-1}{2}\left(\frac{x_i - \mu_k}{\sigma_k}\right)^2}\Big)^{z_{ik}} \Big).
\end{aligned}
$$

The log-likelihood function from complete data

The log-likelihood function from complete data L_c is given by

$$\begin{aligned} L_c(x_1, \ldots, x_N, z_1, \ldots, z_N; \Theta) &= \sum_{i=1}^{N} \sum_{k=1}^{K} z_{ik} \log \left(\pi_k N_k(x_i; \sigma_k, \mu_k) \right) \\ &= \sum_{i=1}^{N} \sum_{k=1}^{K} z_{ik} \log \left(\pi_k \frac{1}{\sigma_k \sqrt{2\pi}} e^{\frac{-1}{2} \left(\frac{x_i - \mu_k}{\sigma_k} \right)^2} \right) \end{aligned}$$

where $\sum_{k=1}^{K} z_{ik} = 1$.

The conditional expectation of the complete-data log likelihood

$$\begin{aligned}
\mathbb{Q}(\Theta||\Theta^{(l)}) &= \mathbb{E}\left(L_c\left(x_1,..,x_N,z_1,..,z_N;\Theta^{(l)}\right)|x_1,..,x_N\right) \\
&= \sum_{i=1}^{N}\sum_{k=1}^{K} {\tau_{ik}}^{(l)} \log\left(\pi_k N_k(x_i;\mu_k,\sigma_k)\right). \\
&= \sum_{i=1}^{N}\sum_{k=1}^{K} {\tau_{ik}}^{(l)} \log\left(\pi_k \frac{1}{\sigma_k\sqrt{2\pi}} e^{\frac{-1}{2}\left(\frac{x_i-\mu_k}{\sigma_k}\right)^2}\right)
\end{aligned}$$

Where

The conditional expectation of the complete-data log likelihood

The posterior probability

$$\begin{aligned}
\tau_{ik}{}^{(l)} &= \mathbb{E}\left(Z_{ik}|\Theta^{(l)}, X_1 = x_1, .., X_N = x_N\right) \\
&= \mathbb{E}\left(Z_{ik} = z_{ik}|X_i = x_i, \Theta^{(l)}\right) \\
&= \frac{\pi_k^{(l)} N_k(x_i; \mu_k^{(l)}, \sigma_k^{(l)})}{\sum_{k=1}^{K} \pi_k^{(l)} N_k(x_i; \mu_k^{(l)}, \sigma_k^{(l)})},
\end{aligned}$$

The conditional expectation of the complete-data log likelihood

and

$\Theta = (\mu_1, .., \mu_K, \pi_1, .., \pi_K, \sigma_1, .., \sigma_K)$ is the parameter vector of the Gaussian mixture models in the l^{th} iteration.

The conditional expectation of the complete-data log likelihood

Challenge

Maximization of the conditional expectation log likelihood from complete data $\mathbb{Q}(\Theta||\Theta^{(l)})$.

Plan

The EM Algorithm

Definition

The EM algorithm is an iterative method for finding maximum likelihood or maximum a posteriori (MAP) estimates of parameters in statistical models involving incomplete data.

The EM algorithm

The EM algorithm steps

- **Step** 1 : **Initialization**: $\Theta^{(0)}$
- **Step** 2 : **Expectation(E)**

We calculate the posterior probability

$$\tau_{ik}^{(l)} = \frac{\pi_k^{(l)} N_k(x_i; \mu_k^{(l)}, \sigma_k^{(l)})}{\sum_{k=1}^{K} \pi_k^{(l)} N_k(x_i; \mu_k^{(l)}, \sigma_k^{(l)})} \tag{3}$$

The EM algorithm

The EM algorithm steps

and we calculate

$$\mathbb{Q}(\Theta||\Theta^{(l)}) = \sum_{i=1}^{N}\sum_{k=1}^{K} \tau_{ik}{}^{(l)} \log\left(\pi_k N_k(x_i;\mu_k,\sigma_k)\right). \quad (4)$$

- **Step 3 : Maximization (M)**

$$\Theta^{(l+1)} = \arg\max_{\Theta} \mathbb{Q}(\Theta||\Theta^{(l)})$$

The EM algorithm

The EM algorithm steps

The maximum likelihood estimators of μ_k and π_k are given respectively by

$$\mu_k^{(l+1)} = \frac{\sum_{i=1}^{N} \tau_{ik}^{(l)} x_i}{\sum_{i=1}^{N} \tau_{ik}^{(l)}} \tag{5}$$

$$\pi_k^{(l+1)} = \frac{1}{N} \sum_{i=1}^{N} \tau_{ik}^{(l)} \tag{6}$$

The EM algorithm

The EM algorithm steps

The maximum likelihood estimator of σ_k is given by

$$\sigma_k{}^{(l+1)} = \sqrt{\frac{\sum_{i=1}^{N} \tau_{ik}^{(l)} (x_i - \mu_k^{(l+1)})^2}{\sum_{i=1}^{N} \tau_{ik}^{(l)}}} \tag{7}$$

Plan

The EM algorithm properties

Proposition 1

At each iteration of the EM algorithm,

$$L\left(x_1, x_2, ..., x_N; \Theta^{(l+1)}\right) \geqslant L\left(x_1, x_2, ..., x_N; \Theta^{(l)}\right), \tag{8}$$

where, $\Theta^{(l+1)}$ is the parameter vector in the $(l+1)$ iteration.

Then,

$$\mathbb{Q}(\Theta^{(l+1)}||\Theta^{(l)}) \geqslant \mathbb{Q}(\Theta^{(l)}||\Theta^{(l)}) \tag{9}$$

Proof of Proposition 1

Let $\mathbb{Z} = (Z_1 = z_1, \ldots, Z_N = z_N)$ the missing data vector and $\mathbb{X} = (X_1 = x_1, \ldots, X_N = x_N)$ the observed data vector.

We can see that

$$\begin{aligned}
\log l\left(\mathbb{X}|\Theta^{(l)}\right) &= \sum_{\mathbb{Z}} l\left(\mathbb{Z}|\mathbb{X}, \Theta^{(l)}\right) \log \frac{l_c\left(\mathbb{Z}, \mathbb{X}; \Theta^{(l)}\right)}{l\left(\mathbb{Z}|\mathbb{X}, \Theta^{(l)}\right)} \\
&= \sum_{\mathbb{Z}} l\left(\mathbb{Z}|\mathbb{X}, \Theta^{(l)}\right) \log l_c\left(\mathbb{Z}, \mathbb{X}|\Theta^{(l)}\right) \\
&- \sum_{\mathbb{Z}} l\left(\mathbb{Z}|\mathbb{X}, \Theta^{(l)}\right) \log l\left(\mathbb{Z}|\mathbb{X}, \Theta^{(l)}\right) \\
&= \mathbb{E}\left(\log l_c\left(\mathbb{Z}, \mathbb{X}|\Theta^{(l)}\right)|\mathbb{X}, \Theta^{(l)}\right) \\
&- \mathbb{E}\left(\log l\left(\mathbb{Z}|\mathbb{X}, \Theta^{(l)}\right)|\mathbb{X}, \Theta^{(l)}\right).
\end{aligned}$$

Proof of Proposition 1

So, we have

$$\begin{aligned}\mathbb{Q}(\Theta^{(l)}||\Theta^{(l)}) &= \mathbb{E}\big(\log l_c(\mathbb{Z},\mathbb{X}|\Theta^{(l)})|\mathbb{X},\Theta^{(l)}\big)\\ &= \mathbb{H}(\Theta^{(l)}||\Theta^{(l)}) + \log l\big(\mathbb{X}|\Theta^{(l)}\big),\end{aligned}$$

Proof of Proposition 1

where

$$\mathbb{H}(\Theta^{(l)}||\Theta^{(l)}) = \mathbb{E}\big(\log l\big(\mathbb{Z}|\mathbb{X}, \Theta^{(l)}\big)|\mathbb{X}, \Theta^{(l)}\big). \tag{10}$$

Proof of Proposition 1

and further we have

$$\begin{aligned}
\log l(\mathbb{X};\Theta) &= \log\Big(\sum_{\mathbb{Z}} l_c(\mathbb{Z},\mathbb{X}|\Theta)\Big) \\
&= \log\Big(\sum_{\mathbb{Z}} l_c(\mathbb{Z},\mathbb{X}|\Theta)\Big)\frac{l(\mathbb{Z}|\mathbb{X},\Theta^{(l)})}{l(\mathbb{Z}|\mathbb{X},\Theta^{(l)})} \\
&= \log\mathbb{E}\Big[\frac{l_c(\mathbb{Z},\mathbb{X}|\Theta)}{l(\mathbb{Z}|\mathbb{X},\Theta^{(l)})}|\mathbb{X},\Theta^{l}\Big].
\end{aligned}$$

Proof of Proposition 1

As though, the log function is concave so, by Jensen's inequality, we obtain

$$\begin{aligned}\log l(\mathbb{X}|\Theta) = \log l(\Theta) &\geq \mathbb{E}[\log\left(\frac{l_c(\mathbb{Z},\mathbb{X}|\Theta)}{l(\mathbb{Z}|\mathbb{X},\Theta^{(l)})}\right)|\mathbb{X},\Theta^{(l)}] \\ &\geq \mathbb{E}\big[\log l_c(\mathbb{Z},\mathbb{X}|\Theta)|\mathbb{X},\Theta^{(l)}\big] \\ &- \mathbb{E}\big[\log\left(l(\mathbb{Z}|\mathbb{X},\Theta^{(l)})\right)|\mathbb{X},\Theta^{(l)}\big].\end{aligned}$$

Proof Proposition 1

Hence, we get the following result

$$\log l\left(\mathbb{X}|\Theta\right) \geq \mathbb{Q}(\Theta||\Theta^{(l)}) - \mathbb{H}(\Theta^{(l)}||\Theta^{(l)}). \tag{11}$$

Since, we have

$$\Theta^{(l+1)} = \arg\max_{\Theta} \mathbb{Q}(\Theta||\Theta^{(l)}).$$

Then, as result

$$\mathbb{Q}(\Theta^{(l+1)}||\Theta^{(l)}) \geqslant \mathbb{Q}(\Theta^{(l)}||\Theta^{(l)}) \tag{12}$$

Proof Proposition 1

we can conclude that

$$\begin{aligned} & \log l\left(\mathbb{X}|\Theta^{(l+1)}\right) - \log l\left(\mathbb{X}|\Theta^{(l)}\right) \\ \geq & \ \mathbb{Q}(\Theta^{(l+1)}||\Theta^{(l)}) - \mathbb{H}(\Theta^{(l)}||\Theta^{(l)}) - \mathbb{Q}(\Theta^{(l)}||\Theta^{(l)}) + \mathbb{H}(\Theta^{(l)}||\Theta^{(l)}) \\ \geq & \ \mathbb{Q}(\Theta^{(l+1)}||\Theta^{(l)}) - \mathbb{Q}(\Theta^{(l)}||\Theta^{(l)}) \geq 0 \end{aligned}$$

which completes this proof.

The EM algorithm properties

Theorem 1

Suppose that, we have

1) $\lim_{l \to +\infty} \Theta^{(l)} = \widehat{\Theta}$.

2) $\frac{\partial \mathbb{Q}(\Theta^{(l+1)} || \Theta^{(l)})}{\partial \Theta} = 0$.

and

3) $\frac{\partial^2 \mathbb{Q}(\Theta^{(l+1)} || \Theta^{(l)})}{\partial \Theta^2}$ is negative definite with eigenvalues bounded away from zero.

Then,

$\frac{\partial l(\widehat{\Theta})}{\partial \Theta} = 0$ and $\frac{\partial^2 \mathbb{Q}(\widehat{\Theta} || \widehat{\Theta})}{\partial \Theta^2}$ is negative definite.

The EM algorithm properties

According to Theorem 1

$\widehat{\Theta}$ is the maximum likelihood estimator parameter vector of $\mathbb{Q}$ function.

and

The EM likelihood sequence $l(\Theta^{(l)})$ converges to some local maximum $l(\widehat{\Theta})$.

The EM algorithm properties

Theorem 2

If $L(\Theta)$ is unimodal and satisfies the same differentiability conditions in Theorem 1 then, $\Theta^{(l)}$ converges to the unique maximum likelihood of Θ.

The EM algorithm properties

Dempster et al. (1977) defined an EM mapping

$$M : \Theta \longrightarrow \Theta \tag{13}$$

which verified at each iteration (l),

$$\Theta^{(l+1)} = M(\Theta^{(l)}). \tag{14}$$

Theorem 3

If $\Theta^{(l)}$ converges to some point $\widehat{\Theta}$, then $\widehat{\Theta}$ is a fixed point of the EM algorithm.