EM Variants

Qinliang Su （苏勤亮）

Sun Yat-sen University

suqliang@mail.sysu.edu.cn

Review of the EM Algorithms

- To use EM algorithms, the key steps below are required
  1) Computing the posteriori distribution

$$p(\boldsymbol{z}|\boldsymbol{x};\boldsymbol{\theta}^{(t)})$$

  2) Evaluating the expectation of $\log p(\boldsymbol{x},\boldsymbol{z};\boldsymbol{\theta})$ *w.r.t.* the posteriori $p(\boldsymbol{z}|\boldsymbol{x};\boldsymbol{\theta}^{(t)})$, *i.e.*,

$$\mathcal{Q}\left(\boldsymbol{\theta};\boldsymbol{\theta}^{(t)}\right)=\mathbb{E}_{p(\boldsymbol{z}|\boldsymbol{x};\boldsymbol{\theta}^{(t)})}[\log p(\boldsymbol{x},\boldsymbol{z};\boldsymbol{\theta})]$$

  3) Maximizing

$$\boldsymbol{\theta}^{(t+1)}=\arg\max_{\boldsymbol{\theta}}\mathcal{Q}\left(\boldsymbol{\theta};\boldsymbol{\theta}^{(t)}\right)$$

However, not all of them are always achievable

Two Issues in the EM

- Issue one

The maximization is not achievable

$$\boldsymbol{\theta}^{(t+1)} = \arg\max_{\boldsymbol{\theta}} \mathcal{Q}\left(\boldsymbol{\theta}; \boldsymbol{\theta}^{(t)}\right)$$

- Issue two

1) The posteriori $p(\boldsymbol{z}|\boldsymbol{x}; \boldsymbol{\theta}^{(t)})$ cannot be derived analytically

2) Even if $p(\boldsymbol{z}|\boldsymbol{x}; \boldsymbol{\theta}^{(t)})$ can be obtained, we still cannot derive the close-form expression for the expectation

$$\mathbb{E}_{p(\boldsymbol{z}|\boldsymbol{x};\boldsymbol{\theta}^{(t)})}[\log p(\boldsymbol{x}, \boldsymbol{z}; \boldsymbol{\theta})]$$

Outline

- Addressing Issue One
- Addressing Issue Two

Generalized EM

- It is quite often in training LVMs that the optimization $\max_{\boldsymbol{\theta}} \mathcal{Q}\left(\boldsymbol{\theta}; \boldsymbol{\theta}^{(t)}\right)$ cannot be solved

How to address this issue?

- Maximizing $\mathcal{Q}\left(\boldsymbol{\theta}; \boldsymbol{\theta}^{(t)}\right)$ is not necessary. Increasing $\mathcal{Q}\left(\boldsymbol{\theta}; \boldsymbol{\theta}^{(t)}\right)$ is sufficient to guarantee the EM algorithm to work

- That is, if we adopt SGD to update the parameter as

$$\boldsymbol{\theta}^{(t+1)} = \boldsymbol{\theta}^{(t)} + \gamma \cdot \left.\frac{\partial \mathcal{Q}\left(\boldsymbol{\theta}; \boldsymbol{\theta}^{(t)}\right)}{\partial \boldsymbol{\theta}}\right|_{\boldsymbol{\theta}=\boldsymbol{\theta}^{(t)}}$$

we can also guarantee the monotonic increase of log-likelihood

$$\log p(\boldsymbol{x}; \boldsymbol{\theta}^{(t+1)}) \geq \log p(\boldsymbol{x}; \boldsymbol{\theta}^{(t)})$$

- Sketch of proof

  - First, after the SGD update, it can be easily seen that

$$\mathcal{Q}\left(\boldsymbol{\theta}^{(t+1)}; \boldsymbol{\theta}^{(t)}\right) \geq \mathcal{Q}\left(\boldsymbol{\theta}^{(t)}; \boldsymbol{\theta}^{(t)}\right)$$

  - From $\mathcal{L}\left(p\left(\boldsymbol{z}|\boldsymbol{x}; \boldsymbol{\theta}^{(t)}\right), \boldsymbol{\theta}\right) = \int p\left(\boldsymbol{z}|\boldsymbol{x}; \boldsymbol{\theta}^{(t)}\right) \log \frac{p(\boldsymbol{x}, \boldsymbol{z}; \boldsymbol{\theta})}{p\left(\boldsymbol{z}|\boldsymbol{x}; \boldsymbol{\theta}^{(t)}\right)} d\boldsymbol{z} = \mathcal{Q}\left(\boldsymbol{\theta}; \boldsymbol{\theta}^{(t)}\right) - \int p\left(\boldsymbol{z}|\boldsymbol{x}; \boldsymbol{\theta}^{(t)}\right) \log p\left(\boldsymbol{z}|\boldsymbol{x}; \boldsymbol{\theta}^{(t)}\right) d\boldsymbol{z}$, we further have

$$\mathcal{L}\left(p\left(\boldsymbol{z}|\boldsymbol{x}; \boldsymbol{\theta}^{(t)}\right); \boldsymbol{\theta}^{(t+1)}\right) \geq \underbrace{\mathcal{L}\left(p\left(\boldsymbol{z}|\boldsymbol{x}; \boldsymbol{\theta}^{(t)}\right); \boldsymbol{\theta}^{(t)}\right)}_{=\log p(\boldsymbol{x}; \boldsymbol{\theta}^{(t)})}$$

  - Due to

$$\log p(\boldsymbol{x}; \boldsymbol{\theta}^{(t+1)}) = \underbrace{\mathcal{L}\left(p(\boldsymbol{z}|\boldsymbol{x}; \boldsymbol{\theta}^{(t)}), \boldsymbol{\theta}^{(t+1)}\right)}_{\geq \log p(\boldsymbol{x}|\boldsymbol{\theta}^{(t)})} + \underbrace{KL(p(\boldsymbol{z}|\boldsymbol{x}; \boldsymbol{\theta}^{(t)})||p(\boldsymbol{z}|\boldsymbol{x}; \boldsymbol{\theta}^{(t+1)}))}_{\geq 0}$$

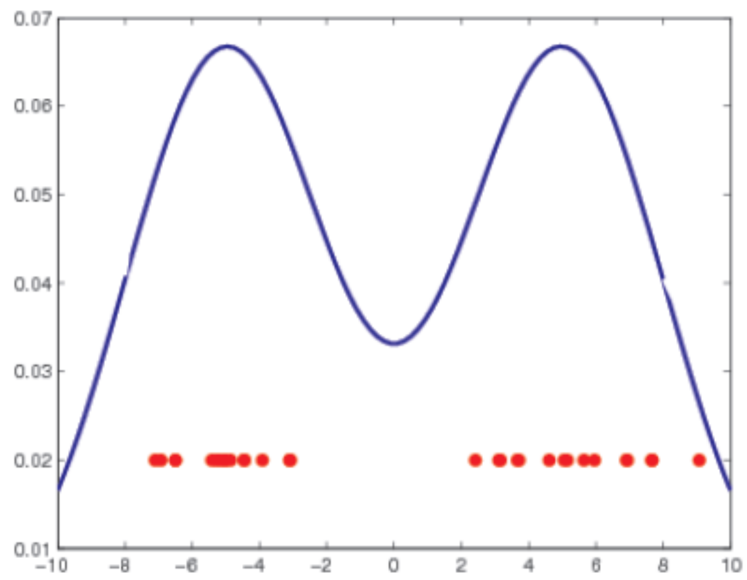
$$\Longrightarrow \log p(\boldsymbol{x}; \boldsymbol{\theta}^{(t+1)}) \geq \log p(\boldsymbol{x}; \boldsymbol{\theta}^{(t)})$$

Outline

- Addressing Issue One
- Addressing Issue Two

MCMC EM

- For any probability distributions, we can always draw samples from it, *e.g.*, using Markov chain Monte Carlo (MCMC) methods



- Although the exact expression of the posteriori $p(\boldsymbol{z}|\boldsymbol{x};\boldsymbol{\theta}^{(t)})$ is not known, we can use samples drawn from it to approximate it

- Thus, we can draw lots of samples $\boldsymbol{z}_s$ for $s = 1, \cdots, S$ from the posteriori distribution $p(\boldsymbol{z}|\boldsymbol{x}; \boldsymbol{\theta}^{(t)})$ such that

$$\boldsymbol{z}_s \sim p(\boldsymbol{z}|\boldsymbol{x}; \boldsymbol{\theta}^{(t)})$$

- Then, the expectation $\mathbb{E}_{p(\boldsymbol{z}|\boldsymbol{x};\boldsymbol{\theta}^{(t)})}[\log p(\boldsymbol{x}, \boldsymbol{z}; \boldsymbol{\theta})]$ can be approximated as

$$\mathcal{Q}\left(\boldsymbol{\theta}; \boldsymbol{\theta}^{(t)}\right) \approx \frac{1}{S} \cdot \sum_{s=1}^{S} \log p(\boldsymbol{x}, \boldsymbol{z}_s; \boldsymbol{\theta})$$

- We can optimize the approximate $\mathcal{Q}\left(\boldsymbol{\theta}; \boldsymbol{\theta}^{(t)}\right)$ with SGD algorithm

The two sub-problems in the Issue Two are both solved. Thus, latent-variable models can always be trained with MCMC EM

VB-EM

- Drawing samples from a distribution is *computationally expensive*

- An alternative approach is to use a simple distribution $q(\boldsymbol{z};\boldsymbol{\phi})$ to approximate the exact posterior distribution $p(\boldsymbol{z}|\boldsymbol{x};\boldsymbol{\theta}^{(t)})$

How to get the approximate simple distribution $q(\boldsymbol{z};\boldsymbol{\phi})$?

- Idea

1) Assuming a simple form for $q(\boldsymbol{z};\boldsymbol{\phi})$, *e.g.*,

$$q(\boldsymbol{z};\boldsymbol{\phi}) = \prod \mathcal{N}\left(z_i;\mu_i,\sigma_i^2\right)$$

2) Finding the best $\boldsymbol{\phi}$ that minimizes the KL-divergence

$$KL\left(q(\boldsymbol{z};\boldsymbol{\phi})\middle\|p(\boldsymbol{z}|\boldsymbol{x};\boldsymbol{\theta}^{(t)})\right)$$

- Steps to update the model parameter $\boldsymbol{\theta}$

1) Finding the best approximate $q(\boldsymbol{z};\boldsymbol{\phi})$ such that

$$\boldsymbol{\phi}^{(t)} = \arg\min_{\boldsymbol{\phi}} KL\left(q(\boldsymbol{z};\boldsymbol{\phi}) \| p(\boldsymbol{z}|\boldsymbol{x};\boldsymbol{\theta}^{(t)})\right)$$

2) Using $q\left(\boldsymbol{z};\boldsymbol{\phi}^{(t)}\right)$ to compute expectation $\mathbb{E}_{p(\boldsymbol{z}|\boldsymbol{x};\boldsymbol{\theta}^{(t)})}[\log p(\boldsymbol{x},\boldsymbol{z};\boldsymbol{\theta})]$ approximately as

$$\tilde{Q}\left(\boldsymbol{\theta};\ \boldsymbol{\phi}^{(t)}\right) = \mathbb{E}_{q\left(\boldsymbol{z};\boldsymbol{\phi}^{(t)}\right)}[\log p(\boldsymbol{x},\boldsymbol{z};\boldsymbol{\theta})]$$

3) Obtaining the new value $\boldsymbol{\theta}^{(t+1)}$ as

$$\boldsymbol{\theta}^{(t+1)} = \arg\max_{\boldsymbol{\theta}} \tilde{Q}\left(\boldsymbol{\theta};\ \boldsymbol{\phi}^{(t)}\right)$$

- The two optimization problems can be equivalently written as

$$\min_{\boldsymbol{\phi}} \; KL\left(q(\boldsymbol{z};\boldsymbol{\phi}) \| p(\boldsymbol{z}|\boldsymbol{x};\boldsymbol{\theta}^{(t)})\right) \Leftrightarrow \max_{\boldsymbol{\phi}} \int q(\boldsymbol{z};\boldsymbol{\phi}) \log \frac{p(\boldsymbol{z}|\boldsymbol{x};\boldsymbol{\theta}^{(t)})}{q(\boldsymbol{z};\boldsymbol{\phi})} d\boldsymbol{z}$$

$$\Leftrightarrow \max_{\boldsymbol{\phi}} \int q(\boldsymbol{z};\boldsymbol{\phi}) \log \frac{p(\boldsymbol{x},\boldsymbol{z};\boldsymbol{\theta}^{(t)})}{q(\boldsymbol{z};\boldsymbol{\phi})} d\boldsymbol{z}$$

$$\max_{\boldsymbol{\theta}} \mathbb{E}_{q(\boldsymbol{z};\boldsymbol{\phi}^{(t)})}[\log p(\boldsymbol{x},\boldsymbol{z};\boldsymbol{\theta})] \Leftrightarrow \max_{\boldsymbol{\theta}} \int q\left(\boldsymbol{z};\boldsymbol{\phi}^{(t)}\right) \log \frac{p(\boldsymbol{x},\boldsymbol{z};\boldsymbol{\theta})}{q\left(\boldsymbol{z};\boldsymbol{\phi}^{(t)}\right)} d\boldsymbol{z}$$

- The algorithm to optimize $\boldsymbol{\theta}$ and $\boldsymbol{\phi}$ can be understood as solving the following optimization problem in an alternative way

$$\max_{\boldsymbol{\phi},\boldsymbol{\theta}} \mathcal{L}(\boldsymbol{x};\boldsymbol{\theta},\boldsymbol{\phi})$$

with

$$\mathcal{L}(\boldsymbol{x};\boldsymbol{\theta},\boldsymbol{\phi}) \triangleq \int q(\boldsymbol{z};\boldsymbol{\phi}) \log \frac{p(\boldsymbol{x},\boldsymbol{z};\boldsymbol{\theta})}{q(\boldsymbol{z};\boldsymbol{\phi})} d\boldsymbol{z}$$

- Instead of updating $\boldsymbol{\theta}, \boldsymbol{\phi}$ alternatively, we can also update them *simultaneously* with the SGD algorithm, that is,

$$\boldsymbol{\theta}^{(t+1)} = \boldsymbol{\theta}^{(t)} + \gamma \cdot \left. \frac{\partial \mathcal{L}(\boldsymbol{x}; \boldsymbol{\theta}, \boldsymbol{\phi})}{\partial \boldsymbol{\theta}} \right|_{\boldsymbol{\theta}=\boldsymbol{\theta}^{(t)}}$$

$$\boldsymbol{\phi}^{(t+1)} = \boldsymbol{\phi}^{(t)} + \gamma \cdot \left. \frac{\partial \mathcal{L}(\boldsymbol{x}; \boldsymbol{\theta}, \boldsymbol{\phi})}{\partial \boldsymbol{\phi}} \right|_{\boldsymbol{\phi}=\boldsymbol{\phi}^{(t)}}$$

The method is dubbed *variational Bayesian EM* (VB-EM)

- In general, we optimize $\mathcal{L}(\boldsymbol{x}; \boldsymbol{\theta}, \boldsymbol{\phi})$ *w.r.t.* the two parameters $\boldsymbol{\theta}, \boldsymbol{\phi}$ simultaneously

- Actually, it can be proved that $\mathcal{L}(\boldsymbol{x};\boldsymbol{\theta},\boldsymbol{\phi})$ is a *lower bound* of the log-likelihood $\ln p(\boldsymbol{x};\boldsymbol{\theta})$ for any $\boldsymbol{\theta}$ and $\boldsymbol{\phi}$, that is,

$$\ln p(\boldsymbol{x};\boldsymbol{\theta}) \geq \mathcal{L}(\boldsymbol{x};\boldsymbol{\theta},\boldsymbol{\phi})$$

(Proof can be found in the next slide)

When the log-likelihood $\ln p(\boldsymbol{x};\boldsymbol{\theta})$ cannot be directly maximized, we can seek to optimize its lower bound

$$\mathcal{L}(\boldsymbol{x};\boldsymbol{\theta},\boldsymbol{\phi}) \triangleq \int q(\boldsymbol{z};\boldsymbol{\phi}) \log \frac{p(\boldsymbol{x},\boldsymbol{z};\boldsymbol{\theta})}{q(\boldsymbol{z};\boldsymbol{\phi})} d\boldsymbol{z}$$

where $q(\boldsymbol{z};\boldsymbol{\phi})$ can be set as any simple distribution forms, *e.g.*,

$$q(\boldsymbol{z};\boldsymbol{\phi}) = \prod \mathcal{N}\left(z_i;\mu_i,\sigma_i^2\right)$$

Proof of $\ln p(\boldsymbol{x};\boldsymbol{\theta}) \geq \mathcal{L}(\boldsymbol{x};\boldsymbol{\theta},\boldsymbol{\phi})$

$$\ln p_{\boldsymbol{\theta}}(\boldsymbol{x}) = \int_{\boldsymbol{z}} q_{\boldsymbol{\phi}}(\boldsymbol{z}) \ln p_{\boldsymbol{\theta}}(\boldsymbol{x})\, d\boldsymbol{z} = \int_{\boldsymbol{z}} q_{\boldsymbol{\phi}}(\boldsymbol{z}) \ln \frac{p_{\boldsymbol{\theta}}(\boldsymbol{x},\ \boldsymbol{z}) q_{\boldsymbol{\phi}}(\boldsymbol{z})}{q_{\boldsymbol{\phi}}(\boldsymbol{z}) p_{\boldsymbol{\theta}}(\boldsymbol{z}|\boldsymbol{x})} d\boldsymbol{z}$$

$$= \int_{\boldsymbol{z}} q_{\boldsymbol{\phi}}(\boldsymbol{z}) \ln \frac{p_{\boldsymbol{\theta}}(\boldsymbol{x},\boldsymbol{z})}{q_{\boldsymbol{\phi}}(\boldsymbol{z})} d\boldsymbol{z} + \int_{\boldsymbol{z}} q_{\boldsymbol{\phi}}(\boldsymbol{z}) \ln \frac{q_{\boldsymbol{\phi}}(\boldsymbol{z})}{p_{\boldsymbol{\theta}}(\boldsymbol{z}|\boldsymbol{x})} d\boldsymbol{z}$$

$$= \int_{\boldsymbol{z}} q_{\boldsymbol{\phi}}(\boldsymbol{z}) \ln \frac{p_{\boldsymbol{\theta}}(\boldsymbol{x},\boldsymbol{z})}{q_{\boldsymbol{\phi}}(\boldsymbol{z})} d\boldsymbol{z} + KL(q_{\boldsymbol{\phi}}(\boldsymbol{z}) || p_{\boldsymbol{\theta}}(\boldsymbol{z}|\boldsymbol{x}))$$

$$\geq \int_{\boldsymbol{z}} q_{\boldsymbol{\phi}}(\boldsymbol{z}) \ln \frac{p_{\boldsymbol{\theta}}(\boldsymbol{x},\boldsymbol{z})}{q_{\boldsymbol{\phi}}(\boldsymbol{z})} d\boldsymbol{z}$$

$$\triangleq \mathcal{L}(\boldsymbol{x};\boldsymbol{\theta},\boldsymbol{\phi})$$