

---

# No Coding Farmer

---

**Tao Hu**

Department of Computer Science  
Peking University  
No.5 Yiheyuan Road Haidian District, Beijing, P.R.China  
taohu@pku.edu.cn

## Abstract

Some Miscellaneous Summary.

## 1 Expectation Maximization Introduction

### 1.1 EM Induction

$$L(\theta) = \sum_{i=1}^M \log p(X; \theta) = \sum_{i=1}^m \log \sum_z p(X, Z; \theta)$$

let  $\theta_i$  be some distribution over  $z$ 's ( $\sum_z \theta_i(z) = 1, \theta_i(z) \geq 0$ )

$$\begin{aligned} & \sum_i \log p(X^{(i)}; \theta) \\ &= \sum_i \log \sum_{Z^{(i)}} \theta_i(Z^{(i)}) \frac{p(X^{(i)}, Z^{(i)}; \theta)}{\theta_i(Z^{(i)})} \\ &\geq \sum_i \sum_{Z^{(i)}} \theta_i(Z^{(i)}) \log \frac{p(X^{(i)}, Z^{(i)}; \theta)}{\theta_i(Z^{(i)})} \quad (f(x) = \log x \text{ is concave.}) \end{aligned}$$

$$\text{let } \frac{p(X^{(i)}, Z^{(i)}; \theta)}{\theta_i(Z^{(i)})} = C$$

the equality can be only reached when  $\frac{p(X^{(i)}, Z^{(i)}; \theta)}{\theta_i(Z^{(i)})}$  is a constant.

we can get:  $\sum_i \frac{p(X^{(i)}, Z^{(i)}; \theta)}{C} = 1$  namely:  $\sum_i p(X^{(i)}, Z^{(i)}; \theta) = C$

further induction:  $\theta_i(Z^{(i)}) = \frac{p(X^{(i)}, Z^{(i)}; \theta)}{\sum_i p(X^{(i)}, Z^{(i)}; \theta)} = p(Z^{(i)} | X^{(i)}; \theta)$

so the procedure of EM algorithm is:

Repeat Until Convergence:

- E-step: for each  $i$ , get  $Q_i(Z^{(i)}) = p(Z^{(i)} | X^{(i)}; \theta)$
- M-step:  $\theta := \operatorname{argmax}_{\theta} \sum_i \sum_{Z^{(i)}} i(Z^{(i)}) \log \frac{p(X^{(i)}, Z^{(i)}; \theta)}{Q_i(Z^{(i)})}$

### 1.2 EM convergence proof

$$\text{let } l(\theta^{(t)}) = \sum_i \sum_{Z^{(i)}} Q_i^{(t)}(Z^{(i)}) \log \frac{p(X^{(i)}, Z^{(i)}; \theta)}{Q_i^{(t)}(Z^{(i)})}$$

then, we have the following inequality:

$$l(\theta^{(t+1)})$$

$$\begin{aligned}
&\geq \sum_i \sum_{Z^{(i)}} Q_i^{(t)}(Z^{(i)}) \log \frac{p(X^{(i)}, Z^{(i)}, \theta^{(t+1)})}{Q_i^{(t)}(Z^{(i)})} \\
&\geq \sum_i \sum_{Z^{(i)}} Q_i^{(t)}(Z^{(i)}) \log \frac{p(X^{(i)}, Z^{(i)}, \theta^{(t)})}{Q_i^{(t)}(Z^{(i)})} \\
&\geq l(\theta^{(t)})
\end{aligned}$$

the first inequality is because  $l(\theta) \geq \sum_i \sum_{Z^{(i)}} Q_i^{(t)}(Z^{(i)}) \log \frac{p(X^{(i)}, Z^{(i)}, \theta)}{Q_i^{(t)}(Z^{(i)})} \forall \theta, Q_i$

the second inequality is because of the maximum of the M-step.

Hence, EM causes the likelihood to converge monotonically.

### 1.3 Different Writing Style of EM Algorithm

There are many writing style of EM algorithm. here I just mention the book <Statistics Learning Method> by LiHang who is very famous in China.

EM algorithm from LiHang(Li-version):

---

#### Algorithm 1 EM from LIHang

---

**Require:** observation  $X$ , hidden variable  $Z$ , joint distribution  $P(X, Z|\theta)$ , conditional distribution  $P(Z|X, \theta)$   
**while** Not convergence **do**  
    E-Step: let  $\theta^{(i)}$  is the  $i$ -th estimate of  $\theta$ ,  
     $Q(\theta, \theta^{(i)}) = E_z[\log P(X, Z|\theta)|X, \theta^{(i)}] = \sum_Z \log P(X, Z|\theta) P(Z|X, \theta^{(i)})$   
    M-step:  $\theta^{(i+1)} = \operatorname{argmax}_{\theta} Q(\theta, \theta^{(i)})$   
**end while**  
output model parameter  $\theta$

---

it seems that Li-version is different from the above version. however, they are the same. because:

- the above version just consider every data, so that it include subscript  $i$ . however Li-version only consider one data.
- the above version can be transformed to Li-version.

$$\begin{aligned}
&\sum_Z Q(Z) \log \frac{P(X, Z; \theta)}{Q(Z)} \\
&= \sum_Z P(Z|X; \theta^{(t)}) \log \frac{p(X, Z; \theta)}{p(Z|X; \theta^{(t)})} \\
&= \sum_Z P(Z|X; \theta^{(t)}) \log P(X, Z; \theta) - \sum_Z P(Z|X; \theta^{(t)}) \log P(Z|X; \theta^{(t)})
\end{aligned}$$

as the variable is  $\theta$ , so  $\sum_Z P(Z|X; \theta^{(t)}) \log P(Z|X; \theta^{(t)})$  can be removed.

- $Q(\theta, \theta^{(i)}) = \sum_Z \log P(X, Z|\theta) P(Z|X, \theta^{(i)})$  can be also written as  $Q(\theta, \theta^{(i)}) = \sum_Z \log P(X, Z|\theta) P(Z, X, \theta^{(i)})$ , because  $X$  is a observation.

## 2 EM applications

### 2.1 Gaussian Mix Model

GMM can be solved by EM. notice here we use the expectation of EM:

$$\begin{aligned}
&Q(\theta, \theta^{(i)}) \\
&= E_{\gamma}[\log P(y, \gamma|\theta)|y, \theta^{(i)}] \\
&= E[\sum_{k=1}^K [n_k \log \alpha_k + \sum_{j=1}^N \gamma_{jk} [\log \frac{1}{\sqrt{2\pi}} - \log \sigma_k - \frac{1}{2\sigma_k^2} (y_j - \mu_k)^2]]] \\
&= \sum_{k=1}^K [(E\gamma_{jk}) \log \alpha_k + \sum_{j=1}^N (E\gamma_{jk}) [\log \frac{1}{\sqrt{2\pi}} - \log \sigma_k - \frac{1}{2\sigma_k^2} (y_j - \mu_k)^2]]
\end{aligned}$$

here  $(E\gamma_{jk})$  can be easily calculated.

$\hat{\mu}_k, \hat{\sigma}_k^2$  can be acquired by derivation.

$\hat{\alpha}_k$  can be acquired by the derivation on the Lagrangian ( $\sum_i^K \alpha_k = 1$ ).

## 2.2 Hidden Markov Model

HMM Learning Method is also called Baum-Welch algorithm. the target is learning  $\lambda = (A, B, \pi)$ .

Q function is:

$$Q(\lambda, \bar{\lambda}) = \sum_I \log P(O, I|\lambda) P(O, I|\bar{\lambda})$$

$$P(O, I, \lambda) = \pi_{i_1} b_{i_1}(o_1) a_{i_1 i_2} b_{i_2}(o_2) \dots a_{i_{T-1} i_T} b_{i_T}(o_T)$$

so the Q function can also be written as:

$$Q(\lambda, \bar{\lambda}) = \sum_I \log \pi_{i_1} P(O, I|\bar{\lambda}) + \sum_I (\sum_{t=1}^{T-1} \log a_{i_t, i_{t+1}}) P(O, I|\bar{\lambda}) + \sum_I (\sum_{t=1}^T \log b_{i_t}(o_t)) P(O, I|\bar{\lambda})$$

**note here: I is not only one state. it includes state length from 1 to T, which all start from  $i_1$**

so we can solve the maximum of Q function by derivation on the Lagrangian polynomial (because exists these limitations:  $\sum_{i=1}^N \pi_i = 1, \sum_{j=1}^N a_{ij} = 1, \sum_{i=1}^M b_i = 1$ )

## 2.3 Naive Bayesian

## 2.4 other papers

We can use softmax to model transition probability, normal distribution to model emission probability.

it's a good example in Car that Knows Before You Do: Anticipating Maneuvers via Learning Temporal Driving Models, the AIO-HMM can be more complicated, which can be enriched by the graphic model by M.I Jordon.

## 3 VAE

here is a complete VAE tutorial [1]

$$\begin{aligned} \max \quad & \log P(x) \\ \text{lhs} = & \log \int P(x, z) dz \\ = & \log \int P(x/z) p(z) dz \\ = & \log \int \frac{P(x/z)}{q(z/x)} q(z/x) p(z) dz \\ = & \log E_{q(z/x)} \left[ \frac{p(x/z)}{q(z/x)} p(z) \right] \end{aligned}$$

jenson's inequality, we can know:  $\geq E_{q(z/x)} [\log \frac{p(x/z)}{q(z/x)} p(z)]$

$$\begin{aligned} &= E_{q(z/x)} [\log p(x/z)] + E_{q(z/x)} [\log \frac{p(z)}{q(z/x)}] \\ &= E_{q(z/x)} [\log p(x/z)] - E_{q(z/x)} [\log \frac{q(z/x)}{p(z)}] \\ &= E_{q(z/x)} [\log p(x/z)] - KL(q(z/x) || p(z)) \end{aligned}$$

## Acknowledgments

## References

[1] Carl Doersch. Tutorial on variational autoencoders. *arXiv preprint arXiv:1606.05908*, 2016.