Expectation Maximization

2021年6月17日 0:57

P61 ELBO + KL Divergence

X: observed data Z: unobserved data (latent (隐藏) variable)

(X,Z): complete data θ: parameter $\theta_{MLE} = \arg\max_{\theta} p(x|\theta)$

EM: $\theta^{(t+1)} = \arg\max_{\theta} \int \underset{\text{constant}}{p(z|x,\theta^{(t)})} \log \underset{\text{variable}}{p(x,z|\theta)} dz$

E-step: $p(z|x,\theta^{(t)}) \Rightarrow E_{z|x,\theta^{(t)}}[\log p(x,z|\theta)] = ELBO$

M-step: $\theta^{(t+1)} = \arg\max_{\theta} E_{z|x,\theta^{(t)}}[\log p(x,z|\theta)]$

Proof

$\log p(x|\theta) = \log p(x,z|\theta) - \log p(z|x,\theta) = \log \frac{p(x,z|\theta)}{q(z)^{(t)}} - \log \frac{p(z|x,\theta)}{q(z)^{(t)}}$ $q(z)^{(t)}$ is the probability distribution of z, $\int q(z)^{(t)} dz = 1$

Direct calculation is too difficult ⇒ Find the expectation # $E_x[g(x)] = \int p(x)g(x)dx$ or $\sum_x p(x)g(x)$

Left side: $\int q(z)^{(t)} \log p(x|\theta) dz = \log p(x|\theta) \int q(z)^{(t)} dz = \log p(x|\theta)$

Right side: $\underbrace{\int q(z)^{(t)} \log \frac{p(x,z|\theta)}{q(z)^{(t)}} dz}_{\text{ELBO: evidence lower bound}} - \underbrace{\int q(z)^{(t)} \log \frac{p(z|x,\theta)}{q(z)^{(t)}} dz}_{KL(q(z)^{(t)} \| p(z|x,\theta))}$ ⇒ relative entropy $KL(P\|Q) = \int p(x) \log \frac{P(x)}{Q(x)} dx$ or $\sum p(x) \log \frac{P(x)}{Q(x)}$

∴ $\log p(x|\theta) = ELBO + KL(q\|P)$ $KL(q\|P) \geq 0$ $\log(x|\theta) \geq ELBO$ Equality holds if and only if $q = p$

When the equlity holds $q(z)^{(t)} = p(z|x,\theta^{(t)})$ → posterior $\log(x|\theta) = ELBO$

$\therefore \hat{\theta} = \arg\max_{\theta} ELBO = \arg\max_{\theta} \int p(z|x,\theta^{(t)}) \log \frac{p(x,z|\theta)}{p(z|x,\theta^{(t)})} dz = \arg\max_{\theta} \int p(z|x,\theta^{(t)}) [\log p(x,z|\theta) - \log p(z|x,\theta^{(t)})] dz$

$= \arg\max_{\theta} \int p(z|x,\theta^{(t)}) \log p(x,z|\theta) dz$

P62 ELBO + Jensen's inequlity

The secant line (割线) of a convex function ($f''>0$) lies above the graph of the function.

Jensen's inequality: $f(tx_1 + (1-t)x_2) \leq tf(x_1) + (1-t)f(x_2)$

φ is convex function, $\varphi(E[X]) \leq E[\varphi(X)]$ $\varphi(\sum_{i=1}^{N} \lambda_i x_i) \leq \sum_{i=1}^{N} \lambda_i \varphi(x_i)$ s.t. $\sum_{i=1}^{N} \lambda_i = 1$ Jensen gap: $E[\varphi(X)] - \varphi(E[X])$

Equilty holds if and and only if $x_1 = x_2 = \cdots = x_N$ or φ is linear on domain (域) containing $x_1, x_2, \cdots, x_N$

Proof (finite form)(mathematical induction)

$N = 1$ or 2 $\varphi(\lambda_1 x_1 + \lambda_2 x_2) \leq \lambda_1 \varphi(x_1) + \lambda_2 \varphi(x_2)$ is clearly true

Assume the induction hypothesis that for a single case $N = n$ is true: $\varphi(\sum_{i=1}^{n} \lambda_i x_i) \leq \sum_{i=1}^{n} \lambda_i \varphi(x_i)$

It follows that: $\varphi((1-\lambda_{n+1}) \sum_{i=1}^{n} \frac{\lambda_i}{1-\lambda_{n+1}} x_i + \lambda_{n+1} x_{n+1})$ Let $\frac{\lambda_i}{1-\lambda_{n+1}} = \eta_i$ ∵ $\sum_{i=1}^{n+1} \lambda_i = 1$ ∴ $\sum_{i=1}^{n} \frac{\lambda_i}{1-\lambda_{n+1}} = 1$ ∴ $\sum_{i=1}^{n} \eta_i = 1$

∴ $= \varphi((1-\lambda_{n+1}) \sum_{i=1}^{n} \eta_i x_i + \lambda_{n+1} x_{n+1}) \leq (1-\lambda_{n+1}) \varphi(\sum_{i=1}^{n} \eta_i x_i) + \lambda_{n+1} \varphi(x_{n+1})$

∵ $\varphi(\sum_{i=1}^{n} \eta_i x_i) \leq \sum_{i=1}^{n} \eta_i \varphi(x_i)$ ∴ $\varphi(\sum_{i=1}^{n+1} \lambda_i x_i) \leq (1-\lambda_{n+1}) \sum_{i=1}^{n} \eta_i \varphi(x_i) + \lambda_{n+1} \varphi(x_{n+1}) = \sum_{i=1}^{n} \lambda_i \varphi(x_i) + \lambda_{n+1} \varphi(x_{n+1})$

∴ $\varphi(\sum_{i=1}^{n+1} \lambda_i x_i) \leq \sum_{i=1}^{n+1} \lambda_i \varphi(x_i)$

⇓

$\varphi(\int x \, d\mu_n(x)) \leq \int \varphi(x) \, d\mu_n(x)$ $\mu_n(x) = \sum_{i=1}^{n} \lambda_i \delta_{x_i}$?

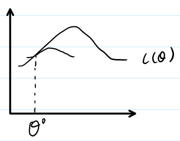
EM: $\log p(x|\theta) = \log \int p(x,z|\theta) dz = \log \int \frac{p(x,z|\theta)}{q(z)^{(t)}} q(z)^{(t)} dz \geq \int q(z)^{(t)} \log \frac{p(x,z|\theta)}{q(z)^{(t)}} dz$ → ELBO

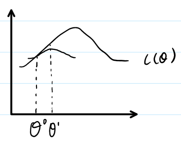
$f(E[\frac{p(x,z|\theta)}{q(z)^{(t)}}]) \geq E[f(\frac{p(x,z|\theta)}{q(z)^{(t)}})]$

When the equlity holds: $\frac{p(x,z|\theta^{(t)})}{q(z)^{(t)}} = c$

∴ $q(z)^{(t)} = \frac{1}{c} p(x,z|\theta^{(t)})$ ∴ $1 = \int q(z)^{(t)} dz = \frac{1}{c} \int p(x,z|\theta^{(t)}) dz$ ∴ $p(x|\theta^{(t)}) = c$

∴ $q(z)^{(t)} = \frac{p(x,z|\theta^{(t)})}{p(x|\theta^{(t)})} = p(z|x,\theta^{(t)})$





① initialize θ^0 ② find the expectation $\log \frac{p(x,z|\theta)}{p(z|x,\theta^0)}$ (ELBO) ③ find θ' that maximizes the expectation ④ repeat ②③ until θ converges

P60 EM Convergence

$\log(x|\theta) = \log(x,z|\theta) - \log(z|x,\theta)$ ∴ $q(z)^{(t)} \log(x|\theta) = q(z)^{(t)} \log(x,z|\theta) - q(z)^{(t)} \log(z|x,\theta)$

$\log(x|\theta) = \underbrace{\int p(z|x,\theta^{(t)}) \log p(x,z|\theta) dz}_{Q(\theta,\theta^{(t)})} - \underbrace{\int p(z|x,\theta^{(t)}) \log p(z|x,\theta) dz}_{H(\theta,\theta^{(t)})}$

∵ $\theta^{(t+1)} = \arg\max_{\theta} \int p(z|x,\theta^{(t)}) \log p(x,z|\theta) dz = \arg\max_{\theta} Q(\theta,\theta^{(t)})$

∴ $Q(\theta^{(t+1)},\theta^{(t)}) \geq Q(\theta^{(t)},\theta^{(t)})$

∵ $H(\theta^{(t+1)},\theta^{(t)}) - H(\theta^{(t)},\theta^{(t)}) = \int p(z|x,\theta^{(t)}) \log \frac{p(z|x,\theta^{(t+1)})}{p(z|x,\theta^{(t)})} dz \overset{\text{Jensen}}{\leq} \log \underbrace{\int p(z|x,\theta^{(t+1)}) dz}_{1} = \log 1 = 0$

∴ $H(\theta^{(t+1)},\theta^{(t)}) \leq H(\theta^{(t)},\theta^{(t)})$

∴ $\log(x|\theta^{(t+1)}) \geq \log(x|\theta^{(t)})$ ∴ EM converges

P63 EM Review

EM mainly solves problems arising from probabilities generative models:

MLE: $\hat{\theta} = \arg\max_{\theta} P(x|\theta)$

But in some cases, P(X) is difficult to find. Because the observable variable X does satify a certain distribution. So we assume that X is generated by the latent variable Z that satisfies a certain distribution.

$P(X) = \int_z P(X,Z) dz = \frac{P(X,Z)}{P(Z|X)}$

P64 Generalized EM

Narrow EM is a special case of generalized EM

The EM algorithm is designed to solve the parameter estimation problems

P64 Generalized EM

Narrow EM is a special case of generalized EM

The EM algorithm is designed to solve the parameter estimation problems

$\log P(X|\theta) = ELBO + KL(q||P)$ $\quad \begin{cases} ELBO = E_{q(z)}\left[\log \frac{P(x,z|\theta)}{q(z)}\right] = \mathcal{L}(q,\theta) \\ KL(q||P) = \int q(z) \log \frac{q(z)}{P(z|x,\theta)} dz \end{cases}$

$KL \geq 0$ Equality holds if and only if $q = P$ $\Rightarrow$ Narrow EM : $q(z) = P(z|X,\theta)$

But in some cases, $P(X|z,\theta)$ is hard to find $\Rightarrow$ Gerenalized EM : Fix θ, then $p(X|\theta)$ is fied $\Rightarrow \hat{q} = \arg\min_q KL(q||P) = \arg\max_q \mathcal{L}(q,\theta)$

$\therefore$ Gerenalized EM :

$$\begin{cases} \text{E-step}: & q^{(t+1)} = \arg\max_q \mathcal{L}(q, \theta^{(t)}) \\ \text{M-step}: & \theta^{(t+1)} = \arg\max_\theta \mathcal{L}(q^{(t+1)}, \theta) \end{cases}$$

$$\mathcal{L}(q,\theta) = E_q[\log P(x,z|\theta) - \log q(z)]$$
$$= E_q[\log p(x,z|\theta)] \underbrace{- E_q[\log q(z)]}_{H(q) \text{ entropy}}$$

Narrow EM : $H(q) = 0$

Gerenalized EM : $H(q) \neq 0$