Principle of Expectation Maximization (EM)

what is the problem: in GMM it is hard to compute the gradient

We want to maximize (over θ) $\log p(x|\theta)$

$\boxed{1.}$ we will introduce a latent variable z

$$\underbrace{p(x,z|\theta) = p(x|\theta)}\cdot p(z|x)$$

$\boxed{2.}$ $$\log p(x|\theta) = \log\left[\frac{p(x,z|\theta)}{p(z|x,\theta)}\right] = \log p(x,z|\theta) - \log(p(z|x,\theta))$$

$\boxed{3.}$ we will define an arbitrary dist: $q(z)$

I will multiply (2.) by q and sum over z

$$p(x,\theta) = \sum_z q(z)\, p(x|\theta) = \sum_z q(z) \log\left[\frac{p(x,z|\theta)}{p(z|x,\theta)}\right]$$

bcoz: $\sum_z q(z) = 1$

$$= \underbrace{\sum_z q(z) \log\left[\frac{p(x,z|\theta)}{q(z)}\right]}_{\mathcal{L}(q,\theta)} - \underbrace{\sum_z q(z) \log\left[\frac{p(z|x,\theta)}{q(z)}\right]}_{D_{KL}[q \| p(z|x,\theta)]}$$

D_{KL} ↓ Kullback-Leibler divergence / distance

$D_{KL}(q\|p) = 0$ if $q = p$

$\geq 0 \ \forall q,p$

Properties of D_{KL}: $$D_{KL}(p\|q) := \sum_i p(i) \log\left[\frac{p(i)}{q(i)}\right]$$

$\bullet \sum_i p(i) \lg\left(\frac{q(i)}{p(i)}\right) \leq \lg\left(\sum_i \frac{\cancel{p(i)}\, q(i)}{\cancel{p(i)}}\right) = 0$

≤ 0

$\Longrightarrow \boxed{D_{KL} \geq 0}$

$\bullet\ D_{KL} = 0$ if $p = q$

$$\log p(x \mid \theta) = \mathcal{L}(q, \theta) + D_{KL}\left(q \,\|\, p(z \mid x, \theta)\right)$$

$\boxed{4.}$ Let's imagine that at iteration t,

I have the parameters $\underline{\theta^{(t)}}$

$\longrightarrow$ I want to find $\theta^{(t+1)}$ such that: $\lg p(x \mid \theta^{(t+1)}) \geq \lg p(x \mid \theta^{(t)})$

(i) The $\underline{E}$-step consists in using $\boxed{q(z) = p(z \mid x, \theta^{(t)})}$

you see that at time t: $D_{KL}\left(q \,\|\, p(z \mid x, \theta^{(t)})\right) = 0$

(ii) The $\underline{M}$-step: $\theta^{(t+1)} = \arg\max_\theta \mathcal{L}\left[p(z \mid x, \theta^{(t)}), \theta\right]$

⚠ We max over θ

but $\theta^{(t)}$ is fixed

We thus have: $\mathcal{L}\left(p(z \mid x, \theta^{(t)}), \theta^{(t+1)}\right) \geq \mathcal{L}\left(p(z \mid x, \theta^{(t)}), \theta^{(t)}\right)$

and therefore: $p(x \mid \theta^{(t+1)}) \geq p(x \mid \theta^{(t)})$

$$\underline{\text{detail}}: \quad \lg p(x|\theta^{(t)}) = \sum_z p(z|x,\theta^{(t)}) \lg\left[\frac{p(x,z|\theta^{(t)})}{p(z|x,\theta^{(t)})}\right]$$

choosing $q(z)$ as earlier

$$\theta^{(t+1)} = \arg\max_{\theta}\left[\sum_z p(z|x,\theta^{(t)}) \lg\left[\frac{p(x,z|\theta)}{p(z|x,\theta^{(t)})}\right]\right]$$

$$\underbrace{\le \sum_z p(z|x,\theta^{(t)}) \lg\left[\frac{p(x,z|\theta^{(t+1)})}{p(z|x,\theta^{(t)})}\right] + \underbrace{D_{KL}\left[p(z|x,\theta^{(t)}) \,\|\, p(z|x,\theta^{(t+1)})\right]}_{\text{we just add a positive term}}}$$

$$= p(x|\theta^{(t+1)})$$

Application of EM to GMM

For a given data x_i, we define the vector

$$z_{ih} = \begin{cases} 1 & \text{if } i \in \text{cluster } h \\ 0 & \text{otherwise} \end{cases}$$

one-hot vector

$$\boxed{\sum_h z_{ih} = 1}$$

$$p(x|\theta) = \sum_z \left[\rho_h \cdot \mathcal{N}(x, \theta_h)\right]^{z_h} = \sum_{h=1}^{c} \rho_h \mathcal{N}(x, \theta_h)$$

$$p(x|\theta) = \sum_z p(x, z|\theta)$$

$$\underbrace{p(x, z|\theta)}_{\text{proba to have } x \text{ and clus } h} = p(x, z) \cdot p(z)$$

$$p(x|z) = \prod_{k=1}^{c} \mathcal{N}(x, \theta_h)^{z_h}$$

$$p(z) = \prod_{h=1}^{c} \rho_h^{z_h}$$

We can define $p(z_h=1|x) = \frac{p(x|z_h=1).p(z_h=1)}{\sum_\ell p(z_\ell=1).p(x|z_\ell=1)}$

$$= \frac{\rho_h . \mathcal{N}(x|\Theta_h)}{\sum_\ell \rho_\ell \mathcal{N}(x|\Theta_\ell)}$$: the responsability

We will use $q(z) = p(z|x,\Theta)$

Now we compute $\mathbb{E}_{p(z|x)}\left[\log p(x,z|\Theta)\right]$

$$= \mathbb{E}_{p(z|x)}\left[\sum_h z_h \left\{\log \rho_h + \log \mathcal{N}(x|\mu_h,\sigma_h)\right\}\right] \quad (*)$$

1) $\mathbb{E}_{p(z|x)}[z_h] = \frac{\sum_{z_h} z_h \left[\rho_h^{(t)} . \mathcal{N}(x|\mu_h^{(t)},\sigma_h^{(t)})\right]^{z_h}}{\sum_\ell \left[\rho_\ell^{(t)} \mathcal{N}(x|\mu_\ell^{(t)},\sigma_\ell^{(t)})\right]}$

$$= \frac{\rho_h^{(t)} \mathcal{N}(x|\mu_h^{(t)},\sigma_h^{(t)})}{\sum_\ell \rho_\ell^{(t)} \mathcal{N}(x|\mu_\ell^{(t)},\sigma_\ell^{(t)})} = p(z_h=1|x,\theta^{(t)})$$

Find the update eq for μ:

$$\frac{\partial(*)}{\partial \mu_k} = 0 \iff \sum_k p(z_k=1 \mid x, \theta^{(t)}) \cdot \frac{(x-\mu_k)}{\cancel{\sigma_k}^2} = 0$$

$$\Rightarrow \mu_k = \frac{p(z_k=1 \mid x, \theta^{(t)}) \cdot x}{p(z_k=1 \mid x, \theta^{(t)})}$$

resp. r_{ki}

for many data:

$$\boxed{\mu_k^{(t+1)} = \frac{\sum_i p(z_{ik}=1 \mid x_i, \theta^{(t)})\, x_i}{\sum_j p(z_{jk}=1 \mid x_j, \theta^{(t)})}}$$

for the variance & the density you get the EM eqs