DATA.STAT.840 Statistical Methods for Text Data Analysis

Exercises for Lecture 4: Vector spaces and document clustering

Answered to problems 4.1(a,b only), 4.3

4.3

EM is supposed to optimize $\log(p(\underline{x}|\theta))$

EM updates parameters $\underline{\theta}$ such that:

$$p(\underline{x}|\underline{\theta}_{t+1}) \geq p(\underline{x}|\underline{\theta}_t)$$

a) $\log(p(\underline{x}|\underline{\theta}_{t+1})) - \log(p(\underline{x}|\underline{\theta}_t)) = a - b$

b) $E_{z|\underline{x},\underline{\theta}_t}[a-b] = a-b$, since z doesn't appear in $a-b$ subtraction

c) $\log(p(\underline{x}|\underline{\theta}_{t+1})) - \log(p(\underline{x}|\underline{\theta}_t))$

$$= \log\left(\frac{p(\underline{x}|\underline{\theta}_{t+1})p(\underline{z}|\underline{x},\underline{\theta}_{t+1})}{p(\underline{z}|\underline{x},\underline{\theta}_{t+1})}\right) - \log\left(\frac{p(\underline{x}|\underline{\theta}_t)p(\underline{z}|\underline{x},\underline{\theta}_t)}{p(\underline{z}|\underline{x},\underline{\theta}_t)}\right)$$

d)

$$E\left[\log\frac{p(\underline{x}|\underline{\theta}_{t+1})p(\underline{z}|\underline{x},\underline{\theta}_{t+1})}{p(z|x,\theta_{t+1})} - \log\frac{p(x|\theta_t)p(z|x,\theta_t)}{p(z|x,\theta_t)}\right)$$

$$= \sum\left(\log\frac{p(x|\theta_{t+1})p(z|x,\theta_{t+1})}{p(z|x,\theta_{t+1}} - \log\frac{p(x|\theta_t)p(z|x,\theta_t)}{p(z|x,\theta_t)}\right)p(z|x,\theta_t)$$

$$= \sum p(z|x,\theta_t)\log\frac{p(x|\theta_{t+1})p(z|x,\theta_{t+1})}{p(z|x,\theta_{t+1})} - \sum p(z|x,\theta_t)\log\frac{p(x|\theta_t)p(z|x,\theta_t)}{p(z|x,\theta_t)} =:$$

$$\left[\sum p(z|x,\theta_t)\log\frac{p(x|\theta_{t+1})p(z|x,\theta_{t+1})}{p(z|x,\theta_{t+1})} = \sum p(z|x,\theta_t)\log\frac{p(z,x|\theta_{t+1})}{p(z|x,\theta_{t+1})}\right.$$

$$= \sum p(z|x,\theta_t)\log p(z,x,\theta_{t+1}) - \sum p(z|x,\theta_t)\log p(z|x,\theta_{t+1})$$

$$\left. = Q(\theta_{t+1}|\theta_t) - \sum p(z|x,\theta_t)\log p(z|x,\theta_{t+1})\right]$$

$$\vdots = Q(\theta_{t+1}|\theta_t) - \sum p(z|x,\theta_t)\log p(z|x,\theta_{t+1}) - \left(Q(\theta_t|\theta_t) - \sum p(z|x,\theta_t)\log p(z|x,\theta_t)\right)$$

$$= Q(\theta_{t+1}|\theta_t) - Q(\theta_t|\theta_t) - \underbrace{\sum p(z|x,\theta_t)\log p(z|x,\theta_{t+1}) + \sum p(z|x,\theta_t)\log(p(z|x,\theta_t)}_{C}$$

$$C = \sum p(z|x,\theta_t)\left[\log p(z|x,\theta_t) - \log(z|x,\theta_{t+1})\right]$$

$$= \sum p(z|x,\theta_t)\log\frac{p(z|x,\theta_t)}{p(z|x,\theta_{t+1})}$$

e.g.) V: $Q(\theta_{t+1}|\theta_t) - Q(\theta_t|\theta_t) + \underbrace{\sum p(Z|X,\theta_t) \log \frac{p(Z|X,\theta_t)}{p(Z|X,\theta_{t+1})}}_{a}$

$Q(\theta_{t+1}|\theta_t) - Q(\theta_t|\theta_t) + \underbrace{a}_{\geq 0} = \log p(X|\theta_{t+1}) - \log p(X|\theta_t)$

$\log p(X|\theta_{t+1}) - \log p(X|\theta_t) \geq Q(\theta_{t+1}|\theta_t) - Q(\theta_t|\theta_t) \geq 0$

By the definition of the M-step

$$\theta_{t+1} = \underset{\theta}{\operatorname{argmax}}\ Q(\theta|\theta_t)$$

Therefore $Q(\theta_{t+1}|\theta_t) \geq Q(\theta_t|\theta_t)$

$\therefore Q(\theta_{t+1}|\theta_t) - Q(\theta_t|\theta_t) \geq 0$

So we get

$\log p(X|\theta_{t+1}) - \log p(X|\theta_t) \geq 0$