EM algorithm

A training set $D = \{x^{(1)}, x^{(2)}, \cdots, x^{(m)}\}$ m i.i.d. samples.

For each $x^{(i)}$, $i \in [m]$, there is a latent variable $y^{(i)}$ associated with it.

The log-likelihood of D is given by

$$\ell(\theta) = \sum_{i=1}^{m} \log p_r(x^{(i)};\theta)$$

$$= \sum_{i=1}^{m} \log \sum_{y} p_r(x^{(i)}, y^{(i)} = y; \theta).$$

Explicitly find $\theta^* = \max_\theta \ell(\theta)$ may be hard. Instead in EM, we optimize over a lower bound of $\ell(\theta)$.

Iterate until convergence {

(E-step) For each $i \in [m]$, set

$$p(y|i) = P_r(y^{(i)} = y \mid x^{(i)}; \theta)$$

(M-step) Set

$$\theta := \arg\max_\theta \sum_i \sum_y p(y|i) \log \frac{p(x^{(i)}, y; \theta)}{p(y|i)} \quad \}$$

Let $\hat{\ell}(\theta) = \sum_i \sum_y Q(y|i) \log \frac{p(x^{(i)}, y; \theta)}{Q(y|i)}$. We'll show

where $Q(\cdot|i)$ is some dist over $y^{(i)}$

① $\ell(\theta) \geq \hat{\ell}(\theta)$, i.e. $\hat{\ell}(\theta)$ is a lower bound on $\ell(\theta)$ and $\ell(\theta) = \hat{\ell}(\theta)$ if $Q(y|i) = p(y|i)$.

② Let $\theta^{(t)}$ be parameters obtained after the t-th iteration of the EM algorithm.

Then $\ell(\theta^{(t+1)}) \geq \ell(\theta^{(t)}) \geq \ell(\theta^{(t-1)}) \cdots \geq \ell(\theta^{(0)})$. That is, the log-likelihood $\ell(\theta)$ is monotonely increasing as EM iterates.

We have

$$\ell(\theta) = \sum_{i=1}^{m} \log \sum_y Q(y|i) \frac{P_r(x^{(i)}, y; \theta)}{Q(y|i)}$$

$$\geq \sum_{i=1}^{m} \sum_y Q(y|i) \log \frac{P_r(x^{(i)}, y; \theta)}{Q(y|i)}$$

$$= \hat{\ell}(\theta)$$

(we used the fact $\log E[X] \geq E[\log X]$)

Check Jensen's inequality

$E[f(X)] \geq f(E[X])$ for convex $f(x)$.

Here $-\log x$ is convex, then

$E[-\log X] \geq -\log(E[X])$

$\Rightarrow E[\log X] \leq \log(E[X])$

which gives ①. The equality holds if

$$\frac{P_r(x^{(i)}, y; \theta)}{Q(y|i)} = \text{const}$$

$$\Rightarrow Q(y|i) \propto P_r(x^{(i)}, y; \theta)$$

$$\Rightarrow Q(y|i) = \frac{P_r(x^{(i)}, y; \theta)}{\sum_y P_r(x^{(i)}, y; \theta)} = \frac{P_r(x^{(i)}, y; \theta)}{P_r(x^{(i)}; \theta)} = p(y|i).$$

Hence $\ell(\theta) = \hat{\ell}(\theta)$ if $Q(y|i) = p(y|i)$

Now we show ②. In this case $p(y|i) = \Pr(y^{(i)}=y \mid x^{(i)};\theta^{(t)})$.

$$\ell(\theta^{(t+1)}) \ge \sum_i \sum_y p(y|i) \log \frac{p(x^{(i)}, y; \theta^{(t+1)})}{p(y|i)} \quad (\text{by ①})$$

$$\ge \sum_i \sum_y p(y|i) \log \frac{p(x^{(i)}, y; \theta^{(t)})}{p(y|i)} \quad (\text{by the maximization in M step})$$

$$= \ell(\theta^{(t)}) \quad (\text{by } \ell(\theta^{(t)}) = \hat{\ell}(\theta^{(t)}) \text{ if } Q(y|i) = p(y|i))$$

$$\Rightarrow \ell(\theta^{(t+1)}) \ge \ell(\theta^{(t)}).$$

Hence $\ell(\theta^{(t)})$ is monotonely increasing and bounded by 0 from above, $\ell(\theta^{(t)})$ is guaranteed to converge. However, convergence to the global max of $\ell(\theta)$ is not guaranteed.

In the application, run EM several times using randomly initialized points and then choose the best one.

(show Lagrange multipliers)

Example: problem 2 and 3 of part 2 in project 3

Recall

$$\Pr(x_{obs}, z \mid \pi, \alpha) = \prod_{i=1}^{n} \Pr(z^{(i)} \mid \pi) \prod_{d=1}^{D} \Pr(x_d^{(i)} \mid \alpha_{d, z^{(i)}})^{[[x_d^{(i)} \text{ is not missing}]]}$$

E step

$$p(z^{(i)} \mid x^{(i)}, \pi, \alpha) = \frac{\pi_{z^{(i)}} \prod_{d=1}^{D} (\alpha_{d, z^{(i)}, x_d^{(i)}})^{[[x_d^{(i)} \text{ is not missing}]]}}{\sum_{j=1}^{K} \pi_j \prod_{d=1}^{D} (\alpha_{d,j,x_d^{(i)}})^{[[x_d^{(i)} \text{ is not missing}]]}}$$

M step:

$$J(\alpha, \pi) = \sum_{\vec{k}} p(z = \vec{k} \mid x) \log (x_{obs}, z \mid \alpha, \pi)$$

$$= \sum_{i=1}^{n} \sum_{k=1}^{K} p(z^{(i)} = k \mid x^{(i)}) \log p(z^{(i)} = k \mid \pi) + \sum_{i=1}^{n} \sum_{k=1}^{K} \sum_{d=1}^{D} p(z^{(i)} = k \mid x^{(i)}) \log p(x_d^{(i)} \mid z^{(i)} = k, \alpha, \pi)^{[[x_d^{(i)} \text{ is not missing}]]}$$

$$= \sum_{i=1}^{n} \sum_{k=1}^{K} p(z^{(i)} = k \mid x^{(i)}) \log \pi_k + \sum_{i=1}^{n} \sum_{k=1}^{K} \sum_{d=1}^{D} p(z^{(i)} = k \mid x^{(i)}) \log \alpha_{d,k,x_d^{(i)}}^{[[x_d^{(i)} \text{ is observed}]]}$$

in M step, we obtain

$$\max_{\alpha,\pi} J(\alpha,\pi)$$

s.t. $\sum_{k=1}^{K} \pi_k = 1$ --- ①

$\sum_{\ell=1}^{L} \alpha_{d,k,\ell} = 1$, $\forall d \in [D]$, $k \in [K]$. --- ②

$\Leftrightarrow$

$$\max_{\substack{\alpha,\pi \\ \beta,\eta_{d,k}}} \underbrace{J(\alpha,\pi) + \beta\left(\sum_{k=1}^{K}\pi_k - 1\right) + \sum_{d=1}^{D}\sum_{k=1}^{K}\eta_{d,k}\left(\sum_{\ell=1}^{L}\alpha_{d,k,\ell} - 1\right)}_{F(\alpha,\pi,\beta,\eta_{d,k};\ \forall d\in[D],\,k\in[K])}$$

necessary condition.

$$\frac{\partial F}{\partial \alpha_{d,k,\ell}} = 0 \quad \forall d\in[D],\ k\in[K],\ \ell\in[L] \qquad ③$$

$$\frac{\partial F}{\partial \pi_k} = 0 \quad \forall k\in[K] \qquad ④$$

$$\frac{\partial F}{\partial \beta} = 0 \Rightarrow ①$$

$$\frac{\partial F}{\partial \eta_{d,k}} = 0 \quad \forall d\in[D],\ k\in[K] \Rightarrow ②.$$

From ④ and ①, we have

$$\left.\begin{array}{l} \sum_{i=1}^{n} \frac{p(z^{(i)}=k \mid x^{(i)})}{\pi_k} + \beta = 0 \quad \forall k\in[K] \\ \sum_{k=1}^{K} \pi_k = 1 \end{array}\right\} \Rightarrow \pi_k = \frac{1}{n}\sum_{i=1}^{n} p(z^{(i)}=k \mid x^{(i)}).$$

From ③ and ②, we have

$$\sum_{i=1}^{n} \frac{p(z^{(i)}=k \mid x^{(i)})}{\alpha_{d,k,\ell}} [[x_d^{(i)} = \ell]] + \eta_{d,k} = 0 \quad \text{for } \forall \ell\in[L]$$

$$\sum_{\ell=1}^{L} \alpha_{d,k,\ell} = 1$$

$$\Rightarrow \alpha_{d,k,\ell} = \frac{\sum_{i=1}^{n} p(z^{(i)}=k \mid x^{(i)})\,[[x_d^{(i)}=\ell]]}{\sum_{i=1}^{n} p(z^{(i)}=k \mid x^{(i)})\,[[x_d^{(i)} \text{ is not missing}]]}$$