Expectation Maximization for Multinomial Mixture Model

Setting :-

- N 'documents'
- nth document contains W_n 'words'
- x_{nj} :- jth word in nth document
  :- a category out of V words.

Assumption :- K multinomial distribution
:- Each document generated from 1 of the K distributions.

Generative model :-

Given K multinomial distributions $\theta_1, \ldots, \theta_K$ and their rate vector Π

- for each document $n : 1 \rightarrow N$
  - Sample $z_n \sim \text{Multinomial}(\Pi)$ (1-out-of-K)
  - for each word $j : 1 \rightarrow W_n$
    - Sample $x_{nj} \mid z_n \sim \text{Multinomial}(\theta_{z_n})$ (1-out-of-V)

Recall :-

$$\log P(X \mid \theta) \geq E_q \log P(X, Z \mid \theta) - E_q \log q(Z)$$

For above model, $P(X, Z \mid \theta)$:-

$$= \prod_{n=1}^{N} P(x_{n,1}, \ldots, x_{n,w_n}, z_n \mid \theta_1, \ldots, \theta_k, \pi)$$

$$= \prod_{n=1}^{N} P(z_n \mid \pi) \prod_{j=1}^{w_n} P(x_{n,j} \mid \theta_{z_n})$$

So,

$$P(z_n \mid \pi) = \prod_{k=1}^{K} \pi_k^{z_{n,k}} \quad , \log P(z_n \mid \pi) = \sum_{k=1}^{K} z_{n,k} \log \pi_k$$

and,

$$P(x_{n,j} \mid \theta_{z_n}) = \prod_{v=1}^{V} \theta_{z_n, v}^{x_{n,j,v}}$$

$$= \prod_{k=1}^{K} \left\{ \prod_{v=1}^{V} \theta_{k,v}^{x_{n,j,v}} \right\}^{z_{n,k}}$$

$$\log P(x_{n,j} \mid \theta_{z_n}) = \sum_{k=1}^{K} z_{n,k} \left(\sum_{v=1}^{V} x_{n,j,v} \log \theta_{k,v} \right)$$

therefore,

$$\log P(X, Z \mid \theta) = \sum_{n=1}^{N} \left\{ \sum_{k=1}^{K} z_{n,k} \log \pi_k + \right.$$

$$w_n \quad K \qquad V$$

$$\sum_{j=1}^{} \left\{ \sum_{k=1} z_{n,k} \left(\sum_{v=1}^{V} x_{n,j,v} \log \theta_{k,v} \right) \right\}$$

→ (A)

Now compute $E_{q(z)} \log P(X, z \mid \theta)$

- What is $q(z)$?
- Recall $q(z)$ is chosen to 'best' capture our beliefs about latent variables

What about $p(z \mid x, \theta)$?

→ Captures all uncertainty about z given the data and model parameters θ

Can $q(z) \stackrel{?}{=} p(z \mid x, \theta)$

↳ Can $P(z \mid x, \theta)$ be computed easily?

→ If yes, proceed

→ If no, further assumptions necessary (subject matter of variational inference)

Comment:- also depends on what is designated as latent variables.

For above model,

$$P(Z|X,\theta,\pi) = P(z_1,\dots,z_n \mid X,\theta,\pi)$$

$$= \frac{P(X \mid z_1,\dots,z_N,\theta)\cdot P(z_1,\dots,z_N|\pi)}{P(X\mid\theta,\pi)}$$

$$= \frac{P(x_1,\dots,x_N \mid z_1,\dots,z_N,\theta)\prod_{n=1}^{N} P(z_n|\pi)}{P(X|\theta,\pi)}$$

$$P(z_1,\dots,z_N|X,\theta) = \frac{\prod_{n=1}^{N} P(x_n|\theta_{z_n})\cdot P(z_n|\pi)}{P(X|\theta,\pi)}$$

What is $P(X|\theta,\pi)$

$$P(X|\theta,\pi) = P(x_1\dots,x_N\mid\theta,\pi)$$

$$= \sum_{z_1,\dots,z_N} P(x_1,z_1,\dots,x_N,z_N\mid\theta,\pi)$$

$$= \sum_{z_1,\dots z_N}\prod_{n=1}^{N} P(x_n,z_n\mid\theta,\pi)$$

$$= \prod_{n=1}^{N}\left(\sum_{z_n} P(x_n,z_n\mid\theta,\pi)\right)$$

$$= \prod_{n=1}^{N} P(x_n \mid \theta, \pi)$$ { our original in-complete data likelihood }

Therefore,

$$P(z_1, \ldots, z_n \mid x_1, \ldots, x_n, \theta, \pi) = \prod_{n=1}^{N} \left[\frac{P(x_n \mid \theta_{z_n}) P(z_n \mid \pi)}{P(x_n \mid \theta, \pi)} \right]$$

$$= \prod_{n=1}^{N} P(z_n \mid x_n, \theta, \pi) \quad \text{— (B)}$$

- Fully factorized posterior.
- Each $P(z_n \mid x_n, \theta, \pi)$ easily computable
- Hence $q(z) = p(z \mid x, \theta)$

Recall in EM :-

- Given $\theta^{(old)}$, we compute $q(z \mid \theta^{old})$
- Compute $E_{q(z \mid \theta^{old})} \log P(x, z \mid \theta) - \underbrace{E_{q(z \mid \theta^{old})} \log q(z \mid \theta^{old})}_{\text{entropy}}$
- maximize above w.r.t 'θ'
- entropy term independent of 'θ'

So, we compute $q(z|\theta^{old}) = p(z|x, \theta^{old})$

$$p(z_n|x_n, \theta^{old}, \pi^{old}) = \frac{p(x_n|\theta^{old}_{z_n}) \cdot p(z_n|\pi^{old})}{\sum_{k=1}^{K} \pi_k^{old} \cdot p(x_n|\theta_k^{old})}$$

$$= [r_{n,1}, \ldots, r_{n,K}]$$

$$p(z_n = \begin{bmatrix}1\\0\\\vdots\\0\end{bmatrix} | x_n, \theta^{old}, \pi^{old})$$

and compute

$$E_{q(z|\theta^{old})} \log p(x, z|\theta)$$

$$= E_{q(z|\theta^{old})} \sum_{n=1}^{N} \left\{ \sum_{k=1} z_{n,k} \log \pi_k + \sum_{j=1}^{w_n} \left\{ \sum_{k=1}^{K} z_{n,k} \left(\sum_{v=1}^{V} x_{n,j,v} \log \theta_{k,v} \right) \right\} \right\}$$

- Because of fully factorized posterior $p(z|x, \theta^{old})$, computing Expectation is trivial

So,

$$= \sum_{n=1}^{N} \left\{ \sum_{k=1}^{K} r_{n,k} \log \pi_k + \sum_{j=1}^{W_n} \left\{ \sum_{k=1}^{K} r_{n,k} \left(\sum_{v=1}^{V} x_{n,j,v} \log \theta_{k,v} \right) \right\} \right\}$$

a little re-arrangement,

$$= \sum_{k=1}^{K} \left(\sum_{n=1}^{N} r_{n,k} \right) \log \pi_k +$$

$$\sum_{k=1}^{K} \sum_{v=1}^{V} \left(\sum_{n=1}^{N} r_{n,k} \sum_{j=1}^{W_n} x_{n,j,v} \right) \log \theta_{k,v}$$

- Recognize this lets us do ML estimates of multinomials $\pi, \theta_1, \ldots, \theta_K$ independently
- We can just read-off result for each multinomial

$$\Longrightarrow \quad \pi_k = \sum_{n=1}^{N} r_{n,k} \; / \; N$$

$\theta_{k,v}$:- vth-component of kth distribution

$$\Longrightarrow \quad \theta_{k,v} = \frac{\sum_{n=1}^{N} r_{n,k} \sum_{j=1}^{w_n} x_{n,j,v}}{\sum_{v=1}^{V} \sum_{n=1}^{N} r_{n,k} \sum_{j=1}^{w_N} x_{n,j,v}}$$

These π and θ_k become π^{old} and θ_k^{old} for next iteration.

Implementation Comments:-

- Compute $E_{q(z)} \log P(x,z|\theta) - E_q \log q(z)$ in every iteration
- Should improve in every iteration.
- Stop when improvement < threshold

Moral

EM decomposes the mixture model in independent ML estimates of model components in each iteration.