Collapsed Sampling

Recap :- Bayesian mixture model

Generative model :-

Given hyper-parameters

(i) Sample mixture components and rate parameter

(ii) For each data point

a) Sample mixture component membership

b) Sample data point given mixture membership and mixture components

prior, likelihood

prior, likelihood pair

Example :- Multinomial mixture model

Given $\gamma = [\gamma_1, \ldots, \gamma_V]$, $\alpha = [\alpha_1, \ldots, \alpha_K]$

V: number of unique items

K: number of mixture components

Sample

$$\theta_k \sim \text{Dir}(\theta_k \mid \gamma) \quad k = 1, \ldots, K$$

$$\pi \sim \text{Dir}(\pi \mid \alpha)$$

$$z_n \mid \pi \sim \text{Multinomial}(z_n \mid \pi) \quad n = 1, \ldots, N$$

$$x_{n,j} \mid z_n, \theta \sim \text{Multinomial}(x_{n,j} \mid \theta_{z_n}) \quad j = 1, \ldots, J_n$$

prior, lik

prior, lik

* Different prior-likelihood give different models

* Of particular importance are "conjugate" prior, likelihood pairs

* In Bayesian setting, we are interested in

$$Pr(z_{new} \mid x_{new}, X_{Train}, \text{hyper-parameters})$$

$$\approx \frac{1}{M} \sum_{m} Pr(z_{new} \mid x_{new}, \theta_1^{(m)}, \ldots, \theta_K^{(m)}, \pi^{(m)})$$

* Sampling based solution

* Sampling involves drawing samples from $P(\theta_1, \ldots, \theta_K, \pi, z_1, \ldots, z_N \mid X_{Train}, h)$

For multinomial mixture :-

* $\theta_k \mid \theta_{-k}, z, x, \pi \sim Dir(\theta_k \mid [\sum_{n:z_n=k} C_{n,1} + \gamma_1, \ldots, \sum_{n:z_n=k} C_{n,V} + \gamma_V])$ (A)

* $\pi \mid z, x, \theta \sim Dir(\pi \mid [\sum_{n=1}^{N} z_{n,1} + \alpha_1, \ldots, \sum_{n=1}^{N} z_{n,K} + \alpha_K])$

* $z_n = k \mid z_{-n}, \theta, \pi, x_{n,1}, \ldots, x_{n,J_n} \sim \frac{\left[\prod_{j=1}^{J_n} \text{Multinomial}(x_{n,j} \mid \theta_k)\right] \cdot \pi_k}{\sum_{m=1}^{K} \prod_{j=1}^{J_n} \text{Multinomial}(x_{n,j} \mid \theta_m)\, \pi_m}$

(A) $C_{n,v} = \sum_{j=1}^{J_n} x_{n,j,v}$

$= \#$ word 'v' in document 'n'

* Notice, Sampling of z_n only depends on $\theta_1, \ldots, \theta_K$ and π (and nth-document)

* z_n and z_m (local states) only influence each other through $\theta_1, \ldots, \theta_K$ and π (global state)

* Intuitively, Speed of convergence depends on how frequently we can update "global state"

* Can we update "global state" more frequently
  - → possibly (expensive)
  - → Can we simply get rid of "global state"
  - → getting rid = marginalizing!

Or :- $Pr(z_n \mid z_{-n}, x_{Train}, \text{hyper-parameter})$?

Our Bayesian prediction now becomes :-

$$Pr(z_{new} \mid x_{new}, x_{Train}, h) \approx \frac{1}{M} \sum_m Pr(z_{new} \mid x_{new}, x_{Train}, z^{(m)}_{Train}, h)$$

$$Pr(z_n | z_{-n}, x_n, x_{-n}, \gamma, \alpha) = \frac{Pr(z_n, z_{-n}, x_n, x_{-n} | \gamma, \alpha)}{\sum_{z_n} Pr(z_n, z_{-n}, x_n, x_{-n} | \gamma, \alpha)}$$

$$= \frac{Pr(z_n | z_{-n}, x_{-n}, \gamma, \alpha) \cdot P(x_n | z_n, z_{-n}, x_{-n}, \gamma, \alpha)}{\sum_{z_n} Pr(z_n | z_{-n}, x_{-n}, \gamma, \alpha) \cdot P(x_n | z_n, z_{-n}, x_{-n}, \gamma, \alpha)}$$

from Sesion 4c

$$Pr(z_n = l | z_{-n}, x_n, \gamma, \alpha) = \int_{\pi} Pr(z_n = l | \pi, z_{-n}, \gamma, \alpha) Pr(\pi | z_{-n}, \gamma, \alpha) \, d\pi$$

$$= \frac{1}{N - 1 + (\sum_k \alpha_k)} \left[\left(\sum_{m \neq n} z_{m,l} \right) + \alpha_l \right]$$

number of assignments to component 'l' (excluding current data point)

Similarly :-

$$Pr(x_n | z_n = l, z_{-n}, x_{-n}, \gamma, \alpha)$$

$$= \int_{\theta_l} Pr(x_n | z_n = l, z_{-n}, x_{-n}, \gamma, \alpha, \theta_l) \cdot Pr(\theta_l | x_{-n}, z_{-n}, \gamma, \alpha) \, d\theta_l$$

why just θ_l?

$$= \int_{\theta_l} d\theta_l \left[\prod_{j=1}^{J_n} \text{Multinomial}(x_{n,j} \mid \theta_l) \right] \text{Dir}\left(\theta_l \mid \left[\sum_{\substack{m: m \neq n \\ \text{and } z_m = l}} c_{m,1} + \gamma_1, \ldots, \sum_{\substack{m: m \neq n \\ \text{and } z_m = l}} c_{m,V} + \gamma_V \right] \right)$$

$$\left[\int d\theta_l \prod_{v=1}^{V} (\theta_{l,v})^{c_{n,v}} \cdot \prod_{v=1}^{V} (\theta_{l,v})^{\sum_{m: m \neq n, z_m = l} c_{m,v} + \gamma_v - 1} \right] \cdot \frac{\Gamma\left(\sum_{v=1}^{V} \left(\sum_{\substack{m: m \neq n \\ z_m = l}} c_{m,v} + \gamma_v \right) \right)}{\prod_{v=1}^{V} \Gamma\left(\sum_{\substack{m: m \neq n \\ z_m = l}} c_{m,v} + \gamma_v \right)}$$

$$= \frac{\prod_{v=1}^{V} \Gamma\left(\sum_{d: z_d = l} c_{d,v} + \gamma_v \right)}{\Gamma\left(\sum_{v=1}^{V} \left(\sum_{d: z_d = l} c_{d,v} + \gamma_v \right) \right)} \cdot \frac{\Gamma\left(\sum_{v=1}^{V} \left(\sum_{\substack{m: m \neq n \\ z_m = l}} c_{m,v} \right) + \gamma_v \right)}{\prod_{v=1}^{V} \Gamma\left(\sum_{\substack{m: m \neq n \\ z_m = l}} c_{m,v} + \gamma_v \right)}$$

number of 'v' words assigned to component 'l' $+ c_{n,v}$

number of words assigned to component l + number of words in document 'n' $(= J_n)$

number of 'v' words assigned to component 'l'

number of words assigned to component 'l':

$$= \frac{\Gamma\left(\sum_{v=1}^{V} \left(\sum_{\substack{m: m \neq n \\ z_m = l}} c_{m,v} + \gamma_v \right) \right)}{\Gamma\left(\sum_{v=1}^{V} \left(\sum_{\substack{m: m \neq n \\ z_m = l}} c_{m,v} + \gamma_v \right) + J_n \right)} \cdot \prod_{v=1}^{V} \frac{\Gamma\left(\sum_{\substack{m: m \neq n \\ z_m = l}} c_{m,v} + \gamma_v + c_{n,v} \right)}{\Gamma\left(\sum_{\substack{m: m \neq n \\ z_m = l}} c_{m,v} + \gamma_v \right)}$$

Using Gamma function recurrance

$$\because \Gamma(x+1) = x\Gamma(x)$$

$$= \left[\frac{1}{\sum_{v=1}^{V} \left(\sum_{\substack{m: m \neq n \\ z_m = l}} C_{m,v} + \gamma_v \right)} \right]^{J_n} \cdot \prod_{v=1}^{V} \left(\sum_{\substack{m: m \neq n \\ z_m = l}} C_{m,v} + \gamma_v \right)^{C_{n,v}}$$

In Summary :-

$$P(z_n = l \mid z_{-n}, x_n, x_{-n}, \alpha, \gamma) \propto \left[\frac{1}{N-1+(\sum_k \alpha_k)} \left(\sum_{m \neq n} z_{m,l} + \alpha_l \right) \right] \times$$

$$\left[\frac{1}{\sum_{v=1}^{V} \left(\sum_{\substack{m: m \neq n \\ z_m = l}} C_{m,v} + \gamma_v \right)} \right] \prod_{v=1}^{V} \left(\sum_{\substack{m: m \neq n \\ z_m = l}} C_{m,v} + \gamma_v \right)^{C_{n,v}}$$

Implementation :- everything is expressed as counts

i-e * Count numbr of assignments of documents to mixture components.

* count number of assignments of words to mixture components.

* Can be very efficiently implemented by maintaining conditional frequency tables.

* entire inference is coupled (not easy to parallelize)

* Collapsed inference reduces dimensionality of the problem a lot
  → resulting in faster 'mixing'
  means MCMC procedures reaches stationary distribution (joint distribution much faster)

* Other benefits :- application to non-parametric models.