Bayesian treatment of Multinomial Mixture

Generative model :-

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

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

for $k = 1 : K$

$\theta_k \sim \text{Dirichlet}(\theta \mid \gamma)$

for document, $n = 1, \ldots, N$

$z_n \sim \text{Multinomial}(z \mid \pi)$

for word, $j = 1, \ldots, J_N$

$x_{n,j} \sim \text{Multinomial}(x \mid \theta_{z_n})$

prior, likelihood

prior, likelihood

$z_n :-$ 1-out of-K , $z_n = k = [0, 0, \ldots, \underset{\uparrow k}{1}, \ldots 0]$

$x_{n,j} :-$ 1-out of-V

Note :-
- We still have parameters without prior i.e K, α, γ
- We can put priors on them
- Those priors will have their own parameters
- We can put priors on them
- and so forth ...

Goal:-

predict $P(z_{new} \mid x_{new}, X_{Train}, \underbrace{hyper\text{-}parameters}_{h})$

$$\int P(z_{new} \mid x_{new}, \theta_1, \ldots, \theta_K, \pi, X_{Train}, h) \cdot P(\theta_1, \ldots, \theta_K, \pi \mid X_{Train}, h) \, d\theta_1 \ldots d\theta_K \, d\pi \longrightarrow (A)$$

$$P(z_{new} \mid x_{new}, \theta_1, \ldots, \theta_K, \pi, X_{Train}, h) = \frac{P(x_{new} \mid z_{new}, \theta_1, \ldots, \theta_K, \pi, X_{Train}, h) \cdot P(z_{new} \mid \theta_1, \ldots, \theta_K, \pi, X_{Train}, h)}{P(x_{new}, \theta_1, \ldots, \theta_K, \pi, X_{Train}, h)}$$

$$= \frac{P(x_{new} \mid \theta_{z_{new}}) \cdot P(z_{new} \mid \pi)}{\sum_{z_{new}} P(x_{new} \mid \theta_{z_{new}}) \cdot P(z_{new} \mid \pi)} \longrightarrow \text{We have seen this many times. } (q(z) \text{ in EM})$$

$$= q(z_{new}; \theta_{z_{new}}, \pi)$$

So in (A),

$$= \int q(z_{new}; \theta_{z_{new}}, \pi) \underbrace{P(\theta_{z_{new}}, \pi \mid X_{Train}, h)}_{\text{We don't need all } \theta\text{'s}} d\theta_{z_{new}} \, d\pi$$

Looks like $\underset{P(\theta_{z_{new}}, \pi \mid X_{Train}, h)}{E} q(z_{new}; \theta_{z_{new}}, \pi)$

$$\approx \frac{1}{M} \sum_{m=1}^{M} q\left(z_{new}; \theta_{z_{new}}^{(m)}, \pi^{(m)}\right) \longrightarrow (\tilde{A})$$

Where $\theta_{z_{new}}^{(m)}$ and $\pi^{(m)}$ are mth-samples from $P(\theta_{z_{new}}, \pi \mid x_{Train}, h)$!

* Hence, instead of finding $P(\theta_{z_{new}}, \pi \mid x_{Train}, h)$ (i.e posterior distribution of parameters of interest) we can just "draw some samples" from $P(\theta_{z_{new}}, \pi \mid x_{Train}, h)$

* We'll call it Approach (1) to approximate (A).

* Approach (2) would be to come up with an answer for $P(\theta_{z_{new}}, \pi \mid x_{Train}, h)$
  - If not possible to find exact form, then possible solutions are
    * Expectation propagation
    * Variational approximation.

Approach (1) :- Sampling from the posterior.

General recipe :-

X :- data, $L = [l_1, \ldots, l_N]$ all latent variables (including $\theta_1, \ldots, \theta_K, \pi$ (model parameters))

h :- hyper-parameters

(i) Step 1 :- Write $P(X, L \mid h)$

(ii) Step 2 :- $l_n \sim P(l_n \mid X, L_{-n}, h) = \dfrac{P(X, L \mid h)}{P(X, L_{-n} \mid h)}$

$L \setminus l_n$

* Should be easy !
* true for models using conjugacy.

for multinomial mixture model,

$$= P(\theta_1, \ldots, \theta_K, \pi, z_1, \ldots, z_N, x_1, \ldots, x_N \mid \gamma, \alpha)$$

$$= \left[\prod_{k=1}^{K} P(\theta_k \mid \gamma)\right] P(\pi \mid \alpha) \cdot \prod_{n=1}^{N} P(z_n \mid \pi) \cdot \prod_{j=1}^{J_n} P(x_{n,j} \mid \theta_{z_n})$$

→ (B)

$$P(z_n \mid x, z_{-n}, h) = \frac{\left[\prod_{k=1}^{K} P(\theta_k \mid \gamma)\right] P(\pi \mid \alpha) \prod_{m=1}^{N} P(z_m \mid \pi) \prod_{j=1}^{J_m} P(x_{m,j} \mid \theta_{z_m})}{\left\{\left[\prod_{k=1}^{K} P(\theta_k \mid \gamma)\right] P(\pi \mid \alpha) \prod_{m \neq n} P(z_m \mid \pi) \prod_{j=1}^{J_m} P(x_{m,j} \mid \theta_{z_m}) \cdot \underbrace{P(x_n \mid \theta_1, \dots, \theta_K, \pi)}\right\}}$$

$$\sum_{z_1} P(x_n \mid \theta_{z_1}) P(z_1 \mid \pi)$$

$$= \frac{P(z_n \mid \pi) \cdot \prod_{j=1}^{J_n} P(x_{n,j} \mid \theta_{z_n})}{\sum_{z_1} P(x_n \mid \theta_{z_1}) \cdot P(z_1 \mid \pi)} \quad \text{——(i)}$$

$$P(\theta_k \mid \theta_{-k}, \pi, z_N, x, h) = \frac{P(\theta_k \mid \gamma) \prod_{n: z_n = k} \prod_{j=1}^{J_n} P(x_{n,j} \mid \theta_k)}{\int_{\theta_k} P(\theta_k \mid \gamma) \prod_{n: z_n = k} \prod_{j=1}^{J_n} P(x_{n,j} \mid \theta_k)}$$

We solved it in Session (4-a)

$$= \text{Dir}\left(\theta \mid \left[\sum_{n: z_n = k} x_{n,1} + \gamma_1, \dots, \sum_{n: z_n = k} x_{n,V} + \gamma_V\right]\right)$$

——(ii)

$$P(\pi \mid \theta_1, \ldots, \theta_k, z_N, \chi, h) = \frac{P(\pi \mid \alpha) \cdot \prod_{n=1}^{N} P(z_n \mid \pi)}{\int_{\pi} P(\pi \mid \alpha) \cdot \prod_{n=1}^{N} P(z_n \mid \pi)\, d\pi}$$

Same as above

$$= \mathrm{Dir}\left(\pi \mid \left[\sum_{n=1}^{N} z_{n,1} + \alpha_1, \ldots, \sum_{n=1}^{N} z_{n,k} + \alpha_k\right]\right)$$

—— (iii)

Algorithm :—

((i), (ii), (iii) would be very easy to write once we cover graph representations)

Initialize $\theta_1^{(0)}, \ldots, \theta_k^{(0)}, \pi^{(0)}$

draw samples $z_N^{(m)}, \theta_1^{(m)}, \ldots, \theta_k^{(m)}, \pi^{(m)}$

Use 'M' samples to compute $\tilde{A}$

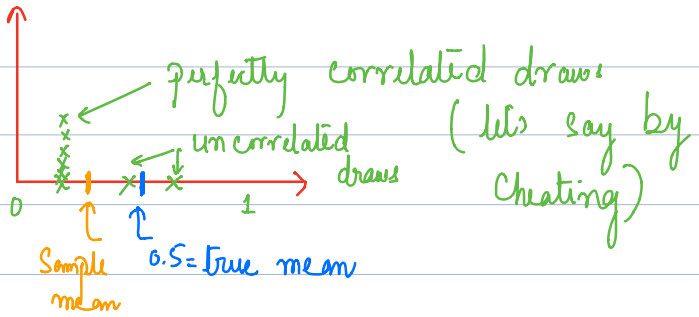
From the samples,

* average value for any distribution can be computed
* Samples from any marginal distribution '$P(l_n)$' can obtained by only keeping samples of l_n
* Sampling in this fashion approximates draws

from the full-joint distribution.

Practical Challenges :-

* (A) only works with independent draws from a distribution.



Solution :- * measure empirical correlation among samples across various intervals.
* Choose an interval among samples with low empirical correlation.

* Initial draws may be coming from a poor approximation of joint-distribution.

  * Solution :- ignore initial draws.

Collapsing

$$P(z_n \mid z_{-n}, \theta_1, \ldots, \theta_k, \pi, h, x)$$

$$= \frac{P(x_n \mid \theta_{z_n}) \, P(z_n \mid \pi)}{\sum_{z_n} P(x_n \mid \theta_{z_n}) \cdot P(z_n \mid \pi)} \quad \{\text{from (ii)}\}$$

If we know z_{-n} (cluster-assignments of all but nth point), do we need $\theta_1, \ldots, \theta_k$ and π?

Or

$$P(z_n \mid z_{-n}, x, h) = \frac{P(x \mid z_n, z_{-n}, h) \cdot P(z_n \mid z_{-n}, h)}{\sum_{z_n} P(x \mid z_n, z_{-n}, h) \cdot P(z_n \mid z_{-n}, h)}$$

$$P(z_n \mid z_{-n}, h) = \int P(z_n \mid z_{-n}, \pi, \alpha) \cdot P(\pi \mid z_{-n}, \alpha) \, d\pi$$

$$= \int \prod_{k=1}^{K} (\pi_k)^{z_{n,k}} \, P(\pi \mid z_{-n}, \alpha) \, d\pi$$

$$P(\pi \mid z_{-n}, \alpha) = \frac{\Gamma\left(\sum_{k=1}^{K} \left(\sum_{m \neq n} z_{m,k} + \alpha_k\right)\right)}{\prod_k \Gamma\left(\sum_{m \neq n} z_{m,k} + \alpha_k\right)} \prod_k (\pi_k)^{\sum_{m \neq n} z_{m,k} + \alpha_k - 1}$$

$$= \frac{\Gamma\left(N - 1 + \sum_k \alpha_k\right)}{\prod_k \Gamma\left(\sum_{m \neq n} z_{m,k} + \alpha_k\right)} \cdot \prod_k \left(\pi_k\right)^{\sum_{m \neq n} z_{m,k} + \alpha_k - 1}$$

$$\frac{\Gamma\left(N - 1 + \sum_k \alpha_k\right)}{\prod_k \Gamma\left(\sum_{m \neq n} z_{m,k} + \alpha_k\right)} \int \prod_k \pi_k^{\left(\sum_m z_{m,k} + \alpha_k - 1\right)} \, d\pi$$

$$= \frac{\Gamma\left(N - 1 + \sum_k \alpha_k\right)}{\prod_k \Gamma\left(\sum_{m \neq n} z_{m,k} + \alpha_k\right)} \cdot \frac{\prod_k \Gamma\left(\sum_m z_{m,k} + \alpha_k\right)}{\Gamma\left(N + \sum_k \alpha_k\right)}$$

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

$$= \frac{\Gamma\left(N - 1 + \sum \alpha_k\right)}{\left(N - 1 + \sum_k \alpha_k\right) \Gamma\left(N - 1 + \sum_k \alpha_k\right)} \prod_k \left[\frac{\Gamma\left(\sum_m z_{m,k} + \alpha_k\right)}{\Gamma\left(\sum_{m \neq n} z_{m,k} + \alpha_k\right)} \right]$$

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

if $z_n = [\, 0 \cdots \underset{\uparrow}{1} \cdots 0 \,]$, l

Similarly $P(x \mid z_n, z_{-n}, h)$ (remaining).

Comment :-

(i) Results in very easy, numerically stable sampler

(ii) One of the major ways, non-parametric models are computed.