Dirichlet Process

1. Bayesian mixture model with K components

$$\Rightarrow \begin{bmatrix} \pi \mid \alpha \sim \mathrm{Dir}\left(\frac{\alpha}{K}, \cdots \frac{\alpha}{K}\right) & \theta_k^* \sim H \\ z_i \mid \pi \sim \mathrm{Mult}(\pi) & x_i \mid z_i, \{\theta_k^*\}_{k=1}^K \sim F(\theta_{z_i}^*) \end{bmatrix}$$

2. Dirichlet Process ($G \sim DP(\alpha, H)$)

: distribution over the probability measure on Θ whose marginals are dirichlet

$\Rightarrow (G(A_1), \cdots, G(A_K)) \sim \mathrm{Dir}(\alpha H(A_1), \cdots, \alpha H(A_K))$ for every finite partitions of Θ

$\Rightarrow {}^\forall A,\ \mathbb{E}[G(A)] = H(A) \longrightarrow H$ is the mean of DP

$\Rightarrow {}^\forall A,\ \mathrm{Var}[G(A)] = H(A)(1-H(A))/(\alpha+1) \longrightarrow \alpha$ is the inverse variance of DP

(As $\alpha \to \infty$, $G \to H$ weakly or pointwise)

① Existence

$\Rightarrow$ Exchangeability : $p(\theta_1, \cdots, \theta_n) = \prod_{i=1}^n p(\theta_i \mid \theta_1, \cdots, \theta_{i-1}) = p(\theta_{\sigma(1)}, \cdots, \theta_{\sigma(n)}) \longrightarrow n$ may be ∞

$\Rightarrow$ De Finetti's theorem : ${}^\exists G$ s.t. $p(\theta_1, \cdots, \theta_n) = \int p(G) \prod_{i=1}^n G(\theta_i)\, dG$

: for infinitely exchangeable sequence, there is a random distribution G s.t. the sequence is composed of i.i.d. draws from it.

② Posterior distribution ($G \mid \theta_1, \cdots, \theta_n \sim DP\left(\alpha+n, \frac{\alpha}{\alpha+n}H + \frac{n}{\alpha+n}\cdot\frac{\sum_{i=1}^n \delta_{\theta_i}}{n}\right)$)

$\Rightarrow \theta_1, \cdots, \theta_n \sim \mathrm{Mult}(G(A_1), \cdots, G(A_K))$ where $A_1, \cdots, A_K \subset \Theta$, $n_k = |\{i : \theta_i \in A_k\}|$

$\Rightarrow (G(A_1), \cdots, G(A_K)) \mid \theta_1, \cdots, \theta_n \sim \mathrm{Dir}(\alpha H(A_1)+n_1, \cdots, \alpha H(A_K)+n_K)$

$\Rightarrow$ As $\alpha \to 0$, prior H becomes non-informative

$\Rightarrow$ As $n \to \infty$, posterior $G \mid \theta_1, \cdots, \theta_n$ is dominated by the empirical distribution

③ Predictive distribution ($\theta_{n+1} \mid \theta_1, \cdots, \theta_n \sim \frac{\alpha}{\alpha+n}H + \frac{n}{\alpha+n}\cdot\frac{\sum_{i=1}^n \delta_{\theta_i}}{n}$)

$$\Rightarrow {}^\forall A,\ p(\theta_{n+1} \in A \mid \theta_1, \cdots, \theta_n) = \int p(G \mid \theta_1, \cdots, \theta_n)\, p(\theta_{n+1} \in A \mid G, \theta_1, \cdots, \theta_n)\, dG$$
$$= \int p(G \mid \theta_1, \cdots, \theta_n)\, G(A)\, dG$$
$$= \mathbb{E}[G(A) \mid \theta_1, \cdots, \theta_n] = \frac{\alpha}{\alpha+n}H(A) + \frac{n}{\alpha+n}\cdot\frac{\sum_{i=1}^n \delta_{\theta_i}(A)}{n}$$

$$\Rightarrow \theta_{n+1} \mid \theta_1, \cdots, \theta_n \sim \frac{\alpha}{\alpha+n}H + \frac{n}{\alpha+n}\cdot\frac{\sum_{i=1}^n \delta_{\theta_i}}{n}$$

(Base function of the posterior DP is the predictive distribution)

$$\Rightarrow \theta_{n+1} \mid \theta_1, \cdots, \theta_n \sim \frac{\alpha}{\alpha+n}H + \frac{n}{\alpha+n}\cdot\frac{\sum_{k=1}^K n_k \delta_{\theta_k^*}}{n}$$ (rich-gets-richer phenomenon)

$\longrightarrow (\theta_1^*, \cdots, \theta_K^*)$ are unique values in $(\theta_1, \cdots, \theta_n) \longrightarrow$ clustering property

$\longrightarrow n_k$ is the number of repeats of θ_k^*

$\longrightarrow \sum_{i=1}^n \frac{\alpha}{\alpha+i} \approx \alpha \log\left(1+\frac{n}{\alpha}\right) = O(\alpha \log n) \longrightarrow$ expectation (# of unique values)

④ Chinese Restaurant Process ($\approx$ Polya urn scheme)

: customers sit on an infinite number of round tables

: balls are colored among an infinite number of colours and put inside an urn

$\Rightarrow$ express an uncertainty about possible number of components

(pick a new table ($\approx$colour) with probability $\frac{\alpha}{\alpha+n}$

pick an old table ($\approx$colour) with probability $\frac{n}{\alpha+n}$

$\longrightarrow$ positive probability on picking the same value

$\longrightarrow$ for a long enough sequence, any draw will be repeated by another draw

$\therefore$ G is composed of a weighted sum of point masses as discrete distribution

(support of G is countably infinite atoms drawn from H)

⑤ stick-breaking construction

$$\begin{bmatrix} \boxed{\beta_k \sim \mathrm{Beta}(1, \alpha)} \\ \pi_k \sim \beta_k \prod_{\ell=1}^{k-1}(1-\beta_\ell) \end{bmatrix} \to \pi \sim GEM(\alpha) \qquad \begin{bmatrix} \theta_k^* \sim H \\ G = \sum_{k=1}^\infty (\pi_k)\delta_{\theta_k^*} \end{bmatrix} \to \sum_{k=1}^\infty \pi_k = 1$$

$\longrightarrow$ sample takes value θ_k^* with probability π_k

$\longrightarrow \theta_k^*$ and π are mutually independent

⑥ Dirichlet Process Mixture Model

$$\begin{bmatrix} \pi \sim GEM(\alpha) & \theta_k^* \sim H \\ z_i \sim \mathrm{Mult}(\pi) & x_i \mid z_i, \{\theta_k^*\}_{k=1}^\infty \sim F(\theta_{z_i}^*) \end{bmatrix}$$

$\longrightarrow$ infinite limit of finite mixture models

3. Hierarchical Dirichlet Process ($\overset{\text{global}}{G_0} \sim DP(\gamma, H)$, $\overset{\text{local}}{G_j} \sim DP(\alpha, G_0)$ for $j = 1, 2, \cdots, J$)

: clusters are shared across multiple, nested groupings of data

$\Rightarrow$ forces the base measure to be discrete as a draw from the dirichlet process

$\Rightarrow$ forces the atoms to be shared among the multiple dirichlet processes

($G_0 = \sum_{k=1}^\infty \beta_k \boxed{\delta_{\theta_k^*}}$ where $\beta \sim GEM(\gamma)$)

($G_j = \sum_{k=1}^\infty \pi_{jk} \boxed{\delta_{\theta_k^*}}$ where $\pi_j \sim DP(\alpha, \beta)$) ($\theta_k^*$: shared ψ_j)

$\longrightarrow (G_j(A_1), \cdots, G_j(A_r)) \sim \mathrm{Dir}(\alpha_0 G_0(A_1), \cdots, \alpha_0 G_0(A_r))$

$\longrightarrow \left(\sum_{k \in k_1} \pi_{jk}, \cdots, \sum_{k \in k_r} \pi_{jk}\right) \sim \mathrm{Dir}\left(\alpha_0 \sum_{k \in k_1}\beta_k, \cdots, \alpha_0 \sum_{k \in k_r}\beta_k\right)$

① Chinese Restaurant Franchise

: customers visit a franchise with J restaurants and sit on an infinite number of round tables, each of which serves the same dishes

($\phi_{ji} \sim G_j$: the dish for the ith customer heading to the jth restaurant

$\psi_{jt} \sim G_0$: the dish for the tth table in the jth restaurant

$\theta_k^* \sim H$: kth dish

(m_{jk} : # of ψ_{jt}'s associated with $\theta_k^* \longrightarrow m_k = \sum_j m_{jk}$

n_{jt} : # of ϕ_{ji}'s associated with ψ_{jt})

$\Rightarrow$ the dish is picked according to its popularity in the whole restaurants

$\longrightarrow \psi_{jt+1} \mid \psi_{j1}, \psi_{j2}, \cdots, \psi_{jt}, \gamma, H \sim \frac{\gamma}{\gamma + \sum_k m_k} H + \sum_{k=1}^k \frac{m_k}{\gamma + \sum_k m_k}\delta_{\theta_k^*}$

$\Rightarrow$ the table is picked according to its popularity in the restaurant

$\longrightarrow \phi_{j,i+1} \mid \phi_{j1}, \phi_{j2}, \cdots, \phi_{ji}, \alpha, G_0 \sim \frac{\alpha}{\alpha+i}G_0 + \sum_{t=1}^{T_j}\frac{n_{jt}}{\alpha+i}\delta_{\psi_{jt}}$

② stick-breaking construction

$$\begin{bmatrix} \beta \sim GEM(\gamma) & \theta_k^* \sim H \\ \pi_{jk} \sim \mathrm{Beta}(\alpha_0\beta_k, \alpha_0(1-\sum_{\ell=1}^k \beta_\ell)) & G_j = \sum_{k=1}^\infty \pi_{jk}\delta_{\theta_k^*} \\ \pi_{jk} = \pi_{jk}' \prod_{\ell=1}^{k-1}(1-\pi_{j\ell}') & \end{bmatrix}$$

$\longrightarrow \left(\sum_{\ell=1}^{k-1}\pi_{j\ell}, \pi_{jk}, \sum_{\ell=k+1}^\infty \pi_{j\ell}\right) \sim \mathrm{Dir}\left(\alpha_0\sum_{\ell=1}^{k-1}\beta_\ell, \alpha_0\beta_k, \alpha_0\sum_{\ell=k+1}^\infty \beta_\ell\right)$

$\longrightarrow \frac{1}{1-\sum_{\ell=1}^{k-1}\pi_{j\ell}}\left(\pi_{jk}, \sum_{\ell=k+1}^\infty \pi_{j\ell}\right) \sim \mathrm{Dir}\left(\alpha_0\beta_k, \alpha_0\sum_{\ell=k+1}^\infty \beta_\ell\right)$

$\therefore \pi_{jk} = \pi_{jk}' \prod_{\ell=1}^{k-1}(1-\pi_{j\ell}')$ where $\pi_{jk}' = \frac{\pi_{jk}}{1-\sum_{\ell=1}^{k-1}\pi_{j\ell}} \sim \mathrm{Beta}\left(\alpha_0\beta_k, \alpha_0(1-\sum_{\ell=1}^k \beta_\ell)\right)$

② Hierarchical Dirichlet Process Mixture Model

$$\begin{bmatrix} \beta \sim GEM(\gamma) & \theta_k^* \sim H \\ \pi_j \sim DP(\alpha, \beta) & x_{ji} \mid z_{ji}, \{\theta_k^*\}_{k=1}^\infty \sim F(\theta_{z_{ji}}^*) \\ z_{ji} \sim \mathrm{Mult}(\pi_j) & \end{bmatrix}$$

4. Nested Chinese Restaurant Process

: customers visit L distinct restaurants among an infinite number of them following the cards on the chosen tables

$\Rightarrow$ prior to be used when modeling topic hierarchies

$\Rightarrow$ express an uncertainty about possible L-level trees

Topic Modeling

: the words in a document are generated according to a mixture model where the mixing proportions are random and document-specific.

(corpus ⟶ documents ⟶ topics ⟶ words)

1. Latent Dirichlet Allocation (LDA)

① For each document in a corpus

① sample a k-dimensional topic proportions : $\theta \sim Dir(\alpha)$

② For each word in a document

① sample a topic : $z \sim Mult(\theta)$

② sample a word from the corresponding topic : $w \sim p(w \mid z, \boxed{\beta})$ ← estimate using variational EM

2. Hierarchical LDA

⇒ corpus : tree , document : path , topic : restaurant , word : table

※ Finite Depth

① For each document in a corpus

① Let c_1 be the root restaurant

② For each level $\ell = \{2, \cdots, L\}$

① Pick a table in restaurant $c_{\ell-1}$ as a way of $CRP(\gamma_{\ell-1})$

② Let c_ℓ be the restaurant referred to by the table

⇒ $c \sim nCRP(\gamma)$

③ Sample an L-dimensional topic proportions : $\theta \sim Dir(\alpha)$

④ For each word in a document

① sample a topic : $z \sim Mult(\theta)$

② sample a word from the corresponding topic : $w \sim CRP(\gamma_z)$

※ Infinite Depth

① For each node in the tree

① sample a topic : $\beta \sim Dir(\eta)$

② For each document in a corpus

① sample a path : $c \sim nCRP(\gamma)$

② Sample a level proportions : $\theta \sim GEM(m, \pi)$

③ For each word in a document

① sample a level : $z \sim Mult(\theta)$

② sample a word from the corresponding level : $w \sim Mult(\beta_z)$

Variational Inference for Dirichlet Process Mixtures

1. Assumptions

⇒ the observed data are drawn from the exponential family distribution

⇒ the base distribution for the DP is the corresponding conjugate prior

⟶ $p(x_n \mid z_n, \theta_1^*, \theta_2^*, \cdots) = \prod_{i=1}^{\infty} \left(h(x_n) \exp\{\theta_i^{*T} x_n - a(\theta_i^*)\} \right)^{\mathbb{1}[z_n = i]}$

⟶ $p(\theta^* \mid \lambda) = h(\theta^*) \exp\{\lambda_1^T \theta^* + \lambda_2 (-a(\theta^*)) - a(\lambda)\}$

2. Mean-field Variational inference

⇒ $\log p(x) \geq \mathbb{E}_q[\log p(w, x) - \log q_\nu(w)]$

⟶ latent distribution : $p(w_i \mid w_{-i}, x) = h(w_i) \exp\{g_i(w_{-i}, x)^T w_i - a(g_i(w_{-i}, x))\}$

⟶ variational distribution : $q_\nu(w) = \prod_{i=1}^{M} \exp\{\nu_i^T w_i - a(w_i)\}$

★ Solution : $\nu_i = \mathbb{E}_q[g_i(w_{-i}, x)]$

⇒ $\log p(x \mid \alpha, \lambda) \geq \mathbb{E}_q[\log p(V \mid \alpha)] + \mathbb{E}_q[\log p(\theta^* \mid \lambda)]$
$+ \sum_{n=1}^{N} \left(\mathbb{E}_q[\log p(z_n \mid V)] + \mathbb{E}_q[\log p(x_n \mid z_n)] \right)$
$- \mathbb{E}_q[\log q(V, \theta^*, z)]$

⟶ latent distribution

⇒ stick length : $V_i \sim Beta(1, \alpha)$

⇒ atoms : $\theta^* \sim H$

⇒ cluster assignment : $z_n \sim Mult(\pi(V))$

⟶ variational distribution ⟶ "Truncated" to the level T

⇒ $q_\phi(V, \theta^*, z) = \prod_{t=1}^{T-1} q_{\gamma_t}(V_t) \prod_{t=1}^{T} q_{\tau_t}(\theta_t^*) \prod_{n=1}^{N} q_{\phi_n}(z_n)$

(q_{γ_t}: Beta, q_{τ_t}: exponential family, q_{ϕ_n}: multinomial)