Латентное размещение Дирихле

Latent Dirichlet Allocation LDA

Тематическая модель

$$p(\theta|\alpha) = Dir(\theta|\alpha) = \underbrace{\frac{\Gamma(\sum_j \alpha_j)}{\prod_j \Gamma(\alpha_j)}}_{= B^{-1}(\alpha_1 \dots \alpha_m)} \prod_{j=1}^{m} \theta_j^{\alpha_j - 1}$$

$\theta \in \mathbb{R}^m$, $\theta_j \geq 0$, $\sum_{j=1}^{m} \theta_j = 1$

$\ln \theta_j$ — достаточные статистики

$$\mathbb{E} \ln \theta_j = -\frac{\partial}{\partial \alpha_j} \ln B^{-1}(\alpha_1 \dots \alpha_m) = -\frac{\partial}{\partial \alpha_j} \ln \Gamma(\sum_j \alpha_j) +$$
$$+ \frac{\partial}{\partial \alpha_j} \ln \Gamma(\alpha_j)) = \psi(\alpha_j) - \psi(\sum_j \alpha_j), \quad \psi(x) = \frac{d}{dx} \ln \Gamma(x)$$

D — документов из N_d — слов

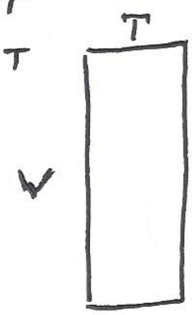
$\{1 \dots W\} \ni w_{d,n}$ — номер n-го ~~документа~~ слова в док. d

$\{1 \dots T\} \ni z_{d,n}$ — номер темы для n-го слова в док. d.

$S^T \ni \theta_d$ — профиль документа

$\varphi_{w,z}$ — вер-ть слова w в ~~док-те~~ теме z

$\Phi \in \mathbb{R}^{W \times T}$



$$p(w, z, \theta | \Phi, \alpha) = \prod_{d=1}^{D} \Big(p(\theta_d | \alpha) \cdot \prod_{n=1}^{N_d} p(z_{dn} | \theta_d) \cdot p(w_{dn} | z_{dn}, \Phi) \Big) =$$

$$= \prod_{d=1}^{D} \Big(Dir(\theta_d | \alpha) \prod_{n=1}^{N_d} \prod_{t=1}^{T} \theta_{dt}^{[z_{dn}=t]} \varphi_{w_{dn} z_{dn}} \{ \varphi_{w_{dn}}^{[z_{dn}=t]} \} \Big) =$$

$$= \prod_{d=1}^{D} \Big(Dir(\theta_d | \alpha) \prod_{n=1}^{N_d} \prod_{t=1}^{T} (\theta_{dt} \varphi_{w_{dn}})^{[z_{dn}=t]} \Big)$$

$p(w | \Phi, \alpha) \to \max_{\Phi, \alpha}$, $\ln p(w | \Phi, \alpha) \to \max_{\Phi}$

E-step $\quad p(Z,\Theta \mid w,\Phi,\alpha) \approx q(Z)q(\Theta)$

M-step $\quad \mathbb{E}_{Z,\Theta} \ln p(w,Z,\Theta \mid \Phi,\alpha)$

$$\ln q(\Theta) = \mathbb{E}_{q(Z)} \ln p(w,Z,\Theta \mid \Phi,\alpha) + C =$$

$$= \mathbb{E}_{q(Z)} \Big(\sum_{d=1}^{D} \sum_{t=1}^{T} (\alpha-1) \ln \Theta_{dt} + \sum_{n=1}^{N_d} \sum_{t=1}^{T} \sum_{d=1}^{D} [Z_{dn}=t] \ln \Theta_{dt} \Big) + C =$$

$$= \sum_{d=1}^{D} \sum_{t=1}^{T} \ln \Theta_{dt} \Big[(\alpha-1) + \sum_{n=1}^{N_d} \underbrace{\mathbb{E}_{Z_{dn}} [Z_{dn}=t]}_{\gamma_{dnt}} \Big] + C =$$

$$= \sum_{d=1}^{D} \sum_{t=1}^{T} \ln \Theta_{dt} \Big[(\alpha-1) + \sum_{n=1}^{N_d} \gamma_{dnt} \Big] + C$$

$$q(\Theta) = \prod_{d=1}^{D} \cancel{\sum_{t=1}^{T}} q(\Theta_{d\cancel{t}}), \quad q(\Theta_{d\cancel{t}}) = \mathrm{Dir}\Big(\Theta_{d\cancel{t}} \mid \alpha + \sum_{n=1}^{N_d} \gamma_{dnt}\Big)$$

$$q(\Theta) = \prod_{d=1}^{D} q(\Theta_d) = \prod_{d=1}^{D} \mathrm{Dir}(\Theta_d \mid \alpha_d)$$

$$\ln q(Z) = \mathbb{E}_{q(\Theta)} \ln p(w, \cancel{Z}Z, \Theta \mid \Phi, \alpha) + C =$$

$$= \mathbb{E}_{q(\Theta)} \Big(\sum_{d=1}^{D} \sum_{t=1}^{T} (\alpha-1) \ln \Theta_{dt} + \sum_{d=1}^{D} \sum_{n=1}^{N_d} \sum_{t=1}^{T} [Z_{dn}=t] (\ln \Theta_{dt} + \ln \Phi_{w_{dn},t}) \Big) + C$$

$$= \sum_{d=1}^{D} \sum_{n=1}^{N_d} \sum_{t=1}^{T} [Z_{dn}=t] \big(\mathbb{E}_{q(\Theta_{dt})} \ln \Theta_{dt} + \ln \Phi_{w_{dn},t} \big) + C$$

$$q(Z) = \prod_{d=1}^{D} \prod_{n=1}^{N_d} q(Z_{dn})$$

$$\ln q(Z_{dn}=t) = \mathbb{E}_{q(\Theta_{dt})} \ln \Theta_{dt} + \ln \Phi_{w_{dn},t} + C$$

$$q(Z_{dn}=t) = \frac{\Phi_{w_{dn},t} \cdot \exp\big(\mathbb{E}_{q(\Theta_{dt})} \ln \Theta_{dt}\big)}{\sum_{s=1}^{T} \Phi_{w_{dn},s} \exp\big(\mathbb{E}_{q(\Theta_{ds})} \ln \Theta_{ds}\big)} = \gamma_{dnt}$$

насколько вероятна тема t в документе d

M-step

$$\mathbb{E}_{Z,\Theta} \ln p(W, Z, \Theta \mid \Phi, \alpha) = \mathbb{E}_{Z,\Theta} \sum_{d=1}^{D} \Big[\ln p(\theta_d \mid \alpha) +$$

$$+ \sum_{n=1}^{N_d} \sum_{t=1}^{T} [z_{dn} = t] \left(\ln \theta_{dt} + \ln \varphi_{w_{dn},t} \right) \Big] = C +$$

$$+ \mathbb{E}_{Z,\Theta} \sum_{d=1}^{D} \sum_{n=1}^{N} \sum_{t=1}^{T} [z_{dn} = t] \ln \varphi_{w_{dn},t} = C +$$

$$\cancel{A} + \sum_{d,n,t} \ln \varphi_{w_{dn},t} \, \gamma_{dnt} \to \max_{\Phi} \qquad \Phi$$

$$\Phi: \sum_{w=1}^{V} \cancel{\varphi} \varphi_{wt} = 1$$

$$\sum_{d,n,t} \gamma_{dnt} \ln \varphi_{w_{dn},t} - \sum_{t=1}^{T} \lambda_t \Big(\sum_{w=1}^{V} \varphi_{wt} - 1 \Big) \to \text{extr}$$

$$\frac{\partial \mathcal{L}}{\partial \varphi_{wt}} = \sum\sum_{(d,n):\, w_{dn} = w} \gamma_{dnt} \frac{1}{\varphi_{wt}} - \lambda_t = 0$$

$$\varphi_{wt} = \frac{1}{\lambda_t} \sum\sum_{(d,n):\, w_{dn} = w} \gamma_{dnt}$$

$$\lambda_t \varphi_{wt} = \sum_{(d,n):\, w_{dn} = w} \gamma_{dnt} \quad \Big| \sum_{w=1}^{V}$$

$$\lambda_t = \sum_{w=1}^{V} \sum_{(d,n):\, w_{dn} = w} \gamma_{dnt} = \sum_{d=1}^{D} \sum_{n=1}^{N} \gamma_{dnt}$$

$$\varphi_{wt} = \frac{\sum\sum_{(d,n):\, w_{dn} = w} \gamma_{dnt}}{\sum_{d=1}^{D} \sum_{n=1}^{N} \gamma_{dnt}}$$

доля вхождения слова в тему

тематический профиль документа

$$q(d_*)\,q(z) \approx p(\Theta_{d_*}, z \mid w_{d_*}, \Phi_\alpha)$$

$$q(d_*) \to Diz$$

$$p(w, z, \Theta \mid \Phi, \alpha) = \prod_d \left(p(\Theta_d \mid \alpha) \prod_{n=1}^{N_d} p(z_{dn} \mid \Theta_d) \cdot p(w_{dn} \mid z_{dn}, \Phi)\right)$$

$$p(\Theta, z, \Phi \mid w, \alpha, \beta) \approx q(\Theta)\,q(z)\,q(\Phi)$$

$p(\Theta_d \mid \alpha)$ — DP, MC

$$p(\Theta_d \mid \Theta_{d-1}, \alpha)$$

18.11.16 бимо сем

смесь Стьюдента

$$p(X, Z, T \mid \mu, \Sigma, \nu, \pi) = \prod_{n,k} \left[\pi_k N(x_n \mid \mu_k, \tfrac{1}{z_n} \Sigma_k) \cdot G(z_n \mid \tfrac{\nu_k}{2}, \tfrac{\nu_k}{2})\right]^{t_{nk}}, \quad q(Z, T) \approx q(Z)\,q(T)$$

$$= q(Z \mid T)\,q(T)$$

$p(Z, T)$, $p(X \mid Z, T)$, сопряжение

$$p(Z, T \mid X) = q(Z, T) \quad \text{аналитический вид}$$

$$p(X \mid Z, T) = \prod_{n,k} \left[N(x_n \mid \mu_k, \tfrac{1}{z_n} \Sigma_k)\right]^{t_{nk}} =$$

$$= \prod_{n,k} \left[\frac{z_n^{d/2}}{(2\pi)^{d/2} \sqrt{\det \Sigma_k}} \exp\left(-\frac{z_n}{2} (x_n - \mu_k)^T \Sigma_k^{-1} (x_n - \mu_k)\right)\right]^{t_{nk}} \neq$$

$$\neq \prod_{n,k} \left[\frac{z_n^c}{c} \exp(-z_n c)\right]^{t_{nk}}$$

$$p(Z, T) = \prod_{n,k} \left[\pi_k\, G(z_n \mid \tfrac{\nu_k}{2}, \tfrac{\nu_k}{2})\right]^{t_{nk}} = \prod_{n,k} \left[\frac{z_n^c}{c} \exp(-z_n c)\right]^{t_{nk}}$$

$$p(w, z, \theta \mid \varphi, \alpha) = \prod_{d=1}^{D} \mathrm{Dir}(\theta_d \mid \alpha) \prod_{n=1}^{N} \prod_{t=1}^{T} (\theta_{dt}\, \varphi_{t w_{dn}})^{[z_{dn}=t]}$$

$$p(w \mid z, \theta) \quad , \quad p(z, \theta) \; ?$$

$$p(w \mid z, \theta) = \prod_{d,n,t} \varphi_{t w_{dn}}^{[z_{dn}=t]} = \prod_{d,n,t} c^{[z_{dn}=t]}$$

$$p(z, \theta) = \prod_{d} \mathrm{Dir}(\theta_d \mid \alpha) \prod_{n,t} \theta_{dt}^{[z_{dn}=t]} =$$

$$= \prod_{d} \frac{\Gamma(\sum_t \alpha_t)}{\prod_t \Gamma(\alpha_t)} \prod_t \theta_{dt}^{\alpha_t - 1} \cdot \prod_{n,t} \theta_{dt}^{[z_{dn}=t]} =$$

$$= \frac{1}{c} \prod_{d,t} \theta_{dt}^{c} \prod_{n,t} \theta_{dt}^{[z_{dn}=t]} \cancel{= \frac{1}{c} \prod_t \Big[\prod_d \theta_{dt}^{c} \prod_n \theta_{dt}^{[z_{dn}=t]} \Big] =}$$

$$\cancel{= \frac{1}{c} \prod_{d,n,t} \theta_{dt}^{c}\, \theta_{dt}^{[z_{dn}=t]} = \frac{1}{c} \prod_{d,n,t} \theta_{dt}^{c + [z_{dn}=t]}}$$

$$\# \quad p(w, z, \theta, \varphi \mid \alpha, \beta) = \prod_d \mathrm{Dir}(\theta_d \mid \alpha) \prod_{n,t} [\theta_{dt}\, \varphi_{t w_{dn}}]^{[z_{dn}=t]} \cdot \prod_t \mathrm{Dir}(\varphi_t \mid \beta)$$

$$p(w \mid z, \theta, \varphi) \quad , \quad p(z, \theta, \varphi) \; ?$$

$$p(z, \theta, \varphi) = \frac{1}{c} \prod_d \prod_t \theta_{dt}^{c} \prod_n \prod_t \theta_{dt}^{[z_{dn}=t]} \prod_t \frac{1}{c} \prod_w \varphi_{tw}^{c}$$

$$p(w \mid z, \theta, \varphi) = \prod_{d,n,t} \varphi_{t w_{dn}}^{[z_{dn}=t]} \qquad q(z)\, q(\theta, \varphi) = q(z)\, q(\theta)\, q(\varphi)$$

$$E: \quad q(z, \theta, \varphi) = q(z) \cdot q(\theta)\, q(\varphi)$$

$$\ln q(\varphi) = \mathbb{E}_{q(z) q(\theta)} \ln q p(w, z, \theta, \varphi \mid \alpha, \beta) + C \;(=)$$

$$\cancel{= \mathbb{E}_{q(z) q(\theta)} \sum_d \sum_t (\alpha - 1)} \qquad \varphi_{t w_{dn}}^{[z_{dn}=t]} = \prod_{w=1}^{V} \varphi_{tw}^{[z_{dn}=t][w_{dn}=w]}$$

$$\ominus \sum_t \sum_w (\beta-1)\ln \varphi_{tw} + \sum_{d,n,t,w} \gamma_{dnt}[w_{dn}=w]\ln \varphi_{tw} + C =$$

$$= \sum_{t,w}\left[\ln \varphi_{tw}\left(\sum_{d,n} \gamma_{dnt}[w_{dn}=w] + \beta - 1\right)\right] + C$$

$$q(\varphi) = \prod_{t} \mathrm{Dir}\left(\varphi_t \mid \beta + \sum_{d,n} \gamma_{dnt}[w_{dn}=w]\right)$$

$$p(w, z, \theta \mid \varphi, \alpha) = \prod_d \arg\max_{\theta_d} \mathrm{Dir}(\theta_d \mid \alpha) \cdot$$

$$\cdot \prod_n \prod_t [\hat{\theta}_{dt}\, \varphi_{t w_{dn}}]^{[z_{dn}=t]}$$

Иерархические процессы Дирихле.

$G \sim DP(G_0, \alpha)$, $G = \sum_{k=1}^{\infty} \pi_k \delta_{\theta_k}$

$\theta_k \sim G_0$, $\pi_k = v_k \prod_{j=1}^{k-1}(1-v_j)$, $v_k \sim Beta(1, \alpha)$

$$p(X, Z, \pi, \theta) = \left[\prod_n \prod_k^{\infty} p(x_n \mid \theta_k)^{[z_n=k]} \pi_n^{[z_n=k]}\right] \underbrace{p(\pi)p(\theta)}_{DP} =$$

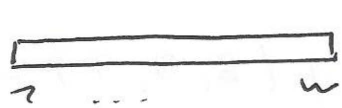
$$= \left[\prod_k^{\infty} p_{G_0}(\theta_k)\, Beta(v_k \mid 1, \alpha)\right] \cdot \prod_n \prod_k^{\infty} \left[p(x_n \mid \theta_k)\, v_k \prod_{j=1}^{k-1}(1-v_j)\right]^{[z_n=k]}$$

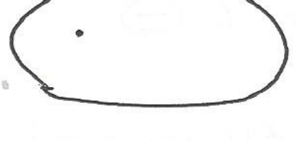
$$p(w, z, \theta, \varphi \mid \alpha, \beta) = \overbrace{\prod_t p(\varphi_t \mid \beta)}^{\sim p(\theta)} \overbrace{\prod_d p(\theta_d \mid \alpha)}^{\sim p(\pi)} \cdot$$

$$\cdot \prod_n \underbrace{p(z_{dn} \mid \theta_d)}_{\sim p(z_n \mid \pi)} \underbrace{p(w_{dn} \mid z_{dn}, \varphi)}_{\sim p(x_n \mid z_n, \theta_{z_n})}$$

слово – точка

$G \sim DP(\alpha, H)$, $H = \mathrm{Dir}(\eta)$ φ_t





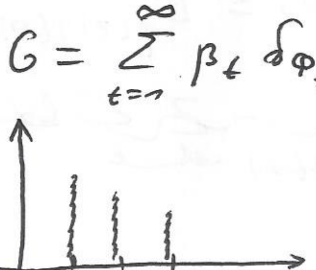
$$\left.\begin{array}{l} \varphi_t \sim \mathrm{Dir}(\varphi_t \mid \eta) \\ v_t \sim Beta(1, \alpha) \\ \beta_t = v_t \prod_{j=1}^{t-1}(1-v_j) \end{array}\right] \rightarrow G = \sum_{t=1}^{\infty} \beta_t \delta_{\varphi_t}$$

$G_d \sim DP(\gamma, G)$ $\Rightarrow$

$\varphi_k \sim G$

$\pi_{dk} \sim Beta(1, \gamma)$

$\theta_{dk} = \pi_{dk} \prod_{j=1}^{k-1}(1-\pi_{dj})$



$G_d = \sum_k \delta_{\varphi_k} \theta_{dk}$

$$\pi_{dt} \sim \mathrm{Beta}\left(\gamma \beta_t, \gamma\left(1 - \sum_{k=1}^{t} \beta_k\right)\right)$$

$$\theta_{dt} = \pi_{dt} \prod_{j=1}^{t-1} (1 - \pi_{dj})$$

$$p(w, z, \theta, \varphi \mid \alpha, \beta) = \prod_t \mathrm{Dir}(\varphi_t \mid \eta)\, \mathrm{Beta}(v_t \mid 1, \alpha) \cdot$$

$$\prod_d \prod_t \mathrm{Beta}\left(\pi_{dt} \mid \gamma \beta_t, \gamma\left(1 - \sum_k \beta_k\right)\right) \cdot \prod_n \prod_t^{\infty} (\theta_{dt})^{[z_{dn}=t]} \cdot$$

$$\cdot \prod_t \prod_w \varphi_{tw}^{[z_{dn}=t][w_{dn}=w]}$$