

预备知识

$$\text{Dirichlet distribution: } \text{Dir}(\vec{p}|\vec{\alpha}) = \frac{1}{\Delta(\vec{\alpha})} \prod_{k=1}^K p_k^{\alpha_k - 1} \quad (1)$$

$$\text{where } \Delta(\vec{\alpha}) = \frac{\prod_{k=1}^K \Gamma(\alpha_k)}{\Gamma\left(\sum_{k=1}^K \alpha_k\right)} \quad (2)$$

$$p(\vec{w}|\vec{z}, \vec{\beta}) = \prod_{z=1}^K \frac{\Delta(\vec{n}_z + \vec{\beta})}{\Delta(\vec{\beta})} \quad (3)$$

$$\text{where } \vec{n}_z = \left\{ n_z^{(t)} \right\}_{t=1}^V \quad (4)$$

$$p(\vec{z}|\vec{\alpha}) = \prod_{m=1}^M \frac{\Delta(\vec{n}_m + \vec{\alpha})}{\Delta(\vec{\alpha})} \quad (5)$$

$$\text{where } \vec{n}_m = \left\{ n_m^{(k)} \right\}_{k=1}^K \quad (6)$$

$$p(\vec{z}, \vec{w}|\vec{\alpha}, \vec{\beta}) = \prod_{z=1}^K \frac{\Delta(\vec{n}_z + \vec{\beta})}{\Delta(\vec{\beta})} \cdot \prod_{m=1}^M \frac{\Delta(\vec{n}_m + \vec{\alpha})}{\Delta(\vec{\alpha})} \quad (7)$$

NOTE: $n_z^{(t)}$ 是词袋中的词 $t \in [1, V]$ 被观察到分配给主题 z 的次数, $n_m^{(k)}$ 表示主题 $k \in [1, K]$ 分配给文档 m 的次数。

Gibbs sampling for LDA

我们用 i 来表示第 \tilde{m} 个文档的第 \tilde{n} 个词位置, 即 $i = (\tilde{m}, \tilde{n})$

我们用 \vec{w} 来表示所有文档的词分别是什么, 即 $\vec{w} = \{w_i = \tilde{t}, \vec{w}_{-i}\}$

我们用 \vec{z} 来表示所有文档的词的主题分别是什么, 即 $\vec{z} = \{z_i = \tilde{k}, \vec{z}_{-i}\}$

$$p(z_i = \tilde{k} | \vec{z}_{-i}, \vec{w}) = \frac{p(\vec{w}, \vec{z})}{p(\vec{w}, \vec{z}_{-i})} = \frac{p(\vec{w} | \vec{z})}{p(\vec{w}_i | \vec{z}_{-i})} \underbrace{\frac{p(w_i)}{p(\vec{z}_{-i})}}_{\text{(evidence)}} \cdot \frac{p(\vec{z})}{p(\vec{z}_{-i})} \quad (8)$$

$$\propto \frac{p(\vec{w} | \vec{z})}{p(\vec{w}_i | \vec{z}_{-i})} \cdot \frac{p(\vec{z})}{p(\vec{z}_{-i})} \quad (9)$$

$$= \frac{\overbrace{\prod_{z=1}^K \frac{\Delta(\vec{n}_z + \vec{\beta})}{\Delta(\vec{\beta})}}^{\text{part 1}}}{\overbrace{\prod_{z=1}^K \frac{\Delta(\vec{n}_{z,-i} + \vec{\beta})}{\Delta(\vec{\beta})}}^{\text{part 2}}} \cdot \frac{\overbrace{\prod_{m=1}^M \frac{\Delta(\vec{n}_m + \vec{\alpha})}{\Delta(\vec{\alpha})}}^{\text{part 3}}}{\overbrace{\prod_{m=1}^M \frac{\Delta(\vec{n}_{m,-i} + \vec{\alpha})}{\Delta(\vec{\alpha})}}^{\text{part 4}}} \quad (10)$$

$$= \prod_{z=1}^K \frac{\Delta(\vec{n}_z + \vec{\beta})}{\Delta(\vec{n}_{z,-i} + \vec{\beta})} \cdot \prod_{m=1}^M \frac{\Delta(\vec{n}_m + \vec{\alpha})}{\Delta(\vec{n}_{m,-i} + \vec{\alpha})} \quad (11)$$

$$= \frac{\overbrace{\Delta(\vec{n}_{\tilde{k}} + \vec{\beta})}^{\text{part 5}}}{\underbrace{\Delta(\vec{n}_{\tilde{k},-i} + \vec{\beta})}_{\text{part 6}}} \cdot \frac{\overbrace{\Delta(\vec{n}_{\tilde{m}} + \vec{\alpha})}^{\text{part 7}}}{\underbrace{\Delta(\vec{n}_{\tilde{m},-i} + \vec{\alpha})}_{\text{part 8}}} \quad (12)$$

$$= \frac{\underbrace{\prod_{t=1}^V \Gamma(n_{\tilde{k}}^{(t)} + \beta_t)}_{\text{part 9}}}{\underbrace{\prod_{t=1}^V \Gamma(n_{\tilde{k},-i}^{(t)} + \beta_t)}_{\text{part 9}}} \cdot \frac{\underbrace{\Gamma(\sum_{t=1}^V n_{\tilde{k},-i}^{(t)} + \beta_t)}_{\text{part 10}}}{\underbrace{\Gamma(\sum_{t=1}^V n_{\tilde{k}}^{(t)} + \beta_t)}_{\text{part 10}}} \cdot \frac{\underbrace{\prod_{k=1}^K \Gamma(n_{\tilde{m}}^{(k)} + \alpha_k)}_{\text{part 11}}}{\underbrace{\prod_{k=1}^K \Gamma(n_{\tilde{m},-i}^{(k)} + \alpha_k)}_{\text{part 11}}} \cdot \frac{\underbrace{\Gamma(\sum_{k=1}^K n_{\tilde{m},-i}^{(k)} + \alpha_k)}_{\text{part 12}}}{\underbrace{\Gamma(\sum_{k=1}^K n_{\tilde{m}}^{(k)} + \alpha_k)}_{\text{part 12}}} \quad (13)$$

$$= \frac{n_{\tilde{k},-i}^{(\tilde{t})} + \beta_{\tilde{t}}}{\sum_{t=1}^V n_{\tilde{k},-i}^{(t)} + \beta_t} \cdot \frac{n_{\tilde{m},-i}^{(\tilde{k})} + \alpha_{\tilde{k}}}{\sum_{k=1}^K n_{\tilde{m},-i}^{(k)} + \alpha_k} \quad (14)$$

$$\propto \frac{\overbrace{n_{\tilde{k},-i}^{(\tilde{t})} + \beta_{\tilde{t}}}^{\text{part 13}}}{\underbrace{\sum_{t=1}^V n_{\tilde{k},-i}^{(t)} + \beta_t}_{\text{part 14}}} \cdot \underbrace{(n_{\tilde{m},-i}^{(\tilde{k})} + \alpha_{\tilde{k}})}_{\text{part 15}} \quad (15)$$

- (8) 中 \vec{w} 和 \vec{z}_{-i} 都是固定的，可以理解为我们是已知的。所以 $p(w_i)$ 是一个 evidence，是一个常量，可以舍去，所以 (8) 正比于 (9)

- 由 (3) 带入 (9) 中得到 part 1 和 part 2 , 由 (5) 带入 (9) 中得到 part 3 和 part 4
- 对 (10) 进行约分后得到(11)
- 观察 (11) 中前一项 $\prod_{z=1}^K \frac{\Delta(\vec{n}_z + \vec{\beta})}{\Delta(\vec{n}_{z,\neg i} + \vec{\beta})}$, 向量 $\vec{n}_z \in \mathcal{R}^V$ 代表的是词袋中的每个词被观察到分配给主题 z 的次数, 因为我们这里假设了文当中位置为 i 的那个词 w_i 的主题为 $z_i = \tilde{k}$, 所以只要 $\prod_{z=1}^K \frac{\Delta(\vec{n}_z + \vec{\beta})}{\Delta(\vec{n}_{z,\neg i} + \vec{\beta})}$ 中 $z \neq \tilde{k}$, 那么就有 $\vec{n}_z = \vec{n}_{z,\neg i}$, 也就是说

$$\frac{\Delta(\vec{n}_z + \vec{\beta})}{\Delta(\vec{n}_{z,\neg i} + \vec{\beta})} = 1, \quad \text{when } z \neq \tilde{k}$$

所以,

$$\prod_{z=1}^K \frac{\Delta(\vec{n}_z + \vec{\beta})}{\Delta(\vec{n}_{z,\neg i} + \vec{\beta})} = \frac{\Delta(\vec{n}_{\tilde{k}} + \vec{\beta})}{\Delta(\vec{n}_{\tilde{k},\neg i} + \vec{\beta})} \quad (11.a)$$

同理对于 (11) 中后一项 $\prod_{m=1}^M \frac{\Delta(\vec{n}_m + \vec{\alpha})}{\Delta(\vec{n}_{m,\neg i} + \vec{\alpha})}$, 有

$$\frac{\Delta(\vec{n}_m + \vec{\alpha})}{\Delta(\vec{n}_{m,\neg i} + \vec{\alpha})} = 1, \quad \text{when } m \neq \tilde{m}$$

所以

$$\prod_{m=1}^M \frac{\Delta(\vec{n}_m + \vec{\alpha})}{\Delta(\vec{n}_{m,\neg i} + \vec{\alpha})} = \frac{\Delta(\vec{n}_{\tilde{m}} + \vec{\alpha})}{\Delta(\vec{n}_{\tilde{m},\neg i} + \vec{\alpha})} \quad (11.b)$$

将 (11.a) 和 (11.b) 带入 (11) 中得到 (12)

- 将 (2) 带入 part 5 得:

$$\Delta(\vec{n}_{\tilde{k}} + \vec{\beta}) = \frac{\prod_{t=1}^V \Gamma(n_k^{(t)} + \beta_t)}{\Gamma(\sum_{t=1}^V n_k^{(t)} + \beta_t)} \quad (12.a)$$

将 (2) 带入 part 6 得:

$$\Delta(\vec{n}_{\tilde{k},\neg i} + \vec{\beta}) = \frac{\prod_{t=1}^V \Gamma(n_{k,\neg i}^{(t)} + \beta_t)}{\Gamma(\sum_{t=1}^V n_{k,\neg i}^{(t)} + \beta_t)} \quad (12.b)$$

将 (2) 带入 part 7 得:

$$\Delta(\vec{n}_{\tilde{m}} + \vec{\alpha}) = \frac{\prod_{k=1}^K \Gamma(n_{\tilde{m}}^{(k)} + \alpha_k)}{\Gamma(\sum_{k=1}^K n_{\tilde{m}}^{(k)} + \alpha_k)} \quad (12.c)$$

将 (2) 带入 part 8 得:

$$\Delta(\vec{n}_{\tilde{m}, \neg i} + \vec{\alpha}) = \frac{\prod_{k=1}^K \Gamma(n_{\tilde{m}, \neg i}^{(k)} + \alpha_k)}{\Gamma(\sum_{k=1}^K n_{\tilde{m}, \neg i}^{(k)} + \alpha_k)} \quad (12.d)$$

将12.a, 12.b, 12.c, 12.d 带入 (12) 中得 (13)

• 对于 part 9 :

$$\underbrace{\frac{\prod_{t=1}^V \Gamma(n_{\tilde{k}}^{(t)} + \beta_t)}{\prod_{t=1}^V \Gamma(n_{\tilde{k}, \neg i}^{(t)} + \beta_t)}}_{\text{part 9}} = \prod_{t=1}^V \frac{\Gamma(n_{\tilde{k}}^{(t)} + \beta_t)}{\Gamma(n_{\tilde{k}, \neg i}^{(t)} + \beta_t)}$$

有

$$\frac{\Gamma(n_{\tilde{k}}^{(t)} + \beta_t)}{\Gamma(n_{\tilde{k}, \neg i}^{(t)} + \beta_t)} = \begin{cases} 1 & t \neq \tilde{t} \\ n_{\tilde{k}, \neg i}^{(\tilde{t})} + \beta_{\tilde{t}} & t = \tilde{t} \end{cases}, \quad \text{where } t \neq \tilde{t}$$

所以

$$\underbrace{\frac{\prod_{t=1}^V \Gamma(n_{\tilde{k}}^{(t)} + \beta_t)}{\prod_{t=1}^V \Gamma(n_{\tilde{k}, \neg i}^{(t)} + \beta_t)}}_{\text{part 9}} = \prod_{t=1}^V \frac{\Gamma(n_{\tilde{k}}^{(t)} + \beta_t)}{\Gamma(n_{\tilde{k}, \neg i}^{(t)} + \beta_t)} = n_{\tilde{k}, \neg i}^{(\tilde{t})} + \beta_{\tilde{t}} \quad (13.a)$$

对于 part 10 :

$$\underbrace{\frac{\Gamma(\sum_{t=1}^V n_{\tilde{k}, \neg i}^{(t)} + \beta_t)}{\Gamma(\sum_{t=1}^V n_{\tilde{k}}^{(t)} + \beta_t)}}_{\text{part 10}} = \frac{1}{\sum_{t=1}^V n_{\tilde{k}, \neg i}^{(t)} + \beta_t} \quad (13.b)$$

对于 part 11:

$$\underbrace{\frac{\prod_{k=1}^K \Gamma(n_{\tilde{m}}^{(k)} + \alpha_k)}{\prod_{k=1}^K \Gamma(n_{\tilde{m}, \neg i}^{(k)} + \alpha_k)}}_{\text{part 11}} = \prod_{k=1}^K \frac{\Gamma(n_{\tilde{m}}^{(k)} + \alpha_k)}{\Gamma(n_{\tilde{m}, \neg i}^{(k)} + \alpha_k)}$$

有

$$\frac{\Gamma(n_{\tilde{m}}^{(k)} + \alpha_k)}{\Gamma(n_{\tilde{m}, \neg i}^{(k)} + \alpha_k)} = \begin{cases} 1 & k \neq \tilde{k} \\ n_{\tilde{m}, \neg i}^{(\tilde{k})} + \alpha_{\tilde{k}} & k = \tilde{k} \end{cases}$$

所以

$$\underbrace{\frac{\prod_{k=1}^K \Gamma(n_{\tilde{m}}^{(k)} + \alpha_k)}{\prod_{k=1}^K \Gamma(n_{\tilde{m}, \neg i}^{(k)} + \alpha_k)}}_{\text{part 11}} = \prod_{k=1}^K \frac{\Gamma(n_{\tilde{m}}^{(k)} + \alpha_k)}{\Gamma(n_{\tilde{m}, \neg i}^{(k)} + \alpha_k)} = n_{\tilde{m}, \neg i}^{(\tilde{k})} + \alpha_{\tilde{k}} \quad (13.c)$$

对于 part 12:

$$\underbrace{\frac{\Gamma\left(\sum_{k=1}^K n_{\tilde{m}, \neg i}^{(k)} + \alpha_k\right)}{\Gamma\left(\sum_{k=1}^K n_{\tilde{m}}^{(k)} + \alpha_k\right)}}_{\text{part 12}} = \frac{1}{\sum_{k=1}^K n_{\tilde{m}, \neg i}^{(k)} + \alpha_k} \quad (13.d)$$

将 13.a, 13.b, 13.c, 13.d 带入 (13) 中得 (14)

- (14) 中 $\sum_{k=1}^K n_{\tilde{m}, \neg i}^{(k)} + \alpha_k$ 与 \tilde{k} 无关, 是一个常量, 可以舍去, 得 (15)