预备知识

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

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

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

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

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

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

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

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 $\vec{z} = \left\{z_i = \tilde{k}, \vec{z}_{\neg i}\right\}$

$$p\left(z_i=\tilde{k} \mid \vec{z}_{\neg i}, \vec{w}\right)=\frac{p(\vec{w}, \vec{z})}{p\left(\vec{w}, \vec{z}_{\neg i}\right)}=\frac{p(\vec{w} \mid \vec{z})}{p\left(\vec{w}_i \mid \vec{z}_{\neg i}\right) \underbrace{p\left(w_i\right)}_{\text{(evidence)}}} \cdot \frac{p(\vec{z})}{p\left(\vec{z}_{\neg i}\right)} \tag{8}$$

$$\propto \frac{p(\vec{w} \mid \vec{z})}{p\left(\vec{w}_i \mid \vec{z}_{\neg i}\right)} \cdot \frac{p(\vec{z})}{p\left(\vec{z}_{\neg i}\right)} \tag{9}$$

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

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

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

$$\begin{aligned} = & \underbrace{\frac{\prod_{t=1}^V \Gamma\left(n_{\tilde{k}}^{(t)}+\beta_t\right)}{\prod_{t=1}^V \Gamma\left(n_{\tilde{k}, \neg i}^{(t)}+\beta_t\right)}}_{\text{part 9}} \cdot \underbrace{\frac{\Gamma\left(\sum_{t=1}^V n_{\tilde{k}, \neg i}^{(t)}+\beta_t\right)}{\Gamma\left(\sum_{t=1}^V n_{\tilde{k}}^{(t)}+\beta_t\right)}}_{\text{part 10}} \cdot \\ & \underbrace{\frac{\prod_{k=1}^K \Gamma\left(n_{\tilde{m}}^{(k)}+\alpha_k\right)}{\prod_{k=1}^K \Gamma\left(n_{\tilde{m}, \neg i}^{(k)}+\alpha_k\right)}}_{\text{part 11}} \cdot \underbrace{\frac{\Gamma\left(\sum_{k=1}^K n_{\tilde{m}, \neg i}^{(k)}+\alpha_k\right)}{\Gamma\left(\sum_{k=1}^K n_{\tilde{m}}^{(k)}+\alpha_k\right)}}_{\text{part 12}} \end{aligned} \tag{13}$$

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

$$\propto \frac{n_{\tilde{k}, \neg i}^{(\tilde{t})}+\beta_{\tilde{t}}}{\sum_{t=1}^V n_{\tilde{k}, \neg i}^{(t)}+\beta_t} \cdot\left(n_{\tilde{m}, \neg i}^{(\tilde{k})}+\alpha_{\tilde{k}}\right) \tag{15}$$

- (8) 中 $\vec{w}$ 和 $\vec{z}_{\neg i}$ 都是固定的，可以理解为我们是已知的。所以 $p\left(w_i\right)$ 是一个 evidenve，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 $\prod_{z=1}^{K}\frac{\Delta\left(\vec{n}_z+\vec{\beta}\right)}{\Delta\left(\vec{n}_{z,\neg i}+\vec{\beta}\right)}$，向量 $\overrightarrow{n_z}\in\mathcal{R}^V$ 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 $\prod_{z=1}^{K}\frac{\Delta\left(\vec{n}_z+\vec{\beta}\right)}{\Delta\left(\vec{n}_{z,\neg i}+\vec{\beta}\right)}$ 中 $z\neq\tilde{k}$，那么就有 $\overrightarrow{n_z}==\overrightarrow{n_{z,\neg i}}$，也就是说

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

  所以，

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

  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

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

  所以

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

  将 $(11.a)$ 和 $(11.b)$ 带入 (11) 中得到 (12)

- 将 (2) 带入 part 5 得：

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

  将 (2) 带入 part 6 得:

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

  将 (2) 带入 part 7 得：

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

  将 (2) 带入 part 8 得：

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

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

- 对于 part 9 :

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

有

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

$$,\qquad \text{where } t\neq\tilde{t}$$

所以

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

对于 part 10 :

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

对于 part 11:

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

有

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

所以

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

对于 part 12:

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

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)