预备知识

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

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

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

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

$$p(\vec{z}|\vec{lpha}) = \prod_{m=1}^{M} rac{\Delta \left(\vec{n}_m + \vec{lpha}
ight)}{\Delta (\vec{lpha})}$$
 (5)

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

$$p(\vec{z}, \vec{w} | \vec{\alpha}, \vec{\beta}) = \prod_{z=1}^{K} \frac{\Delta \left(\vec{n}_z + \vec{\beta} \right)}{\Delta (\vec{\beta})} \cdot \prod_{m=1}^{M} \frac{\Delta \left(\vec{n}_m + \vec{\alpha} \right)}{\Delta (\vec{\alpha})}$$
(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 $ec{w}$ 来表示所有文档的词分别是什么,即 $ec{w} = \{w_i = ilde{t}\,, ec{w}_{-i}\}$

我们用 $ec{z}$ 来表示所有文档的词的主题分别是什么,即 $ec{z} = \left\{ z_i = ilde{k}, ec{z}_{\lnot i}
ight\}$

$$p\left(z_{i} = \tilde{k}|\vec{z}_{\neg i}, \vec{w}\right) = \frac{p(\vec{w}, \vec{z})}{p\left(\vec{w}, \vec{z}_{\neg i}\right)} = \frac{p(\vec{w}|\vec{z})}{p\left(\vec{w}_{i}|\vec{z}_{\neg i}\right)} \underbrace{p\left(w_{i}\right)}_{\text{(evidence)}} \cdot \frac{p(\vec{z})}{p\left(\vec{z}_{\neg i}\right)}$$
(8)

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

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

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

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

$$= \underbrace{\frac{\prod_{t=1}^{V} \Gamma\left(n_{\tilde{k}}^{(t)} + \beta_{t}\right)}{\prod_{t=1}^{V} \Gamma\left(n_{\tilde{k}, \neg i}^{(t)} + \beta_{t}\right)}}_{\text{part 9}} \cdot \underbrace{\frac{\Gamma\left(\sum_{t=1}^{V} n_{\tilde{k}, \neg i}^{(t)} + \beta_{t}\right)}{\Gamma\left(\sum_{t=1}^{V} n_{\tilde{k}}^{(t)} + \beta_{t}\right)}}_{\text{part 10}}.$$

$$\underbrace{\frac{\prod_{k=1}^{K} \Gamma\left(n_{\tilde{m}}^{(k)} + \alpha_{k}\right)}{\prod_{k=1}^{K} \Gamma\left(n_{\tilde{m},\neg i}^{(k)} + \alpha_{k}\right)}}_{\text{part }11} \cdot \underbrace{\frac{\Gamma\left(\sum_{k=1}^{K} n_{\tilde{m},\neg i}^{(k)} + \alpha_{k}\right)}{\Gamma\left(\sum_{k=1}^{K} n_{\tilde{m}}^{(k)} + \alpha_{k}\right)}}_{\text{part }12} \tag{13}$$

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

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

$$(15)$$

- (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 $\overrightarrow{n_z} = = n_{z,\neg i}^{\rightarrow}$,也就是说

$$rac{\Delta \left(ec{n}_z + ec{eta}
ight)}{\Delta \left(ec{n}_{z,
eg i} + ec{eta}
ight)} = 1, \qquad ext{when} \quad z
eq ilde{k}$$

所以,

$$\prod_{z=1}^{K} rac{\Delta \left(ec{n}_z + ec{eta}
ight)}{\Delta \left(ec{n}_{z,
eg i} + ec{eta}
ight)} = rac{\Delta \left(ec{n}_{ ilde{k}} + ec{eta}
ight)}{\Delta \left(ec{n}_{ ilde{k},
eg i} + ec{eta}
ight)}$$
 (11.a)

同理对于 (11) 中后一项 $\prod_{m=1}^{M} rac{\Delta(ec{n}_m + ec{lpha})}{\Delta(ec{n}_{m - ec{\imath}} + ec{lpha})}$,有

$$rac{\Delta\left(ec{n}_{m}+ec{lpha}
ight)}{\Delta\left(ec{n}_{m,
egin{array}{c}i}+ec{lpha}
ight)}=1, \qquad ext{when } m
eq ilde{m}$$

所以

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

将 (11.a) 和 (11.b) 带入 (11) 中得到 (12)

● 将(2)带入part 5得:

$$\Delta \left(ec{n}_{ ilde{k}} + ec{eta}
ight) = rac{\prod_{t=1}^{V} \Gamma \left(n_k^{(t)} + eta_t
ight)}{\Gamma \left(\sum_{t=1}^{V} n_k^{(t)} + eta_t
ight)}$$
 (12.a)

将 (2) 带入 part 6 得:

$$\Delta \left(ec{n}_{ ilde{k}, \lnot i} + ec{eta}
ight) = rac{\prod_{t=1}^{V} \Gamma \left(n_{k, \lnot i}^{(t)} + eta_{t}
ight)}{\Gamma \left(\sum_{t=1}^{V} n_{k, \lnot i}^{(t)} + eta_{t}
ight)}$$
 (12.b)

将(2)带入part 7得:

$$\Delta \left(ec{n}_{ ilde{m}} + ec{lpha}
ight) = rac{\prod_{k=1}^{K} \Gamma \left(n_{ ilde{m}}^{(k)} + lpha_{k}
ight)}{\Gamma \left(\sum_{k=1}^{K} n_{ ilde{m}}^{(k)} + lpha_{k}
ight)}$$
 (12.c)

将 (2) 带入 part 8 得:

$$\Delta \left(ec{n}_{ ilde{m},
eg i} + ec{lpha}
ight) = rac{\prod_{k=1}^{K} \Gamma \left(n_{ ilde{m},
eg i}^{(k)} + lpha_k
ight)}{\Gamma \left(\sum_{k=1}^{K} n_{ ilde{m},
eg i}^{(k)} + lpha_k
ight)}$$
 (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

$$rac{\Gamma\left(n_{ ilde{k}}^{(t)}+eta_{t}
ight)}{\Gamma\left(n_{ ilde{k},
egin{subarray}{c} -i \ } t = ilde{t}
ight)} = egin{cases} 1 & t
eq ilde{t} \ n_{ ilde{k},
egin{subarray}{c} -i \ } t = ilde{t} \end{cases}$$

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

所以

$$\underbrace{\frac{\prod_{t=1}^{V}\Gamma\left(n_{\tilde{k}}^{(t)}+\beta_{t}\right)}{\prod_{t=1}^{V}\Gamma\left(n_{\tilde{k},\neg i}^{(t)}+\beta_{t}\right)}}_{\text{part 0}} = \prod_{t=1}^{V}\frac{\Gamma\left(n_{\tilde{k}}^{(t)}+\beta_{t}\right)}{\Gamma\left(n_{\tilde{k},\neg i}^{(t)}+\beta_{t}\right)} = n_{\tilde{k},\neg i}^{(\tilde{t})} + \beta_{\tilde{t}} \tag{13.a}$$

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

$$rac{\Gamma\left(n_{ ilde{m}}^{(k)}+lpha_{k}
ight)}{\Gamma\left(n_{ ilde{m},
egin{subarray}{c} -i \ } +lpha_{k}
ight)} = egin{cases} 1 & k
eq ilde{k} \ n_{ ilde{m},
egin{subarray}{c} -i \ } +lpha_{ ilde{k}} & k = ilde{k} \end{cases}$$

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)