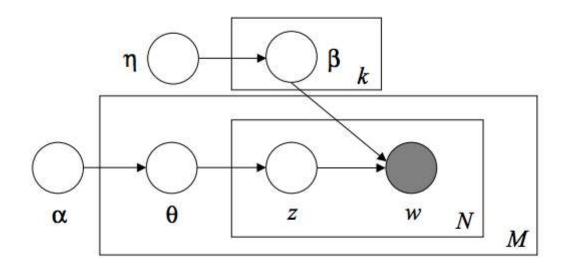
1. 联合分布



$$egin{aligned} p(heta,z,w,eta|lpha,\eta) &= p(heta,z,w|eta,lpha)p(eta|\eta) \ &= p(eta|\eta)\prod_{d=1}^D p(heta_d,z_d,w_d|eta,lpha) \ &= \prod_{k=1}^K p(eta_k|\eta)\prod_{d=1}^D [p(heta_d|lpha)\prod_{i=1}^L p(z_{di}| heta_d)p(w_{di}|z_{di},eta)] \ &= \prod_{k=1}^K Dir(eta_k|\eta)\prod_{d=1}^D [Dir(heta_d|lpha)\prod_{i=1}^L Cat(z_d| heta_d)Cat(w_{di}|eta_{z_{di}})] \end{aligned}$$

2. Gibbs Sampling

迪利克雷先验, Multinomial似然时的后验。

假设有i.i.d.的样本 X_1, X_2, \cdots, X_N ,

$$p(X) \sim Cat(X|\pi)$$

 π 的先验分布为 $Dir(\alpha)$

则 π 后验分布为?

$$p(\pi|D,lpha) = rac{p(\pi|lpha)\prod_{i=1}^N p(X_i|\pi)}{p(D)} \ \propto \prod_k^K \pi_k^{lpha_k-1} \prod_i^N \prod_k^K \pi_k^{I(X_i=k)}$$

$$egin{aligned} \prod_{k=1}^{K} \pi_k^{lpha_k-1} \prod_{i=1}^{K} \pi_k^{N(X_i=k)} \ &= \prod_{k=1}^{K} \pi_k^{N(X_i=k)+lpha_k-1} \end{aligned}$$

所以

$$p(\pi|D, lpha) \sim \sim Dir(\mathbf{N} + lpha - 1)$$

其中

$$N_k = N(X_i = k)$$

回到LDA

$$egin{aligned} p(heta_d|\cdot) &= Dir(\{lpha_k + \sum_{i=1}^{L_d} I(z_{di} = k)\}) \ p(eta_k|\cdot) &= Dir(\{\eta_v + \sum_{d=1}^D \sum_{i=1}^{L_d} I(w_{di} = v, z_{di} = k)\}) \ p(z_{di} = k|\cdot) \propto heta_{dk} eta_{kw_{di}} \end{aligned}$$

3. Collapsed Gibbs sampling

通过积分消去 θ , β

$$egin{aligned} p(z|lpha) &= \prod_{d=1}^D p(z_d|lpha) = \prod_{d=1}^D \int p(z_d| heta_d) p(heta_d|lpha) d heta_d \ &= \prod_{d=1}^D \int p(heta_d|lpha) \prod_{i=1}^{L_d} p(z_{di}| heta_d) d heta_d \ &= \prod_{d=1}^D \int Dir(heta_d|lpha) \prod_{i=1}^{L_d} Cat(z_{di}| heta_d) d heta_d \end{aligned}$$

由2中讨论设

$$D=\{z_{di},i=1,2,\cdots,L_d\}$$

则

$$egin{split} p(heta_d|D) &= rac{p(D| heta_d)p(heta_d)}{p(D)} \ &\sim Dir(C_d+lpha) \ B(C_d+lpha) \end{split}$$

$$egin{aligned} p(D) &= rac{\sum_{k \in \mathcal{X}} F(\Delta)}{B(lpha)} \ &= rac{\mathcal{T}(\sum_{k} lpha_{k})}{\prod_{k=1}^{K} \mathcal{T}(lpha_{k})} rac{\prod_{k=1}^{K} \mathcal{T}(C_{dk} + lpha_{k})}{\mathcal{T}(\sum_{k} (C_{dk} + lpha_{k}))} \end{aligned}$$

其中 C_{dk} 为文档d中主题为k(即 $z_{di}=k$)的单词个数。

实现中,向量 α 各维度取为一样,设为常数 α 。

则

$$p(D) = rac{\mathcal{T}(Klpha)}{\mathcal{T}(lpha)^K} rac{\prod_{k=1}^K \mathcal{T}(C_{dk} + lpha)}{\mathcal{T}(L_d + Klpha)}$$

故

$$p(z|lpha) = (rac{\mathcal{T}(Klpha)}{\mathcal{T}(lpha)^K})^D \prod_{d=1}^D rac{\prod_{k=1}^K \mathcal{T}(C_{dk}+lpha)}{\mathcal{T}(L_d+Klpha))}$$

同理

$$p(w|z,\eta) = (rac{\mathcal{T}(V\eta)}{\mathcal{T}(\eta)^V})^K \prod_{k=1}^K rac{\prod_{v=1}^V \mathcal{T}(C_{vk} + \eta)}{\mathcal{T}(C_k + V\eta))}$$

其中 C_{vk} 为所有文档中单词v主题为k的个数。 C_k 为所有文档中主题为k的单词个数。

最终

$$egin{aligned} p(z_{di} = k | z_{-di}, w, lpha, \eta) &= rac{p(z, w | lpha, \eta)}{p(z_{-di}, w | lpha, \eta)} \ &\propto rac{\mathcal{T}(C_{vk} + \eta)}{\mathcal{T}(C_{vk}^- + \eta)} rac{\mathcal{T}(C_{ik}^- + V \eta)}{\mathcal{T}(C_k^- + V \eta)} rac{\mathcal{T}(C_{ik} + \eta)}{\mathcal{T}(C_{ik}^- + \eta)} \ &= rac{C_{vk}^- + \eta}{C_k^- + V \eta} (C_{dk}^- + lpha) \end{aligned}$$

其中 $v=w_{di}$