GSDMM推导

贺成 2017.6.23

论文核心思想：提出GSDMM用于short text clustering

MGP（用于帮助理解短文本聚类和GSDMM）：
学生（documents）被用一个短的电影（words）列表表示，然后根据以下规则选择桌子（cluster）：

- Rule 1: Choose a table with more students. ← （completeness）
- Rule 2: Choose a table whose students share similar interests

(i.e., watched more movies of the same) with him. ← （homogeneity）

目标： $p\left(z_d = z \middle| \vec{z}_{\neg d}, \vec{d}\right) = \frac{p\left(\vec{d},\vec{z}|\vec{\alpha},\vec{\beta}\right)}{p\left(\vec{d},\vec{z}_{\neg d}|\vec{\alpha},\vec{\beta}\right)}$

$$p\left(\vec{d},\vec{z}|\vec{\alpha},\vec{\beta}\right) = p\left(\vec{d}|\vec{z},\vec{\alpha},\vec{\beta}\right)p\left(\vec{z}\middle|\vec{\alpha},\vec{\beta}\right)$$
$$= p\left(\vec{d}|\vec{z},\vec{\beta}\right)p(\vec{z}|\vec{\alpha}\ )$$

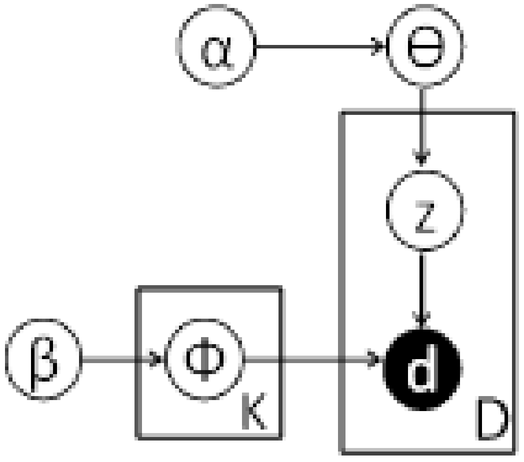

Figure 1: Graphical model of DMM.

求解：$p(\vec{d}|\vec{z},\vec{\beta})$

先求第k个簇对应的文档分布：

$p_{k,w}$为第k个主题（簇）中第w个词的概率，n_k^w为：number of occurrences of word w in cluster k

$$p(\vec{d}|\vec{z}_k,\vec{\beta}) = \int p(\vec{d}|\vec{z}_k,\varphi_k)p(\varphi_k|\vec{\beta})d\varphi_k \qquad (1)$$

$$= \int \prod_{w=1}^{V} p_{k,w}^{n_k^w}\, Dirichlet(\vec{\beta})\, d\varphi_k \qquad (2)$$

$$= \int \prod_{w=1}^{V} p_{k,w}^{n_k^w} \frac{1}{\Delta(\vec{\beta})} \prod_{w=1}^{V} p_{k,w}^{\beta_w - 1}\, d\varphi_k$$

$$= \frac{1}{\Delta(\vec{\beta})} \int \prod_{w=1}^{V} p_{k,w}^{n_k^w + \beta_w - 1}\, d\varphi_k$$

$$= \frac{\Delta(\vec{n}_k + \vec{\beta})}{\Delta(\vec{\beta})} \qquad \vec{n}_k = (n_k^{(1)}, n_k^{(2)}, \dots, n_k^{(V)})$$

$$=> p(\vec{d}|\vec{z},\vec{\beta}) = \prod_{k=1}^{K} \frac{\Delta(\vec{n}_k + \vec{\beta})}{\Delta(\vec{\beta})}$$

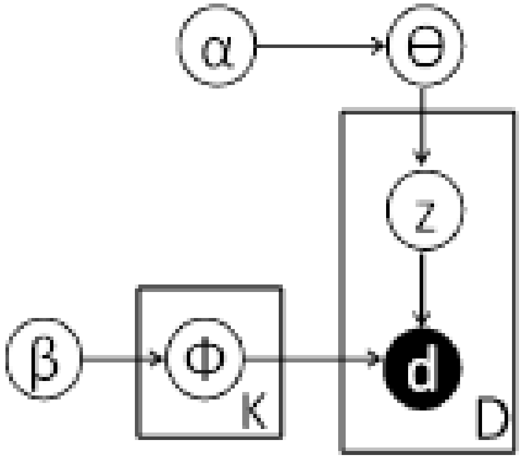

Figure 1: Graphical model of DMM.

(1) ->(2):

the probability of document d generated by cluster k can be derived as follows

$$p(d|z=k) = \prod_{w \in d} p(w|z=k)$$

求$p(\vec{z}|\vec{\alpha})$：

p_k为 第k个主题的概率， m_k 为该主题（簇）中文档个数

$$p(\vec{z}|\vec{\alpha}) = \int p(\vec{z}|\theta)p(\theta|\vec{\alpha})d\theta$$

$$= \int \prod_{k=1}^{K} {p_k}^{m_k}\, Dirichlet(\vec{\alpha})\, d\theta$$

$$= \int \prod_{k=1}^{K} {p_k}^{m_k} \frac{1}{\Delta(\vec{\alpha})} \prod_{k=1}^{K} {p_k}^{\alpha_k - 1} d\theta$$

$$= \frac{1}{\Delta(\vec{\alpha})} \int \prod_{k=1}^{K} {p_k}^{m_k + \alpha_k - 1} d\theta$$

$$= \frac{\Delta(\vec{m} + \vec{\alpha})}{\Delta(\vec{\alpha})}$$

$$\vec{m} = (m_1, m_2, \dots, m_K)$$

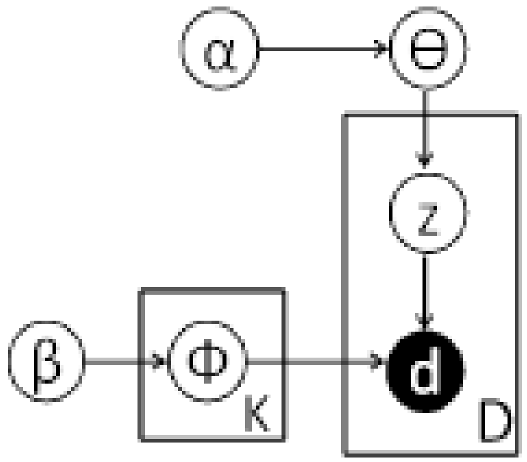

Figure 1: Graphical model of DMM.

因此：$p\left(\vec{d},\vec{z}|\vec{\alpha},\vec{\beta}\right)=p\left(\vec{w}|\vec{z},\vec{\beta}\right)p(\vec{z}|\vec{\alpha}\ )$

$$= \prod_{k=1}^{K}\frac{\Delta(\vec{n}_k+\vec{\beta})}{\Delta(\vec{\beta})} * \frac{\Delta(\vec{m}+\vec{\alpha})}{\Delta(\vec{\alpha})}$$

所以, $p\left(z_d=z\middle|\vec{z}_{\neg d},\vec{d}\right)=\dfrac{p\left(\vec{d},\vec{z}|\vec{\alpha},\vec{\beta}\right)}{p\left(\vec{d},\vec{z}_{\neg d}|\vec{\alpha},\vec{\beta}\right)}=\dfrac{\prod_{k=1}^{K}\frac{\Delta(\vec{n}_k+\vec{\beta})}{\Delta(\vec{\beta})}*\frac{\Delta(\vec{m}+\vec{\alpha})}{\Delta(\vec{\alpha})}}{\prod_{k=1}^{K}\frac{\Delta(\vec{n}_{k,\neg d}+\vec{\beta})}{\Delta(\vec{\beta})}*\frac{\Delta(\vec{m}_{\neg d}+\vec{\alpha})}{\Delta(\vec{\alpha})}}$

$$=\frac{\Delta(\vec{m}+\vec{\alpha})}{\Delta(\vec{m}_{\neg d}+\vec{\alpha})}\cdot\frac{\Delta(\vec{n}_k+\vec{\beta})}{\Delta(\vec{n}_{k,\neg d}+\vec{\beta})}$$

计算 $\dfrac{\Delta(\vec{m}+\vec{\alpha})}{\Delta(\vec{m}_{\neg d}+\vec{\alpha})}$:

$$\frac{\Delta(\vec{m}+\vec{\alpha})}{\Delta(\vec{m}_{\neg d}+\vec{\alpha})}=\frac{\prod_{k=1}^{K}\Gamma(m_k+\alpha_k)}{\Gamma(\sum_{k=1}^{K}m_k+\alpha_k)}\Big/\frac{\prod_{k=1}^{K}\Gamma\left(m_{k,\neg d}+\alpha_k\right)}{\Gamma\left(\sum_{k=1}^{K}m_{k,\neg d}+\alpha_k\right)}$$

$$=\frac{\prod_{k=1}^{K}\Gamma(m_k+\alpha_k)}{\Gamma(D+K\alpha)}\Big/\frac{\prod_{k=1}^{K}\Gamma\left(m_{k,\neg d}+\alpha_k\right)}{\Gamma(D-1+K\alpha)}$$

$$=\frac{\prod_{k=1}^{K}\Gamma(m_k+\alpha_k)}{\prod_{k=1}^{K}\Gamma\left(m_{k,\neg d}+\alpha_k\right)}*\frac{\Gamma(D-1+K\alpha)}{\Gamma(D+K\alpha)}$$

$$=\frac{\Gamma(m_k+\alpha_k)}{\Gamma\left(m_{k,\neg d}+\alpha_k\right)}*\frac{1}{D+K\alpha-1}$$

$$=\frac{m_{k,\neg d}+\alpha}{D+K\alpha-1}$$

注:$m_k=m_{k,\neg d}+1$

V	number of words in the vocabulary
D	number of documents in the corpus
$\bar{L}$	average length of documents
$\vec{d}$	documents in the corpus
$\vec{z}$	cluster labels of each document
I	number of iterations
m_z	number of documents in cluster z
n_z	number of words in cluster z
n_z^w	number of occurrences of word w in cluster z
N_d	number of words in document d
N_d^w	number of occurrences of word w in document d

计算 $\frac{\Delta(\vec{n}_k+\vec{\beta})}{\Delta(n_{k,\neg d}+\vec{\beta})}$ ：

$$\frac{\Delta(\vec{n}_k+\vec{\beta})}{\Delta(n_{k,\neg d}+\vec{\beta})} = \frac{\dfrac{\prod_{w=1}^{V}\Gamma(n_k^w+\beta_w)}{\Gamma(\sum_{w=1}^{V}(n_k^w+\beta_w))}}{\dfrac{\prod_{w=1}^{V}\Gamma(n_{k,\neg d}^w+\beta_w)}{\Gamma(\sum_{w=1}^{V}(n_{k,\neg d}^w+\beta_w))}}$$

$$= \frac{\prod_{w=1}^{V}\Gamma(n_k^w+\beta_w)}{\Gamma(\sum_{w=1}^{V}(n_k^w+\beta_w))} \cdot \frac{\Gamma(\sum_{t=1}^{V}(n_{k,\neg d}^w+\beta_w))}{\prod_{w=1}^{V}\Gamma(n_{k,\neg d}^w+\beta_w)}$$

$$= \frac{\prod_{w=1}^{V}\Gamma(n_k^w+\beta_w)}{\prod_{w=1}^{V}\Gamma(n_{k,\neg d}^w+\beta_w)} \cdot \boxed{\frac{\Gamma(n_{k,\neg d}+V\beta)}{\Gamma(n_k+V\beta)}}$$

$$= \frac{\prod_{w\in d}\Gamma(n_k^w+\beta_w)}{\prod_{w\in d}\Gamma(n_{k,\neg d}^w+\beta_w)} \cdot \frac{1}{\prod_{i=1}^{N_d}(n_{k,\neg d}+V\beta+i-1)}$$

注：$n_k = n_{k,\neg d} + N_d$

注：对于$(n_k^1, n_k^2, \dots\dots, n_k^V)$只有该单词出现在第d篇文档中分子分母才不一样，如若不出现在第d篇中的单词分子分母均一样可以约掉

接上一页：

$$\frac{\Delta(\vec{n}_k+\vec{\beta})}{\Delta(n_{k,\neg d}+\vec{\beta})} = \boxed{\frac{\prod_{w\in d}\Gamma(n_k^w+\beta_w)}{\prod_{w\in d}\Gamma(n_{k,\neg d}^w+\beta_w)}} \cdot \frac{1}{\prod_{i=1}^{N_d}(n_{k,\neg d}+V\beta+i-1)}$$

- 如果假设在一篇文档中每个单词至多出现一次，则有：$n_k^w = n_{k,\neg d}^w + 1$，则：

$$\frac{\prod_{w\in d}\Gamma(n_k^w + \beta_w)}{\prod_{w\in d}\Gamma(n_{k,\neg d}^w + \beta_w)} = \frac{\Gamma(n_k^1+\beta)\cdot\Gamma(n_k^2+\beta)\cdots}{\Gamma(n_{k,\neg d}^1+\beta)\cdot\Gamma(n_{k,\neg d}^2+\beta)\cdots} = \prod_{w\epsilon d}(n_{k,\neg d}^w + \beta)$$

- 如果允许在一篇文档中一个单词可以出现多次，则有：$n_k^w = n_{k,\neg d}^w + N_d^w$,则：

$$\frac{\prod_{w\in d}\Gamma(n_k^w + \beta_w)}{\prod_{w\in d}\Gamma(n_{k,\neg d}^w + \beta_w)} = \frac{\Gamma(n_k^1+\beta)\cdot\Gamma(n_k^2+\beta)\cdots}{\Gamma(n_{k,\neg d}^1+\beta)\cdot\Gamma(n_{k,\neg d}^2+\beta)\cdots}$$

$$= \prod_{w\epsilon d}\prod_{j=1}^{N_d^w}(n_{k,\neg d}^w + \beta + j - 1)$$

因此：

➢ 一篇文档中，每个单词至多允许出现一次：

$$p\left(z_d = z \middle| \vec{z}_{\neg d}, \vec{d}\right) = \frac{m_{k,\neg d} + \alpha}{D + K\alpha - 1} \frac{\prod_{w \epsilon d}(n_{k,\neg d}^{w} + \beta)}{\prod_{i=1}^{N_d}(n_{k,\neg d} + V\beta + i - 1)}$$

➢ 一篇文档中，一个单词可以允许出现多次（论文中用的这个）：

$$p\left(z_d = z \middle| \vec{z}_{\neg d}, \vec{d}\right) = \frac{m_{k,\neg d} + \alpha}{D + K\alpha - 1} \frac{\prod_{w \epsilon d}\prod_{j=1}^{N_d^w}(n_{k,\neg d}^{w} + \beta + j - 1)}{\prod_{i=1}^{N_d}(n_{k,\neg d} + V\beta + i - 1)}$$

$$\varphi_{z,w} = \frac{n_z^w + \beta}{\sum_{w=1}^{V} n_z^w + V\beta}$$

where $\varphi_{z,w}$ corresponds to the probability of word w being generated by cluster z, and can be regarded as the importance of word w to cluster z. As a result, GSDMM can obtain the representative words of each cluster like Topic Models