概率潜在语义分析共现模型的 EM 算法

单词集合 $W=\{w_1,w_2,...,w_M\}$

文本集合 $D=\{d_1,d_2,...,d_N\}$

话题集合 $Z=\{z_1,z_2,...,z_K\}$

单词-文本共现数据 $T=\{n(w_i,d_j)\,|\,i=1,...,M,\ j=1,...,N\}$

每个单词-文本对的概率为

$$P(w_i,d_j)=\sum_{k=1}^{K}P(w_i,z_k,d_j)$$

$$=\sum_{k=1}^{K}P(z_k)P(w_i\,|\,z_k)P(d_j\,|\,z_k)$$

似然函数

$$L=\prod_{i=1}^{M}\prod_{j=1}^{N}P(w_i,d_j)^{n(w_i,d_j)}$$

对数似然函数并构造 Q 函数

$$LL=\sum_{i=1}^{M}\sum_{j=1}^{N}n(w_i,d_j)\log P(w_i,d_j)$$

$$=\sum_{i=1}^{M}\sum_{j=1}^{N}n(w_i,d_j)\log\sum_{k=1}^{K}P(z_k)P(w_i\,|\,z_k)P(d_j\,|\,z_k)$$

$$\geq\sum_{i=1}^{M}\sum_{j=1}^{N}n(w_i,d_j)\sum_{k=1}^{K}P(z_k\,|\,w_i,d_j)\log\left[P(z_k)P(w_i\,|\,z_k)P(d_j\,|\,z_k)\right]$$

$$=\sum_{i=1}^{M}\sum_{j=1}^{N}n(w_i,d_j)\sum_{k=1}^{K}P(z_k\,|\,w_i,d_j)\left[\log P(z_k)+\log P(w_i\,|\,z_k)+\log P(d_j\,|\,z_k)\right]$$

$$=Q$$

模型参数为

（1） $P(z_k),\ k=1,...,K$

（2） $P(w_i\,|\,z_k),\ i=1,...,M,\ k=1,...,K$

（3） $P(d_j\,|\,z_k),\ j=1,...,N,\ k=1,...,K$

给定当前参数的估计值隐变量的后验分布

$$P\left(z_k \mid w_i, d_j\right)=\frac{P\left(z_k, w_i, d_j\right)}{P\left(w_i, d_j\right)}$$

$$=\frac{P\left(z_k, w_i, d_j\right)}{\sum_{l=1}^{K} P\left(z_l, w_i, d_j\right)}$$

$$=\frac{P\left(z_k\right) P\left(w_i \mid z_k\right) P\left(d_j \mid z_k\right)}{\sum_{l=1}^{K} P\left(z_l\right) P\left(w_i \mid z_l\right) P\left(d_j \mid z_l\right)}$$

约束条件

$$\sum_{k=1}^{K} P\left(z_k\right)=1$$

$$\sum_{i=1}^{M} P\left(w_i \mid z_k\right)=1,\ k=1,2,...,K$$

$$\sum_{j=1}^{N} P\left(d_j \mid z_k\right)=1,\ k=1,2,...,K$$

构造 Lagrange 函数

$$\Lambda=Q+\lambda\left(1-\sum_{k=1}^{K} P\left(z_k\right)\right)+\sum_{k=1}^{K} \tau_k\left(1-\sum_{i=1}^{M} P\left(w_i \mid z_k\right)\right)+\sum_{k=1}^{K} \rho_k\left(1-\sum_{j=1}^{N} P\left(d_j \mid z_k\right)\right)$$

$$=\sum_{i=1}^{M}\sum_{j=1}^{N} n\left(w_i, d_j\right)\sum_{k=1}^{K} P\left(z_k \mid w_i, d_j\right)\left[\log P\left(z_k\right)+\log P\left(w_i \mid z_k\right)+\log P\left(d_j \mid z_k\right)\right]$$

$$+\lambda\left(1-\sum_{k=1}^{K} P\left(z_k\right)\right)+\sum_{k=1}^{K} \tau_k\left(1-\sum_{i=1}^{M} P\left(w_i \mid z_k\right)\right)+\sum_{k=1}^{K} \rho_k\left(1-\sum_{j=1}^{N} P\left(d_j \mid z_k\right)\right)$$

分别对各参数求偏导，并置零，得到

（1）

$$\frac{\partial \Lambda}{\partial P\left(z_k\right)}=\sum_{i=1}^{M}\sum_{j=1}^{N} n\left(w_i, d_j\right) P\left(z_k \mid w_i, d_j\right)\frac{1}{P\left(z_k\right)}-\lambda=0$$

$$\Rightarrow P\left(z_k\right)=\frac{1}{\lambda}\sum_{i=1}^{M}\sum_{j=1}^{N} n\left(w_i, d_j\right) P\left(z_k \mid w_i, d_j\right)$$

$$\Rightarrow 1=\sum_{k=1}^{K} P\left(z_k\right)=\frac{1}{\lambda}\sum_{k=1}^{K}\sum_{i=1}^{M}\sum_{j=1}^{N} n\left(w_i, d_j\right) P\left(z_k \mid w_i, d_j\right)=\frac{1}{\lambda}\sum_{i=1}^{M}\sum_{j=1}^{N} n\left(w_i, d_j\right)$$

$$\Rightarrow \lambda=\sum_{i=1}^{M}\sum_{j=1}^{N} n\left(w_i, d_j\right)$$

$$\Rightarrow P\left(z_k\right)=\frac{\sum_{i=1}^{M}\sum_{j=1}^{N} n\left(w_i, d_j\right) P\left(z_k \mid w_i, d_j\right)}{\sum_{i=1}^{M}\sum_{j=1}^{N} n\left(w_i, d_j\right)}$$

（2）

$$\frac{\partial \Lambda}{\partial P(w_i \mid z_k)} = \sum_{j=1}^{N} n(w_i, d_j) P(z_k \mid w_i, d_j) \frac{1}{P(w_i \mid z_k)} - \tau_k = 0$$

$$\Rightarrow P(w_i \mid z_k) = \frac{1}{\tau_k} \sum_{j=1}^{N} n(w_i, d_j) P(z_k \mid w_i, d_j)$$

$$\Rightarrow 1 = \sum_{i=1}^{M} P(w_i \mid z_k) = \frac{1}{\tau_k} \sum_{i=1}^{M} \sum_{j=1}^{N} n(w_i, d_j) P(z_k \mid w_i, d_j)$$

$$\Rightarrow \tau_k = \sum_{i=1}^{M} \sum_{j=1}^{N} n(w_i, d_j) P(z_k \mid w_i, d_j)$$

$$\Rightarrow P(w_i \mid z_k) = \frac{\sum_{j=1}^{N} n(w_i, d_j) P(z_k \mid w_i, d_j)}{\sum_{m=1}^{M} \sum_{j=1}^{N} n(w_m, d_j) P(z_k \mid w_m, d_j)}$$

（3）

$$\frac{\partial \Lambda}{\partial P(d_j \mid z_k)} = \sum_{i=1}^{M} n(w_i, d_j) P(z_k \mid w_i, d_j) \frac{1}{P(d_j \mid z_k)} - \rho_k = 0$$

$$\Rightarrow P(d_j \mid z_k) = \frac{1}{\rho_k} \sum_{i=1}^{M} n(w_i, d_j) P(z_k \mid w_i, d_j)$$

$$\Rightarrow 1 = \sum_{j=1}^{N} P(d_j \mid z_k) = \frac{1}{\rho_k} \sum_{i=1}^{M} \sum_{j=1}^{N} n(w_i, d_j) P(z_k \mid w_i, d_j)$$

$$\Rightarrow \rho_k = \sum_{i=1}^{M} \sum_{j=1}^{N} n(w_i, d_j) P(z_k \mid w_i, d_j)$$

$$\Rightarrow P(d_j \mid z_k) = \frac{\sum_{i=1}^{M} n(w_i, d_j) P(z_k \mid w_i, d_j)}{\sum_{i=1}^{M} \sum_{l=1}^{N} n(w_i, d_l) P(z_k \mid w_i, d_l)}$$

共现模型的 EM 算法

输入：

单词集合 $W=\{w_1,w_2,...,w_M\}$

文本集合 $D=\{d_1,d_2,...,d_N\}$

话题集合 $Z=\{z_1,z_2,...,z_K\}$

单词-文本共现数据 $T=\{n(w_i,d_j)\mid i=1,...,M,\ j=1,...,N\}$

输出：

$P(z_k),\ k=1,...,K$

$P(w_i\mid z_k),\ i=1,...,M,\ k=1,...,K$

$P(d_j\mid z_k),\ j=1,...,N,\ k=1,...,K$

（1）设置参数 $P(z_k)$、$P(w_i\mid z_k)$、$P(d_j\mid z_k)$ 的初始值

（2）迭代执行以下 E 步和 M 步，直到收敛为止

E 步：

$$P(z_k\mid w_i,d_j)=\frac{P(z_k)P(w_i\mid z_k)P(d_j\mid z_k)}{\sum_{l=1}^{K}P(z_l)P(w_i\mid z_l)P(d_j\mid z_l)}$$

M 步：

$$P(z_k)=\frac{\sum_{i=1}^{M}\sum_{j=1}^{N}n(w_i,d_j)P(z_k\mid w_i,d_j)}{\sum_{i=1}^{M}\sum_{j=1}^{N}n(w_i,d_j)}$$

$$P(w_i\mid z_k)=\frac{\sum_{j=1}^{N}n(w_i,d_j)P(z_k\mid w_i,d_j)}{\sum_{i=1}^{M}\sum_{j=1}^{N}n(w_i,d_j)P(z_k\mid w_i,d_j)}$$

$$P(d_j\mid z_k)=\frac{\sum_{i=1}^{M}n(w_i,d_j)P(z_k\mid w_i,d_j)}{\sum_{i=1}^{M}\sum_{j=1}^{N}n(w_i,d_j)P(z_k\mid w_i,d_j)}$$