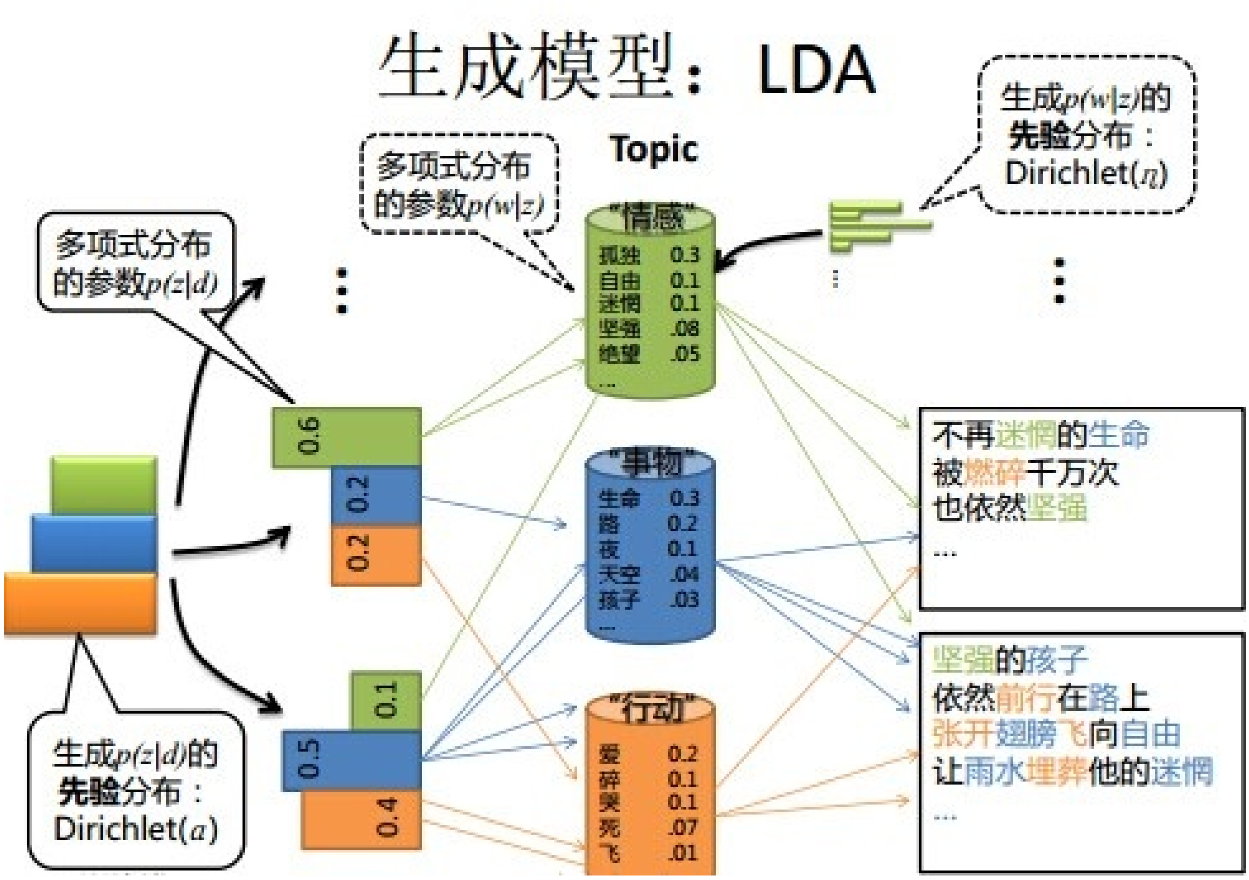


nlp中的主题模型

JayLou
NLP算法工程师

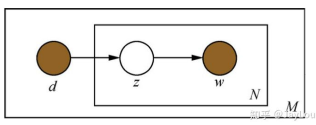
本文对nlp中一个极为重要的模型——主题模型LDA(Latent Dirichlet Allocation)从宏观理解与数学解释两个维度进行介绍。

1、LDA的宏观理解

谈起LDA，自然需要引入pLSA。pLSA是用一个生成模型来建模文章的生成过程。假设有K个主题，M篇文章；对语料库中的任意文章d，假设该文章有N个词，则对于其中的每一个词，我们首先选择一个主题z，然后在当前主题的基础上生成一个词w。

生成主题z和词w的过程遵照一个确定的概率分布。设在文章d中生成主题z的概率为 $p(z|d)$ ，在选定主题的条件下生成词w的概率为 $p(w|z)$ ，则给定文章d，生成词w的概率可以写成：

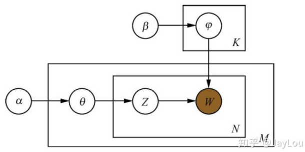
$$p(w|d)=\sum_{z}^{Z}p(w,z|d)=\sum_{z}^{Z}p(w|d,z)p(z|d)=\sum_{z}^{Z}p(w|z)p(z|d)$$



pLSA概率图模型

> LDA可以看作是pLSA的贝叶斯版本，其文本生成过程与pLSA基本相同，不同的是为主题分布和词分布分别加了两个狄利克雷（Dirichlet）先验。为什么要加入狄利克雷先验呢？这就要从频

> 率学派和贝叶斯学派的区别说起。pLSA采用的是频率派思想，将每篇文章对应的主题分布 $p(z|d)$ 和每个主题对应的词分布 $p(w|z)$ 看成确定的未知常数，并可以利用EM算法求解出来；
> 而LDA采用的是贝叶斯学派的思想，认为待估计的参数（主题分布和词分布）不再是一个固定的常数，而是服从一定分布的随机变量。这个分布符合一定的先验概率分布（即狄利克雷分布），并且在观察到样本信息之后，可以对先验分布进行修正，从而得到后验分布。LDA之所以选择狄利克雷分布作为先验分布，是因为它为多项式分布的共轭先验概率分布，后验概率依然服从狄利克雷分布，这样做可以为计算带来便利。——《百面机器学习》



LDA概率图模型

在LDA概率图模型中，α，β分别为两个狄利克雷分布的超参数，为人工设定。

补充：pLSA虽然可以从概率的角度解释了主题模型，却都只能对训练样本中的文本进行主题识别，而对不在样本中的文本是无法识别其主题的。根本原因在于NMF与pLSA这类主题模型方法没有考虑主题概率分布的先验知识，比如文本中出现体育主题的概率肯定比哲学主题的概率要高，这点来源于我们的先验知识，但是无法告诉NMF主题模型。而LDA主题模型则考虑到了这一问题，目前来说，绝大多数的文本主题模型都是使用LDA以及其变体。

2、LDA的数学基础

2.1 概率基础

（1）二项分布与多项分布

二项分布： $P(K=k)=\binom{n}{k}p^k(1-p)^{n-k}$

多项分布： $P(x_1,x_2,\ldots,x_k;n,p_1,p_2,\ldots,p_k)=\frac{n!}{x_1!\ldots x_k!}{p_1}^{x_1}\ldots {p_k}^{x_k}$

（2）Gamma函数

$$\Gamma(x)=\int_0^\infty t^{x-1}e^{-t}dt$$

Gamma函数如有这样的性质：$\Gamma(x+1)=x\Gamma(x)$

Gamma函数可以看成是阶乘在实数集上的延拓： $\Gamma(n)=(n-1)!$

（3）Beta分布和Dirichlet分布

Beta分布的概率密度函数为：

$$f(x;\alpha,\beta)=\frac{1}{B(\alpha,\beta)}x^{\alpha-1}(1-x)^{\beta-1}\quad \text{其中，}\ B(\alpha,\beta)=\frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)}$$

Dirichlet分布的概率密度函数为：

$$f(x_1,x_2,\ldots,x_k;\alpha_1,\alpha_2,\ldots,\alpha_k)=\frac{1}{B(\alpha)}\prod_{i=1}^{k}{x_i}^{\alpha^i-1}$$

其中， $B(\alpha)=\frac{\prod_{i=1}^{k}\Gamma(\alpha^i)}{\Gamma(\sum_{i=1}^{k}\alpha^i)},\sum_{i=1}^{k}x^i=1$

这说明，对于Beta分布的随机变量，其均值可以用 $\frac{\alpha}{\alpha+\beta}$ 来估计。

Dirichlet分布也有类似的结论，如果 $\vec{p} \sim Dir(\vec{t}|\vec{\alpha})$ ，同样可以证明：

$$E(p) = \left(\frac{\alpha^1}{\sum_{i=1}^K \alpha_i}, \frac{\alpha^1}{\sum_{i=2}^K \alpha_i}, \cdots, \frac{\alpha^K}{\sum_{i=1}^K \alpha_i} \right)$$

（4）共轭先验分布

在贝叶斯概率理论中，如果后验概率 $P(\theta|x)$ 和先验概率 $p(\theta)$ 满足同样的分布律，那么，先验分布和后验分布被叫做共轭分布，同时，先验分布叫做似然函数的共轭先验分布。Beta分布是二项式分布的共轭先验分布，而狄利克雷(Dirichlet)分布是多项式分布的共轭先验分布。

2.2 MCMC及Gibbs Sampling

（1）MCMC简介

> MCMC采样法主要包括两个MC，即蒙特卡洛法（Monte Carlo）和马尔可夫链（Markov Chain）。蒙特卡洛法是指基于采样的数值型近似求解方法，而马尔可夫链则用于进行采样。MCMC采样法基本思想是：针对待采样的目标分布，构造一个马尔可夫链，使得该马尔可夫链的平稳分布就是目标分布；然后，从任何一个初始状态出发，沿着马尔可夫链进行状态转移，最终得到的状态转移序列会收敛到目标分布，由此可以得到目标分布的一系列样本。在实际操作中，核心点是如何构造合适的马尔可夫链，即确定马尔可夫链的状态转移概率，这涉及一些马尔可夫链的相关知识点，如时齐性、细致平衡条件、可遍历性、平稳分布等。——《百面机器学习》

在现实应用中，我们很多时候很难精确求出精确的概率分布，常常采用近似推断方法。近似推断方法大致可分为两大类：第一类是采样(Sampling), 通过使用随机化方法完成近似；第二类是使用确定性近似完成近似推断，典型代表为变分推断(variational inference)。在很多任务中，我们关心某些概率分布并非因为对这些概率分布本身感兴趣，而是要基于他们计算某些期望，并且还可能进一步基于这些期望做出决策。采样法正式基于这个思路。

蒙特卡洛法（Monte Carlo）是指基于采样的数值型近似求解方法，具体来说，假定我们的目标是计算函数f(x)在概率密度函数p(x)下的期望：

$$E_p[f] = \int f(x)p(x)dx$$

根据 $p(x)$ 进行样本采样 $x_1, x_2, \cdots, x_N$ ，最终可计算f(x)在这些样本上的均值：

$$\hat{f} = \frac{1}{N}\sum_{i=1}^N f(x_i)$$

若概率密度函数 $p(x)$ 很复杂，则构造服从p分布的独立同分布样本也很困难。MCMC方法的关键在于通过构造“平稳分布为 $p(x)$ 的马尔可夫链”来产生样本：若马尔科夫链运行时间足够长，即收敛到平稳状态，则此时产出的样本X近似服从分布p。细致平衡条件为：

$$p(x^t)T(x^{t-1} \mid x^t) = p(x^{t-1})T(x^t \mid x^{t-1})$$

（2）Metropolis-Hastings算法采样过程:

对于目标分布 $p(x)$ ，首先选择一个容易采样的参考条件分布 $q(x^*|x)$ ，并令

$$A(x,x^*) = min[\frac{p(x^*)q(x|x^*)}{p(x)q(x^*|x)}, 1]$$

然后根据如下过程进行采样：

1）随机选一个初始样本 $x^{(0)}$ ；

2）For t = 1, 2, 3, ... :

根据参考条件分布 $q(x^*|x^{t-1})$ 抽取一个样本 x^* ；

根据均匀分布U(0,1)产生随机数 u ；

若 $u < A(x^{t-1}, x^*)$ ，则令 $x^t = x^*$ ，否则令 $x^t = x^{t-1}$ 。

（3）Gibbs Sampling算法采样过程:

吉布斯采样法是Metropolis-Hastings算法 $A(x^{t-1}, x^*) = 1$ 时的一个特例，其核心思想是每次只对样本的一个维度进行采样和更新。对于目标分布p(x)，按如下过程进行采样：

（1）随机选择初始状态 $x^{(0)} = (x_1^{(0)}, x_2^{(0)}, \ldots, x_d^{(0)})$。

（2）For t = 1, 2, 3, ... :

- 对于前一步产生的样本 $x^{(t-1)} = (x_1^{(t-1)}, x_2^{(t-1)}, \ldots, x_d^{(t-1)})$，依次采样和更新每个维度的值，即依次抽取分量 $x_1^{(t)} \sim p(x_1 \mid x_2^{(t-1)}, x_3^{(t-1)}, \ldots, x_d^{(t-1)})$，$x_2^{(t)} \sim p(x_2 \mid x_1^{(t)}, x_3^{(t-1)}, \ldots, x_d^{(t-1)})$，...，$x_d^{(t)} \sim p(x_d \mid x_1^{(t)}, x_2^{(t)}, \ldots, x_{d-1}^{(t)})$；
- 形成新的样本 $x^{(t)} = (x_1^{(t)}, x_2^{(t)}, \ldots, x_d^{(t)})$。

3、pLSA中的参数估计：EM求解

（1）通过极大似然估计建立目标函数：

极大似然估计：w_j在d_i中出现的次数 $n(d_i, w_j)$

$$L = \prod_{i=1}^{N}\prod_{j=1}^{M} P(d_i, w_j) = \prod_i \prod_j P(d_i, w_j)^{n(d_i, w_j)}$$

$$P(d_i, w_j) = P(w_j \mid d_i)P(d_i)$$

$$P(w_j \mid d_i) = \sum_{k=1}^{K} P(w_j \mid z_k)P(z_k \mid d_i)$$

$$\begin{aligned} l &= \sum_i \sum_j n(d_i, w_j) \log P(d_i, w_j) \\ &= \sum_i \sum_j n(d_i, w_j) \log P(w_j \mid d_i)P(d_i) \\ &= \sum_i \sum_j n(d_i, w_j) \log\left(\sum_{k=1}^{K} P(w_j \mid z_k)P(z_k \mid d_i)\right)P(d_i) \\ &= \sum_i \sum_j n(d_i, w_j) \log\left(\sum_{k=1}^{K} P(w_j \mid z_k)P(z_k \mid d_i)P(d_i)\right) \end{aligned}$$

（2）EM求解-E步：

确定后验概率：

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

并带入新的期望目标函数中：

$$E(l_{new}) = \sum_i \sum_j n(d_i, w_j) \sum_{k=1}^{K} P(z_k \mid d_i, w_j) \log P(w_j, z_k \mid d_i)$$
$$= \sum_i \sum_j n(d_i, w_j) \sum_{k=1}^{K} P(z_k \mid d_i, w_j) \log P(w_j \mid z_k) P(z_k \mid d_i)$$

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

（3）EM求解-M步：

$$\begin{cases} P(w_j \mid z_k) = \dfrac{\sum_i n(d_i, w_j) P(z_k \mid d_i, w_j)}{\sum_{m=1}^{M} \sum_i n(d_i, w_j) P(z_k \mid d_i, w_j)} \\ P(z_k \mid d_i) = \dfrac{\sum_j n(d_i, w_j) P(z_k \mid d_i, w_j)}{\sum_{k=1}^{K} \sum_j n(d_i, w_j) P(z_k \mid d_i, w_j)} \end{cases}$$

4、LDA中的参数估计：Gibbs Sampling

本节中通过Gibbs Sampling对进行参数估计，需要特别指出的是，**Gibbs Sampling其实不是求解的过程，而是通过采样去求后验分布的期望，从而估计最终参数。**

通过Gibbs Sampling对进行参数估计分为3个步骤：1）确定联合分布；2）求解后验概率Gibbs updating rule；3）确立后验分布并求期望估计参数；

（1）确定联合分布：

$$p(\vec{w}, \vec{z} \mid \vec{\alpha}, \vec{\beta}) = p(\vec{w} \mid \vec{z}, \vec{\beta}) p(\vec{z} \mid \vec{\alpha}) = \prod_{k=1}^{K} \frac{\Delta(\vec{\phi}_K + \vec{\beta})}{\Delta(\vec{\beta})} \prod_{m=1}^{M} \frac{\Delta(\vec{\theta}_m + \vec{\alpha})}{\Delta(\vec{\alpha})}$$

（2）根据（1）求出的联合分布可以求解Gibbs updating rule

$$\begin{aligned} p(z_i{=}k \mid \vec{z}_{\neg i}, \vec{w}) &= \frac{p(\vec{w}, \vec{z})}{p(\vec{w}, \vec{z}_{\neg i})} = \frac{p(\vec{w} \mid \vec{z})}{p(\vec{w}_{\neg i} \mid \vec{z}_{\neg i}) p(w_i)} \cdot \frac{p(\vec{z})}{p(\vec{z}_{\neg i})} \\ &\propto \frac{\Delta(\vec{n}_z + \vec{\beta})}{\Delta(\vec{n}_{z,\neg i} + \vec{\beta})} \cdot \frac{\Delta(\vec{n}_m + \vec{\alpha})}{\Delta(\vec{n}_{m,\neg i} + \vec{\alpha})} \\ &= \frac{\Gamma(n_k^{(t)} + \beta_t)\, \Gamma(\sum_{t=1}^{V} n_{k,\neg i}^{(t)} + \beta_t)}{\Gamma(n_{k,\neg i}^{(t)} + \beta_t)\, \Gamma(\sum_{t=1}^{V} n_k^{(t)} + \beta_t)} \cdot \frac{\Gamma(n_m^{(k)} + \alpha_k)\, \Gamma(\sum_{k=1}^{K} n_{m,\neg i}^{(k)} + \alpha_k)}{\Gamma(n_{m,\neg i}^{(k)} + \alpha_k)\, \Gamma(\sum_{k=1}^{K} n_m^{(k)} + \alpha_k)} \\ &= \frac{n_{k,\neg i}^{(t)} + \beta_t}{\sum_{t=1}^{V} n_{k,\neg i}^{(t)} + \beta_t} \cdot \frac{n_{m,\neg i}^{(k)} + \alpha_k}{[\sum_{k=1}^{K} n_m^{(k)} + \alpha_k] - 1} \\ &\propto \frac{n_{k,\neg i}^{(t)} + \beta_t}{\sum_{t=1}^{V} n_{k,\neg i}^{(t)} + \beta_t} (n_{m,\neg i}^{(k)} + \alpha_k) \end{aligned}$$

（3）确立后验分布并求期望估计参数：

每个文档上Topic的后验分布和每个Topic下的词的后验分布分别如下（据上文可知：其后验分布跟它们的先验分布一样，也都是Dirichlet 分布）：

$$p(\vec{\vartheta}_m|\vec{z}_m,\vec{\alpha}) = \frac{1}{Z_{\vartheta_m}} \prod_{n=1}^{N_m} p(z_{m,n}|\vec{\vartheta}_m) \cdot p(\vec{\vartheta}_m|\vec{\alpha}) = \mathrm{Dir}(\vec{\vartheta}_m|\vec{n}_m + \vec{\alpha})$$

$$p(\vec{\varphi}_k|\vec{z},\vec{w},\vec{\beta}) = \frac{1}{Z_{\varphi_k}} \prod_{\{i:z_i=k\}} p(w_i|\vec{\varphi}_k) \cdot p(\vec{\varphi}_k|\vec{\beta}) = \mathrm{Dir}(\vec{\varphi}_k|\vec{n}_k + \vec{\beta})$$

根据Dirichlet 分布参数估计：

$$\varphi_{k,t} = \frac{n_k^{(t)} + \beta_t}{\sum_{t=1}^{V} n_k^{(t)} + \beta_t}$$

$$\vartheta_{m,k} = \frac{n_m^{(k)} + \alpha_k}{\sum_{k=1}^{K} n_m^{(k)} + \alpha_k}$$

5、LDA的训练和预测过程：

（1）训练过程

1）选择合适的主题数K，选择合适的超参数向量$\vec{\alpha}, \vec{\eta}$

2）对应语料库中每一篇文档的每一个词，随机的赋予一个主题编号z

3）重新扫描语料库，对于每一个词，利用Gibbs采样公式更新它的topic编号，并更新语料库中该词的编号。

4）重复第3步的基于坐标轴轮换的Gibbs采样，直到Gibbs采样收敛。

5）统计语料库中的各个文档各个词的主题，得到文档主题分布θ_d，统计语料库中各个主题词的分布，得到LDA的主题与词的分布β_k。

（2）预测过程：LDA的各个主题的词分布 β_k 已经确定:

1）对应当前文档的每一个词，随机的赋予一个主题编号z

2）重新扫描当前文档，对于每一个词，利用Gibbs采样公式更新它的topic编号。

3）重复第2步的基于坐标轴轮换的Gibbs采样，直到Gibbs采样收敛。

4）统计文档中各个词的主题，得到该文档主题分布。

6、LDA主题数目选择及评估标准

在LDA中，主题的个数K是一个预先指定的超参数。对于模型超参数的选择，实践中的做法一般是将全部数据集分成训练集、验证集、和测试集3部分，然后利用验证集对超参数进行选择。例如，在确定LDA的主题个数时，我们可以随机选取60%的文档组成训练集，另外20%的文档组成验证集，剩下20%的文档组成测试集。在训练时，尝试多组超参数的取值，并在验证集上检验哪一组超参数所对应的模型取得了最好的效果。最终，在验证集上效果最好的一组超参数和其对应的模型将被选定，并在测试集上进行测试。

为了衡量LDA模型在验证集和测试集上的效果，需要寻找一个合适的评估指标。一个常用的评估指标是困惑度（perplexity）。在文档集合D上，模型的困惑度被定义为：

$$\mathrm{perplexity}(D) = \exp\left\{-\frac{\sum_{d=1}^{M} \log p(\boldsymbol{w}_d)}{\sum_{d=1}^{M} N_d}\right\}$$

其中M为文档的总数， w_d 为文档d中单词所组成的词袋向量，p(w_d)为模型所预测的文档d的生成概率， N_d 为文档d中单词的总数。

References

[1] 《百面机器学习：算法工程师带你去面试》

[2] 通俗理解LDA主题模型

[3] LDA求解之Gibbs采样算法