学员分享： LDA实践中的几个问题

squirrel_d@126.com

提　纲

- LDA中主题数量的确定方法
- LDA中超参数的意义
- LDA的假设及其在公式推导中的意义

- **LDA中主题数量的确定方法**

方法一. Likelihood

也即p(W|T)，W是语料库中出现的所有词(不去重)，T是设置的主题数量

方法二. perplexity（困惑度）

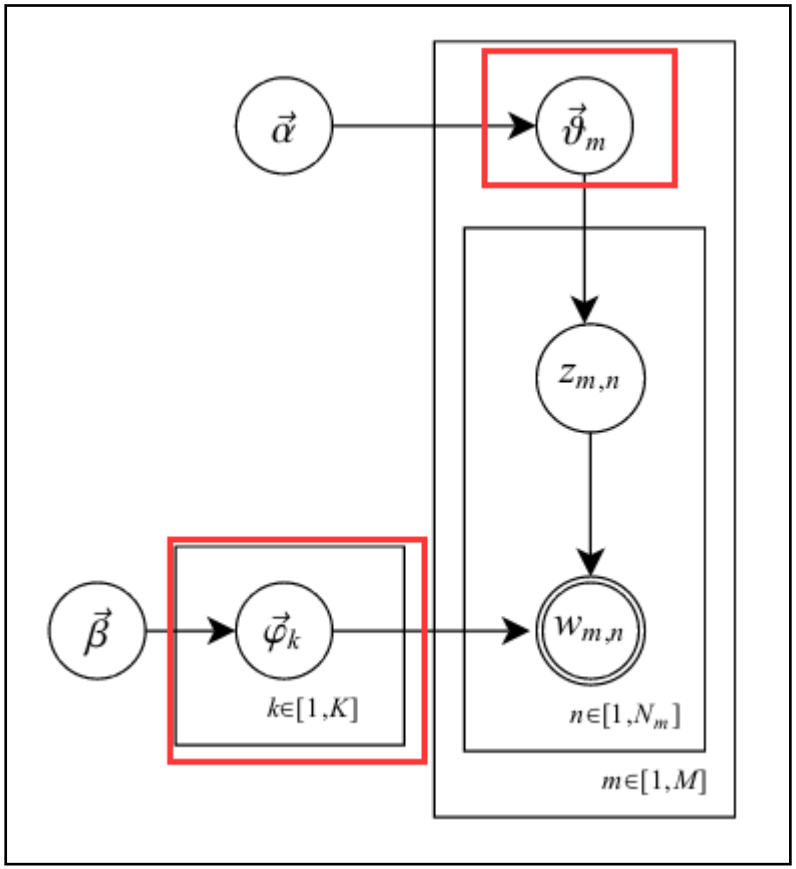
计算公式： $perplexity(D_{\text{test}}) = \exp\left\{-\frac{\sum_{d=1}^{M}\log p(\mathbf{w}_d)}{\sum_{d=1}^{M}N_d}\right\}$

一种常用的聚类质量评价标准；
随主题数量增多而单调减少，困惑度越低，模型越好；
计算的是测试集上所有单词出现概率的几何均值的倒数[1]；

直观上来讲，困惑度描述的是在测试集上产生一个token所需词表的期望大小，这个词表的单词符合均值分布[2]。

- LDA中主题数量的确定方法

Likelihood：已知theta和phi，怎么计算p(W|T)？



- # LDA中主题数量的确定方法

Likelihood：怎么计算p(W|T)？

回到问题的出发点,根据log最大似然估计法，我们要计算：

$$\arg\max_{\theta,\varphi} L(\theta,\varphi) = \arg\max_{\theta,\varphi} \log P(\mathrm{D},\mathrm{W};\theta,\varphi)$$

$$= \arg\max_{\theta,\varphi} \sum_{d,w} n(d,w) \log P(d,w;\theta,\varphi) \qquad \textbf{(1)}$$

根据LDA的图模型，我们可以得到[1]:

$$p(\mathcal{W}|\underline{\Theta},\underline{\Phi}) = \prod_{m=1}^{M} p(\vec{w}_m|\vec{\vartheta}_m,\underline{\Phi}) = \prod_{m=1}^{M}\prod_{n=1}^{N_m} p(w_{m,n}|\vec{\vartheta}_m,\underline{\Phi})\ . \qquad \textbf{(2)}$$

观察两个公式，对于公式(2)中，$p(w_{m,n}|\vec{\vartheta}_m,\underline{\Phi})$ 暗含了其所在文档dm的信息，因为生成m文档中词汇是在确定文档dm的文档-主题概率 $\vec{\vartheta}_m$ 的前提下 。也即，两个公式描述的是一回事。

• LDA中主题数量的确定方法

Likelihood:怎么计算p(W|T)

在LDA运行时需要设定T，算出theta和phi后，我们用公式(1)来求p(W|T)

$$\arg\max_{\theta,\varphi} \sum_{d,w} n(d,w) \log P(w|d;\varphi,\theta)P(d)$$

$$= \arg\max_{\theta,\varphi} \left\{ \sum_{d,w} n(d,w) \log P(w|d;\varphi,\theta) + \sum_{d,w} n(d,w) \log P(d) \right\}$$

红框内容与theta，phi值无关，可以略去

$$\propto \arg\max_{\theta,\varphi} \sum_{d,w} n(d,w) \log P(w|d;\theta)$$

$$= \arg\max_{\theta,\varphi} \sum_{d,w} n(d,w) \log \sum_{z} P(w|z)P(z|d)$$

红框内容在LDA运行后都已得到，因而这就是求训练集likelihood的公式

- # LDA中主题数量的确定方法

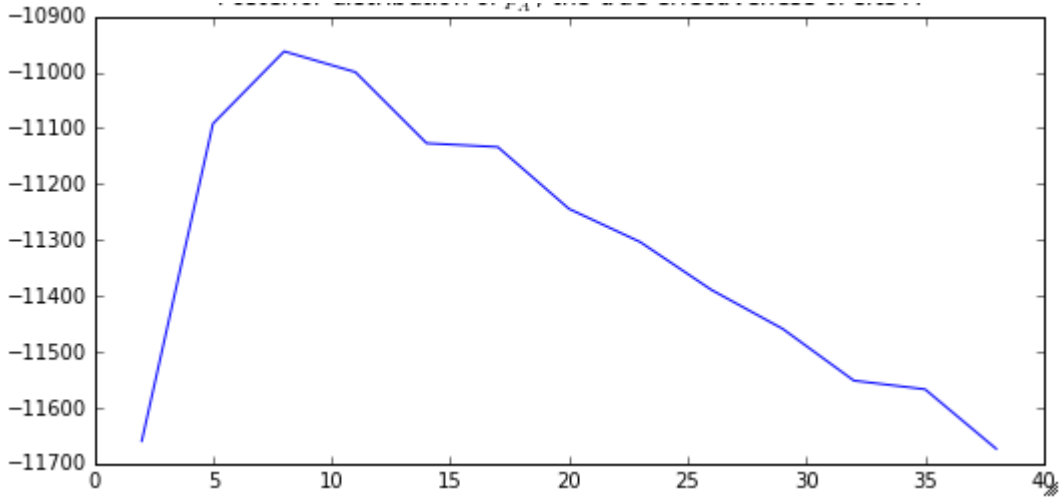
Likelihood：已知theta和phi，怎么计算p(W|T)？

示例：

(1) 收集50个文档

(2) 主题数设置为range(2,40,3)，也即从5到40之间，每3个数取一个值所形成的数列

(3) 对每个主题数，跑一次lda，计算该主题数对应的似然值



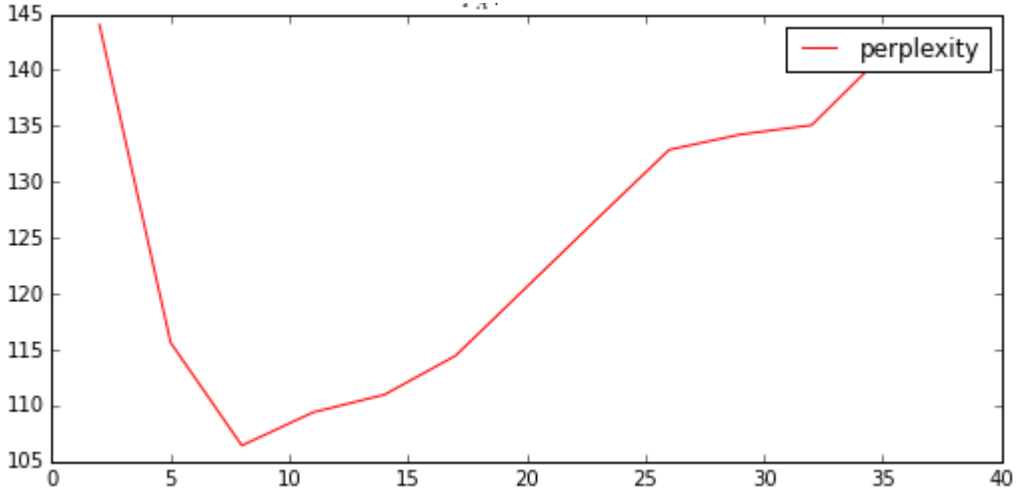
• LDA中主题数量的确定方法

Perplexity方法

我找的LDA代码中这部分的实现有误。建议大家Debug下这部分代码，加深LDA的理解

代码地址：https://github.com/shuyo/iir/blob/master/lda/

Debug前perplexity随主题数量的变化情况

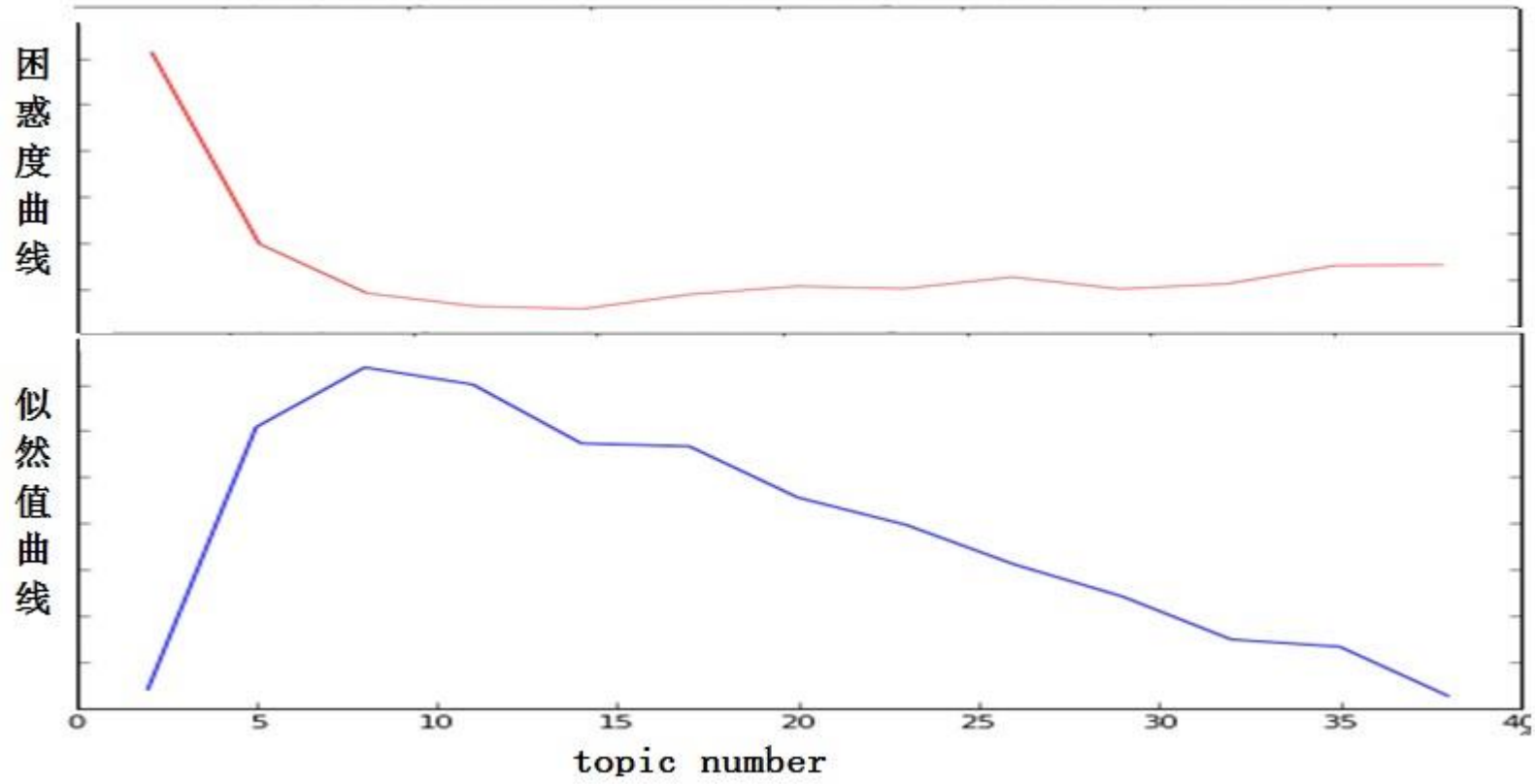


• LDA中主题数量的确定方法

Perplexity方法

Debug后perplexity随主题数量的变化情况

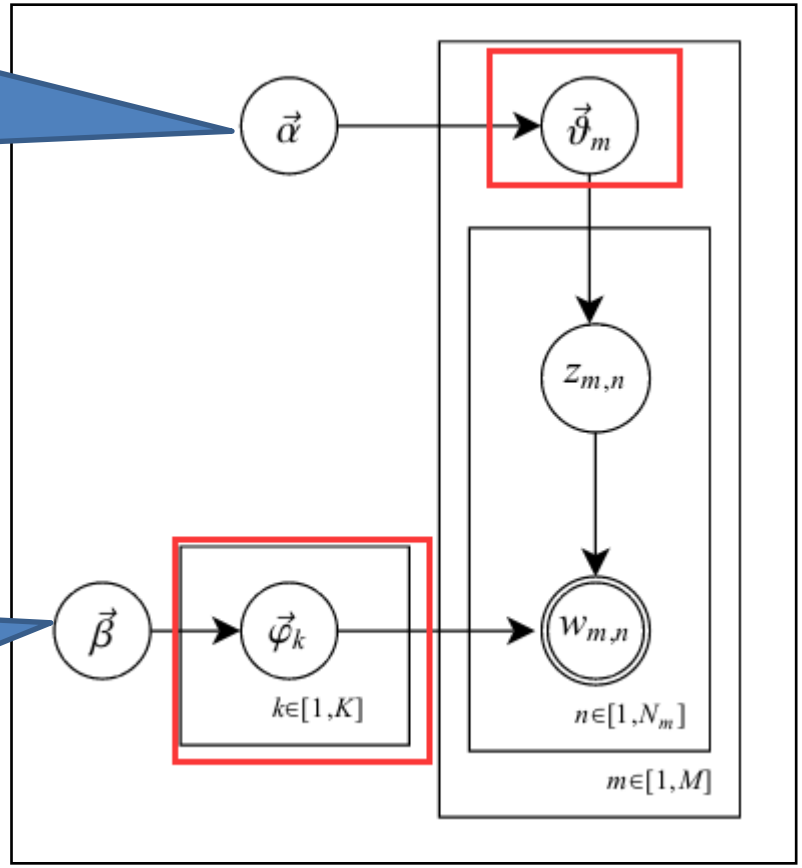
So,根据困惑度曲线选择主题数量时候，采用elbow方法



- LDA中超参数的意义(对称狄氏分布)

控制文档-主题多项分布的超参数，alpha越小，文档越集中于少数几个主题

控制文档-主题多项分布的超参数，beta越小，主题越集中于少数几个词汇

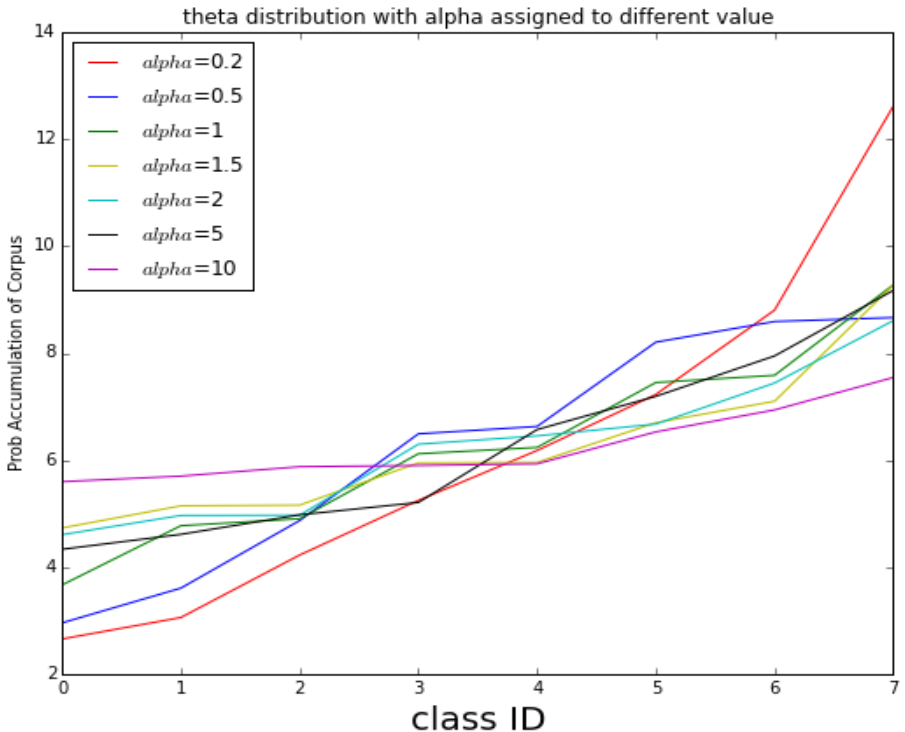


- **LDA中超参数的意义(对称狄氏分布)**

以α为例:

(1) 在设定主题数量和β不变的情况下， α分别取值[0.2, 0.3, 0.5, 1, 1.5, 2, 5, 10]

(2) 对于每个α，训练得到每个文档上的主题分布概率，对于每个主题，计算所有文档在它上面的累积概率



- **LDA的假设及其在公式推导中的意义**

LDA的假设：

(1) 给定主题的条件下，词汇和文档独立；

(2) 同一主题下，词汇独立同分布；

(3) 同一文档下，主题独立同分布。

提几个问题供大家思考：

(1) 语料库中，每个词汇wi最终都分配了主题zi，那么zi与zj(i!=j)是否独立？

(2) wi与zj(i!=j)是否独立？

(3) $p(w_i|\mathbf{w}_{\neg i}, \mathbf{z}_{\neg i}, \alpha, \beta)$ =? $p(w_i|\alpha, \beta)$

(4) 如何把p(W,D)写成wi,dj的形式？其中W表示语料库所有词，D表示语料库所有文档。

参考引文：

[1] Blei D,latent Dirichlet Allocation

[2] Heinrich G,parameter estimation in text mining