主题模型

解决一词多义/多词一义的不足

① 应用：语义分析，文档分类/聚类，文章摘要，社区挖掘

对时空间图像聚类，目标识别，生物信息数据应用，

用主题模型做特征降维：e.g. N个词的特征空间 —主题模型→ 20个主题的特征空间

（文档在20个主题的命重情况 → one hot编码）

带权重的主题模型，当样本通常包含1-2个少数主题时，认为是一个soft聚类

② 输入输出（20个主题为例）

输入：从N篇文档提出的N个样本，每个样本都是一个词袋（用TF-IDF等预处理得到）

输出：每篇文档在20个主题上的分布 [.1,.01,....] ⇐ 文档i

每种文档类型在20个主题上的分布 [.1,.01,...] ⇐ 武侠

20个主题，以及代表每个主题的关键词到表

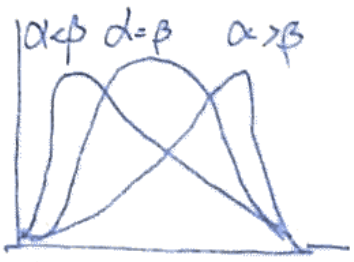
③ 用到的数学知识：

* Γ函数（阶乘在实数域上的推广）：见概率

* Beta分布：$f(x)=\begin{cases}\frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}x^{\alpha-1}(1-x)^{\beta-1}, & x\in[0,1]\\ 0 & \text{其它}\end{cases}$ ：见概率：

是两点/二项分布的共轭先验分布.

$$E(x)=\frac{\alpha}{\alpha+\beta}$$



* 共轭先验分布，先验概率，后验概率，"证据"，似然概率（贝叶斯概率范畴）

$$\underbrace{P(\theta|x)}_{\text{后验概率}}=\underbrace{P(x|\theta)}_{\text{似然概率}}\cdot\underbrace{P(\theta)}_{\text{先验概率}}/\underbrace{P(x)}_{\text{"证据"(归一化因子)}}$$

参数 样本

<a> <b>

后验概率：从样本计算参数θ的概率

似然概率：给定参数θ计算样本发生的概率

先验概率：θ取值的概率

"证据"：对θ影响时样本发生的概率（与θ无关可当作常数）

当后验概率与先验概率分布相同时 $P(\theta|x)$ $P(\theta)$ "同一种分布率"

<a> $P(\theta|x)$ 与 $P(\theta)$ 是共轭分布

<b> $P(\theta)$ 先验分布被叫作 $P(x|\theta)$（似然函数）的共轭先验分布

用"二项分布"建模，似然概率就是二项分布

用"贝塔分布"建模，……

④ 共轭先验分布应用于二分类问题

后验 $P(\theta|x)$ = 似然 $P(x|\theta)$ · 先验 $P(\theta)$ / 证据 $P(x)$ $\propto$ $P(x|\theta)P(\theta)$

$P(x)$ 与 θ 无关，将 $P(x)$ 当作常数消去

二分类问题用二项分布建模

$$P(x|\theta) = C_n^k \theta^k (1-\theta)^{n-k}$$

用Beta分布来建模

$$p(\theta|\alpha,\beta) = \begin{cases} \frac{1}{B(\alpha,\beta)} \theta^{\alpha-1}(1-\theta)^{\beta-1} & \theta \in [0,1] \\ 0 & \text{其它} \end{cases}$$

后验分布也是Beta分布，推导如下

$$P(\theta|x) = P(x|\theta)\cdot P(\theta)/P(x) \propto P(x|\theta)P(\theta)$$

$$= \left(C_n^k \theta^k (1-\theta)^{n-k}\right)\left(\frac{1}{B(\alpha,\beta)}\theta^{\alpha-1}(1-\theta)^{\beta-1}\right)$$

$$= \frac{C_n^k}{B(\alpha,\beta)} \theta^{(k+\alpha)-1}(1-\theta)^{(n-k+\beta)-1}$$

成正比，因为是两个常数

$$\propto \frac{1}{B(k+\alpha, n-k+\beta)} \theta^{(k+\alpha-1)}(1-\theta)^{(n-k+\beta-1)}$$

共轭分布：

- 先验分布：服从Beta分布，数学期望是 $\frac{\alpha}{\alpha+\beta}$ → 是二项分布的共轭先验分布
- 后验分布：服从Beta分布，数学期望是 $\frac{k+\alpha}{n+\alpha+\beta}$

* 如何理解上面的概率

先验分布：例如男生占总人数比例为50%，即 $\alpha=500, \beta=500$（是事先给定的超参数，是伪计数）（学校总共1000人） $\frac{\alpha}{\alpha+\beta}=50\%$

似然概率：作了 n 次独立实验，共观察到 k 个男生，例如 $n=200, k=150$

后验分布：数学期望就是 $\frac{k+\alpha}{n+\alpha+\beta} = \frac{150+500}{200+500+500}$（来自样本 / 来自经验）

* 在似然概率中，样本越多，最终后验分布越接近由 n 次试验得到的概率
样本越少，越接近来自经验的先验概率

⑤ 共轭先验分布用于多分类问题

后验 $P(\theta|X)$ = 似然 $P(X|\theta)$ · 先验 $P(\theta)$ / 证据 $P(X)$ $\propto P(X|\theta)\cdot P(\theta)$

多分类问题用多项分布建模

$$P(X_1=n_1,\cdots,X_k=n_k)=\begin{cases}n!\prod_{i=1}^{k}\frac{p_i^{n_i}}{n_i!}, & \sum_{i=1}^{k}n_i=n\\ 0, & \text{otherwise}\end{cases}$$

将样本观察似然函数，得到在k个类别上样本出现的次数

$$\vec{C}=\{C_1,C_2,\cdots C_k\}$$

用Dirichlet分布来建模

$$Dir(\vec{p}|\vec{\alpha})=\begin{cases}\frac{1}{\Delta(\vec{\alpha})}\prod_{k=1}^{K}p_k^{\alpha_k-1}, & p_k\in[0,1]\\ 0, & \text{其它}\end{cases}$$

$\vec{p}$：k维向量

其中 $\Delta(\vec{\alpha})=\prod_{k=1}^{K}\Gamma(\alpha_k)\Big/\Gamma\left(\sum_{k=1}^{K}\alpha_k\right)$

数学期望

$$E(p_i)=\alpha_i\Big/\sum_{i=1}^{k}\alpha_k$$

记为 $Dir(K,\vec{\alpha})$

Dirichlet分布是多项分布的共轭先验分布

因此后验分布也服从Dirichlet，略去推导过程，得到后验分布

$P(\theta|X)$ 服从 $Dir(K,\vec{\alpha}+\vec{C})$

数学期望

$$E(p_i)=(\alpha_i+C_i)\Big/\sum_{i=1}^{k}(\alpha_k+C_k)$$

$$Dir(K,\vec{\alpha}+\vec{C})$$
$$=Dir(K,C_1+\alpha_1,\cdots,C_k+\alpha_k)$$

⑥ 二项分布 ⟶ 多项分布

Beta分布 ⟶ Dirichle分布

是从2分类到多分类的扩展

$\vec{\alpha}$调参（见PPT）

将$\vec{\alpha}$简化为对称Dirichlet分布的情况，即$(\alpha_1=\alpha_2=\cdots=\alpha_k)$作函数图，得到：

① α_i $(i=1,\cdots k)$ 值越小，主题越集中（小于1，等于1相当于没先验，大于1越均匀）

另外：根据后验概率数学期望

② $\vec{\alpha}$ 越不均一，越容易在某些主题聚集

③ α 越大，先验参数向量越大

④ α 越小，越聚集

⑦ LDA的算法~~框架~~理解.

每一篇文章都有一个主题分布 $D_i \sim (\theta_{i1}, \theta_{i2}, \cdots \theta_{ik})$: 超参数记为 $\vec{\alpha}$ 设为01

↓

根据①，后验都是Dirichlet分布

- 先验概率：建模为Dirichlet分布，$\vec{\alpha} = \{\alpha_1, \cdots \alpha_k\}$ {方法1：事先给定如果 / 方法2：计算(吉伯斯)
- 似然概率：建模为多项分布.

↑

每个主题都有一个词分布 $T_j \sim (\theta_{w1}, \theta_{w2}, \cdots \theta_{wv})$ 超参数记为 $\vec{\beta}$ 设为01

- 先验概率：建模为Dirichlet分布，$\vec{\beta} = \{\beta_1, \cdots, \beta_n\}$ {方法1：事先给定 / 方法2：
- 似然概率：建模为多项分布

遍历 查词位 t：$0 \sim \max\{length(D_j)\}$.

遍历文章序号：$j: 0 \sim m$

拿到二元组 $<j, t>$，即第j篇文章第t个词的位置标记，向这个位置填一个词

如何填这个词？

文章 D_j ⟶ 主题 T_k ⟶ 词 W_{jt}

根据 D_j 的主题分布采样一个主题

$\theta_{D_j} = (\theta_{j1}, \theta_{j2}, \cdots \theta_{jK})$

根据 T_k 的词分布采样一个词.

$\varphi_{T_k} = (\varphi_{k1}, \varphi_{k2}, \cdots \varphi_{kV})$

上面来理解，假定我们知道 $\vec{\theta}$ 和 $\vec{\varphi}$，如何生成M篇文档的词 $\{w_{ji}\}$

而实际中，词 w_{mn} 是能从样本观察到的已知变量，$\vec{\alpha}, \vec{\beta}$ 是事先给定的先验参数，$\vec{\theta}, \vec{\varphi}$ 是未知的隐变量，是我们要求的.

① 如何从已知的词 $w_{m,n}$ (文档m，词n) 背后主题是 z_k 的概率？(算主题的词

用Gibbs采样：<a> 初始给每个 $w_{m,n}$ 随机指定 $z_k^{(0)}$

<b> 每一轮迭代：a. 统计每个主题下出词t的数量

b. 统计每个文档m下出现主题z的数量

c. 用计算 $p(z | z_{-i}, d, w)$：即排除当前词的主题分布 (d: 文档词, w: 给定的)

d. 用新的分布更新 $z_k^{(t)}$

<d> 直到收敛或迭到最大迭代次数

理解下面的指导思想

词w是主题为k的概率.

$$p(z_i = k | z_{-i}, \overset{\text{词}}{w}) = \cdots \propto \frac{n_{k-i}^{(t)} + \beta_i}{\sum_{t=1}^{V} n_{k-i}^{(t)} + \beta_i} (n_{m-i}^{(k)} + \alpha_k)$$

词w是主题k的概率 (文章中) × 这是主题k的概率 (文章中)

排除词 其他词是主题k的概率 (文章中)

⑥ 结论：

主题的词分布：$\varphi_{k,t} = \frac{n_k^{(t)} + \beta_t}{\sum_{t=1}^{V} n_k^{(t)} + \beta_t} = \cdots$ （见PPT）

文章的主题分布：$\vartheta_{m,k} = \frac{n_m^{(k)} + \alpha_k}{\sum_{k=1}^{K} n_m^{(k)} + \alpha_k} = \cdots$ （见PPT）

⑦ 代码：见PPT.

⑧ 主题个数（超参数K）的确定

方法1：主题间相似度.

K个主题词分布向量，计算向量间的两两相似度 求均值

方法2：使用概率分布困惑度（Perplexity）

① 原理：困惑度定义 $Perplexity = e^{-\frac{1}{N}\sum_{i=1}^{n} \ln q(x_i)}$

样本数 N，样本 i，$q(x_i)$ 概率分布

$= e^{H(q)}$ ← 信息熵

越大越接近均匀分布

越小越退化成一种确定性质的度量.

（越小越好）

如果已知样本真实分布 p，这个指数部分也可以换成交叉熵.

$H(p,q) = -\sum_x p(x) \ln q(x)$

也可定义为 $a^{-\frac{1}{N}\sum_{i=1}^{n} \log_a q(x_i)}$，$a$ 个任意整数

② 现计算LDA模型的困惑度：

算出 $q(x)$ 代入 Perplexity 公式 就可以

$q(x) = P(W|Model) = \prod_{i=1}^{V} p(w_i|Model)^{-\frac{1}{V}} = e^{-\frac{1}{V}\sum_{i=1}^{V} \log_e p(w_i|Model)}$

V 个样数

$p(w_i|Model)$ 么

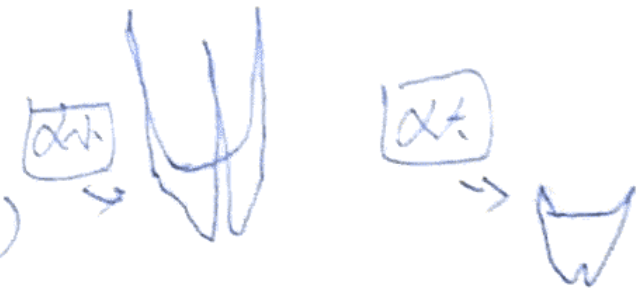
$p(w_i|Model) = \prod_{n=1}^{N_m} \sum_{i=1}^{K} p(z=k|d=m)\, p(w=t|z=k)$

$= \prod_{t=1}^{V} \left(\sum_{k=1}^{K} \theta_{m,k} \varphi_{k,t} \right)$

③ 使用：为不同 K 训练出 LDA 模型，计算困惑度，

选困惑度小的（实践效果并非理论那么好）

⑨ α, β 超参数确定：



α：越小越能明（文章主题的超参数）

β：（主题的词分布超参数）越小主题越突出，越不受无关词的影响

e.g. β小 → 可能只有大理、段誉这样的词影响武侠这个主题

β大 → 金庸、洛阳这样的词也会差距到武侠上.

<a> 直接设置（不具普遍适用性）← 但实践中往往这么设就够了

$\alpha = 50/k$

$\beta = 200/W$

<d> 用 Di-gamma（双Gamma）函数迭代求解

$$\alpha_k = \frac{\left[\sum_{m=1}^{M} \psi(n_m^{(k)} + \alpha_k)\right] - M\,\psi(\alpha_k)}{\left[\sum_{m=1}^{M} \psi\left[n_m + \sum_{j=1}^{k} \alpha_j\right]\right] - M \cdot \psi\left(\sum_{j=1}^{k} \alpha_j\right)} \cdot \alpha_k$$

↑ 先指定 然后一轮轮迭代

k：主题编号

$n_m^{(k)}$：主题k在文档m中出现多少次. (?)

⑩ TextRank

模仿 PageRank 的算法，从文章中提取代表句子.

⑪ word2vec → 用来发现近义词（给定个窗口长度）

把文章中的每个词，映射到一个向V维向量（V维空间一个点）

维度减小时（如减到200）→ 对应200维空间的一个点上.（词潜入）

☆如果[词A]与[词B]所在的语境相近，就会映射到接近的两个向量

词A与词B就当"近义词"（相邻词相近）

☆多种使用方法（见PPT）（实践课手写）