大数据管理方法与应用
第04章　概率潜在语义分析

西安交通大学管理学院
信息管理与电子商务系
智能决策与机器学习研究中心
刘佳鹏

2021 年 3 月 25 日

简介

- 概率潜在语义分析（probabilistic latent semantic analysis, PLSA），也称概率潜在语义索引（probabilistic latent semantic indexing, PLSI），是一种利用概率生成模型对文本集合进行话题分析的无监督学习方法
- 包含以下特点
  - 用隐变量表示话题
  - 整个模型表示文本生成话题，话题生成单词，从而得到单词-文本共现数据的过程
  - 每个文本由一个话题分布决定，每个话题由一个单词分布决定
- 概率潜在语义分析受潜在语义分析的启发，1999 年由Hofmann提出，前者基于概率模型，后者基于非概率模型
- 概率潜在语义分析最初用于文本数据挖掘，后来扩展到其他领域

概率潜在语义分析模型

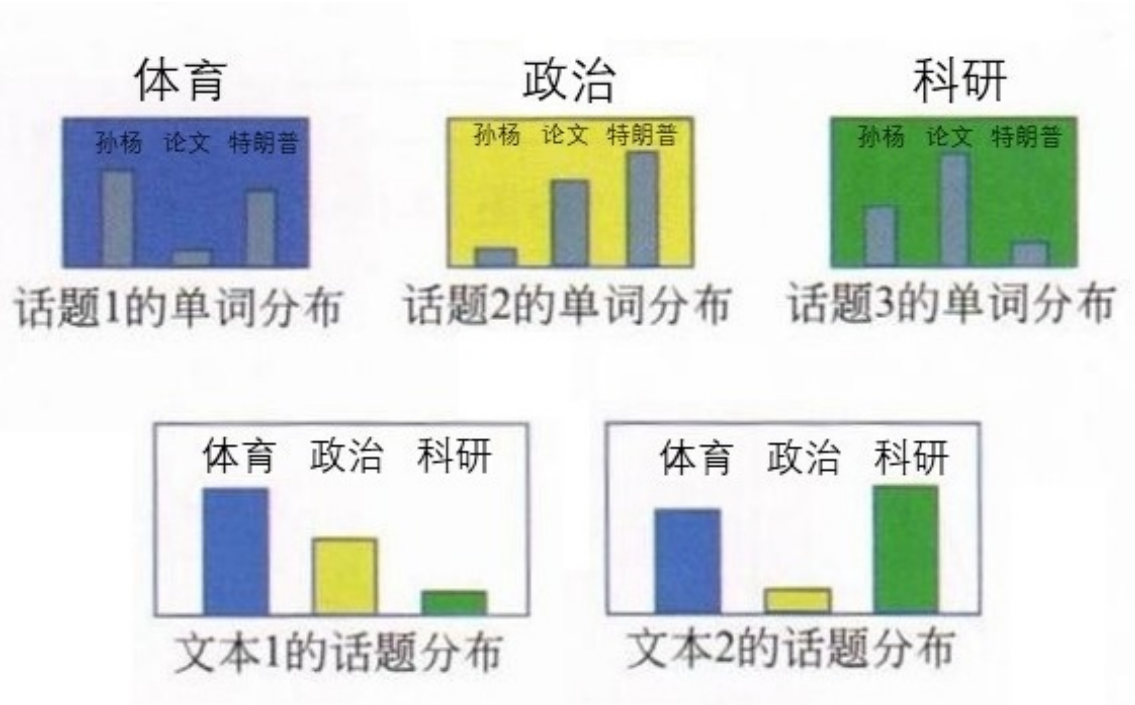
- 概率潜在语义分析模型有生成模型，以及等价的共现模型。先介绍生成模型，然后介绍共现模型，最后讲解模型的性质
- **问题界定：**
  - 单词集合$W = \{w_1, w_2, \cdots, w_M\}$，其中$M$是单词个数
  - 文本集合$D = \{d_1, d_2, \cdots, d_N\}$，其中$N$是文本个数
  - 话题集合$Z = \{z_1, z_2, \cdots, z_K\}$，其中$K$是预先设定的话题数
  - 随机变量w取值于单词集合
  - 随机变量d取值于文本集合
  - 随机变量z取值于话题集合

概率潜在语义分析模型

- **三个概率分布：**
- $P(d)$表示生成文本d的概率
- $P(z \mid d)$表示文本d生成话题z的概率
- $P(w \mid z)$表示话题z生成单词w的概率
- 概率分布$P(d)$、条件概率分布$P(z \mid d)$、条件概率分布$P(w \mid z)$皆属于多项分布

概率潜在语义分析模型

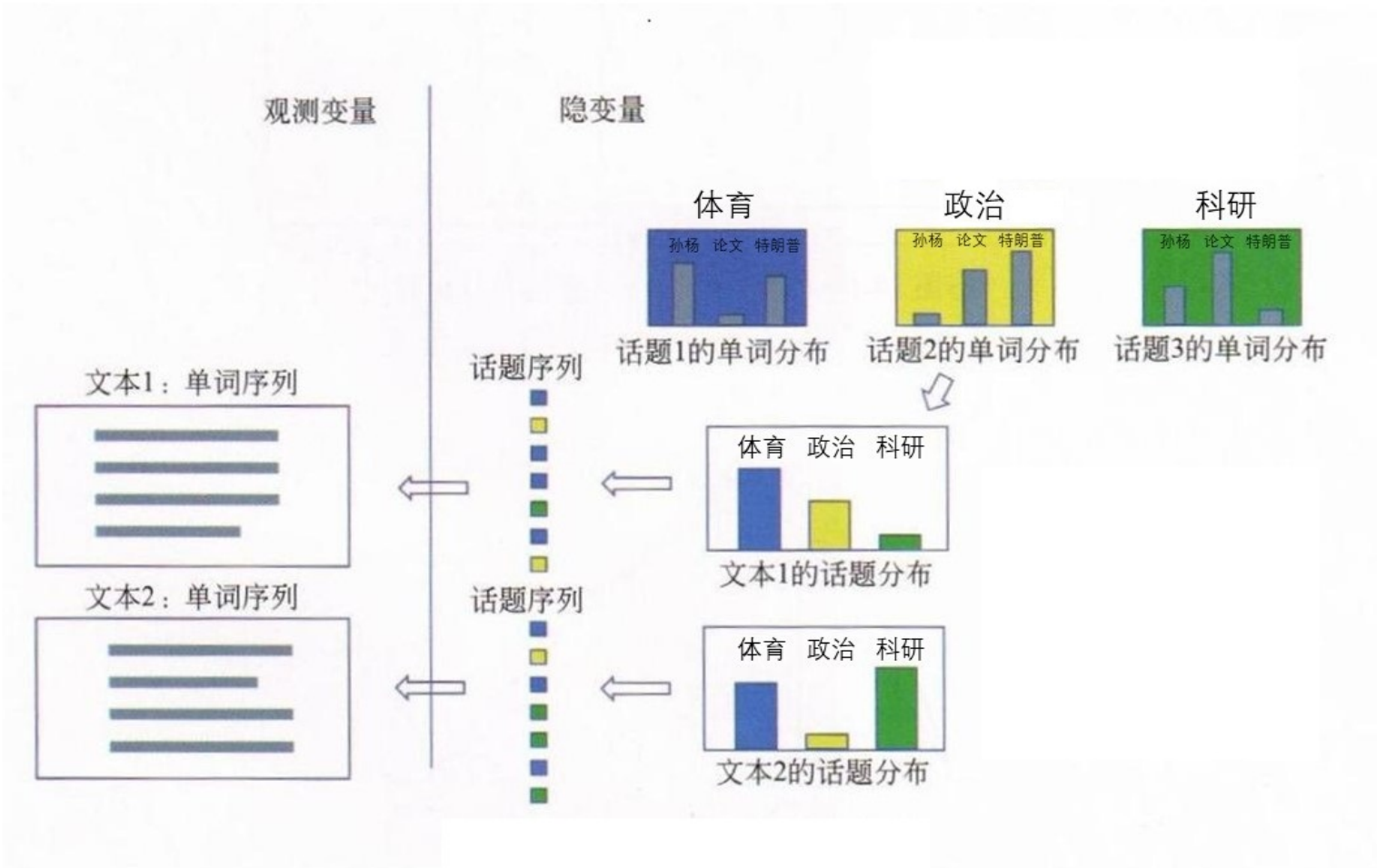
- **首先介绍生成模型**
- **模型要点：**
- 每个文本d拥有自己的话题概率分布$P(z \mid d)$，这意味着一个文本的内容由其相关话题决定
- 每个话题z拥有自己的单词概率分布$P(w \mid z)$，这意味着一个话题的内容由其相关单词决定



概率潜在语义分析模型

- 生成模型通过以下步骤生成文本-单词共现数据：
- （1）依据概率分布$P(d)$，从文本（指标）集合中随机选取一个文本d，共生成N个文本；针对每个文本，执行以下操作
- （2）在文本d给定条件下，依据条件概率分布$P(z \mid d)$，从话题集合随机选取一个话题z，共生成L个话题，这里L是文本长度
- （3）在话题z给定条件下，依据条件概率分布$P(w \mid z)$，从单词集合中随机选取一个单词w
- 注意这里为叙述方便，假设文本都是等长的，现实中不需要这个假设

概率潜在语义分析模型


观测变量
隐变量
体育
政治
科研
孙杨 论文 特朗普
孙杨 论文 特朗普
孙杨 论文 特朗普
话题1的单词分布
话题2的单词分布
话题3的单词分布
文本1：单词序列
话题序列
体育 政治 科研
文本1的话题分布
文本2：单词序列
话题序列
体育 政治 科研
文本2的话题分布

概率潜在语义分析模型

- **模型说明：**
- 生成模型中，单词变量w与文本变量d是观测变量，话题变量z是隐变量
- 模型生成的是单词-话题-文本三元组(w, z, d)的集合，但观测到的是单词-文本二元组(w, d)的集合
- 观测数据表示为单词-文本矩阵T的形式，矩阵T的行表示单词，列表示文本，元素表示单词-文本对(w, d)的出现次数

概率潜在语义分析模型

- 从数据的生成过程可以推出，文本-单词共现数据T的生成概率为所有单词-文本对(w, d)的生成概率的乘积，

$$P(T) = \prod_{(w,d)} P(w, d)^{n(w,d)}$$

这里$n(w, d)$表示(w, d)的出现次数，单词-文本对出现的总次数是$N \times L$

概率潜在语义分析模型

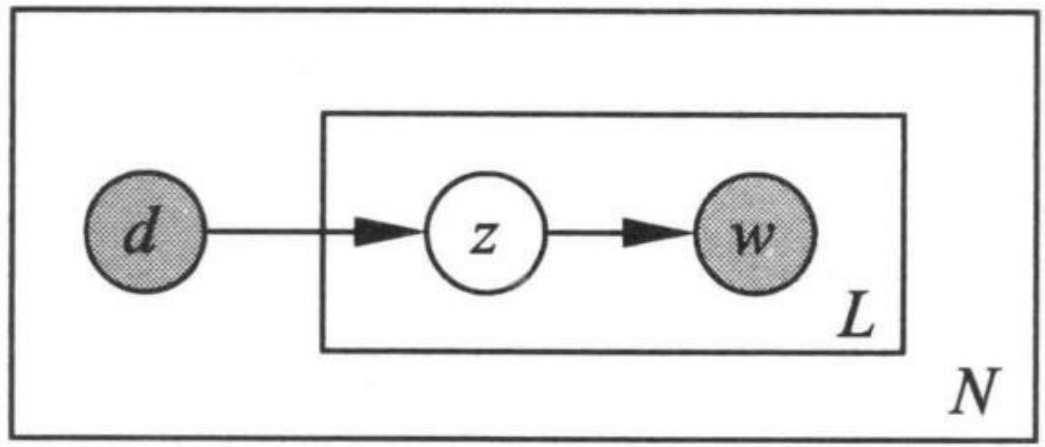
- 每个单词-文本对(w, d)的生成概率由以下公式决定：

$$\begin{aligned} P(w, d) &= P(d)P(w \mid d) \\ &= P(d) \sum_z P(w, z \mid d) \\ &= P(d) \sum_z P(z \mid d)P(w \mid z) \end{aligned}$$

  这就是生成模型的定义
- 生成模型假设在话题z给定条件下，单词w与文本d条件独立，即

$$P(w, z \mid d) = P(z \mid d)P(w \mid z)$$

概率潜在语义分析模型



概率潜在语义分析的生成模型

- 生成模型属于概率有向图模型，可以用有向图（directed graph）表示
- 实心圆表示观测变量，空心圆表示隐变量，箭头表示概率依存关系，方框表示多次重复，方框内数字表示重复次数
- 文本变量d是一个观测变量，话题变量z是一个隐变量，单词变量w是一个观测变量

概率潜在语义分析模型

- **接着介绍与以上生成模型等价的共现模型：**
- 文本-单词共现数据T的生成概率为所有单词-文本对(w, d)的生成概率的乘积：

$$P(T) = \prod_{(w,d)} P(w, d)^{n(w,d)}$$

- 每个单词-文本对(w, d)的概率由以下公式决定

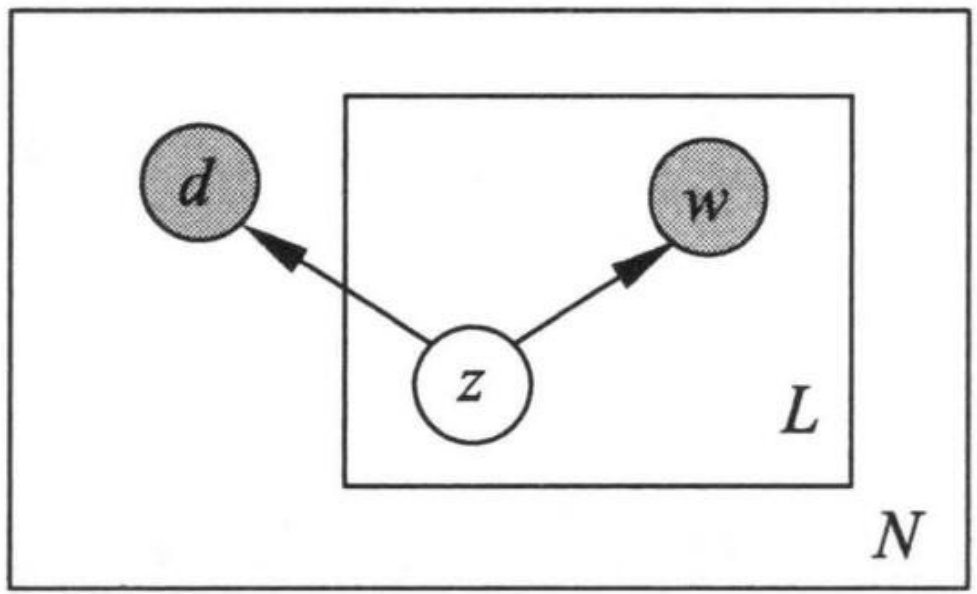
$$P(w, d) = \sum_{z \in Z} P(z)P(w \mid z)P(d \mid z)$$

此即共现模型的定义

概率潜在语义分析模型

- 共现模型假设在话题z给定条件下，单词w与文本d是条件独立的，即

$$P(w, d \mid z) = P(w \mid z)P(d \mid z)$$



概率潜在语义模型的共现模型

概率潜在语义分析模型

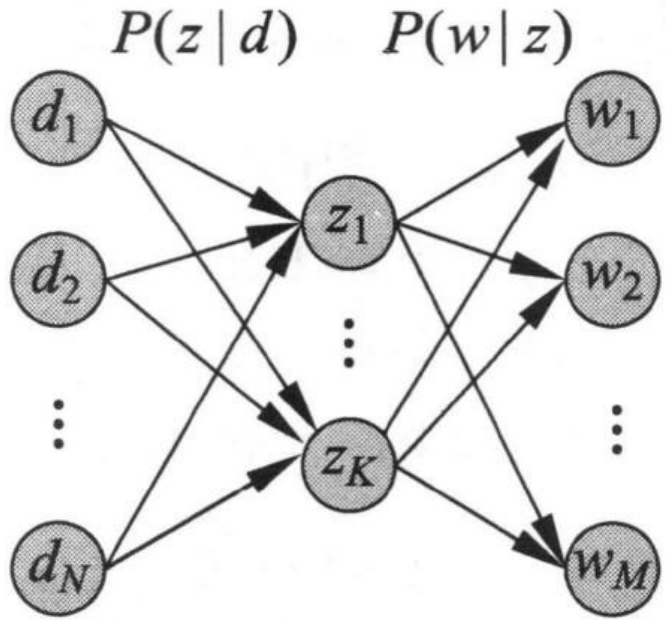
- **两种模型的比较：**
- 虽然生成模型与共现模型在概率公式意义上是等价的，但是拥有不同的性质：
- （1）生成模型刻画文本-单词共现数据生成的过程，共现模型描述文本-单词共现数据拥有的模式
- （2）生成模型中单词变量w与文本变量d是非对称的，而共现模型中单词变量w与文本变量d是对称的
- （3）由于两个模型的形式不同，其学习算法的形式也不同

概率潜在语义分析模型

- **接下来分别从“模型参数”、“几何解释”以及“与潜在语义分析的关系”这三个方面分别讨论概率潜在语义分析模型的性质：**
- **（1）模型参数**
- 如果直接定义单词与文本的共现概率$P(w, d)$，模型参数的个数是$O(M \cdot N)$，其中M是单词数，N是文本数
- 概率潜在语义分析的生成模型和共现模型的参数个数是$O(M \cdot K + N \cdot K)$，其中 K 是话题数
- 现实中 $K \ll M$，所以概率潜在语义分析通过话题对数据进行了更简洁地表示，减少了学习过程中过拟合的可能性

概率潜在语义分析模型

- 下图展示了模型中文本、话题、单词之间的关系



概率潜在语义分析中文本、话题、单词之间的关系

概率潜在语义分析模型

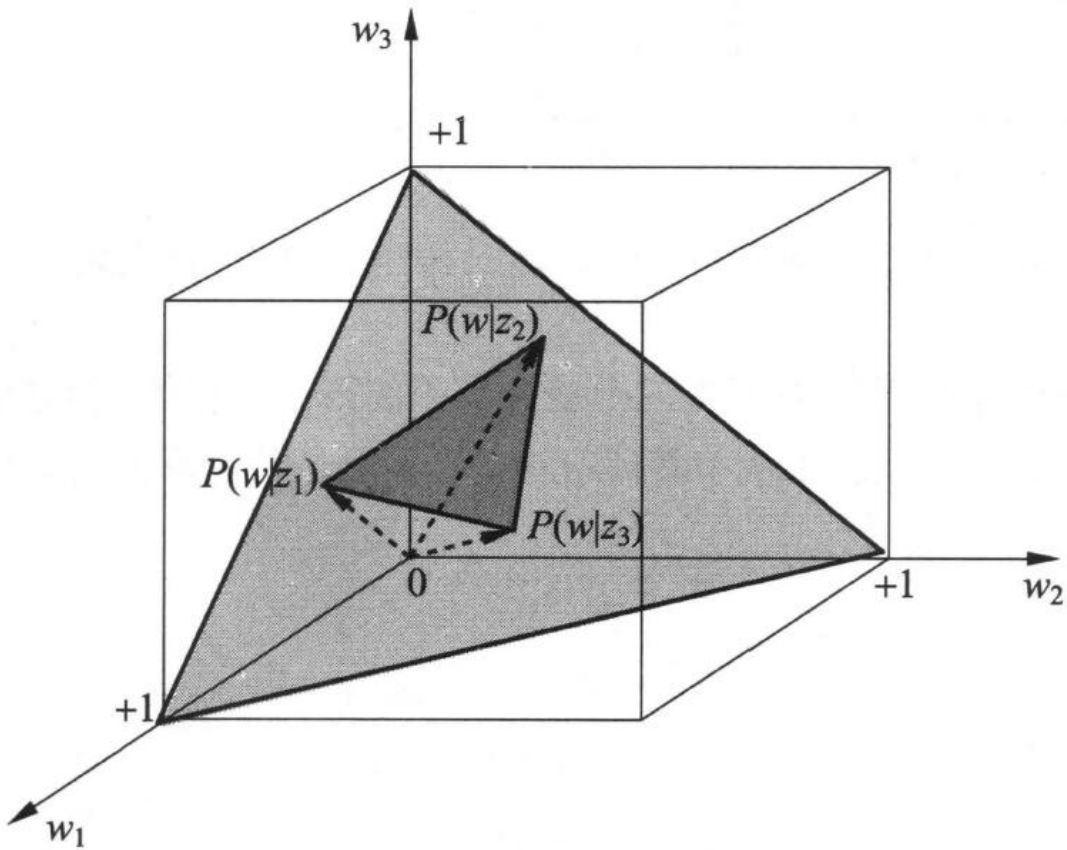
- **（2）几何解释**
- 概率分布$P(w \mid d)$表示文本d生成单词w的概率，

$$\sum_{i=1}^{M} P\left(w_i \mid d\right)=1, \quad 0 \leqslant P\left(w_i \mid d\right) \leqslant 1, \quad i=1, \cdots, M$$

可以由M维空间的$(M-1)$单纯形（simplex）中的点表示

概率潜在语义分析模型

- 单纯形上的每个点表示一个分布$P(w \mid d)$（分布的参数向量），所有的分布$P(w \mid d)$（分布的参数向量）都在单纯形上，称这个$(M-1)$单纯形为单词单纯形



概率潜在语义分析模型

- 概率潜在分析模型（生成模型）中的文本概率分布$P(w \mid d)$有下面的关系成立

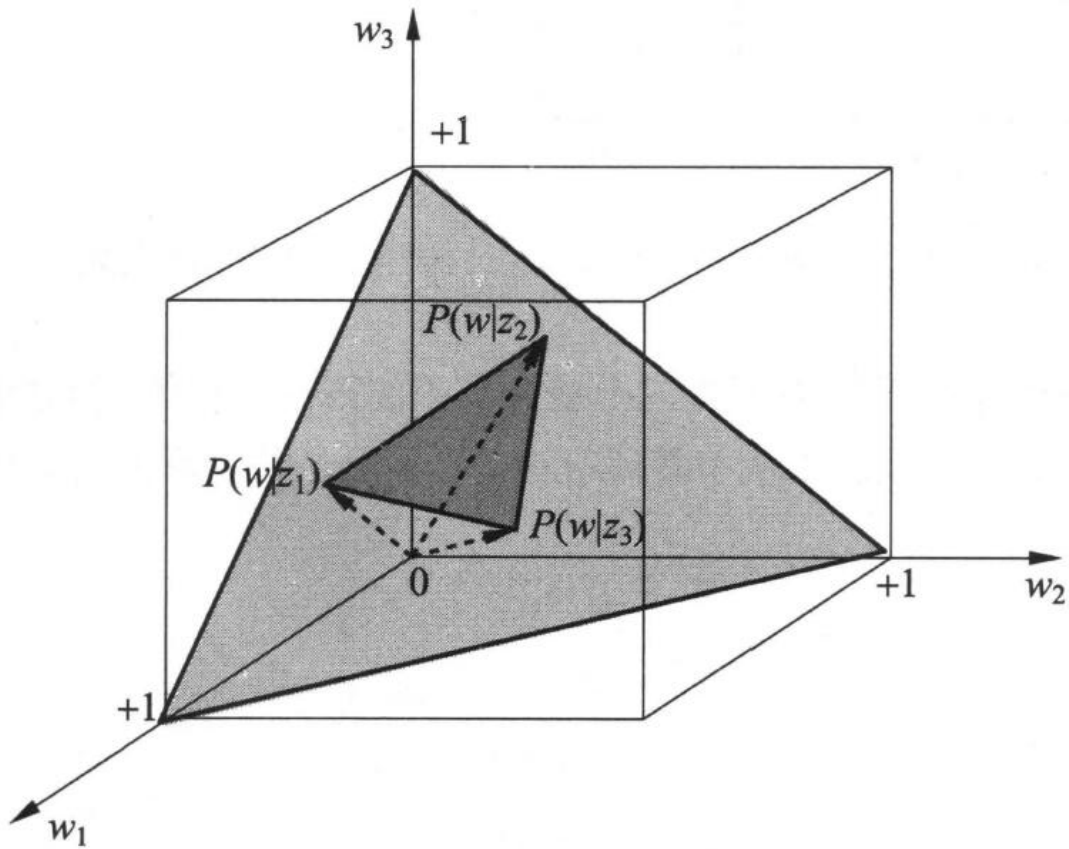
$$P(w \mid d) = \sum_{z} P(z \mid d) P(w \mid z)$$

这里概率分布$P(w \mid z)$表示话题z生成单词w的概率

- 概率分布$P(w \mid z)$也存在于M维空间中的$(M-1)$单纯形之中

概率潜在语义分析模型

- 如果有K个话题，那么就有K个概率分布$P\left(w \mid z_k\right), k=1,2, \cdots, K$，由$(M-1)$单纯形上的$K$个点表示
- 以这K个点为顶点，构成一个$(K-1)$单纯形，称为话题单纯形

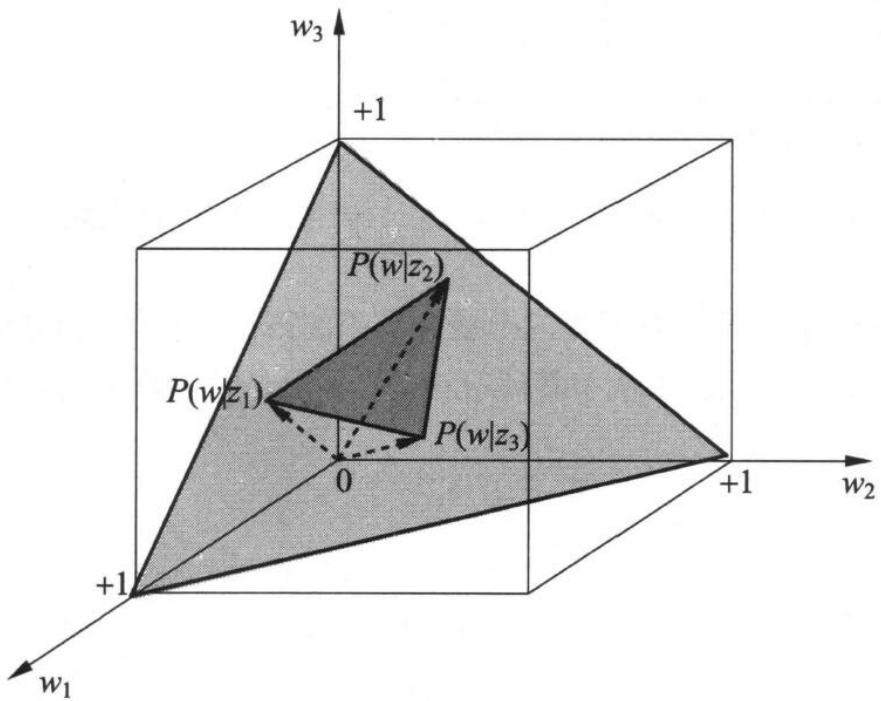


- 话题单纯形是单词单纯形的子单纯形

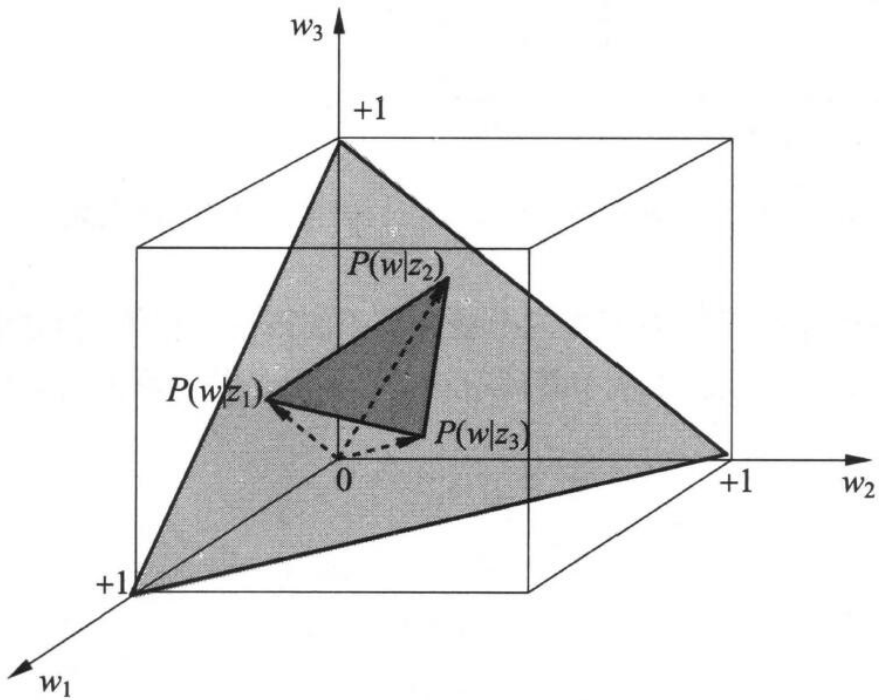
概率潜在语义分析模型

$$P(w \mid d) = \sum_{z} P(z \mid d) P(w \mid z)$$

- 由上式可知，生成模型中文本的分布$P(w \mid d)$可以由K个话题的分布$P\left(w \mid z_k\right), k = 1, \cdots, K$,的线性组合表示，文本对应的点就在$K$个话题的点构成的$(K-1)$话题单纯形中。这就是生成模型的几何解释



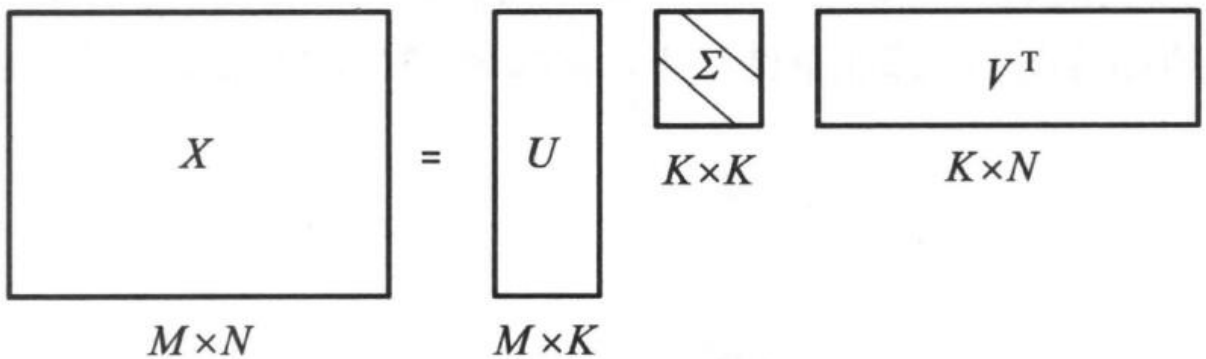
概率潜在语义分析模型



- 通常$K \ll M$，概率潜在语义模型存在于一个相对很小的参数空间中
- 上图显示的是$M = 3, K = 3$时的情况
- 当$K = 2$时话题单纯形是一个线段，当$K = 1$时话题单纯形是一个点

概率潜在语义分析模型

- **（3）与潜在语义分析的关系**
- 概率潜在语义分析模型（共现模型）可以在潜在语义分析模型的框架下描述
- 下图显示潜在语义分析，对单词-文本矩阵进行奇异值分解得到$\boldsymbol{X} = \boldsymbol{U}\Sigma\boldsymbol{V}^{\mathrm{T}}$，其中$\boldsymbol{U}$和$\boldsymbol{V}$为正交矩阵，$\Sigma$为非负降序对角矩阵



概率潜在语义分析模型

- 概率潜在语义分析模型的共现模型也可以表示为三个矩阵乘积的形式

$$
\begin{aligned}
\boldsymbol{X}' &= \boldsymbol{U}'\boldsymbol{\Sigma}'\boldsymbol{V}'^{\mathrm{T}} \\
\boldsymbol{X}' &= [P(w,d)]_{M\times N} \\
\boldsymbol{U}' &= [P(w \mid z)]_{M\times K} \\
\boldsymbol{\Sigma}' &= [P(z)]_{K\times K} \\
\boldsymbol{V}' &= [P(d \mid z)]_{N\times K}
\end{aligned}
$$

概率潜在语义分析模型

- **两者之间的对比：**
- 潜在语义分析模型中的矩阵$\boldsymbol{U}$和$\boldsymbol{V}$是正交的，未必非负，并不表示概率分布
- 概率潜在语义分析模型中的矩阵$\boldsymbol{U}'$和$\boldsymbol{V}'$是非负的、规范化的，表示条件概率分布

概率潜在语义分析模型的学习算法

- 概率潜在语义分析模型是含有隐变量的模型，其学习通常使用EM算法
- EM算法是一种迭代算法，每次迭代包括交替的两步：E步，求期望；M步，求极大
- E步是计算Q函数，即完全数据的对数似然函数对不完全数据的条件分布的期望
- M步是对Q函数极大化，更新模型参数

概率潜在语义分析模型的学习算法

- 下面叙述生成模型的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
  $T = \{n(w_i, d_j) | i = 1, ..., M,\ j = 1, ..., N\}$

概率潜在语义分析模型的学习算法

- 每个单词-文本对的生成概率为

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

- 模型参数
  - 各文本的概率分布
    $P(d_j), \quad j = 1, ..., N$
  - 各话题下单词的概率分布
    $P(w_i|z_k), \quad i = 1, ..., M, \quad k = 1, ..., K$
  - 各文本中话题的概率分布
    $P(z_k|d_j), \quad k = 1, ..., K, \quad j = 1, ..., N$

概率潜在语义分析模型的学习算法

- 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

$$\begin{aligned} LL &= \sum_{i=1}^{M} \sum_{j=1}^{N} n(w_i, d_j) \log P(w_i, d_j) \\ &= \sum_{i=1}^{M} \sum_{j=1}^{N} n(w_i, d_j) \log \left(P(d_j) \sum_{k=1}^{K} P(w_i, z_k | d_j) \right) \\ &= \sum_{i=1}^{M} \sum_{j=1}^{N} n(w_i, d_j) \left(\log P(d_j) + \log \sum_{k=1}^{K} P(w_i, z_k | d_j) \right) \\ &= \sum_{i=1}^{M} \sum_{j=1}^{N} n(w_i, d_j) \log P(d_j) \\ &\quad + \sum_{i=1}^{M} \sum_{j=1}^{N} n(w_i, d_j) \log \sum_{k=1}^{K} P(w_i, z_k | d_j) \end{aligned}$$

概率潜在语义分析模型的学习算法

- 构造Q函数

$$
\begin{aligned}
LL &= \sum_{i=1}^{M}\sum_{j=1}^{N} n(w_i, d_j) \log P(d_j) \\
&\quad + \sum_{i=1}^{M}\sum_{j=1}^{N} n(w_i, d_j) \log \sum_{k=1}^{K} P(w_i, z_k | d_j) \\
&\geqslant \sum_{i=1}^{M}\sum_{j=1}^{N} n(w_i, d_j) \log P(d_j) \\
&\quad + \sum_{i=1}^{M}\sum_{j=1}^{N} n(w_i, d_j) \sum_{k=1}^{K} P(z_k | w_i, d_j) \log P(w_i, z_k | d_j) \\
&= Q
\end{aligned}
$$

概率潜在语义分析模型的学习算法

- 可以从数据中直接统计得出各文本的概率分布

$$P(d_j) = \frac{n(d_j)}{\sum_{l=1}^{N} n(d_l)}, \quad j = 1, ..., N$$

- 实际上，我们并不关心文本的概率分布 $P(d_j), \quad j = 1, ..., N$，只关心

$$P(w_i, z_k|d_j) = P(w_i|z_k) P(z_k|d_j)$$

- 因此，Q函数可以简化为

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

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

概率潜在语义分析模型的学习算法

- 观察

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

其中

$$\log P(w_i, z_k|d_j) = \log \left(P(w_i|z_k) P(z_k|d_j)\right)$$

是完全数据的对数似然函数，而

$$P(z_k|w_i, d_j)$$

是不完全数据的条件分布（给定参数的当前估计值隐变量的后验分布）。因此，Q函数是完全数据的对数似然函数对不完全数据的条件分布的期望

概率潜在语义分析模型的学习算法

- 在EM算法的E步中，需要根据参数的当前估计值计算出隐变量的后验分布

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

概率潜在语义分析模型的学习算法

- 在EM算法的M步中，目的是极大化Q函数，因此通过约束最优化求解 Q' 函数的极大值，这时 $P(z_k \mid d_j)$ 和 $P(w_i \mid z_k)$ 是变量。因为变量 $P(w_i \mid z_k), P(z_k \mid d_j)$ 形成概率分布，满足约束条件

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

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

概率潜在语义分析模型的学习算法

- 应用拉格朗日法，引入拉格朗日乘子 τ_k 和 ρ_j，定义拉格朗日函数 Λ

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

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

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

概率潜在语义分析模型的学习算法

- 将拉格朗日函数 Λ 对 $P(w_i \mid z_k)$ 求偏导，并令其等于 0，得到

$$\frac{\partial \Lambda}{\partial P(w_i|z_k)} = \frac{1}{P(w_i|z_k)} \sum_{j=1}^{N} n(w_i, d_j) P(z_k|w_i, d_j) - \tau_k = 0, \quad i = 1, ..., M, \quad k = 1, ..., K$$

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

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

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

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

概率潜在语义分析模型的学习算法

- 将拉格朗日函数 Λ 对 $P\left(z_k \mid d_j\right)$ 求偏导，并令其等于 0，得到

$$\frac{\partial \Lambda}{\partial P\left(z_k|d_j\right)} = \frac{1}{P\left(z_k|d_j\right)} \sum_{i=1}^{M} n\left(w_i, d_j\right) P\left(z_k|w_i, d_j\right) - \rho_j = 0, \quad j = 1, ..., N, \quad k = 1, ..., K$$

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

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

$$= \frac{\sum_{i=1}^{M} n\left(w_i, d_j\right) \sum_{k=1}^{K} P\left(z_k|w_i, d_j\right)}{\rho_j} = \frac{n\left(d_j\right)}{\rho_j}$$

$$\Rightarrow \rho_j = n\left(d_j\right)$$

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

概率潜在语义分析模型的学习算法

概率潜在语义模型参数估计的 EM 算法

输入：设单词集合为 $W=\{w_1,w_2,\cdots,w_M\}$，文本集合为 $D=\{d_1,d_2,\cdots,d_N\}$，话题集合为 $Z=\{z_1,z_2,\cdots,z_K\}$，共现数据 $\{n(w_i,d_j)\},i=1,2,\cdots,M,j=1,2,\cdots,N$；

输出：$P(w_i|z_k)$ 和 $P(z_k|d_j)$。

（1）设置参数 $P(w_i|z_k)$ 和 $P(z_k|d_j)$ 的初始值。

（2）迭代执行以下 E 步，M 步，直到收敛为止。

E 步：

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

M 步：

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

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