**第十八章读书报告**

09118223 吴亦珂

读书进度：读完18章

1. **问题列表**

（我提出）概率潜在语义分析模型的共现模型与17章中SVD算法一个很重要的区别是不是就是概率潜在语义分析模型的值是非负的，因而可以表示概率分布而SVD不能表示概率分布因为可能出现负值。

讨论结果：区别应该不仅仅局限于此，因为无论是SVD还是非负矩阵分解，都是伪概率分布，并不是真正意义上的概率。而概率潜在语义分析模型则假设话题、文本、单词之间具有生成关系。因此生成的就是一个概率分布。

（我提出）概率潜在语义分析模型的共现模型与17章中非负矩阵分解有什么关系？非负矩阵分解得到结果均非负，是不是能用这种方法求解共现模型的参数？

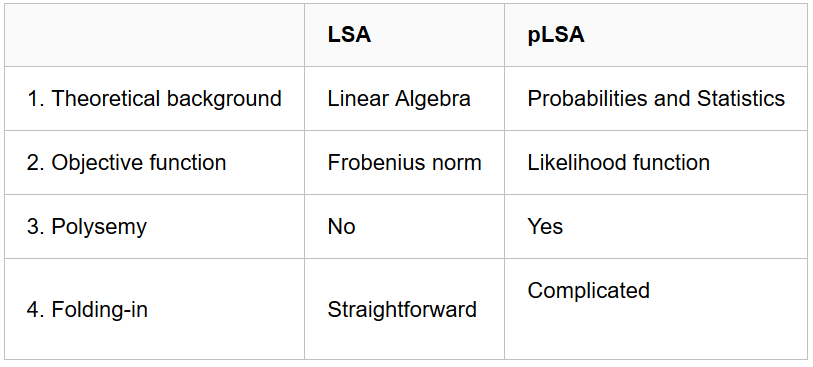
讨论结果：这两者都具有较好的可解释性，但是非负矩阵分解求解的应该是一种伪概率分布，并不是真正的概率，而共现模型则是假设了由话题生成文本以及单词，因此共现模型求出的是在一个假设条件下的概率分布。

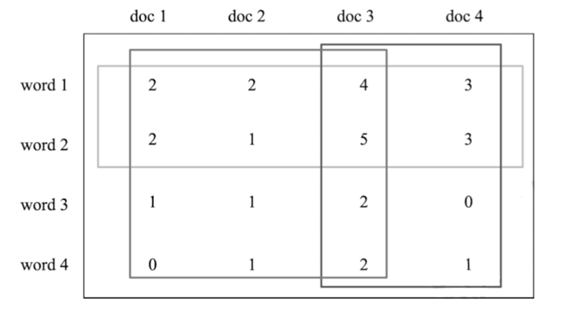
（别人提出）如何理解生成模型和共现模型等价性下的差异，两种模型的学习算法思路有何不同？

讨论结果：等价性我觉得主要是因为两者的结果是一样的，只不过是由不同的条件概率组成。这就导致生成模型是文本到话题到单词。而贡献模型是由话题到单词，话题到文本。学习思路应该就是导致对应的生成关系不同。

（别人提出）PLSA和LSA的相同点和区别分别是什么？两种算法哪一种使用的更加广泛。

讨论结果：相同点我觉得都是一种对于文本话题的选择判断。求出了话题的分布关系。但是感觉LSA是直接通过矩阵的分解求得，但是PLSA则是通过一个概率模型求出所需要的的概率分布。具体区别可以用下面这个表格呈现：

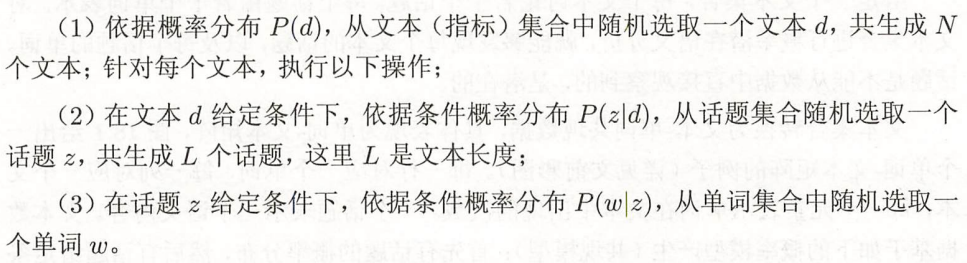


（别人提出）为什么每个话题也可以由若干个文本表示？  
  
讨论结果：不同的文本可以是属于同一个话题的，比如都属于体育这个话题，也就是对于不同文本可以归属到几个话题之中。

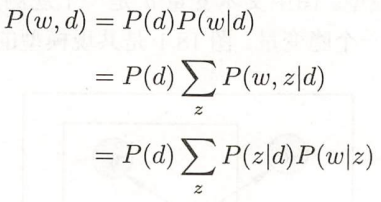
1. 读书收获

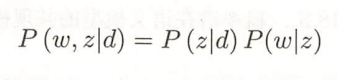
18.1 ：介绍了概率潜在语义分析的两种模型：生成模型和共现模型。

生成模型：生成模型主要是基于以下步骤生成共现数据：

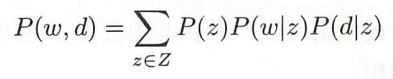


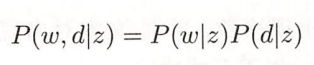
因此该模型的基本想法是由文本到话题到单词。所以用概率表示：





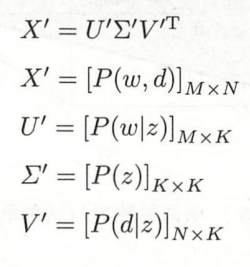
共现模型：共现模型是对概率式子进行了适当地修正，从话题生成文本和单词的角度构建模型，用概率表示：





几何解释：文本生成单词的概率P（w|d）可用M维空间的（M-1）单纯形中的点表示。话题生成单词的概率P（w|z）也存在于M维空间的（M-1）单纯形中。话题生成单词的概率分布P（w|d）在K个话题的点构成的（k-1）话题单纯性中。

与潜在语义分析的关系：可以将PLSA的各种概率分布于LSA矩阵奇异值分解的结果进行一个对照：



18.2：介绍了概率潜在语义分析的算法（生成模型）

考虑到话题是一个隐变量，因此采用的是EM算法。具体步骤如下：

