读书报告

09118119 黄一凡

# 自己提出的问题

#### 1.共现数据模型的学习算法是否只需在生成模型的EM算法中做一些修改即可，比如概率式子上的不同。

我认为是这样的，考虑到生成模型和共现模型在概率公式意义上是等价的，只是其模型的推导形成过程不同。因此可以想到可以对生成模型的概率表达进行相应的修改，即可实现共现模型的学习算法。

# 别人提出的问题

#### 1.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。

SVD和非负矩阵分解实际上都不能很好的表示概率分布，因为它们本质上并不是概率模型，而是用分解和加权的思想来找一个合理的矩阵组合。而PLSA一个很重要的地方在于它是基于一个很强的假设， 也就是话题，文本，单词的生成关系，有了这个关系，才能在其限定下用EM算法估计概率关系产生概率模型，而这个假设不一定合理，可能还有更复杂的关系， 而SVD等是不受任何假设制约的，所以它们应该区别还是比较明显的.

#### 2.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。

相同点为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。

#### 3.如何理解概率潜在语义分析和潜在语义分析的关系，感觉潜在语义分析像是提出一个大的方法，而概率潜在语义分析像是提供一种具体的解法。

它们是基于不同的思路来进行语义分析，分析的基础都是针对单词文本矩阵来进行，一个用的是非概率的思路，也就是分解， 另一个是用概率和假设的思路进行分析，所以应该是从两个不同的角度来分析。

#### 4.相比于SVD等分解算法，基于概率估计的EM算法的优势是什么？

EM算法是通过迭代以及概率估计得到的，而不论是SVD还是非负矩阵分解，都只是通过数学方法对矩阵进行压缩，并不能完全表示概率分布。

# 读书计划

#### 本周所读：

《统计学习方法》18章

#### 下周计划：

《统计学习方法》19章