统计学习读书报告

09017244 郑健雄

1. 自己提出的问题的理解：
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。

讨论后的理解：

两种算法的相同点在于都是根据文本和单词来估计话题。不同之处在于LSA基于线性代数，比较直截了当，也没有概率假设的限制，所以并不是概率模型。PLSA是基于概率假设的模型，目标函数是似然函数，求解的过程比较复杂，使用了EM算法，是典型的概率模型。对于一些没有背景知识的应用，可能LSA更加方便一点，而PLSA是基于概率的，有着更好的解释性，所以如果有足够的背景知识，可以用PLSA。

1. 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？

讨论后的理解：

从概率角度来看，这两种模型的意义是等价的，只不过是同一种关系的两个观察角度。生成模型的思路是生成单词的过程，是一个线性的过程，而共现模型是假设话题生成了文本和单词，也就是出发点不同。但不论如何，话题是隐变量，都需要使用EM算法来估计，只是概率公式有区别。

1. 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？

讨论后的理解：

SVD模型直截了当，是用代数分解的思想来分析单词文本的关系，使用了加权和的思路。其优势是直接并且不需要思考关系，但是缺点很明显，没有可解释性，也没有融入人的认知，只是一种纯数学的处理，分解出相应的主成分。而EM算法的优势在于其是确实产生了一个概率模型，并且这个模型是基于人们对语言生成的认知来产生的，是基于三种元素的相互关系的，那么首先，产生了一个概率模型，其能够被人们解释，此外，该模型融入了人们的看法和假设，产生的结果更有意义。

1. 别人提出的问题的理解
2. 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。

自己的理解:

我的理解是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，尽管这个假设并不一定完美并且可以改进，但其考虑了人的认知，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。

1. 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贡献模型的参数？

自己的理解：

非负矩阵分解本质上还是代数分解，是没有考虑概率关系的。plsa和非负矩阵分解应该都有比较好的解释性，但其区别在于其模型是基于概率还是非概率以及是不是受到概率假设的限制，非负矩阵比较随机，并且不唯一，没有相互的制约关系，所以其不一定和概率模型对等。

1. 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。

自己的理解:

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，所以应该是两个角度，而不是包含的关系，虽然其分析基础一致，并且分析过程也有很多相同点。

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。

自己的理解：

生成和共现是等价的，其区别在于概率等式表达的方式不太一样，而分析等式的思路是一致的，所以遵循一样的思路进行即可。EM算法的流程是一样的，将共现模型的概率公式带入推导即可。

1. 读书计划

1、本周完成的内容章节：复习18章，阅读完Speech and Language Processing第四章。

2、下周计划：复习19章，阅读部分第五章内容。

四、读书摘要及理解

1. 概率潜在语义分析：概率潜在语义分析受到LSA的思想启发，是从概率角度对文本进行分析。其基于话题是潜在隐变量的基本假设，对文本进行分析。其分析的基础仍然是文本单词矩阵。

2. 生成模型：该模型认为文本生成话题，话题生成单词，以极大似然函数为目标函数，话题要预先设定。

3. 共现模型：该模型认为话题是隐变量，而单词和文本在话题给定的情况下是条件独立的。

4. 模型比较：两个模型在数学上等价，但是刻画了不同方面，一个是生成，一个是共现数据遵循的模式，前者是非对称的，后者是对称的。基于此，两者学习算法的形式有所区别。其学习都用到了EM算法，来估计局部最优解。

5. 与LSA的关系：两者结构实际上很类似，PLSA最后也是三个矩阵的乘积，只不过估计矩阵参数用到了基本假设和EM算法，是一种概率角度的思路。而LSA则直接进行了分解。区别在于LSA的矩阵是正交的，并不是概率分布，而PLSA是非负规范化的，有更好的解释性。

思考：

机器学习算法在很多情况下都可以分为概率和非概率两个角度，实际上这也是对一个问题思考的两个角度。比如像SVM和PCA这样的方法，结果是很确定的，而概率分析比如PLSA和EM算法，其结果只是遵循规律。这些方法的好坏并不能简单评价，但是从不同角度思考问题是很有意义的。