统计学习读书报告

09017244 郑健雄

1. 自己提出的问题的理解：
2. LDA与潜在语义分析和概率潜在语义分析之间有什么区别和联系？

讨论后的理解：

LDA和PLSA以及LSA的思路是很像的，都是引入了话题这个中间的内容，而LDA和PLSA的流程也非常类似，都是通过文本生成话题，话题生成单词，而区别在于LDA假设了一个先验分布，其价值在于可以更有效的限制生成概率公式的形式，从而有更好的效果，防止过拟合。

1. LDA中，狄利克雷分布的重要性是什么？能否使用其他分布来充当先验分布？

讨论后的理解：

狄利克雷分布可以用来拟合多项分布，并且其拟合出的先验分布和后验分布都是狄利克雷分布，也就是共轭分布。共轭分布的性质较好，两者之间也存在推导关系，方便产生公式，再加上有着较好的拟合效果，所以狄利克雷分布的重要性不可忽视。而其他分布可能并没有这么好的性质。

1. 吉布斯抽样和变分EM推理分别适合在什么情况使用？

讨论后的理解：

变分EM算法比较复杂，但是收敛速度较快，同时准确性也有保证，对于大规模比较好。而吉布斯抽样是基于蒙特卡洛模拟的，其优势是实现简单，缺陷是迭代次数过多，可能会慢一些，适合小规模数据量且精度要求不高的情况。

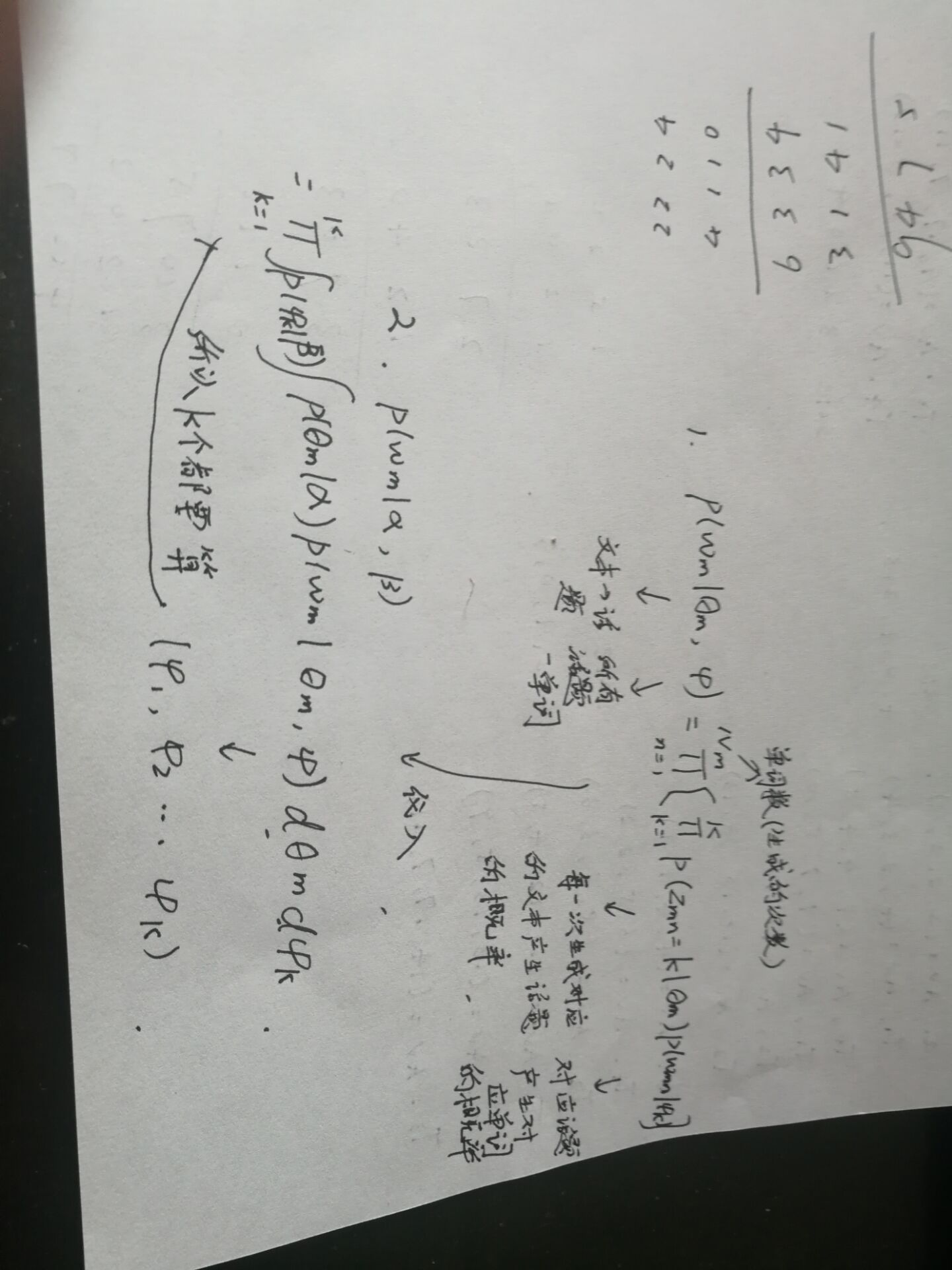
1. 别人提出的问题的理解
2. P391页，为什么说PLSA使用的先验分布是均匀分布？从哪个地方可以看出LDA是基于贝叶斯的学习？

自己的理解：

我的理解是这里说的先验分布是均匀分布可能是说PLSA的概率分布并没有固定的形式或者说不是根据某个基本分布产生的，它们的最终结果是通过学习来得到的。LDA是贝叶斯学习可能主要是因为其涉及的大部分概率公式全部用到了狄利克雷分布以及其性质，同时引入了超参数来计算先验分布和后验分布，而狄利克雷分布本身的性质和贝叶斯公式有着紧密的联系，比如20.11，20.12的推导

1. P395（20.17）和（20.18）是如何推导的？

自己的理解：



1. 怎么理解'狄利克雷分布θ存在于(k-1)维单纯形上'?

自己的理解：

狄利克雷分布的参数有着和为1的限制，这就导致会降一维，也就是k-1单纯形。

1. 为什么LDA的狄利克雷先验比起PLSA的可以达到更好的效果？

自己的理解：

我的理解是PLSA主要是没有一个固定的先验分布，所以其产生的概率分布的形式是完全依赖数据的，而LDA引入了迪利克雷分布作为基础的先验分布，而迪利克雷分布又是一个非常合理的先验分布，其优势类似书上说的，可以更好的防止过拟合，提升学习的效果。

1. 狄利克雷分布的参数a代表什么，先验分布的a怎么选取？

自己的理解：

这个a是超参数，可以看一下书上390的讲解，a可以理解为引入的一种不存在的假设计数，类似于在没有实验数据的情况下的一种假设，其选择很多都是根据实际应用的经验来确定的。

1. 共轭分布与共轭函数的共轭是否代表同样的意思？

自己的理解：

我不太确定是否有共轭函数的概念，不过这里的共轭分布应该和共轭复数不太一样，共轭复数是有规律的一对复数，而共轭分布似乎是说先验分布和后验分布的分布类别一致即可，比如说都是服从狄利克雷分布的，所以可能不是特别相似的概念

1. 读书计划

1、本周完成的内容章节：复习20章。

2、下周计划：复习21章，阅读speech and processing第六章内容。

四、读书摘要及理解

1. 狄利克雷分布：

狄利克雷分布是一种多元来纳许随机变量的概率分布，其特例是贝塔分布。其可以作为多项分布的共轭先验(先验后验的类别一样)，而贝塔分布可以作为二项分布的先验分布。狄利克雷分布此时就可以作为一个先验概率帮助分析和生成实际的多项分布，从而让学习更具理论性。狄利克雷分布的参数可以理解为一种伪计数，一种假设，通常根据经验选定，来提升学习的效果。

2. LSA模型：

与PLSA类似，也是文本生成话题，话题生成单词，其不同点是使用了狄利克雷分布作为先验分布。不同于PLSA,这一次每个话题生成某个单词或者某个文本生成某个话题的概率都服从狄利克雷分布及其背后的超参数，然后在此基础上学习，其特征是加入了生成概率分布的先验分布的步骤，这与PLSA不同。

3. 吉布斯抽样与变分EM算法：

LDA模型的学习和推理是不能直接求解的，其通常使用吉布斯抽样和变分推理。吉布斯抽样通过模拟和积分运算来进行模型的求解，估算出相应的概率。变分EM算法的思路基于EM算法，其通过寻找近似概率分布并且KL散度最小化来将问题转换为证据下界最大化问题，通过交替进行E步和M步，来分别估计变分参数和模型参数，从而得出最终结果。

思考：

LSA是PLSA之后的一个很有价值的突破，其突破了以往的思路，从数学上引入了狄利克雷分布，同时借用了PLSA的思路，提升了文本分析的效果。其缺点在于求解过程繁琐，无论是吉布斯抽样还是变分EM算法都是比较复杂或者比较费时间的。LSA作为一种全局算法，可以将全局隐藏的模式挖掘出来，同时有着比PLSA更好的效果，虽然计算量和效率都有所限制，但有着广泛的应用场景，特别是挖掘文本主题等领域。