读书报告 7.6-7.12

61518426 周之遥

·读书进度

学习统计学习方法第20章：潜在狄利克雷分布。

LDA是一个生成概率模型。使用BOW模型，假设一个话题中单词呈多项分布，一个文本中话题呈多项分布。文本的单词序列即通过生成话题、生成单词的步骤生成。

而话题中单词的多项分布、文本中话题的多项分布的参数则由假设的先验分布——狄利克雷分布生成，其参数为超参数。

LDA的学习有吉布斯抽样和变分EM算法。

·提出问题

1. 采用狄利克雷分布作为均匀分布的先验分布有什么好处？

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. LDA和PLSA有什么区别和联系？

两者对文本、话题、单词的关系有相同的假设，即两个多项分布。不同点在于PLSA没有对先验分布做出假设，而LDA假设先验分布为狄利克雷分布。

·解决问题

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？  
这里a是超参数，事先给定，描述了狄利克雷分布

2.怎么理解'狄利克雷分布θ存在于(k-1)维单纯形上?

狄利克雷分布中对随机变量有和为1的约束（见P386），因此其概率密度函数的定义域可以看做一个单纯形

·下周计划

完成21章PageRank算法。