统计学习方法读书笔记

61518122　丁自民

读书内容

第 18 章

读书收获

1、概率潜在语义分析，也称概率潜在语意索引。是一种利用概率生成模型对文本集合进行话题划分的无监督学习方法。

2、该模型用隐变量表示话题。

3、概率潜在语义分析模型

4、生成模型与共现模型

生成模型：主要用联合概率密度来进行分析。一个文本的内容由其话题决定，一个文本的话题由其单词决定。

其实还是不知道而这什么区别，姑且理解为因果不同吧。

5、直接建立单词和文本之间的关系，则复杂度较大。中间引入话题，整合一下单词，算法复杂度就会下降很多。

6、单纯形：k 维单纯形是指包含 k+1 个节点的凸多面体。

1 维单纯形是线段，2 维单纯形是三角形，3 维单纯形是四面体。

7、所有条件概率簇都可以表示成单纯形

8、用 EM 算法求解

读书疑问

P343 为什么这个概率可以由单纯形表示?

疑问解答

回答郑健雄问题 2:

殊途同归。两种算法连优化的目标函数都一样，说明应该大差不差。生成模型是“单词-话题-文章”层层递进，而共现模型是“话题-单词+文章”，认为话题是根本。两种理解应该都有道理，但是这种差异没有体现在优化上，说明本质上差不多。

回答殷春锁问题 1:

单纯形是可以张成全空间的图形的推广。

1 维单纯形是线段，2 维单纯形是三角形，3 维单纯形是四面体。

他们都是可以张成本空间的图形。

回答吴亦珂问题 3:

不完全数据一般包括四类数据：截断数据(truncated data)，删失数据(censored data)，既截断又删失数据(truncated and censored data)以及缺失数据(missing data).

截断数据(truncated data). 在现实生活中，一般常见的是左截断数据(left-truncated data). 偶尔，也会存在右截断数据(right-truncated data).

指在采样时，刻意避开某些特征得到的数据。详见知乎。