统计学习读书报告

09017244 郑健雄

1. 自己提出的问题的理解：
2. 求解单词话题矩阵可以使用单词文本矩阵的奇异值分解，除此之外，是否可以使用其他类型的方法直接构造单词话题矩阵关系而不使用分解的方法？

讨论后的理解：

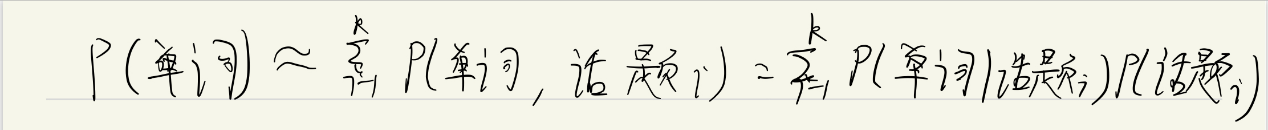
其实从概率层面来思考，概率潜在语义分析和LDA都用到了概率模型，概率潜在语义分析基于潜在语义分析，使用了条件概率的思路和生成模型，但是其也使用了分解的方法。除此之外还有LDA方法，是使用贝叶斯思想的一种提取主题的方法，效果比较好。初次之外，我的理解是潜在语义分析的分解其实并不是特别直观，是比较抽象的一种分解，而如果可以人工给出话题，并且有足够的数据，那么深度学习其实也是一种很好的选择，其产生的单词-话题向量可能更贴近于人们自己的理解，有更好的解释性。

1. 潜在语义分析与主成分分析之间存在什么关系？研究思路是否有差异？

讨论后的理解：

主成分分析与潜在语义分析的相似性在于其都是使用了降维的思想，从而让数据的结构得以呈现。我自己的理解是可以将潜在语义分析作为主成分分析的一种应用，就是将语义的概念从单词上升到话题这样的降维后的结构，从而更好的表示出语义特征，这和PCA的思路非常类似，也是保留重要信息的特征。PCA和LSA属于思想层面上的产物，SVM是这两种方法的一种实现手段。

1. 别人提出的问题的理解
2. P326页上说X约等于TY，是否可以这么理解



自己的理解:

其实这个公式的思想感觉更类似于概率潜在语义分析，其思路就是生成叠加式的一种处理方式，虽然细节部分不太一样，而潜在语义分析使用的是非概率的思路，所以这样的解释可能直观上讲和潜在语义分析不太匹配。不过其实其思想是和潜在语义分析一致的，所以这个理解是很合理的，不过潜在语义分析的层面可能不是到了概率，只是一种词频上的表示。

1. 为什么通过奇异值分解的方法就能够直接将左奇异矩阵作为话题矩阵，后面两项作为文本在话题空间的表示？这么做有什么根据吗？

自己的理解：

潜在语义分析的目标其实就是把一个矩阵变成两个矩阵的乘积，SVD其实就是产生了相应形式的两个矩阵，但是这样有一个缺点， svd产生的话题向量感觉比较抽象，是一串数字，不一定可以和现实中的话题对应，所以感觉语义上没什么依据。但是如果不去考虑这个话题向量语义的意思，只是想找符合潜在语义分析的两个相应的矩阵，这种处理是可行的，虽然很难有直观的理解。它本质上还是类似主成分分析，其实话题应该有很多，它只保存了最重要的k个话题，而这些话题实际上表征了数据内部的组织结构，但是其是否符合人的认知就不一定了。

1. 觉得这一章节所谓单词和话题上的相似某种意义上来说就是与文本其它部分的相对关系相似，可以做这样的理解吗？

自己的理解:

这样理解是合理的。其实衡量相似性的指标是内积，那么两个文本单词分布差不多，其结构其实也差不多，也就是很相似，而除此之外话题是更深入挖掘的文本结构，其相似性也更能说明文本内容和结构上的相似性。但是这样的处理也有一个问题，也就是文本的表意除了所谓的一词多义性和多词一义性外，文本的组织结构和行文顺序其实并不能被观察到，所以严格来说，对应部分的比较是很难实现的，这也是文本分析的一个问题，只能说明它们是类似的文章，但是其主题或者思想这样却无法挖掘。

1. 为什么奇异值分解出来的单词-话题矩阵选择的单词表示同一个话题？

自己的理解：

其实SVD的问题就在于它得出来的是计算的结果，但是不一定很好解释。它产生的话题本身也是一个抽象的表示，只是在矩阵分解的过程中这样的组合占比重很大，具有重要性，所以就作为了话题，话题的单词组合有各种各样的，最后呈现的一般是最重要的几个话题。

1. 怎么选择话题数，即k的值？

自己的理解：

可以根据经验，也可以使用主成分分析的量化指标来分析。

1. 读书计划

1、本周完成的内容章节：复习17章并阅读Speech and Language Processing第三章内容上半部分。

2、下周计划：准备复习18章并阅读完第三章内容。

四、读书摘要及理解

1. 潜在语义分析：潜在语义分析的目的是挖掘文本的内部结构，从而表征出文本的更加抽象的语义，其是一种无监督学习方法，主要思路是降维来暴露内部的结构，从而使得文本的语义被挖掘表示出来。其思路是通过挖掘单词和文本的关系，抽取其中的话题，和PCA很类似。

2. 潜在语义分析的思想：其最终的目的是找出一种表示方法，使得X=TY，X是单词文本矩阵，T是单词话题矩阵，Y是话题文本矩阵，TY中间的话题向量矩阵应该是由于正交的性质被消去了。最初的X是一种比较粗糙的分析方法，其基于词频和独特性，使用单词来对文本进行表示，由于单词众多，实际上分析效果不是很好。为了解决这种问题，认为文本其实具有多个话题，而话题是可以用单词表示的，话题又是文本的子结构，通过这个中间媒介的存在，可以更好的对文本进行近似和分析。话题的数目一般远远小于文本的数目，更能表现出文本的共性，也往往更有意义。

3. 求解的方法：方法有SVD和非负矩阵分解。一般直接得到的只有单词文本矩阵，可以用分解的方法挖掘其中的重要成分作为话题来分析，其思路类似于主成分分析。

思考：

LDA是一种比较好的挖掘方法，在使用LDA之后可以使用层次聚类或者均值聚类的方法对话题进行进一步分析，从而找出其层次结构，方便更好的分析。但是该模型也有着严重的局限性，其完全基于数据，也没有使用概率分布的思想，并且没有考虑文本本身的组织结构等重要信息，所以其分析结果有很大的局限性，后续的LDA分解方法以及很多深度学习网络可以更好地进行这样的处理。