087 | 基础文本分析模型之二：概率隐语义分析

2018-04-23 洪亮劼

讲述：初明明
时长 04:34　大小 2.10M

在上一篇的分享里，我们展开了文本分析这个方向，讨论了“隐语义分析”（Latent Semantic Indexing）这个模型。隐语义分析的核心是基于矩阵分解的代数方法。这种方法的好处自然是能够直接利用代数计算方法对文本进行分析，而短板则是无法很好地解释结果。而“解释性”是很多概率模型的一大优势，因此，自然就有很多研究者想到是否能够把概率的语言移植到隐语义分析上。

今天，我们就来分享“**概率隐语义分析**”（Probabilistic Latent Semantic Indexing）的一些基本内容。概率隐语义分析有时候又被简称为 **PLSA**（Probability Latent Semantic Analysis）。

隐语义分析核心思想

上周我们介绍过隐语义分析的核心思想，首先来简要回顾一下。

隐语义分析的核心其实就是用无监督的方法从文本中提取特性，而这些特性可能会对原来文本的深层关系有着更好的解释。

简单来说，隐语义分析就是利用了“矩阵分解”的概念，从而对“词 - 文档矩阵”（Term-Document Matrix）进行分解。

概率隐语义分析

既然概率隐语义分析是利用概率的语言，那么我们就来看看概率隐语义分析是如何对文档进行建模的。

首先，**PLSA 是对文档和里面单词的联合分布进行建模。**这个文档和单词的联合分布其实就是类似隐语义分析中的那个文档和单词的矩阵。只不过，在 PLSA 里，我们不是直接对数据进行建模，而是认为数据是从某个分布中产生的结果。那么，对于这个联合分布该如何建模呢？

一种方法就是对这个联合分布**直接进行建模**，但是这往往会遇到数据不足的情况，或者无法找到一个合适的已知参数的分布来直接描述我们需要建模的这个联合分布。另外一种经常使用的方法就是**简化这个联合分布**，从而找到我们可以建模的形式。

那么，如何简化联合分布呢？一种方法就是**把一个联合分布进行分解。**

一种分解分布的方法就是假定一些隐含的变量，然后数据又是从这些隐含变量中产生而来。在我们现在的情况里，我们从文档和单词的联合分布入手，可以做出这样的假设：这个文档和单词的联合分布是，我们首先从文档出发，产生了当前所需要的主题（比如金融、运动等），然后再从主题出发，产生相应的单词。很明显，这里的主题是我们并不知道的隐含变量，是需要我们从数据中估计出来的。这就是 PLSA 模型的基本假设。

PLSA 还有一个**等价的模型描述**，也是对文档单词联合分布的另外一种分解，那就是，我们首先假设有一个主题的先验概率，然后根据这个主题的分布，产生了一个文档，同时，也产生了这个文档里面的所有单词。这种假设观点非常类似我们之前在介绍高级的主题模型时谈到的“下游方法”（Down-Stream）。这里，文档变量和单词变量都成为了隐变量，也就是主题变量的下游变量。

通过一定的代数变形，我们可以得到这两种方法其实就是等价的。

如果我们按照第一种分解方法来认识文档单词分布，有一种更加通俗的解释：我们其实是给每一个单词都联系了一个未知的主题变量，这个主题变量是从一个文档级别的主题分布得来的，实际上，这是一个多项分布（Multinomial Distribution）；然后，根据这个主题变量，我们又从相应的一个语言模型中，抽取出了当前的单词，这又是另外的一个多项分布。如果从这个角度来看待这个模型，你会发现，**PLSA 其实和 LDA 非常相似。**

实际上，从模型的根本特征上来说，PLSA 和 LDA 都是对文档单词分布的一种分解，或者叫作产生解释。只不过，LDA 针对刚才我们所说的两个多项分布，一个是每个文档的主题分布，另外一个是 K 个语言模型，都外加了先验分布，使得整个模型更加符合贝叶斯统计的观点。然而**在建模的核心思想上，PLSA 和 LDA 是一样的。**

关于如何学习 PLSA 这样的隐变量模型，我将会在后面的分享中和你详细讨论。

总结

今天我为你介绍了基于概率模型的隐语义模型的相关知识。

一起来回顾下要点：第一，我们简要回顾了隐语义模型的重要性；第二，我们讨论了基于概率语言的隐语义模型的核心思想，以及 PLSA 和 LDA 的联系和区别。

最后，给你留一个思考题，PLSA 的建模流程有没有什么局限性？

欢迎你给我留言，和我一起讨论。

上一篇　086 | 基础文本分析模型之一：隐语义分析

下一篇　088 | 基础文本分析模型之三：EM算法

精选留言 (1)

写留言

林彦

2018-04-27

PLSA是从现有的数据简化的联合分布估计出来的，会不会导致对训练集的数据过拟合？