【统计学方法及其应用】第1次大作业

混合模型　EM算法

注意：

1. 计算和证明题要写出推导的详细过程，只写答案要扣一半的分数
2. 尽量上传pdf文件，如果手写答案则扫描成pdf文件
3. 本次作业截止日期为2023年11月28日
4. 作业中附上运行结果截图，与源代码一并打包上传
5. 本次作业严禁抄袭，包括往年的作业！！代码和报告部分应独立完成！！

基本任务

1. 阅读《统计推断》（George Casella, Roger L. Berger著）第7章
2. 两个大作业内容二选其一，选自选题的同学【务必】找助教讨论选题的合理性

潜在语义分析（Latent Semantic Analysis, LSA）

在信息检索和自然语言处理领域，潜在语义分析是一种发现文本数据集中潜在主题结构的技术。通过这种分析，我们可以更好地理解和处理大量的文档集合，使得主题发现和文档分类等任务变得更加高效。

问题描述

在多个学科的文献中，作者往往隐含地表达了其研究的核心主题。识别这些潜在的主题不仅有助于理解单篇文献的核心思想，也有助于在更宏观的层面上理解学科的发展动态。本研究旨在应用概率潜在语义分析模型（PLSA），从给定的跨学科文献摘要中提取和识别这些隐含的主题。

概率潜在语义分析模型（Probablistic Latent Semantic Analysis, PLSA）介绍

PLSA模型是一种统计模型，它假设文档生成过程中有潜在的类别因素影响。这个模型通过观察文档和单词的共现信息，推断文档和潜在类别、潜在类别和单词之间的关系。具体地，模型假定每篇文档是由一个潜在主题集合生成的，而每个潜在主题又生成特定的词汇。相关文献如Hofmann, T. (1999). Probabilistic Latent Semantic Analysis. In Proceedings of the Fifteenth conference on Uncertainty in artificial intelligence (pp. 289-296). Morgan Kaufmann Publishers Inc. 提供了PLSA的详细介绍和理论基础。

符号和模型定义

在本研究中，我们定义以下符号来形式化PLSA模型，并对文档生成过程进行建模：

- X：数据集，包含 N 篇文档，即 $X = \{x_1, x_2, \ldots, x_N\}$。
- x_i：第 i 篇文档，表示为一个词频向量，即 $x_i = \{n_{i1}, n_{i2}, \ldots, n_{i|W|}\}$，其中 n_{ij} 表示词 w_j 在文档 x_i 中的出现次数。

- W：词汇表，包含 $|W|$ 个唯一词汇。
- Z：潜在主题集合，包含 $|Z|$ 个唯一主题。
- $\theta_{z,w}$：在主题 z 下词 w 的生成概率。
- Θ：所有$\theta_{z,w}$参数的集合，即$\Theta = \{\theta_{z,w}|z = 1, \ldots, |Z|; w = 1, \ldots, |W|\}$ 。
- $p(w|z)$：在给定主题 z 的条件下词 w 的条件概率。主题生成每个词是独立同分布的多类分布，即 $W|Z \sim multinomial(\theta_{Z,1}, \ldots, \theta_{Z,|W|})$。
- $p(z|d)$：在给定文档 d 的条件下主题 z 的条件概率。
- $p(d)$：文档 d 在数据集中出现的概率。

在PLSA模型中，每个文档 d 被假设为由一个潜在的主题分布生成，每个主题 z 又对应一个词的分布。文档中的每个词 w 被视为以下概率生成过程的结果：首先从文档的主题分布中选择一个主题 z，然后从该主题对应的词分布中选择一个词 w。这一过程可以用以下概率模型表示：

$$p(w|d) = \sum_{z} p(z|d)p(w|z) \tag{1}$$

其中 $p(w|d)$ 表示词 w 在文档 d 中出现的概率。这些符号和参数构成了PLSA模型的基础，使我们能够在后续章节中详细推导模型的似然函数、EM算法的各个步骤以及算法的实现。

数据描述

本研究使用的数据集来自附件 `Task-Corpus.csv`，其中包含14004篇文献摘要，涉及计算机科学、数学、物理学和统计学四个主要学科。每篇文献进一步细分到25个子领域，如“计算机视觉与模式识别”、“数据结构与算法”等。

必做要求

1. 似然函数推导

推导在每篇文章的主题未知和已知的情况下，相应的非完全数据的似然函数 $L(\Theta|X)$ 以及完全数据的似然函数 $L(\Theta|X, Z)$。这里要求详细展示推导过程和结果。

2. EM算法步骤推导

详细推导EM算法中的E步和M步的迭代公式。特别关注于Q函数的推导，即计算给定数据 X 和当前估计的参数集合 $\hat{\Theta}$ 下，完全数据对数似然的期望 $Q(\Theta) = E(\log L(\Theta|X, Z)|\hat{\Theta})$。

3. 编写程序实现推导的EM算法

根据推导出的EM算法步骤，编写实现这一算法的程序。明确指出程序在何种条件下会停止迭代，并详细阐述程序是如何避免EM算法陷入局部最优解的。

4. 参数推断与主题发现

计算与推断

基于所提供的语料库，首先计算每篇文章的词频向量。然后，使用你编写的程序根据这些词频向量推断模型参数。

主题与学科领域对应性分析

如果假定潜在语义的数量为4，分析并观察推断得到的参数，找出每个潜在语义下出现频次最高的词汇，并评估这些潜在语义是否与计算机科学、数学、物理学和统计学这四个领域相对应。

潜在语义个数的确定

如果潜在语义的数量未知，提出一个合理的方法来推断可能的潜在语义个数。请注意，罕见词和停用词应从词频向量中排除。你需要在网上搜索或自行定义停用词表。

选做部分

1. 生成仿真数据，系统评价方法的正确性

2. 设计一个MCMC方法求解该问题

自选题

作业描述

根据课上已学过的内容，自行设计大作业的选题和内容，并利用包含课上所涉及过的知识，完成所设计的大作业。

要求

1. 开始大作业前，请【务必】跟助教讨论题目的合理性，在得到助教允许后方可开始进行大作业
2. 不能与其他课程大作业内容重复！