LDA 模型文本建模

彭海琅 ZY2203322

plunck@buaa.edu.cn

摘要

LDA（Latent Dirichlet Allocation，隐含狄利克雷分布）主题模型主要用于推测文档的主题分布，可以将文档集中每篇文档的主题以概率分布的形式给出根据主题进行主题聚类或文本分类。LDA 主题模型不关心文档中单词的顺序，通常使用词袋特征（bag-of-word feature）来代表文档。本文实现了从给定的语料库中均匀抽取 200 个段落（每个段落大于 500 个词）， 每个段落的标签就是对应段落所属的小说。利用 LDA 模型对于文本建模，并把每个段落表示为主题分布后使用 SVM 支持向量机进行分类。验证与分析了分类结果在不同数量的主题个数下分类性能的变化和分别以"词"和以"字"为基本单元下分类结果的差异。

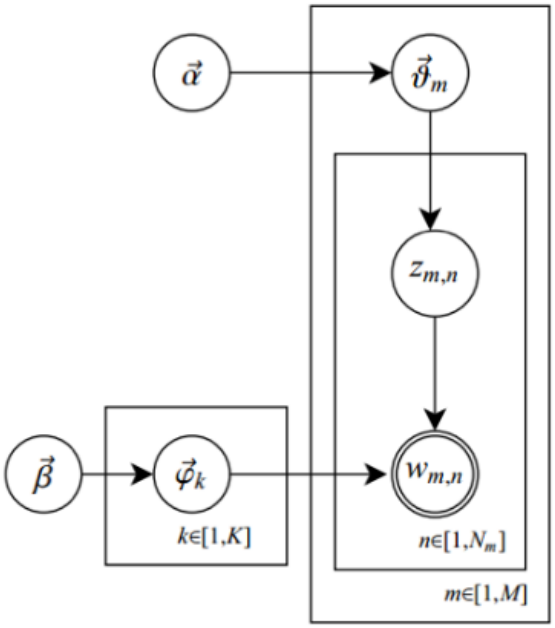
导论

一、LDA 主题模型

LDA 是一种典型的词袋模型，即它认为一篇文档是由一组词构成的一个集合，词与词之间没有顺序以及先后的关系。一篇文档可以包含多个主题，文档中每一个词都由其中的一个主题生成。

另外，正如 Beta 分布是二项式分布的共轭先验概率分布，狄利克雷分布作

为多项式分布的共轭先验概率分布。因此正如 LDA 贝叶斯网络结构中所描述的，在 LDA 模型中一篇文档生成的方式如下:



从狄利克雷分布α中取样生成文档 i 的主题分布θ_i

从主题的多项式分布θ_i中取样生成文档 i 第 j 个词的主题$z_{i,j}$

从狄利克雷分布β中取样生成主题$z_{i,j}$的词语分布$\phi_{z_{i,j}}$

从词语的多项式分布$\phi_{z_{i,j}}$中采样最终生成词语$\omega_{i,j}$

因此整个模型中所有可见变量以及隐藏变量的联合分布是

$$p(\omega_i, z_i, \theta_i, \phi|\alpha, \beta) = \prod_{j=1}^{N} p(\theta_i|\alpha)p(z_{i,j}|\theta_i)p(\phi|\beta)p(\omega_{i,j}|\phi_{z_{i,j}})$$

最终一篇文档的单词分布的最大似然估计可以通过将上式的θ_i以及ϕ进行积分和对z_i进行求和得到

$$p(\omega_i|\alpha, \beta) = \int_{\theta_i} \int_{\phi} \sum_{z_i} p(\omega_i, z_i, \theta_i, \phi|\alpha, \beta)$$

根据$p(\omega_i|\alpha, \beta)$的最大似然估计，最终可以通过吉布斯采样等方法估计出模型中的参数。

LDA 模型的训练过程是一个迭代的过程，具体步骤如下：

1) 初始化每个词语的主题，可以随机初始化或根据先验知识初始化。

2) 遍历每篇文档中的每个词语，计算每个词语属于每个主题的概率分布。

3) 根据计算得到的概率分布，重新分配每个词语的主题。

4) 重复步骤 2 和 3，直到主题分布收敛为止。

二、支持向量机

支持向量机（Support Vector Machine, SVM）是一类按监督学习方式对数据进行二元分类的广义线性分类器，其决策边界是对学习样本求解的最大边距超平面。SVM 使用铰链损失函数计算经验风险并在求解系统中加入了正则化项以优化结构风险，是一个具有稀疏性和稳健性的分类器。SVM 可以通过核方法进行非线性分类，是常见的核学习方法之一。

实验内容

一、问题描述

从给定的语料库中均匀抽取 200 个段落（每个段落大于 500 个词），每个段落的标签就是对应段落所属的小说。利用 LDA 模型对于文本建模，并把每个段落表示为主题分布后进行分类。验证与分析分类结果，（1）在不同数量的主题个数下分类性能的变化；（2）以"词"和以"字"为基本单元下分类结果有什么差异?

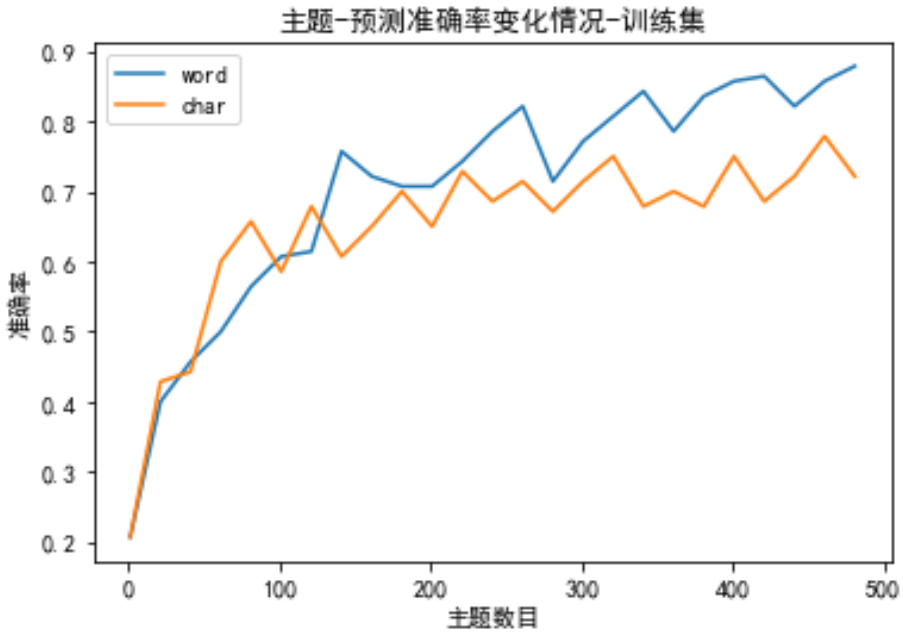
二、实现步骤

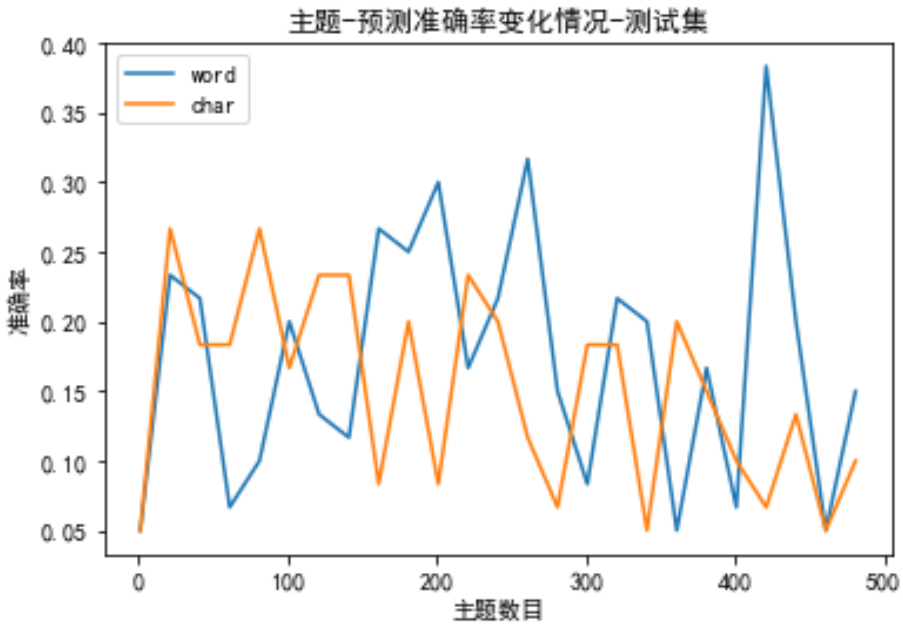
1、读取文本，对超过 500 字的段落进行筛选，并从中均匀抽取 200 个段落，然后对这些段落进行停词处理，并且按比例分为训练集和测试集。

2、分别按字和词对段落进行分割，生成 LDA 学习中的词典和文本向量。

3、按主题 topics 数目迭代，用 LDA 模型输出的主题分布作为 SVM 的输入进行分类学习，验证不同数量的主题个数下分类性能的变化，并且对比分别以字和词为基本单元时分类结果的差异。本文使用 gensim 库中的 LdaModel 函数进行模型训练。

实验结果





1、在不同主题数目下分类性能的变化

随着主题数目的增加，训练集的分类准确率显著提高，但是测试集的准确率却呈现出极大的波动，可以看到不同主题数对于分类准确率有着很大的影响。这是因为主题数过少时，SVM 分类器获得的主题分布的特征向量维度太小，有用的信息太少，SVM 分类器训练欠拟合，提高主题数目，分类器可能产生过拟合造成分类准确率的波动。选择一个合适的主题数目对于分类性能非常重要。

2、以"词"和以"字"为基本单元下分类结果的差异

主题数目比较小的时候，以字为基本单位分类的准确率高于以词为基本单位分类，主题数目比较大的时候，以词为基本单位分类的准确率在大部分时候高于以字为基本单位分类。一方面，以词作为基本单元的 LDA 模型则可能会更加关注语义信息，但可能会忽略一些细节信息，另一方面，基于词的 LDA 模型可以更加准确地丰富地获得文章的信息，从而更好地获得主题分布，更有利于 SVM 分类。具体选择什么基本单元需要具体分析。

结论

本文通过 LDA 主题模型对金庸武侠小说进行主题建模，并探究了不同的基本词元和主题数对于分类准确率的影响。随着主题数目的增加，训练集的分类准确率显著提高，但是测试集的准确率却呈现出极大的波动。主题数目比较小的时候，以字为基本单位分类的准确率高于以词为基本单位分类，主题数目比较大的时候，以词为基本单位分类的准确率在大部分时候高于以字为基本单位分类。