深度学习与自然语言处理作业 2 报告

段沛东
ZY2303106
19231209@buaa.edu.cn

摘要

本文为深度学习与自然语言处理课程第二次作业的报告。潜在狄利克雷分配（latent Dirichlet allocation，LDA）是一种基于贝叶斯学习的话题模型，是潜在语义分析、概率潜在语义分析的扩展，在文本数据挖掘、图像处理、生物信息处理等领域被广泛使用。本文使用了 LDA 模型在给定的金庸小说语料库上进行文本建模，把每个段落表示为主题分布后使用支持向量机进行分类，并且分析了选择不同的 Token 和 Topic 数量对分类性能的影响。

问题描述

本实验的主要任务为从给定的语料库中均匀抽取 1000 个段落作为数据集（每个段落可以有 K 个 token，K 可以取 20、100、500、1000、3000），每个段落的标签就是对应段落所属的小说。利用 LDA 模型在给定的语料库上进行文本建模，主题数量为 T，并把每个段落表示为主题分布后进行分类，分类结果使用 10 次交叉验证，并讨论以下几个方面：（1）在设定不同的主题个数 T 的情况下，分类性能是否有变化；（2）以“词”和以“字”为基本单元的分类结果有什么差异；（3）不同的取值的 K 的短文本和长文本，主题模型性能上是否有差异。

模型原理和实现

LDA 主题模型的原理与概率潜在语义分析（PLSA）模型相似，它们都在文本和词之间加了一层主题（Topic），先让文本和主题产生关联，再在主题中寻找词的概率分布，这样可以将具有相同意义的词总结在同一个主题下，降低了同词不同义和同义不同词的影响。PLSA 和 LDA 的主要区别在于：PLSA 模型生成文本的过程可以看作两个词袋模型的组合，第一个模型确定主题，第二个模型重复独立做等同于文本词数的次数，确定文本；LDA 模型则可以看成将两个词袋模型替换为两个贝叶斯词袋模型的 PLSA，贝叶斯词袋模型的概率分布是一个狄利克雷同轴分布，则 LDA 的整个过程实际就是两个狄利克雷同轴分布的组合，其生成文本集的具体过程如下：

（1）首先生成随机生成一个文本的话题分布；

（2）之后在该文本的每个位置，依据该文本的话题分布随机生成一个话题；

（3）然后每个位置根据话题的词分布随机生成一个词，直至文本的最后一个位置；

（4）重复以上过程生成所有文本。

LDA 模型的学习过程则为在文本集已给定的条件下逆向估计其各分布参数的过程。

本实验调用了 gensim 库中的 LDA 方法训练了 LDA 模型，并调用 scikit-learn 库训练了 SVM 模型并进行交叉验证。

实验结果

为了验证问题描述中的(1)(3)问题，我们首先固定每个段落 Token 的数量 K 为 100，只改变主题数量 T，观察模型性能的变化情况，结果如表 1 所示。之后我们固定主题数量 T 为 200，只改变每个段落 Token 的数量 K，观察模型性能的变化情况，结果如表 2 所示。同时为了验证(2)问题，表 1 和表 2 都给出了其他条件相同时以“词”和以“字”为基本单元的性能对比。

表 1 固定 K，T 不同取值时模型分类准确率表

Topics	50	100	200	400	800
Word	0.244	0.274	0.291	0.318	0.37
Char	0.464	0.54	0.594	0.653	0.691

表 2 固定 T，K 不同取值时模型分类准确率表

Tokens	20	100	500	1000	3000
Word	0.146	0.291	0.698	0.856	0.969
Char	0.389	0.594	0.817	0.892	0.979

结果分析与结论

根据表 1 和表 2 可以观察到主题个数 T 和文本长度 K 的变化都对主题模型的性能有显著的影响。随着 K 的增大，模型的性能得到了显著的提升，模型训练所需的时间也大幅增加，可以认为 LDA 模型对长文本的分类性能更好；随着 T 的增大，模型的性能也获得了一定程度的提升，但提升的幅度不如增大 K；所有条件下以“字”为基本单元的分类性能都优于以“词”为基本单元，使用 LDA 进行文本建模时，应当优先考虑使用“字”作为基本单元。