DNLP 作业 2

LDA 主题分布与分类

张乃天 SY2302516

freshznt@buaa.edu.cn

Abstract

本次作业旨在利用 LDA（Latent Dirichlet Allocation，隐狄利克雷分布）算法，将从语料库中随机抽取的段落表示成其对应的主题概率分布。根据概率分布特征对段落进行分类，分类器选择 SVM（Support Vector Machine，支持向量机）。分类结果与段落所处小说作对比，观察并分析分类结果的优劣。

In this assignment, we use the Latent Dirichlet Allocation (LDA) algorithm to represent paragraphs randomly sampled from a corpus as their corresponding probability distributions over topics. Paragraphs are classified based on the probability distribution features with a Support Vector Machine (SVM) . The classification results are compared with the novels from which the paragraphs originate, further analyzing the classification performance.

Introduction

本次作业的具体要求如下：从给定语料库中均匀抽取 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（分类器自由选择），分类结果使用 10 次交叉验证（i.e. 900 做训练，剩余 100 做测试循环十次）。实现和讨论如下的方面：（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"为基本单元下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？

由上述要求可见，核心模块为“LDA 模型”和“分类器”。本次作业选择 SVM（Support Vector Machine，支持向量机）作为分类器。下面介绍 LDA 模型和 SVM 的原理。

- LDA 模型

LDA 是一种生成式概率模型，用于主题建模，在 NLP 领域广泛应用。它假设每个文档都是主题的混合体，而每个主题又是词汇的混合体，通过为出现在文档和主题中的词汇分配概率来揭示语料库中的隐藏主题结构。

LDA 模型通过一定的层级结构随机生成一篇文档，并通过最大化该随机文档与原文档匹配的概率来获取最优模型。具体来说，LDA 模型生成文档的步骤如下：

1）固定主题数为 T，从狄利克雷分布α中取样生成文档 d 的主题多项式分布$\theta^{(d)}$；

2）从主题的多项式分布$\theta^{(d)}$ 中取样生成文档 d 第 N_d 个词的主题 z;

3）从狄利克雷分布β中取样生成主题的词语多项式分布$\Phi^{(z)}$；

4）从词语的多项式分布$\Phi^{(z)}$ 中采样最终生成词语 w。

如此循环，生成 T 个主题下，D 篇 N_d 长的文档。上述步骤也就形成了如图 1 所示的层级结构：

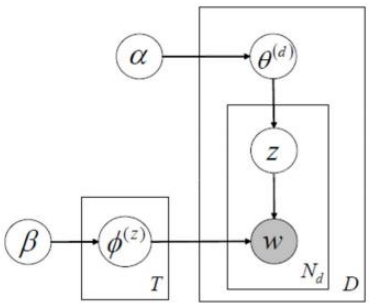

图 1 LDA 流程图

狄利克雷分布是多项式分布的共轭先验，就像 Beta 分布和二项式分布的关系一样。当选择狄利克雷分布作为多项式分布参数的先验时，观测采样数据后的后验概率将也呈现狄利克雷分布，且后验狄利克雷分布的参数与先验狄利克雷分布的参数有简单的闭式表达关系。这样在训练时就可以避免处理最大后验概率，而是直接处理先验和似然函数。

- SVM

SVM 是一种监督学习算法，用于分类和回归任务。其主要思想是找到一个最优的超平

面，能够将不同类别的样本点分开，并且使得分类边界与最近的样本点之间的间隔最大化。具体来说，SVM 致力于找到一个能够在不同类别之间形成最大间隔的超平面。这个间隔是指距离超平面最近的训练样本点到超平面的距离，通过最大化这个间隔来提高模型的泛化能力。除了简单的“线性”分割超平面外，SVM 通过核函数将样本从原始特征空间映射到更高维的特征空间，从而使得非线性问题也能够被线性分类器解决。常用的核函数包括线性核、多项式核和高斯核等。

为了简便，本次作业采用线性核。

Methodology

本次作业的实现方法步骤如下：

1）语料库预处理：去掉“广告”、停用词，并根据 token 类别进行切分（词需要使用 jieba 分词）；

2）段落采样：根据所要求的段落长度，将每篇小说划分成若干个段落。再根据每篇小说段落数量的多少，成比例的随机抽取段落个数，使得段落数量总和等于要求的 1000 个；

3）交叉训练：将 1000 个段落分成 10 组，其中 9 组作为训练集，1 组作为测试集，循环 10 次；

4）训练 LDA：采用 gensim.models.LdaModel 建模 LDA 模型的训练过程。值得一提的是，需要利用 gensim.copora.Dictionary 命令将训练集段落中的 token 保存成字典，再利用 doc2bow 方法将 token 以“词袋”形式保存。经过上述操作后才是 LdaModel 训练数据的默认格式。

5）分类器训练与测试：采用 sklearn.svm 建模分类器 SVC（Support Vector Classifier）。

Experimental Studies

（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？

在固定 token 为“字”、段落长度为 3000 的基础上，探究不同主题个数对于分类准确程度的影响，主题个数取值范围设置为 5~300 之间。结果如表 1 所示。

表 1 探究主题数量的影响

主题数量 T	Train Accuracy	Test Accuracy
5	0.445	0.473

10	0.660	0.662
20	0.731	0.729
30	0.754	0.749
50	0.793	0.789
100	0.802	0.805
150	0.790	0.780
200	0.803	0.790
300	0.791	0.785

从表 1 中可以看出，当主题数量较小时，分类器的性能较低；当主题数量增加后，分类的准确性逐渐增加；当主题数量达到 100 左右之后，性能达到饱和。这说明，用主题概率分布表示段落时，如果向量维度较小（主题数少），则段落在主题分布特征上会混叠，难以分类。当向量维度较大（主题数增多）才能将段落的特征可分辨。若在此基础上继续增加维度，增益不明显。

（2）以"词"和以"字"为基本单元下分类结果有什么差异?

固定段落长度为 1000，结果如表 2 所示。

表 2 探究不同基本单元的影响

主题数量 T	以“字”为单位		以“词”为单位	
	Train Accuracy	Test Accuracy	Train Accuracy	Test Accuracy
5	0.391	0.371	0.254333333	0.298
10	0.476	0.487	0.296444444	0.328
20	0.577	0.587	0.338666667	0.378
30	0.608	0.627	0.371333333	0.411
50	0.609	0.627	0.383111111	0.375
100	0.626	0.615	0.402555556	0.401
150	0.591	0.598	0.434555556	0.457
200	0.611	0.631	0.448111111	0.517
300	0.599	0.61	0.433222222	0.468

从上述结果上看，以“字”为 token 的分类精准度略高于用“词”做为 token 的分类精准度，但差别不是很明显。本人认为这和本次仿真条件中段落长度不够长，以及段落个数较少有关。

直觉上，若段落长度较长，且段落数较多时，以“词”为单位表示与主题的关系应该十分准确，进而使得主题分布特征准确。但受限于仿真条件，以“词”为单位的分类性能较差。

（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?

此在固定主题数量 T 为 200、token 为“字”的基础上，探究不同段落长度对于分类准确性的影响。具体结果如表 3 所示。

表 3 探究段落长度的影响

段落长度 K	Train Accuracy	Test Accuracy
20	0.367	0.214
100	0.381	0.308
500	0.461	0.433
1000	0.611	0.631
3000	0.804	0.799

由上面的结果可知，当段落长度越长时，分类的准确度越高。这是因为，当段落中的 token 足够多时，才能更充分的囊括和主题强相关的词语，使得其主题的特征表示更准确；当段落中的 token 很少时，主题强相关词语较少，弱相关词语等无关信息干扰较严重，导致分类性能下降。

值得一提的是，上述表格 1 到 3 中的分类准确度没有出现特别高的性能。除了和超参数相关之外，分类器的选择也可能有较大的影响。本次作业采用了最为简单的“线性核”SVM，如果采用其他更加复杂的分类器，可能会使得分类性能进一步变好。

Conclusions

本次作业利用 LDA 模型对语料库中的段落进行建模，提取主题分布特征，并利用了 SVM 进行分类。结果表明，对于给定的语料库以及段落个数 1000，当主题数达到 100 个即以上，且段落长度较长时，分类性能较好。

References

[1] https://www.bilibili.com/video/BV123411G7Z9