Report of Deep Learning for Natural Langauge Processing

LDA

Jiayu Cui
cuijiayu_2001@163.com

Abstract

从下面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？

Methodology

M1: LDA Model

隐含狄利克雷分布（Latent Dirichlet Allocation，简称 LDA），是一种概率主题模型。LDA 可以将文档集中每篇文档的主题以概率分布的形式给出，通过分析一批文档集，抽取出它们的主题分布，就可以根据主题分布进行主题聚类或文本分类。同时，它是一种典型的词袋模型，即一篇文档是由一组词构成，词与词之间没有先后顺序关系。此外，一篇文档可以包含多个主题，文档中每个词都由其中的一个主题生成。

LDA 是一种无监督学习方法，在训练时不需要手工标注的训练集，需要的仅仅是文档集以及指定主题数量为 T 即可。此外，LDA 的另一个优点是，对于每一个主题均可找出一些词语来描述它。

一篇文章的每个词都是以一定概率选择某个主题，并从这个主题中以一定概率选择某个词语而组成的。用公式表示为：

$$P(word|doc) = P(word|topic) * P(topic|doc)$$

从公式来看，$P(word|doc)$可以通过文档中该词语出现的次数除以文档中词语总数计算出来。

M2: SVM Classifier

支持向量机分类器，是在数据空间中找出一个超平面作为决策边界，利用这个决策边界来对数据进行分类，并使分类误差尽量小的模型。决策边界是比所在

数据空间小一维的空间，在三维数据空间中就是一个平面，在二维数据空间中就是一条直线。以二维数据为例，图中的数据集有两个特征，标签有两类，一类为紫色，一类为红色。对于这组数据，我们找出的决策边界被表达为$wx+b=0$，决策边界把平面分成了上下两部分，决策边界以上的样本都分为一类，决策边界以下的样本被分为另一类。以图 1 为例，红色实线上部分为一类（全部都是红色点），下部分为另一类（全都是蓝色点）。

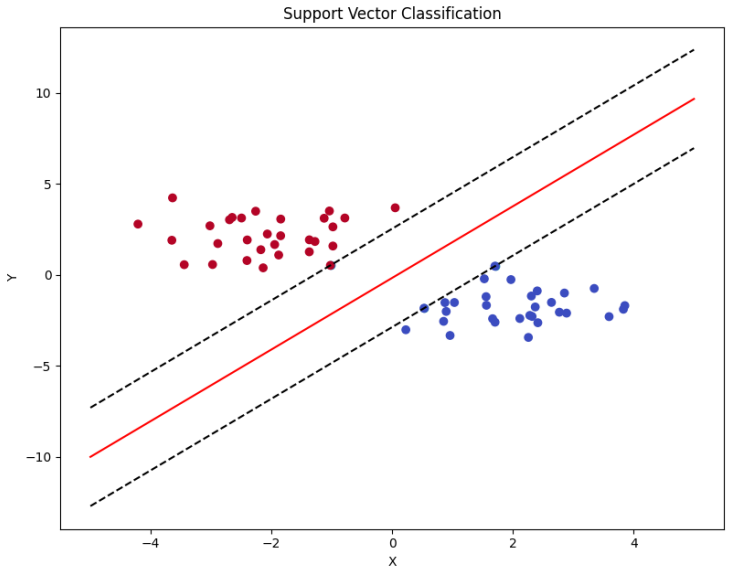

图 1 SVM 分类图

SVM 中最核心的是核函数的选取和参数选择。高斯径向基函数是一种局部性强的核函数，其可以将一个样本映射到一个更高维的空间内，该核函数是应用最广的一个，无论大样本还是小样本都有比较好的性能，而且其相对于多项式核函数参数要少。其公式为：

$$\kappa(x,x_i)=exp(-\frac{||x-x_i||^2}{\delta^2})$$

M3:训练过程

1. 获取语料库：停用词文件中读取停用词列表。然后，遍历指定目录下的所有文件，读取其中的文本内容。

2. 文本预处理：对于每个文本文件，进行以下预处理步骤：去除指定的特定字符串，如"本书来自..."和"更多更新..."。使用正则表达式去除英文字符、数字和特殊符号。使用 jieba 库对文本进行分词。去除停用词，将剩余的词加入新的词列表中。打印每个文本文件的总词数。将处理后的文本数据以字典的形式返回，其中键为文件名，值为经过预处理后的词列表。根据 use_character 参数，将文本数据转换为字符级别或词级别。在本示例中，选择使用字符级别。

3. 段落抽取：从每篇文章中抽取一定数量的段落，以构建训练数据。段落的数量和长度由 paragraph_num=1000 和 token 值 K（paragraph_length）参数控制。这里使用一个循环来实现，确保从每篇文章中抽取相同数量的段落，同时保证总段落数达到设定的 paragraph_num=1000。

4. 标签处理：将标签从字符串形式映射为整数形式，以便于模型处理。

5. 数据划分：将数据集划分为训练集和测试集，按照设定的训练-测试比例

train_p=0.9 进行划分。

6. 构建词典和语料库：使用 corpora.Dictionary 构建词典，并将文本数据转换为词袋表示形式。

7. 训练 LDA 模型：使用 models.LdaModel 训练 LDA 模型，其中参数 num_topics 指定了主题的数量。打印每个主题的主要词分布。

8. 特征提取：获取训练集和测试集的每个段落的主题分布，作为模型的特征输入。

9. 模型训练：使用 SVM 分类器（采用 RBF 核函数）进行模型训练，并对训练集和测试集进行测试。

10. 性能评估：计算模型在训练集和测试集上的准确率，并输出结果。

Experimental Studies

S1：不同主题数目 T

给定 K=500，在设定不同的主题个数 T 的情况下，分类性能如表 1 和图 2 所示。

表 1 主题个数 T 对分类性能的影响

主题数目 T	训练集准确率	测试集准确率
1	6.84%	1.00%
5	18.72%	23.00%
10	23.54%	18.00%
20	31.39%	34.00%
30	34.87%	37.00%
40	40.81%	46.00%
50	40.47%	25.00%
100	51.12%	50.00%
200	59.30%	34.00%
300	63.68%	43.00%
400	69.73%	38.00%
500	72.53%	37.00%

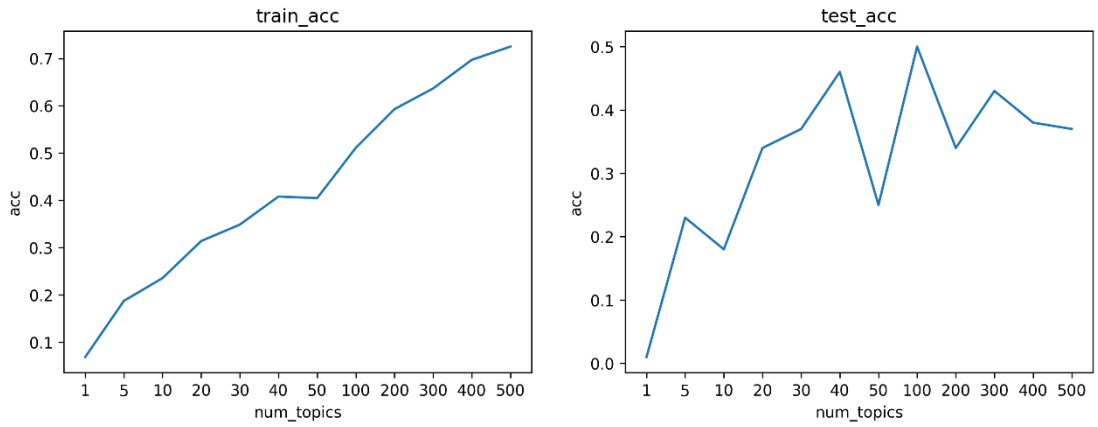

图 2 主题个数 T 对分类性能的影响

随着主题数目 T 的增加，训练集准确率和测试集准确率并不是单调递增或者单调递减的。在低主题数目时，模型的拟合能力不足，导致训练集和测试集的准确率都较低。当主题数目增加到一定程度时，模型开始提升性能，训练集和测试集的准确率逐渐增加。然而，当主题数目进一步增加时，模型可能会出现过拟合现象，导致测试集准确率下降，而训练集准确率上升。在测试集准确率最高的情况下，主题数目为 100，但这并不意味着模型效果最佳，因为这可能是因为模型在训练集上过度拟合而导致的。

因此，需要综合考虑模型在训练集和测试集上的表现，选择一个合适的主题数目，以在测试集上取得良好的泛化性能。

S2：不同 token 值 K

给定 T=50，在设定不同的 Token 值 K 的情况下，分类性能如表 2 和图 3 所示。

表 2 Token 值 K 对分类性能的影响

Token 值 K	训练集准确率	测试集准确率
20	18.50%	14.00%
100	24.55%	14.00%
500	36.66%	30.00%
1000	47.42%	50.00%
3000	70.96%	69.00%

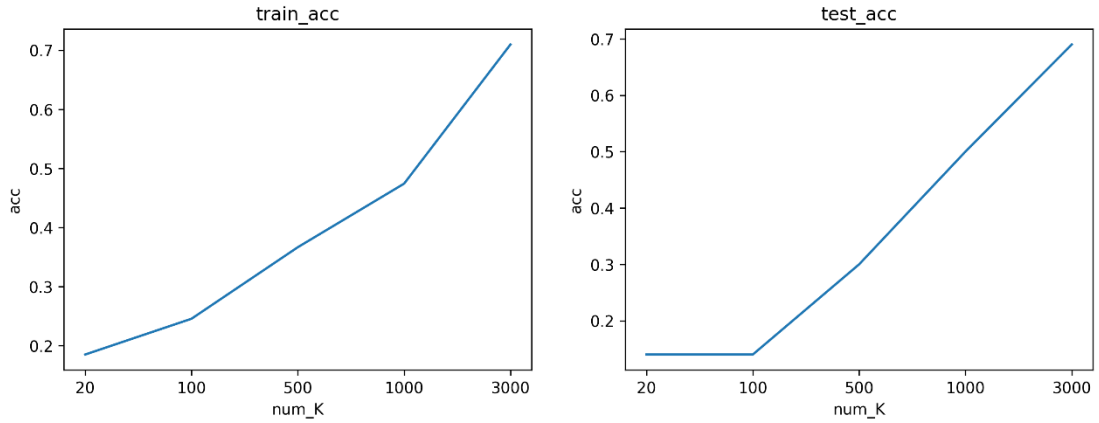

图 3 Token 值 K 对分类性能的影响

不同 Token 值 K 对于分类准确有着一定的影响。K 值过小时，LDA 模型获取的信息太少，无法准确地归纳出主题分布，而且 SVM 分类器训练集太少，最终会导致 SVM 分类器欠拟合，测试集准确率较低。随着 K 值增加，测试集准确率逐渐上升，说明 K 值增加在一定程度上帮助了 LDA 模型更好地获得主题分布，测试集准确率提高。

S3：以"词"和以"字"为基本单元下分类

当 T = 100，K=1000 时，训练集和测试集以"词"和以"字"为基本单元下分类的准确率如表 3 所示。

表 3 以"词"和以"字"为基本单元对分类性能的影响

分词	训练集准确率	测试集准确率
字	53.59%	40.00%
词	57.40%	55.00%

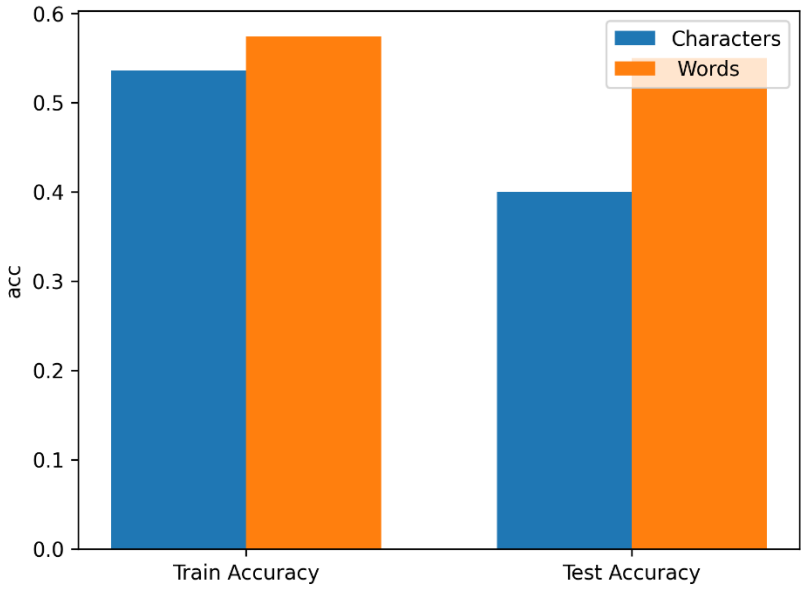

图 4 以"词"和以"字"为基本单元对分类性能的影响

以字为单位的训练集准确率为 53.59%，测试集准确率为 40.00%，而以词为单位的训练集准确率为 57.40%，测试集准确率为 55.00%，较字相比有所提升，因为基于词的 LDA 模型可以更加准确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。

References

[1] https://samperson1997.github.io/2018/06/24/svm/
[2] https://zhuanlan.zhihu.com/p/31470216