Report of Deep Learning for Natural Langauge Processing

Jiajie Wu
ZY2343402

Abstract

本文基于中文语料库中金庸的 16 部作品所有文本作为数据输入，利用 LDA 模型在指定语料库上进行文本建模；在此基础上使用随机森林分类算法实现文本分类。基于此，本文探讨了主题个数 T、分词的基本单元(“词”和“字”)，以及每个段落中 token 的数量 K 对分类性能和主题模型性能的差异。

Introduction

LDA 是一种文档主题生成模型，包含词、主题和文档三层结构。LDA 中文翻译为：潜在狄利克雷分布。LDA 主题模型是一种文档生成模型，是一种非监督机器学习技术。它认为一篇文档是有多个主题的，而每个主题又对应着不同的词。一篇文档的构造过程，首先是以一定的概率选择某个主题，然后再在这个主题下以一定的概率选出某一个词，这样就生成了这篇文档的第一个词。不断重复这个过程，就生成了整篇文章(当然这里假定词与词之间是没有顺序的，即所有词无序的堆放在一个大袋子中，称之为词袋，这种方式可以使算法相对简化一些)。LDA 的使用是上述文档生成过程的逆过程，即根据一篇得到的文档，去寻找出这篇文档的主题，以及这些主题所对应的词。

朴素贝叶斯算法是一种基于概率统计和特征条件独立假设的分类方法。它常被用于文本分类、垃圾邮件过滤、情感分析等领域。这个算法的核心思想是利用贝叶斯定理来计算在给定输入的情况下，各个类别的概率，并选择具有最高概率的类别作为输出。在文本分类任务中，由于朴素贝叶斯假设特征之间相互独立，因此其处理高维稀疏数据效果好。朴素贝叶斯算法只需要估计各个特征的概率分布，而不需要对参数进行复杂的优化过程，因此计算效率高。同时朴素贝叶斯算法对于噪声数据具有一定的鲁棒性，因为它在计算类别概率时考虑了所有的特征。朴素贝叶斯算法基于贝叶斯定理。

本文使用 LDA 主题模型对指定语料库进行文本建模，主题数量为 T。其中，指定语料库为从金庸小说集中均匀抽取的 1000 个段落(每个段落有 K 个 token)，段落的标签为其所属的小说。主题模型通过学习一系列的文档，根据统计规律发现抽象主题。基于主题模型获得的主题分布进行文本分类，分类器选择朴素贝叶斯模型，分类结果使用 10 次交叉验证（900 做训练，剩余 100 做测试循环十次）。

Methodology

Part 1:LDA 模型

对于语料库中的每篇文档，LDA 生成过程如下：

1.对每一篇文档，从主题分布中抽取一个主题；

2.从上述被抽到的主题所对应的单词分布中抽取一个单词；

3.重复上述过程直至遍历文档中的每一个单词。

定义文档集合 D，主题（topic)集合 T。语料库中的每一篇文档与 T（通过反复试验等方法事先给定）个主题的一个多项分布相对应，将该多项分布记为 θ。每个主题又与词汇表中的 V 个单词的一个多项分布相对应，将这个多项分布记为 φ。

D 中每个文档 d 看作一个单词序列$< w_1, w_2, \dots, w_n >$，w_i表示第i个单词，设 d 有 n 个单词。D 中涉及的所有不同单词组成一个大集合，LDA 以文档集合 D 作为输入，希望训练出的两个结果向量：对每个 D 中的文档 d，对应到不同 T 的概率$\theta_d =< p_{t1}, p_{t2}, \dots, p_{tk} >$，其中$p_{ti}$表示 d 对应 T 中第 i 个 topic 的概率,$p_{ti} = \frac{\mathrm{n_{ti}}}{\mathrm{n}}$,其中$n_{ti}$表示 d 中对应第 i 个 topic 的词的数目，n 是 d 中所有词的总数。对每个 T 中的 topic，生成不同单词的概率$\varphi_t =< p_{w1}, p_{w2}, \dots, p_{wm} >$，其中，$p_{wt}$表示 t 生成 VOC 中第 i 个单词的概率。计算方法同样很直观，$p_{wi} = \frac{N_{wi}}{N}$,其中$N_{wi}$表示对应到 topic 的 VOC 中第 i 个单词的数目，N 表示所有对应到 topic 的单词总数。

LDA 的核心公式如下：

$$p(w|d) = p(w|t) * p(t|d)$$

以 Topic 作为中间层，可以通过当前的θ_d和φ_t给出了文档 d 中出现单词 w 的概率。其中$p(t|d)$利用θ_d计算得到，$p(w|t)$利用φ_t计算得到。

利用当前的θ_d和φ_t，我们可以为一个文档中的一个单词计算它对应任意一个 Topic 时的$p(w|d)$，然后根据这些结果来更新这个词应该对应的 topic。然后，如果这个更新改变了这个单词所对应的 Topic，就会反过来影响θ_d和φ_t。

LDA 算法开始时，先对所有的 d 和 t 随机地给θ_d和φ_t赋值，不断重复，最终收敛到的结果就是 LDA 的输出。

Part 2: 朴素贝叶斯模型

朴素贝叶斯算法基于贝叶斯定理。

$$P(y \mid x_1, x_2, \dots, x_n) = \frac{P(x_1, x_2, \dots, x_n \mid y) \cdot P(y)}{P(x_1, x_2, \dots, x_n)}$$

其中：$P(y \mid x_1, x_2, \dots, x_n)$是在给定输入$x_1, x_2, \dots, x_n$的情况下，类别 y的概率。$P(y)$是类别 y的先验概率。$P(x_1, x_2, \dots, x_n \mid y)$是在类别$y$下，输入特征$x_1, x_2, \dots, x_n$的条件概率。$P(x_1, x_2, \dots, x_n)$ 是输入特征的先验概率。

Experimental Studies

Table 1: 不同 K,T 以及字词基本单位下分类准确度

基本单元	T	K	Test Accuracy	基本单元	T	K	Test Accuracy
字	50	20	0.701	词	50	20	0.915
		100	0.927			100	0.920
		500	0.917			500	0.924
		1000	0.917			1000	0.928
		3000	0.917			3000	0.931
	10	500	0.910		10	500	0.915
	20		0.915		20		0.921
	50		0.917		50		0.924
	60		0.918		60		0.920
	80		0.912		80		0.915
	100		0.906		100		0.909

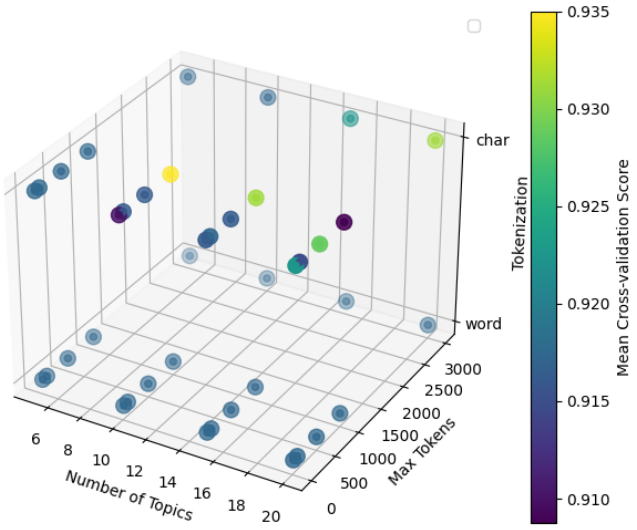

Figure 1: 不同 K,T 以及字词基本单位下分类情况

对于基于字符的朴素贝叶斯分类器，训练的准确率在所有不同的最大标记长度下，训练准确率都保持在 91.75%的水平。这表明模型对训练数据的拟合效果很好，能够准确地预测训练样本的标签，测试的准确率在所有不同的最大标记长度下都稳定在 91.5%。这说明模型具有很好的泛化能力，能够在未见过的数据上进行准确的预测。

对于基于单词的朴素贝叶斯分类器，当最大标记长度为 20 时，训练准确率较低，仅为 70.125%。然而，随着最大标记长度的增加，训练准确率显著提高，

达到了 92.75%（最大标记长度为 100）。在最大标记长度为 500、1000 和 3000 时，训练准确率保持在 91.75%的高水平，表明模型已经收敛并且不再随最大标记长度的增加而提高。与训练准确率类似，测试准确率在最大标记长度为 20 时较低，为 73%，而在最大标记长度为 100 时达到了 93%。随着最大标记长度的增加，测试准确率保持在 91.5%的水平，这表明模型在不同最大标记长度下都能够保持较高的泛化能力。

基于字符的朴素贝叶斯模型表现稳定，在训练和测试数据上都表现良好。基于单词的朴素贝叶斯模型在较低的最大标记长度下表现较差，但随着最大标记长度的增加，性能得到了显著提升，并在较高的最大标记长度下达到了稳定的性能水平。

Conclusions

从 Table1 中可以看出，当主题数 T 和 token 取值 K 相同时，以词为基本单元的主题模型性能优于以字为基本单元的主题模型。以词为基本单元时，其分类模型准确率相较于以字为基本单元时更高。以词为基本单元的主题模型中主题的独特性更高，能够更好地体现不同小说的特点，而以字为单位时独特性较低，因此以词为基本单元的分类性能更高，符合常理认知。

随着 K 值的增加，不同基本单元的主题模型性能变化呈现出相同的趋势，其分类结果的准确度也随之增加，主题模型表现出更好的性能。说明长文本相较于短文本的独特性更高，使用长文本进行训练能够提取出更为有效的文本主题。

当 K 固定为 500 时，随着 T 从 10 上升到 50，测试集准确度稳定上升，分类模型性能逐渐提高。当 T 继续增大时，测试集准确度逐渐下降，分类模型性能下降。在选择主题模型时，应当选择合适的主题数进行训练，若主题数过高，则主题间过于分散，不利于文本分类。