机器学习工程师纳米学位

开题报告

Flyzhg
2018/9/20

文档分类

项目背景

自然语言处理（Natural Language Processing, 简称 NLP）是目前机器学习技术的主要的应用范畴之一。从手机上语言识别成文字，再到语义理解，推理，，这些都不离不开 NLP，说的直接一点就是让机器能和人类快速有效的沟通，达到类似于人和人沟通一样的极致体验效果。

从国际上看，当前主要的热点研究在美国，主要以英语的 NLP 为主要方向，对于英文有许多要考虑的问题，比如要考虑区分单词的大小写，是否要对同一个词不同形式（单复数）进行统一处理等。

从国内发展情况来看，由于国内使用中文，而中文含有上万个汉字，不同的汉字又可以进行各种不同的组合，不同的组合代表不同的意思，不同的汉字也可能代表相同的意思，所以在国内研究自然语言处理 NLP 这个领域相对而言更复杂一些。

背景从两方面来说，最简单的背景就是我们学习机器学习纳米学位工程师必须要通过的一个锻炼过程。从另一方面讲，这个具有广泛应用领域的项目，也是在实际生活和工作中能经常接触到的一个项目；需要我们好好认真对待

问题描述

今天的问题，是对文档进行分类。目前有 20 个新闻组数据集，大约 20000 个新闻组文档的集合；我的问题就是要通过对 20 个新闻组集合（中的一部分，大约 80%）的学习训练，获得一个好的模型，通过获得的模型，希望能对那些未分类的文档进行很好的预测和分类。

，

数据和输入

分类文本数据可以使用经典的 20 类新闻包，里面大约有 20000 条新闻，比较均衡地分成了 20 类，是比较常用的文本数据之一。[参考 udacity 文档]

此外，词向量的训练也需要大量数据，如果感觉 20 类新闻数据样本量不足以训练出较好的词向量模型，可以采用 Mikolov 曾经使用过的 text8 数据包进行训练。[参考 udacity 文档]

解决办法

解决办法主要是分为如下几步：

（1） 探索文本的表示方式

使用词袋子模型表示每篇文档，也就是将一个文本文件分成单词的集合，建立词典。每篇文档表示成特征词的频率向量或者加权词频 TF-IDF 向量，这样就可以得到熟悉的特征表。

利用 Word2Vec 方式即词向量模型表示每篇文档， 这里面包含两部分工作

I， 利用文本数据对词向量进行训练，将每个次表示成向量形式。 词向量训练后需要进行简单评测， 比如检验一些单词之间的相似性是否符合逻辑等。

II， 探讨怎样用文档中每个词的向量来表达整个文档。

（2）分别在词袋子，词向量表达的基础上采用你认为适当的模型对文本分类，优化模型并分析其稳健性。[参考 udacity’s document_classification]

基准模型

毕业项目采用决策树模型。 关于决策树模型，决策树(decision tree)是一种基本的分类与回归方法。决策树模型呈树形结构，在分类问题中，表示基于特征对实例进行分类的过程。它可以认为是 if-then 规则的集合，也可以认为是定义在特征空间与类空间上的条件概率分布。

其主要优点是模型具有可读性，分类速度快。学习时，利用训练数据，根据损失函数最小化的原则建立决策树。预测时，对新的数据，利用决策树模型进行分类。

[https://baike.baidu.com/item/%E5%86%B3%E7%AD%96%E6%A0%91%E7%AE%97%E6%B3%95/8595872?fr=aladdin]

决策树学习通常包括 3 个步骤：特征选择、决策树的生成和决策树的修剪。[李航. 统计学习方法[M]. 清华大学出版社, 2012.]

评估标准

模型经过训练测试做出以后，需要有一个评价指标，这个指标能够衡量我们采用的解决方案的标准，在该项目中，如果对于测试的新闻能够正确的划分类别，正确率达到 80%以上，就可以认为该模型基本符合需要。

项目设计

项目设计参考毕业项目报告。