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。[1]

在实际的过程中，需要了解输入的数据是 20NewsGroups 的新闻数据集。通过代码 log 如下：

```
['alt.atheism',
 'comp.graphics',
 'comp.os.ms-windows.misc',
 'comp.sys.ibm.pc.hardware',
 'comp.sys.mac.hardware',
 'comp.windows.x',
 'misc.forsale',
 'rec.autos',
 'rec.motorcycles',
 'rec.sport.baseball',
 'rec.sport.hockey',
 'sci.crypt',
 'sci.electronics',
 'sci.med',
 'sci.space',
 'soc.religion.christian',
 'talk.politics.guns',
 'talk.politics.mideast',
 'talk.politics.misc',
 'talk.religion.misc']
newsgroups_data category numbers: 20
```

总共有 20 类的数据，输入一个数据后， 我们想得到的结果是通过分类模型，这个数据属于这 20 个数据集中的那一个。就是所谓的多分类的问题。

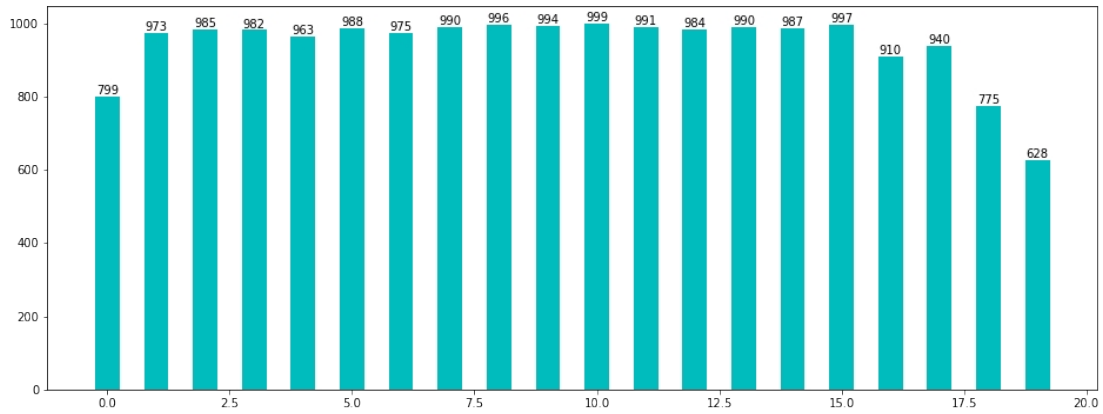
根据输出：当前新闻有 20 个类别；

subset，子类类型，用来设定数据类型：可以设置 train，test，all。

category，类别，就是数据集新闻的类型； 如果设置成 None， 会加载所有的 categories。

remove， 删除新闻的部分内容，主要包括 headers， footers， quotes，这三部分

整个新闻集每一个类的新闻数量都差不多，第一类和最后两类数量能偏少一些，样本分布是均匀的。具体数据通过 log 如下所示：
0 : 799， 1 : 973， 2 : 985， 3 : 982， 4 : 963， 5 : 988， 6 : 975， 7 : 990，
8 : 996， 9 : 994， 10 : 999， 11 : 991， 12 : 984， 13 : 990， 14 : 987， 15 : 997， 16 : 910， 17 : 940， 18 : 775， 19 : 628
更形象化的展示，如下图所示；



在这 20newsgroups 中的训练集中，每一个 category 中，排名靠前的 10 个单词分别如下列所示：

```
alt.atheism: keith it and you in that is to of the
comp.graphics: edu in for it is and graphics of to the
comp.os.ms-windows.misc: file for of and edu is it to the windows
comp.sys.ibm.pc.hardware: card ide is of it drive and scsi to the
comp.sys.mac.hardware: in it is and of edu apple mac to the
comp.windows.x: it mit in motif and is of window to the
misc.forsale: shipping offer of 00 to and edu the for sale
rec.autos: that is you it in of and to car the
rec.motorcycles: dod you it com in of and bike to the
rec.sport.baseball: that is baseball and of in to he edu the
rec.sport.hockey: ca game he team and hockey of in to the
sci.crypt: chip that encryption is and clipper key of to the
sci.electronics: for edu you it in is and of to the
sci.med: edu pitt that it in and is to of the
sci.space: it that is nasa in and to of space the
soc.religion.christian: we it in and is god that to of the
talk.politics.guns: it is you that gun and in of to the
talk.politics.mideast: is you israeli that israel in and to of the
talk.politics.misc: edu it is you and in that of to the
talk.religion.misc: sandvik god you in is that and to of the
```

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。[1]

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。[2]

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。 [3]

我们采用决策树模型，作为基本模型，通过在 newsgroups_train 训练，然后在 newsgroups_test 测试， 得出的结果如下表：

-	Indicater	Value
1	train time	21.662s
2	predict time	0.026s
3	f1_score	0.578
4	log_loss	14.41255632092464
5	accuracy	0.583

决策树分类器的参数如下图:

```
DecisionTreeClassifier(class_weight=None, criterion='gini', max_depth=None,
            max_features=None, max_leaf_nodes=None,
            min_impurity_decrease=0.0, min_impurity_split=None,
            min_samples_leaf=1, min_samples_split=2,
            min_weight_fraction_leaf=0.0, presort=False, random_state=42,
            splitter='best')
train time: 21.662s
predict time:  0.026s
f1_score:   0.578
log_loss: 14.412556320924647
accuracy:   0.583
```

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，远大于基准模型的 58%;　就可以认为该模型基本符合需要。

accuracy_score 这个指标的意思是：　准确度分类得分。
在多标签分类中，此函数计算子集精度：为样本预测的标签集必须与 y_true 中的相应标签集完全匹配。[4]

log_loss 对数损失，又称逻辑损失或交叉熵损失。

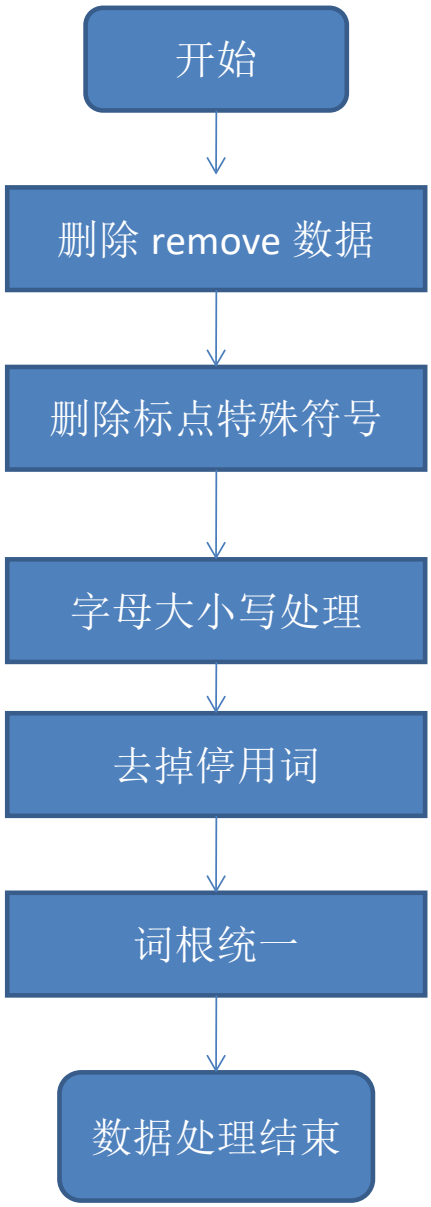
这是（多项式）逻辑回归中使用的损失函数及其扩展，例如神经网络，定义为给定概率分类器预测的真实标签的负对数似然。仅针对两个或更多标签定义了日志丢失。对于具有{0,1}中的真实标签 yt 和 yt = 1 的估计概率 yp 的单个样本，对数损失为

-log P（yt | yp）= -　（yt log（yp）+（1 - yt）log（1 - yp））　[5]

项目设计

项目设计主要包括如下内容：

首先是数据处理，数据处理包含以下的步骤：



接下来，需要对 20newsgroups 数据清理后得到 tokens 进行文本表示，主要 2 种表示方案，词袋子模型表示，和词向量模型表示；

词袋子模型表示，主要的实现函数是 CountVectorizer 和 TfidfVectorizer，这两个函数都来自于 sklearn.feature_extraction.text 这个子模块；该子模块的作用就是收集实用程序从文本文档建立特征向量。

词向量简单的说就是将新闻文档中的单词转化成一个密集向量，对于其中相似的词，其对应的词向量也非常相近。其优点是能表达出词之间的相似关系，还可以包含更多的信息。

目前，转向于 Word2Vec 模型。 Word2Vec 模型是一群用来产生词向量的相关模型， 通过这些模型，实现词向量模型表示每篇文档。操作过程就是利用文本数据对词向量进行训练，将每个单词表示成向量形式。另外，词向量训练后需要进行简单评测，比如检验一些单词之间相似性是否符合逻辑。

采用相关的分类器，对文本模型进行分类；

针对词袋子模型，我们要选用如下 3 种分类器模型

1. 决策树模型
2. 支持向量机 SVM 模型分类器
3. 朴素贝叶斯模型分类器

通过 GridSearchCV 网格搜索调优，GridSearchCV，它存在的意义就是自动调参，只要把参数输进去，就能给出最优化的结果和参数。但是这个方法适合于小数据集，一旦数据的量级上去了，很难得出结果。这个时候就是需要动脑筋了。数据量比较大的时候可以使用一个快速调优的方法——坐标下降。它其实是一种贪心算法：拿当前对模型影响最大的参数调优，直到最优化；再拿下一个影响最大的参数调优，如此下去，直到所有的参数调整完毕。这个方法的缺点就是可能会调到局部最优而不是全局最优，但是省时间省力，巨大的优势面前，还是试一试吧，后续可以再拿 bagging 再优化。[6]

词向量模型分类器主要采用 2 种方案

采用 Keras 的 LSTM 来实现

参考文献

1. 参考 udacity’s document_classification
2. https://baike.baidu.com/item/%E5%86%B3%E7%AD%96%E6%A0%91%E7%AE%97%E6%B3%95/8595872?fr=aladdin
3. 李航. 统计学习方法[M]. 清华大学出版社, 2012
4. http://scikit-learn.org/stable/modules/generated/sklearn.metrics.accuracy_score.html
5. http://scikit-learn.org/stable/modules/generated/sklearn.metrics.log_loss.html
6. https://blog.csdn.net/yanhx1204/article/details/79498626