Machine Learning Engineer Nanodegree

Capstone Proposal

杨铁龙

June 16th, 2018

Proposal

项目背景

自然语言处理(NLP)属于人工智能的分支，作用于将自言语言转化为计算机语言进行处理，以及将计算机语言转化为自然语言进行表达[1]。自然语言在现代人工智能产品中应用广泛，尤其在智能语音方面，自然语言成为人机交互的纽带。当前自然语言处理仍面临诸多挑战，文章的表达分析是其中之一。单词边界、词义以及行为推里都是影响文章推理问题的因素，统计语言中统计概率模型，但是难以确认词与词的关联性，影响到文本分析的准确性[2]。本文将尝试Word2Vec模型分析并表示样本数据中的每篇文章，然后作为输入向量，进行训练，解决文章的表达分析问题。

问题描述

使用Word2Vec模型分析并表示样本数据中的每篇文章，需要考虑到数据的解析问题，准确把握数据的分类特征，以及Word2Vec模型分析中文章样本作为参数输入和调整。另外还需考虑Word2Vec模型分析完文章后得到的词向量需要匹配机器学习的分类模型进行训练和参数调整。最后选择合适的评估模型进行性能评估。

数据或输入

分类的文本数据为经典的20类新闻包，包含约20000条新闻，较均衡的分成了20类，是较常用的文本分类数据之一。项目中可以通过 `sklearn.datasets.fetch_20newsgroups` 方法进行引用，获取数据时通过指定 subset 为 train 或者 test 来指明数据用来训练还是测试，有效且方便训练模型并进行验证。

解决方法描述

本文将基于Word2Vec模型分析并表示样本数据中的每篇文章，然后作为输入向量，选取合适的分类模型对文本分类，并优化模型并分析其稳健性。分类模型主要包括决策树模型、支持矢量机(SVM)模型、朴素贝叶斯模型和神经网络模型，将从这些分类模型中选取合适的模型进行训练并验证[2]。

基准模型

本文将以Bag-of-words模型(BOW)作为基准文本表示模型。BOW模型对于一个文本分析时，会忽略其中的语法和词序，将其分割并建立一个词集。然后基于词典建立tf-idf向量作为机器学习分类模型的输入向量。

评估标准

本文中Word2Vec模型和Bag-of-words模型作为文本表示模型，选取合适的机器学习分类模型进行训练，完成后分别计算各自的准去率进行性能评估，验证Word2Vec是否确实改善了本文分析推理。

项目设计

本文中将针对词、语句以及文章的表达分析问题，尝试Word2Vec模型分析并表示样本数据中的每篇文章，然后作为输入向量，选取合适的机器学习训练模型进行训练并验证测试，以Bag-of-words模型为基准文本表示模型作为对比，分析Word2Vec模型性能表现，从而得出Word2Vec模型能否确实应对文章的表达分析问题

参考文献

[1] https://zh.wikipedia.org/wiki/自然语言处理

[2] https://github.com/nd009/capstone/tree/master/document_classification