PowerAI 人工智能马拉松编程大赛

Hackathon for Finance Industry

作品展示-第五组

（大赛三等奖）

CSDN Power Systems

作品展示

- CNN 在这样的NLP的表现可能会比较好，因为CNN可以对周围数据有的很好的抽取表达，如果卷积范围大了，说不定可以在B中的数据cover A中的对话，再跟上各种pooling 和 激活函数，可能会取得比较好的效果。我们在比赛中，也是用了CNN，在十类的分类中有 83% + 的正确率，然而在 情感分类 中正确率很低，跟取随机数差不多（ 30%+ ）。
- 所以，我们对情感分类使用了贝叶斯（这是一个FLAG），流程上大体是这样，对于情感，只抽取B的对话，连接后进行 JIEBA 分词，之后套用 nltk 中的 NaiveBayesClassifier 进行训练 。

作品展示

预处理： 我们Train data是对每篇文章的每个词的判断是否在文章中。
例如有文章是 你好，那这个文章的特征就是除了你好这个value 是True之外，其他词典的词对于这篇文章的贡献都是 False。可以是文章理解为一个对于词典的稀疏矩阵，value 只有 True 和 False。

训练

1、 我们计算 label_probdist，就是每个 label 文章的频次 或者 频率 。
2、 计算 feature_probdist，即 每个 label 下的词语的先验概率。

测试

之后的测试数据我们也按照词典构建稀疏矩阵，并利用模型进行判别。用贝叶斯我们很惊讶的发现对于十类多分类有 **90+%**正确率，对于情感有**60+%**正确率。
在最后的提交当中，我们本想用深度学习的模型，但是由于前期准备不足...模型只会save, 不会load，那些embedding不会用。。所以及时上了贝叶斯。
十类正确率有 92+， 情感 65+。
不过我们直接用的NLTK包，这个包对于构建稀疏矩阵的冗余性太高了，可以直接用HashMap我感觉。可能是由于这个包已经很老了。。。
我们能达到那么好可能还有一个原因是测试数据的分布太平均了，导致先验概率的贡献非常对，如果测试数据一致偏向某一类，则会导致朴素贝叶斯的判别存在问题，所以也是非常的幸运。

个人评价

- 在我们使用深度学习模块Keras训练中，PowerAI展现了强大的计算能力。在我自己mac电脑需要几分钟的计算中，PowerAI只需要十几秒就完成了。然而，由于我们的准备不足，最后还是上了朴素贝叶斯。
- 在这样的情况下，我们的模型遗憾放弃了流行的Deep Learning算法，从概率的角度出发，结合类别和词频的先验概率，进行了一些简单的数据预处理后，也取得了满意的效果。
- 当然，给我们比较大的感想是对于NLP问题，往往可以结合传统方法与深度学习一起考虑，在深度学习参数不确定的情况下，先用传统方法达到一个满意的效果，再调节深度学习的模型以期突破传统模型的瓶颈，发挥其优势。

团队风采

黄楷: 秒针系统 高级研发工程师
张俊飞：维信金科 高级软件架构师
何海亮：聚淘百恒网络 技术总监
王传健：平安健康 高级软件工程师