影评情感分类

问题描述

给定已标注正面或负面的影评数据 50k 条，使用 Logistics 回归等可以标示每个词情感权重的方法，对影评做情感分类。

方案

数据集中已经给出了很多中间数据和预处理过的数据，分别包括词典、各影评在词典中出现的次数等。

最简单的方案

设 `m` 为字典中单词数量，直接使用词频，则每个影评为一个 `m` 维的向量，`N` 个样本组成 N×m 的矩阵，直接使用 Logistics 回归。

结果

	precision	recall	f1-score	support
neg	0.86	0.88	0.87	12500
pos	0.87	0.86	0.87	12500
micro avg	0.87	0.87	0.87	25000
macro avg	0.87	0.87	0.87	25000
weighted avg	0.87	0.87	0.87	25000

分类准确率为 87%，已经是比较高的准确率了，但是应该有可以优化的空间

词频预处理

想象一下，对于两个长度不一致的文本而言，一个单词相同的词频在两文中的意义并不完全一样，需要对其进行归一化处理，才更加有意义。所以这里对词频除以文本的词总数，得到归一化后的结果，此时词频大小在[0, 1]中。

结果

	precision	recall	f1-score	support
neg	0.86	0.85	0.86	12500
pos	0.85	0.87	0.86	12500
avg / total	0.86	0.86	0.86	25000

准确率为 86%，比预处理之前下降了。

尝试一下别的分类算法

使用 朴素贝叶斯 算法对简单方案中的数据进行的分类

结果

	precision	recall	f1-score	support
neg	0.78	0.88	0.83	12500
pos	0.86	0.75	0.80	12500
avg / total	0.82	0.81	0.81	25000

结果更差。

结论

由于时间限制，没有对 Logistics 做更多的优化尝试，不过即便如此，其分类准确率也比其他几种算法优秀。不过比较不解的是做了词频预处理，准确率反倒下降了，在预期效果之外。