电影评论评分预测作业报告

1401111377 徐威迪　1401214384林萍萍

1　方法框架

给定一条电影评论，要求预测其评分。没有提供用户信息，只有文本信息本身可用。显然这是一个回归/分类问题，关键在于特征的选取和回归模型的选择。而特征的选取中，又以情感词的利用最为关键。

1.1　特征的选取

我们一共使用了8维特征，见“map-not-ext-slt-norm-test”和“mle-not-ext-slt-norm-test”（最后一列是label，未知）。

导演的评分　将movies.xml中每部电影的评分作为导演的评分。我们从主观上可以接受这样的假设：一个导演的不同作品的水平应该相当。

电影的总体评分　将train.xml中每部电影的所有评分的平均作为一维特征。一般来说，电影的总体评分能提供一定的先验知识。

评论的长度　评论的长度从一定程度上能体现评论者的情感。

情感词　我们的模型基于这样一个假设前提：所有的情感词之间是独立不相关的。

– 首先根据信息论中的互信息这一定义来提取可能的情感词。
考虑unigram/bigram/trigram三个粒度。每个元组的互信息定义如等式（1）：

$$
\begin{aligned}
p(m) &= \frac{\sum_{w\in m} f(w,m)}{\sum_{m\in M}\sum_{w\in m} f(w,m)} \\
p(w) &= \frac{\sum_{m\in M} f(w,m)}{\sum_{m\in M}\sum_{w\in m} f(w,m)} \\
p(w,m) &= \frac{f(w,m)}{\sum_{m\in M}\sum_{w\in m} f(w,m)} \\
mutals(w) &= \sum_{m\in M} p(w,m)\log\frac{p(w,m)}{p(w)p(m)} + (p(m)-p(w,m))\log\frac{p(m)-p(w,m)}{(1-p(w))p(m)}
\end{aligned}
\tag{1}
$$

其中m表示一部电影，w表示一个unigram（或bigram或trigram），$f(w,m)$表示w在m中出现的次数。

– 从训练数据中提取形容词。
  形容词也是情感词的一大主要组成部分。
– 对前2步中得到的特征词进行人工过滤：过滤非情感词以及极性情感词。
  极性词本身模棱两可、但可能统计出来的频率呈一边倒，将会导致较大的误差。如“很”、“非常”、“特别”、“太”、“极其”、“最”、“尼玛”的一些词组等。
  “mutals-and-adj”保存了最后得到的情感词。第二列是该词出现在每一种评分中的次数，最后一列为总次数。unigram保留了约50%，bigram保留了约30%。
– 对每一条评论，贪心匹配第3步中过滤剩下的词（贪心策略是选择能匹配到的长度最长的情感词），并根据等式（2）和（3）分别尝试两种方法来利用情感词信息。等式（2）计算该评论属于1-5 各评分分类的最大后验概率，等式（3）计算该评论属于各评分分类的最大似然函数值。等式（2）或（3）均得到的是5 维特征。其中r代表一条评论，s取值为1,2,3,4,5，$f(w,s)(s=1,2,3,4,5)$表示词w在评分为s的评论中出现的总次数。

$$MAP(r,s) \propto \sum_{w \in r} \log \frac{p(s|w)}{p(s)} + \log p(s) \tag{2}$$

$$MLE(r,s) \propto \sum_{w \in r} \log \frac{p(s|w)}{p(s)} \tag{3}$$

  等式（4）定义w 关于五类评分的分布为通过先验概率是Dirichlet 的后验概率进行平滑后的概率分布。

$$p(s|w) = \frac{f(w,s)+1}{\sum\limits_{s=1,2,3,4,5} f(w,s)+5} \tag{4}$$

其他经过尝试但未被采用的特征

– 传统的“one-hot”特征，每一维是对应的词在该评论里的频数，正则化使得和为1。很明显的缺点就是特征向量太稀疏，学习不到信息。
– 利用词向量（word embedding）做近义词扩展，但最后发现效果不如没有进行扩展的。这是因为词向量只考虑词对的共现信息，但不能区分情感倾向，比如“讨厌”和“喜欢”利用词向量计算的相似度也会很大。因此词向量不适合本次作业的任务。
– 目前有一些可以直接训练带情感信息的word embedding，但是考虑到训练数据量不够，无法得到较好的结果。

1.2 算法流程

主要步骤如下：

– 预处理
  去停用词、英文情感词人工翻译成中文（I just love it/very boring/cool/ cute /marvelous等有用的词。人工寻找出训练集里所有的这种词并对应成中文）、利用开源分词工具*结巴分词*进行分词。

– 提取情感词
  如1.1中所述。
– 训练。
  我们主要考虑了决策树分类模型、随机森林回归模型和svm分类。
  决策树模型：找到最大$p(s|r)$对应的分数s，如果和平均评分差别大于2.5，则返回平均评分，否则返回s。十折交叉检验测得MAP的MSE等于0.773，MLE的MSE等于0.823。
  随机森林回归模型：使用开源工具*weka*，设置不同的随机特征数，测得MAP的MSE等于0.535，MLE的MSE等于0.538。
  svm分类：使用*libsvm*，分类函数选择C_SVC，核函数选择RBF，测得MAP的MSE等于0.699，MLE的MSE等于0.703。
– 测试
  根据上述结果，我们决定选用随机森林模型来预测结果。
  MAP的预测结果保存在map.txt中。MLE的预测结果保存在mle.txt中。

2 分工情况

共同讨论出算法。
  徐威迪实现特征提取算法。
  林萍萍完成人工过滤、模型训练、报告撰写。

3 编译/运行环境

python2.7