基于朴素贝叶斯的评论有用性分析

赵海臣

背景

- 为了帮助用户在过载的评论中找到有用评论，需要开发针对评论的预制筛选器，同时，对提供有价值评论的用户进行奖励。
- 已经设计了一套基于规则的评论评分算法，但是人工成分过多，基于大数据的机器学习/统计的算法能够更多地脱离研发人员主观看法而更符合广大用户的客观看法。

基于单词分析的文本分类

- 基于单词分析的文本分类
  - 假设：不同类别的文本之间的单词使用是有明显区别的
  - 途径：文本分类通过分析文本的单词组成来判断
  - 数据来源：将所有评论分解成BOW向量，基于朴素贝叶斯公式计算出对应的类先验概率、类条件概率。
  - 算法类型：朴素贝叶斯
  - 预测方法：在预测时先对评论分解成BOW向量，再调用计算好的类先验概率、类条件概率来预测评论所属类别。

训练数据来源

- 训练数据两种来源：
  - 人工标注：
    - 通过人工的方式审视每一条评论，对评论进行打分，从而获得训练数据。缺点是，人工标注很容易引入标注者的主观色彩，与标注者的性格、心情息息相关。
  - 基于用户反馈的标注：
    - 评论下有“觉得该评论有用”的用户选项，可以基于“觉得有用”的用户数/所有浏览的用户数计算出一个评论的有用性：

$$\text{评论价值性} = \frac{\text{显式觉得有用的用户数}}{\text{所有浏览用户数}}$$

      - 选择评论价值性最高的top20%评论作为正面标签。
      - 选择评论价值性最低的bottom20%评论作为负面标签。
    - 更加客观，但是缺点是：bottom20%很难作为负面标签，因为很可能仅仅因为长度不够，而非用词缘故没有获得用户点赞。

贝叶斯定理

计算某个评论有用性分类的贝叶斯定理：

$$P(c|\vec{x}) = \frac{P(c)P(\vec{x}|c)}{P(\vec{x})}$$

其中，c是对应的评价有用性分类，$\vec{x}$ 是评论对应的BOW(Bag Of Words)向量。

- P(c)是类先验概率
  - 根据大数定律，当训练集包含充足的独立同分布(IID)样本时，P(c)可以通过各类样本出现的频率来进行估计。
- $P(\vec{x}|c)$ 是类条件概率(似然)
  - 由于涉及关于 $\vec{x}$ 所有属性联合概率，不能直接根据样本来估计。

朴素贝叶斯

- 朴素贝叶斯采用"属性条件独立性假设"：对已知类别，假设所有属性相互独立，换言之，假设每个属性独立地对分类结果发生影响。

$$P(\vec{x}\,|\,c)=\prod_{i=1}^{d}P(x_i\,|\,c)$$

- 按照属性条件独立性假设，贝叶斯定理可以写成(朴素贝叶斯)：

$$P(c\,|\,\vec{x})=\frac{P(c)P(\vec{x}\,|\,c)}{P(\vec{x})}=\frac{P(c)}{P(\vec{x})}\prod_{i=1}^{d}P(x_i\,|\,c)$$

- 使用朴素贝叶斯公式计算出 $\vec{x}$ 下各个类别的概率，选出概率最大的一个类别作为 $\vec{x}$ 的类别：

$$c_{\vec{x}}=\arg\max_{c}P(c)\prod_{i=1}^{d}P(x_i\,|\,c)$$

概率值的平滑

若有些属性 x_i 在训练数据的某个类c中未曾出现，则

$$P(\vec{x} \mid c) = \prod_{i=1}^{d} P(x_i \mid c) = 0$$

从而将对应的类条件概率“抹去”，为了避免此类现象，常用拉普拉斯修正(Laplacian correction)来平滑：

- 类先验概率统计修正为：$\hat{P}(c) = \frac{|D_c| + 1}{|D| + N}$
  - 其中，N为训练集D中的总类别数
- 类条件概率统计修正为：$\hat{P}(x_i \mid c) = \frac{|D_{c,x_i}| + 1}{|D_c| + N_i}$
  - 其中，Ni为第i个属性可能的取值数

拉普拉斯修正避免了因训练集样本不充分而导致概率估值为零的问题，在训练集变大时，修正过程导致的偏差也会逐渐变得可忽略不计。

算法过程

- 将所有评论进行分词，构建BOW向量映射模型
- 统计各个类(“有用评论”，“无用评论”，“垃圾评论”)的频次，并存于数据库之中： D_c
- 统计各个类下各个词的频次，并存于数据库之中： D_{c,x_i}
- 对需要预测的评论进行分词，分解成BOW向量
- 基于评论的词调用数据库中现有的频次数据，使用朴素贝叶斯公式预测类

$$c_{\vec{x}} = \arg\max_c \hat{P}(c)\prod_{i=1}^{d}\hat{P}(x_i \mid c)$$

$$= \arg\max_c \frac{|D_c|+1}{|D|+N}\prod_{i=1}^{d}\frac{|D_{c,x_i}|+1}{|D_c|+N_i}$$

  - ✪ 若对应的Dc不存在，或者Dcxi不存在，则Dc=0或Dcxi=0

THE END

THANK YOU!