基于线性回归的评论有效性

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。

基于评论特征的文本有效性

- 基于评论特征的文本有效性
  - 假设：评论是否有效性是基于一系列客观条件决定的，例如评论文本属性、评论这属性和店铺属性等
  - 途径：通过分析判断评论的客观条件来判断该评论的有效性
  - 数据来源：通过数据清洗、数据统计获得各个评论的一系列客观指标
  - 算法类型：多元线性回归模型
  - 预测方法：通过训练获得多元线性回归的参数，新的评论先获得客观指标，再使用线性回归模型进行预测。

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：

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

属性	子属性	描述
评论属性	rev_len	评论长度
	rev_sen_num	评论短句子数
	rev_word_num	评论词数
	rev_matched_word	匹配标签数
	rev_compo_num	评论词性数
	rev_a_num	评论形容词数
	rev_v_num	评论动词数
	rev_n_num	评论名词数
	rev_d_num	评论副词数
用户属性	user_avg_star	用户平均评论价值
	user_rev_num	用户评论数
	user_duration	用户注册时长
	user_reci_star	用户收获的所有星数
商品属性	product_rev_num	产品评论数
	product_price	产品价格
	product_rate	产品好评率

多元线性回归

- 纯线性回归：

$$helpful = \partial_0 * rev\_len + \partial_1 * rev\_sen\_num + ... + \partial_n * product\_rate + \beta + \varepsilon$$

  - 其中 $\partial_0, \partial_1, \partial_2 ... \partial_n, \beta$ 是回归系数
  - ε 是误差项

- 非线性化回归

$$\log(helpful) = \partial_0 * \log(rev\_len) + \partial_1 * \log(rev\_sen\_num) + ...$$
$$+ \partial_n * \log(product\_rate) + \beta + \varepsilon$$

  - 其中 $\partial_0, \partial_1, \partial_2 ... \partial_n, \beta$ 是回归系数
  - ε 是误差项

THE END

THANK YOU!