O2O优惠券使用预测

案例网址：
https://tianchi.aliyun.com/notebook-ai/detail?spm=5176.12281897.0.0.2fa139a97O76ma&postId=4796
代码及数据网盘链接：
https://disk.pku.edu.cn/#/link/3C6E894E2CEC4BAC0AFB999768E33525

汇报人：季佳雯

（1800016635）

目录

- 背景
- 数据集
- 数据预处理方法
- 机器学习算法
- 主要结果
- 总结

案例背景

- 天池新人实战赛：O2O优惠券使用预测
- 本赛题提供用户在2016年1月1日至2016年6月30日之间真实线上线下消费行为，预测用户在2016年7月领取优惠券后15天以内的使用情况。
- 本赛题目标是预测投放的优惠券是否核销，使用优惠券核销预测的平均AUC（ROC曲线下面积）作为评价标准。 即对每个优惠券coupon_id单独计算核销预测的AUC值，再对所有优惠券的AUC值求平均作为最终的评价标准。

数据集

- 赛题提供三个数据集
  - 用户线下消费和优惠券领取行为
  - 用户线上点击/消费和优惠券领取行为
  - 用户O2O线下优惠券使用预测样本
- 选手提交一个数据集
  - 对预测样本的预测结果

数据集——字段表

Table 1: 用户线下消费和优惠券领取行为

Field	Description
User_id	用户ID
Merchant_id	商户ID
Coupon_id	优惠券ID：null表示无优惠券消费，此时Discount_rate和Date_received字段无意义
Discount_rate	优惠率：x \in [0,1]代表折扣率；x:y表示满x减y。单位是元
Distance	user经常活动的地点离该merchant的最近门店距离是x*500米（如果是连锁店，则取最近的一家门店），x\in[0,10]；null表示无此信息，0表示低于500米，10表示大于5公里；
Date_received	领取优惠券日期
Date	消费日期：如果Date=null & Coupon_id != null，该记录表示领取优惠券但没有使用，即负样本；如果Date!=null & Coupon_id = null，则表示普通消费日期；如果Date!=null & Coupon_id != null，则表示用优惠券消费日期，即正样本；

Table 2: 用户线上点击/消费和优惠券领取行为

Field	Description
User_id	用户ID
Merchant_id	商户ID
Action	0 点击， 1购买，2领取优惠券
Coupon_id	优惠券ID：null表示无优惠券消费，此时Discount_rate和Date_received字段无意义。“fixed”表示该交易是限时低价活动。
Discount_rate	优惠率：x \in [0,1]代表折扣率；x:y表示满x减y；“fixed”表示低价限时优惠；
Date_received	领取优惠券日期
Date	消费日期：如果Date=null & Coupon_id != null，该记录表示领取优惠券但没有使用；如果Date!=null & Coupon_id = null，则表示普通消费日期；如果Date!=null & Coupon_id != null，则表示用优惠券消费日期；

数据集——字段表

Table 3：用户O2O线下优惠券使用预测样本

Field	Description
User_id	用户ID
Merchant_id	商户ID
Coupon_id	优惠券ID
Discount_rate	优惠率：x \in [0,1]代表折扣率；x:y表示满x减y.
Distance	user经常活动的地点离该merchant的最近门店距离是x*500米（如果是连锁店，则取最近的一家门店），x\in[0,10]；null表示无此信息，0表示低于500米，10表示大于5公里；
Date_received	领取优惠券日期

Table 4：选手提交文件字段，其中user_id,coupon_id和date_received均来自Table 3,而P

Field	Description
User_id	用户ID
Coupon_id	优惠券ID
Date_received	领取优惠券日期
Probability	15天内用券概率，由参赛选手给出

数据预处理方法

- 1、数据导入与简单分析
- 2、将满减类型优惠券变成折扣率形式，将距离从str转变成int类型（将空值null替换成-1）
- 3、时间：领优惠券的日期和消费的日期
  - 日期的范围
  - 计算每日领取优惠券的数目、以及使用该优惠券消费的数目、该日的优惠券使用率
  - 为每条消费记录新建星期特征，并分成工作日和周末两种类型
  - 数据标注：新增属性label，将数据分成三类——未领取优惠券、领取优惠券日期和消费日期相隔15天以内、其他

机器学习算法

- 随机梯度下降法（Stochastic gradient descent, SGD）-1
  - 使用 discount, distance, weekday 类的特征。（14个）
  - 训练集train/验证集valid 的划分：用20160101到20160515的作为train，20160516到20160615作为valid。
  - 利用sklearn中的SGDClassifier 构造线性模型 ，在训练集和验证集上进行训练与测试

机器学习算法

- 随机梯度下降法-2
  - 对用户进行特征提取，加入用于预测的属性（原始+用户特征）
  - 同样进行训练集/测试集的划分、模型的训练与预测
- 随机梯度下降法-3
  - 对商户进行特征提取，加入用于预测的属性（原始+用户+商户）
  - 同样进行训练集/测试集的划分、模型的训练与预测
- 随机梯度下降法-4
  - 对用户和商户进行联合特征提取，加入用于预测的属性（原始+用户+商户+联合）
  - 同样进行训练集/测试集的划分、模型的训练与预测

机器学习算法

- 随机梯度下降法-5
  - 将特征提取写成函数，重复前面过程，但是重新划分数据
  - trainSub, validSub = train_test_split(train, test_size = 0.2, stratify = train['label'], random_state=100)
  - 模型以及训练、预测过程不变

机器学习算法

- 梯度提升决策树（Gradient Boosting Decision Tree，GBDT）
  - 用于预测的属性包括原始的属性、所提取的用户特征、商户特征以及用户与商户的联合特征
  - 训练集/测试集和“随机梯度下降法-5”的方法一样
  - 使用lightgbm中的LGBMClassifier构造模型

主要结果-AUC得分

模型	描述	AUC
随机梯度下降法-1	使用原始属性进行预测	0.535
随机梯度下降法-2	在1的基础上增加用户特征	0.599
随机梯度下降法-3	在2的基础上增加商户特征	0.602
随机梯度下降法-4	在3的基础上增加用户和商户的联合特征	0.616
随机梯度下降法-5	在4的基础上重新划分训练集和验证集	0.621
梯度提升决策树	在5的基础上更换模型	0.631

总结

- 案例对线上线下消费行为数据进行一定的处理，尝试了不同的数据划分方式、特征提取方式以及预测模型。
- 可以看出，用于预测的特征越多，预测效果越好。
- 对样本分层抽样，比按照时间划分训练集/验证集效果更好。
- 算法未来可以从数据的划分、特征的提取以及不同模型的角度进行改进。（例如使用“线上消费行为”样本提取一些特征作为补充）