解题思路

陈祈宇

目录

[1.缩小商品子集item的范围 1](#_Toc24843)

[2.特征工程设计 2](#_Toc17224)

[1)用户特征（2） 3](#_Toc10505)

[2)三级品类特征（3+5+10=18） 3](#_Toc17206)

[3)用户对三级品类的行为特征（7+5+10=22） 3](#_Toc23760)

[3.模型训练 3](#_Toc14417)

1.缩小商品子集item的范围

item为给出的商品子集，每个商品属于一个三级品类，我们要预测的是用户点击了商品子集中的哪个三级品类。考虑到一些三级品类可能被点击的概率非常低，为了提高预测准确性，我们可以根据行为数据表中三级品类的点击频率将那些不太可能被点击的三级品类进行过滤。

筛选出行为数据中的所有点击数据（act\_type=1），对这些数据中的三级类目出现次数进行统计，得到每一个三级品类被点击了多少次，并进行排名。结果保存为表格third\_cate\_counts.csv，如图：



总共有2567个三级品类，可以发现，出现次数排名前1000的三级品类，其频率已经占到了97.14%，后面的1567个三级品类被点击的概率非常小。因此考虑将三级类目子集进一步缩小，出现次数不在前1000的三级类目在预测时不予考虑。

再对item商品数据表中不同商品的第三品类出现次数进行统计，共有3448个三级品类，比行为数据中的2567多了近一千，这些三级品类在行为数据表时间段内完全没有被点击过，属于冷门品类。

我们根据行为数据表中得到的三级品类出现频率排名，将item数据集进一步缩小，凡是三级品类不属于前1000的予以过滤。过滤后得到1737个商品共369个三级品类（best\_item.csv），这说明了很多行为数据中出现的热门三级品类并没有出现在商品子集item中，而item中的很多三级品类在行为数据中没有被点击或者点击率很低。

由此我们将tiem中的3448个三级品类缩小至369个，后面我们基于这369个三级品类进行预测。

注意：这样做的前提是item中出现的三级品类都有包括在行为数据考虑范围内。

例如行为数据只考虑了ABCD四个品类的数据，但item中是ABCDEF，那么EF将被当做冷门品类过滤掉，但这是不合理的。（咨询过俊栩师兄后确定这种情况是不存在的）

2.特征工程设计

由于最后是对三级品类而不是商品进行预测，所以一个用户和一个三级品类构成一个数据。

特征工程主要包含三个部分：某用户的特征，某三级品类的特征，该用户对该三级品类的操作行为特征，总共42个特征。

1)用户特征（2）

性别，年龄

2)三级品类特征（3+5+10=18）

该三级品类包含的商品数量，该三级品类所属的一二级品类包含商品的数量；

该三级品类在规定时间段内分别被点击、浏览、加车、删车的次数，以及各种操作总次数；

该三级品类所属的一二级品类在规定时间段内分别被点击、浏览、加车、删车的次数，以及各种操作总次数；

3)用户对三级品类的行为特征（7+5+10=22）

在规定时间段（7天）内该用户每天是否对该三级品类进行操作；

规定时间段内该用户对该三级品类的点击、浏览、加车、删车次数，以及各种操作总次数；

在规定时间段内该用户对该三级品类所属的一二级品类的点击、浏览、加车、删车的次数，以及各种操作总次数；

3.模型训练

行为数据中仅保留best\_item.csv中存在的三级品类数据。统计过去一周存在的所有（用户，三级品类）组合，每个组合构成一条数据，根据特征工程提取相应特征。并根据未来一周该组合是否存在点击行为得到标签。

对['2017-07-05','2017-07-12','2017-07-19','2017-07-26','2017-08-02','2017-08-09','2017-08-16']开始的每一周提取训练特征，在训练时将6周作为训练集，1周作为验证集。应用xgboost进行模型训练。

对2017-08-25至2017-08-31提取预测特征，以此预测九月份第一周的点击情况。

在得到预测结果后，根据预先设置的阈值进行过滤，仅保留高于阈值的预测结果，作为最终提交结果。