**阿里天池：商铺定位比赛**

**赛题数据信息**

店铺和商场信息表：店铺ID、店铺类型ID、店铺经纬度（脱敏，相对距离依然可信）、人均消费指数、店铺所在商场ID

用户在店铺内交易表：用户ID、用户所在店铺ID、行为时间戳、行为发生时经纬度、行为发生时周围wifi环境（包括bssid-wifi唯一标识码、signal强度，数值越大表示信号越强，flag是否连接，true表示连接）

赛题业务场景是各大商场中用户使用支付宝消费的常见现象，然而有事存在消费店铺信息丢失等问题。赛题提供了过去一段时间内（2个月）500个商场的消费记录，包括用户行为与商铺信息，要求预测接下来一段时间（2周）内的消费记录对应的真实店铺。显然这是一个分类任务，即可以是多分类，也可以认为是二分类。

评估方式：给出的shop\_id与预测的shop\_id相等

**关于赛题数据的思考：**

* 问题理解：问题可以理解为分类任务。即可以作为作为多分类（每个样本对应的商铺即为真实类别），也可以构造为二分类（构造每个样本对应的商铺为正例，其他商铺为反例），不管哪一种，都需要分商场来做，即可认为每个商场之间是相互独立的。
* 可能需要的数据处理：
  + 首先，样本的本质是消费记录，消费记录可能存在伪造现象：具体体现为某个用户在短时间内在同一个店铺内频繁的消费，处理方法：删除此类数据
  + 第二，由于设备、室内GPS信号不好等原因经纬度可能不准确，例如在同一个商场中存在店铺，其经纬度严重偏离商场经纬度，对于这种经纬度，我们将店铺经纬度与在此店铺内用户消费时的经纬度进行比较，如果整体偏离程度较大，则用用户在此店铺消费时经常出现的经纬度替换店铺的经纬度
  + 第三：Wifi环境复杂，可能存在数值不准，严重偏离正常wifi强度，对于wifi强度异常的wifi采用删除进行处理
  + 第四：在给出的数据中，同一个商场，部分店铺的用户消费记录特别多，部分店铺的用户消费记录特征少，在多分类时造成每类的样本数量相差特别大，多分类时容易把实际消费在稀疏店铺的用户预测在热门店铺
* 特征提取方向：用户属性，时间属性，坐标属性，wifi信息，店铺属性
* 模型选择：作为一个分类任务，典型方法是梯度提升树，工具无非就是xgboost。

**数据划分**

充分利用数据

| **集合** | **特征区间** | **特征区间** |
| --- | --- | --- |
| 线下训练集 | [2017-07-01, 2017-08-18] | [2017-08-19, 2017-08-25] |
| 线下测试集（线上训练集） | [2017-07-01, 2017-08-25] | [2017-08-26, 2017-09-01] |
| 线上测试集 | [2017-07-01, 2017-08-31] | [2017-09-01, 2017-09-14] |

**数据预处理**

wifi特征：首先在同一商场中存在多个bssid、不同商场中有同一bssid

* 对wifi强度为null的数据，直接删除掉，以免造成干扰
* 对于训练集中出现次数小于k此的wifi过滤掉，一定程度可以减少wifi特征的数量
* 对于线上，每个商场会有新出现的wifi，对于此类wifi不做考虑

**构造多分类特征**

**wifi特征：**

* wifi空值处理：取最近经纬度用户的相同wifi来填充
* 记录中是否有连接的wifi记录
* wifi列表排序位置（第一强，第二强，）上bssid出现过的总次数，候选shop在其中的占比
* 连接的bssid在历史出现的总次数，候选shop在其中的占比

**商店特征：**

* 店铺经纬度
* 商场所在mall的总人流量，候选shop人流量在其中的占比
* 商场所在mall每小时人流量，候选shop每小时人流量在其中占比

**用户特征**

* 消费时的经纬度
* 在候选shop的消费次数
* 与候选shop位置的GPS距离（l2）
* 与候选shop历史记录中心位置的GPS距离（l2）

**模型：LR、GBDT、以及基于树的stacking**

**多分类解决方案：**

* 训练集划分：每个mall单独构造训练集
* 候选：所在mall的所有店铺
* 特征：经纬度、wifi强度、wifi强度统计特征、是否连接wifi、是否为空
* 覆盖率（线下）：覆盖率
* 缺点：单个模型样本不多，特征稀疏，模型精度不高
* 优点：样本召回率高，可以作为二分类的候选集

**二分类解决方案：**

**用简单规则构造二分类候选样本**

* 规则1：用户访问过的店铺
* 规则2：wifi有交集的店铺（商店wifi只取出现最多的15个）
* 规则3：和当前gps最近的三个店铺（欧式距离）

**训练集负样本抽样：**

* 规则4：保留全部正样本，负样本随机抽取一定比例加入训练集

**覆盖率：99.1%**

**优点：相对于直接用全部的商场店铺作为候选，用简单规则构造的训练集样本在不降低模型预测精度情况下，大大降低模型训练的负担**

**二分类特征**

**标记特征**

* 记录中是否有连接wifi
* 记录中是否含有null
* 记录中wifi与候选shop出现过的wifi重合的个数

**总量-比例特征**

该mall总历史记录数、候选shop在其中的占比

该user的总历史记录数、候选shop在其中的占比

wifi历史上出现过的总次数，候选shop在其中的占比

在当前排序位置（第一、第二。。）上wifi历史上出现过的总次数、候选shop在其中的比例

连接的wifi出现的总次数、候选shop在其中的占比

将wifi由强到弱进行排序，可生成多个特征

**差值特征**

wifi强度-候选shop历史记录中该wifi的平均强度

wifi强度-候选shop历史记录中该wifi的最大强度

wifi强度-候选shop历史记录中该wifi的最小强度

**距离特征**

与候选shop位置的GPS距离

与候选shop历史记录中心位置的GPS距离

**多分类特征：**

多分类的概率作为特征

**词向量特征**

将wifi按照强度进行排序，并将信号离散化，每个wifi看成一个单词，生成的词向量作为二分类模型的输入

**wifi余弦相似度**

每个样本与候选商店wifi余弦相似度

**LR模型：**

将one-hot过的ID类（包含wifi以及信号强度）作为LR的输入，让LR实现多分类。效果不好、耗时特特别大。原因：对于wifi信号，由于室内wifi互相干扰，wifi信号强度并不是特别准确；另外，对于一个商场做多分类，类别达到100+，由于线性模型本身缺点，对商场中有些店铺并不能明显区分开来（有些店铺用户消费记录中wifi信息很相似，使用线性模型难以进行区分）

**GBDT**

主要使用工具为LightGBM和XGBoost，对于不同商场将店铺预测当成多分类任务训练，两个工具的准确度相近，但LightGBM比XGBoost快得多，后面仅仅利用调参成功地提升了2个百分点（从0.89到0.91级别），所调参数主要为抑制树模型过拟合的参数（例如min\_data\_in\_leaf, feature\_fraction等）

**二阶段stacking：**

stacking分为两个阶段：第一个阶段是多分类方案，旨在构造候选shop，第二阶段是二分类方案，用于模型预测，本方法需要同时兼顾候选的覆盖率和准确率。

* 首先，在第一阶段利用LightGBM实现分mall的多分类任务，并输出各shop概率，得到每个用户最可能去的前5个shop作为候选shop，覆盖率（5个候选shop包含真实标签）大概有96%。
* 然后，在第二阶段，将每一个测试样本扩充5行，每一行对应候选shop，拼接上shop的相关特征，并将真实shop打上正例的标签，作为二分类任务给LightGBM训练。

**数据异常分析**

**恶意刷单**

统计各用户单月消费记录得到，有部分用户在一个月内在同一个店铺内消费了400多单，大多数用户只在某店铺内消费1次到2次，这样的用户被认为是恶意刷单。对于这样的数据是直接去除。

**单值wifi**

部分wifi列表中只有一个wifi bssid，对于这种样本，删除

**个人热点wifi**

在众多WIFI bss中，存在大量只出现过几次的bss，显然不可能是商铺中固有的WIFI热点，而很可能是其他消费者开启的热点（可以到大商场里打开手机WIFI列表看看，确实能搜到很多别人的热点）。**我们的处理是去除出现次数低于20的WIFI bss**（相当于降维）。

**wifi缺失**

WIFI列表中的强度字段为null，这个现象在复赛中特别严重，我们的理解是这些用户都是苹果手机或者没有root权限的安卓手机，支付宝没有权限获取信号强度。**我们的处理是将null替换为均值**。

**用户和商铺的空间关系**

* 原本企图通过空间关系，尝试基于距离还原出WIFI热点的空间分布，然而经过数据的可视化发现经纬度的可信度并不高。
* 部分shop位置非常接近，甚至经纬度重合，推测是上下楼层关系
* 存在远离大部分shop的噪声点，因此经纬度不完全准确

**商铺与用户之间的空间关系**

* 用户在空间上并非呈现围绕着标签shop分布的规律
* 存在大量位置看起来不正常的用户（通过经纬度计算距离店铺上百公里）

**误判原因**

提取的特征不足以对店铺进行区分

* 误判店铺的性质接近（类别、消费水平、历史样本量）
* 大量店铺在训练集中从未出现，而在测试集中大量出现。大部分的错误由此产生。
* **部分店铺样本太少学不到东西，无从判断**

（1）经纬度坐标接近 （2）商店类型相同 （3）训练集样本量接近 （4）样本时间戳分布相近（每天都有，而且每天分布一致） （5）wifi 特征分布接近（bss重合度高） （6）用户冷启动（大多是第一次来，没有历史信息可供参考）

**辨别一个特征是否有用**

* 特征对label是否具有区分度