腾讯社交广告
Tencent Social Ads

2018 腾讯广告算法大赛
参赛手册

题目：相似人群拓展

01 大赛简介

基于社交关系的广告（即社交广告）已成为互联网广告行业中发展最为迅速的广告种类之一。腾讯社交广告平台，依托于腾讯丰富的社交产品，植根于腾讯海量的社交数据，借助强大的数据分析、机器学习和云计算能力打造出了一个服务于千万商家和亿万用户的商业广告平台。我们一直致力于提供精准高效的广告解决方案，而复杂的社交场景，多样的广告形态，以及庞大的用户数据，给实现这一目标带来了不小的挑战。为攻克这些挑战，腾讯社交广告也在不断地寻找出更为优秀的数据挖掘和机器学习算法。

本次算法大赛的题目源于腾讯社交广告业务中的一个真实的广告产品——相似人群拓展（Lookalike）。该产品的目的是基于广告主提供的目标人群，从海量的人群中找出和目标人群相似的其他人群。在实际广告业务应用场景中，Lookalike 能基于广告主已有的消费者，找出和已有消费者相似的潜在消费者，以此有效帮助广告主挖掘新客、拓展业务。目前，腾讯社交广告 Lookalike 相似人群拓展产品以广告主提供的第一方数据及广告投放效果数据（即后文提到的种子包人群）为基础，结合腾讯丰富的数据标签能力，透过深度神经网络挖掘，实现了可在线实时为多个广告主同时拓展具有相似特征的高质潜客的能力。

通过本次大赛，我们旨在挑选出更为优秀 Lookalike 算法以及遴选出杰出的社交广告算法达人。

02

赛制说明

本次大赛分为初赛、复赛和答辩三个环节。

初赛分为初赛 A 和初赛 B 两个阶段，初赛 A 阶段数据可在完成报名后登录个人页面下载，结果提交和初赛 B 与复赛阶段数据下载等功能需要选手通过实名认证后方可操作。每天（中午 12 点开始的 24 小时内，如无特殊约定，涉及的时间均为北京时间，二十四小时制）限提交 3 次结果，系统将实时计算得到此次提交结果的得分，并在个人信息页展示。

初赛开始后，系统将每天进行一次排名。排名基于每天 12:00 前各队伍提交的结果，并按照参赛队伍当前赛事阶段的历史最优成绩从高到低依次排序。排行榜将于每天 15:00 更新，此排行榜仅供参考不作为最终排名计算。

初赛 A 阶段时间为 4 月 18 日 12:00:00 - 5 月 19 日 11:59:59，初赛 B 阶段时间为 5 月 19 日 12:00:00 - 5 月 23 日 11:59:59。AB 阶段的训练集相同，测试集不同，系统将在 5 月 19 日 12:00:00 切换测试集，参赛队伍需要再次下载测试集数据文件。最终初赛成绩排行榜将以初赛 B 阶段各参赛队伍的历史最好成绩进行排名。

初赛结束时，成绩排名前 20%（原则上最多不超过 200 支队伍，但大赛举办方有权根据报名情况等确定最终数量）的队伍进入复赛。

02

赛制说明

复赛阶段（5 月 24 日 12:00:00-6 月 13 日 11:59:59），大赛将会更换一批种子包，同时，加大训练数据量。复赛和初赛类似，区分 A 阶段（5 月 24 日 12:00:00 - 6 月 09 日 11:59:59）和 B 阶段（6 月 09 日 12:00:00 - 6 月 13 日 11:59:59），系统将在 6 月 09 日 12:00:00 切换测试集。

复赛结束时，成绩排名前 10 名（含并列，大赛举办方有权根据复赛情况等确定最终数量）的队伍进入最终答辩环节。

本次大赛将对复赛 B 阶段成绩、答辩成绩和代码进行综合评估，作为最终的比赛成绩。

03 赛题描述

相似人群拓展（Lookalike）基于广告主提供的一个种子人群（又称为种子包），自动计算出与之相似的人群（称为扩展人群）。本题目将为参赛选手提供几百个种子人群、海量候选人群对应的用户特征，以及种子人群对应的广告特征。出于业务数据安全保证的考虑，所有数据均为脱敏处理后的数据。整个数据集分为训练集和测试集。训练集中标定了人群中属于种子包的用户与不属于种子包的用户（即正负样本）。测试集将检测参赛选手的算法能否准确标定测试集中的用户是否属于相应的种子包。训练集和测试集所对应的种子包完全一致。

为了检验参赛选手的算法能否很好地理解用户以及种子人群，本次大赛要求参赛者提交的结果中，提供测试集中各种子包候选用户属于该种子包的得分（得分越高说明候选用户是某个包潜在的扩展用户的可能性越大）。大赛官网的后台算法将自动计算提交结果的得分及排名。详情可参看【评估方式】【提交方式】。

初赛和复赛所提供的种子包除量级有所不同外，其他的设置均相同。

04 数据说明

比赛数据（脱敏后）抽取的时间范围是某连续 30 天的数据。总体而言，数据分为：训练集数据文件、测试集数据文件、用户特征文件以及种子包对应的广告特征文件四部分。

训练集数据文件 train.csv 每行代表一个训练样本，各字段之间由逗号分隔，格式为：“aid,uid,label”。其中，aid 唯一标识一个广告，uid 唯一标识一个用户。样本 label 的取值为 +1 或 -1，其中 +1 表示种子用户，-1 表示非种子用户。为简化问题，一个种子包仅对应一个广告 aid，两者为一一对应的关系。

测试集数据文件 test.csv 每行代表一个训练样本，各字段之间由逗号分隔，格式为：“aid,uid”。字段含义同训练集。

用户特征文件 userFeature.data 每行代表一个用户的特征数据，格式为：“uid|features”，uid 和 features 用竖线“|”分隔。其中 feature 采用 vowpal wabbit（https://github.com/JohnLangford/vowpal_wabbit）格式：“feature_group1|feature_group2|feature_group3|...”。每个 feature_group 代表一个特征组，多个特征组之间也以竖线“|”分隔。一个特征组若包括多个值则以空格分隔，格式为：“feature_group_name fea_name1 fea_name2 …”，其中 fea_name 采用数据编号的格式。用户特征详情见【用户特征说明】。

04 数据说明

广告特征文件 adFeature.csv 格式为：“aid,advertiserId,campaignId,creativeId,creativeSize,adCategoryId,productId,productType”。其中，aid 唯一标识一个广告，其余字段为广告特征，各字段之间由逗号分隔。广告特征详情见【广告特征说明】。

出于数据安全的考虑，我们对 uid、aid、用户特征、广告特征按照如下方式进行加密处理：

- uid：对每个用户 ID 进行 1 到 N 的随机化编号，生成一个不重复的加密 uid，N 为用户总数目（假设用户数为 100w，将所有用户随机打散排列，将其序号作为 uid，取值范围是 [1, 100w]）；
- aid：参考 uid 的加密方式，生成加密后的 aid；
- 用户特征：参考 uid 的加密方式，生成加密后的 fea_name；
- 广告特征：参考 uid 的加密方式，生成加密后的各字段

用户特征说明

用户特征包含以下特征组（feature_group_name），如果具体特征取值未知，均以 0 表示：

- 年龄（age）：分段表示，每个序号表示一个年龄分段
- 性别（gender）：男 / 女
- 婚姻状况（marriageStatus）：单身 / 已婚等状态（多个状态可共存）

数据说明

- 学历（education）：博士 / 硕士 / 本科 / 高中 / 初中 / 小学
- 消费能力（consumptionAbility）：高 / 低
- 地理位置（LBS）：每个序号代表一个地理位置
- 兴趣类目（interest）：由不同数据源挖掘得到的 5 个特征组，分别以 interest1, interest2, interest3, interest4, interest5 表示，每个兴趣特征组包含若干个兴趣 ID
- 关键词（keyword）：较兴趣类目更细粒度地表示用户喜好，由不同数据源挖掘得到的 3 个特征组，分别以 kw1, kw2, kw3 表示，每个关键词特征组包含若干用户感兴趣的关键词
- 主题（topic）：使用 LDA 挖掘的用户喜好主题，由不同数据源挖掘得到的 3 个特征组，分别以 topic1, topic2, topic3 表示
- APP 近期安装行为（appIdInstall）：63 天内安装的 APP，每个 APP 表示为一个唯一的 ID
- APP 活跃（appIdAction）：用户使用的活跃 APP
- 上网连接类型（ct）：WIFI/2G/3G/4G
- 操作系统（os）：Android/IOS，不区分版本号
- 移动运营商（carrier）：移动运营商，移动 / 联通 / 电信 / 其他
- 有房（house）：是否有房

04 数据说明

广告特征说明

- 广告 ID（aid）：广告是指广告主创建的广告创意（或称广告素材）及广告展示相关设置，包含广告的基本信息（广告名称、投放时间等）、广告的推广目标、投放平台、投放的广告规格、所投放的广告创意、广告的受众（即广告的定向设置）以及广告出价等信息
- 广告主 ID（advertiserId）：账户结构分为四级：账户——推广计划——广告——素材，账户和广告主是一一对应关系
- 推广计划 ID（campaignId）：推广计划是广告的集合（类似电脑文件夹功能），广告主可以将推广平台、预算限额、是否匀速投放等条件相同的广告放在同一个推广计划中，方便管理
- 素材 ID（creativeId）：展示给用户直接看到的广告内容，一条广告下可以有多组素材
- 素材大小（creativeSize）：素材大小 ID，标识广告素材不同大小
- 广告类目（adCategoryId）：广告分类 ID，使用广告分类体系
- 商品 ID（productId）：推广的商品 ID，系统中用 product id 来标识
- 商品类型（productType）：广告投放目标对应的商品类型（如京东 -- 商品、app-- 下载）

05 评估方式

对于扩展后的相似用户，如果在广告投放上有相关的效果行为（点击或者转化），则认为是正例；如果不产生效果行为，则认为是负例。

每个待评估的种子包会提供如下信息：种子包对应的广告 aid 及其特征，以及对应的候选用户集合（uid 及其特征）。选手需要为每个种子包计算测试集中用户的得分，比赛会据此计算每个种子包的 AUC 指标，AUC_i 表示第 i 个包的 AUC 值，并以所有待评估的 m 个种子包的平均 AUC 作为最终的评估指标：

$$\frac{1}{m}\sum_{i=1}^{m} AUC_i$$

06 提交方式

选手提交结果为一个 submission.csv 文件，编码采用无 BOM 的 UTF-8，格式如下：每行记录表示该用户在该种子包中的得分，各字段用逗号分隔，中间无空格。注意，score 字段有效数字不得超过 8 位，以防止上传的结果文件过大。

提交文件参考如下示例：

```
aid,uid,score
100,10000000,0.62124588
101,10000000,0.33245682
102,10000001,0.46814249
……
```