玉古路38号_

TAAC
腾讯广告算法大赛
Tencent Advertising Algorithm Competition

Contents

目录

02

赛题理解

赛题描述/问题分析

2.1 赛题描述

赛题题目：广告受众基础属性预估

输入：用户在长度为 91 天（3 个月）时间窗口内的广告点击历史记录

输出：用户的年龄段（共10类）和性别

2.2 问题分析

问题转化：脱敏场景下的文本分类问题

评价指标：年龄段准确率+性别准确率

难点：

- 各用户点击序列程度分布不均，整体呈现长尾分布；
- 广告id词表过大（例如复赛数据中的creative id有40w+）；
- 测试集中出现大量未登录id(例如复赛数据集中的creative id有15.84%的id未在训练集中出现过)。

用户点击历史记录（clock_log.csv）

- user_id:用户id
- creative_id:广告素材id
- time:天粒度时间
- click_times：当天点击次数

广告属性(ad.csv)

- creative_id:广告素材id
- ad_id:广告id
- product_id:广告产品id
- product_category:广告类别id
- advertisr_id:广告主id
- Industry:广告行业id

特征工程

点击序列构造/目标编码

3.1 点击序列构造

原始数据：用户在长度为 91 天（3 个月）的时间窗口内的广告点击历史记录,主要字段：
user_id,time,creative_id,click_times

序列构造：

- 异常数据（click_times > 20）去除；
- 时间粒度为天，天内点击序列无明确时间顺序，按click_times进行排序；
- 各用户每日点击序列长度分布不均，对用户每日点击序列进行截断，截断长度66， 占比95%；
- 各用户91天点击序列长度分布不均，对其进行截断，截断长度512，占比90%。

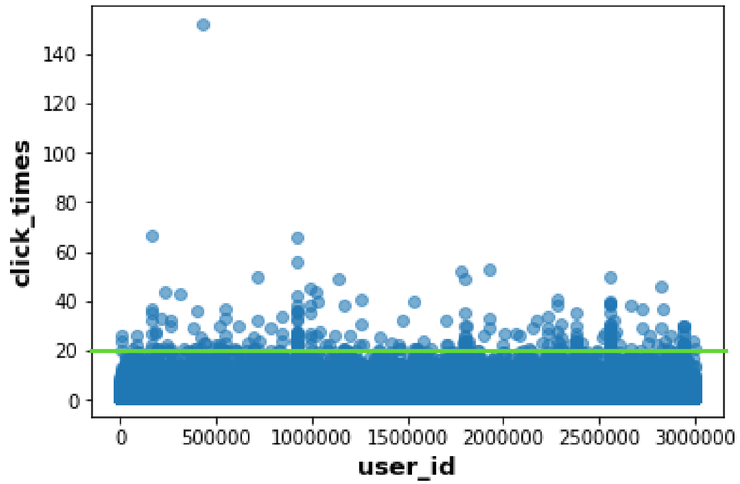


图1：用户单个广告点击次数分布图

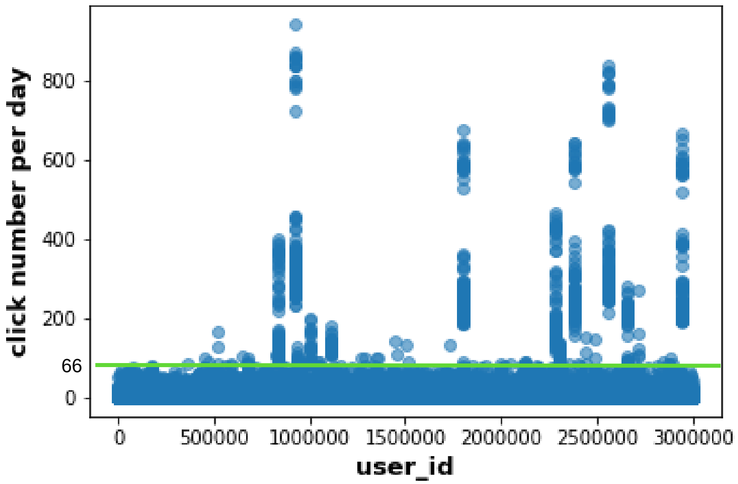


图2：用户单天广告点击数量分布图

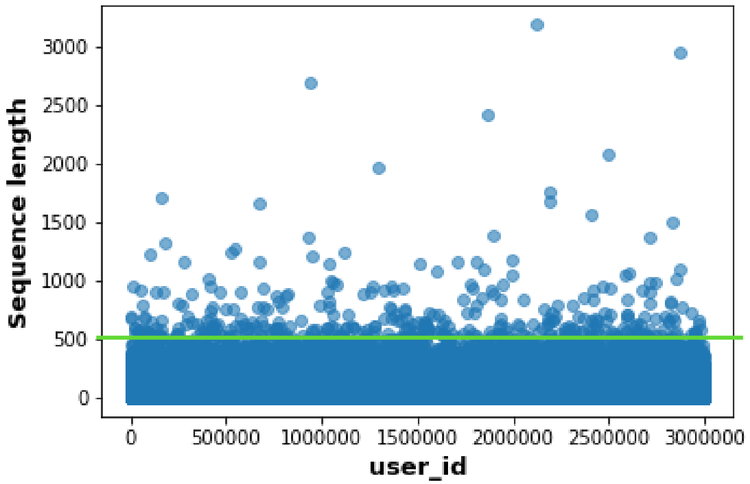


图3：用户单天广告点击历史记录数量分布图

模型介绍

模型结构/模型融合

4.1 模型结构

- Embedding Layer

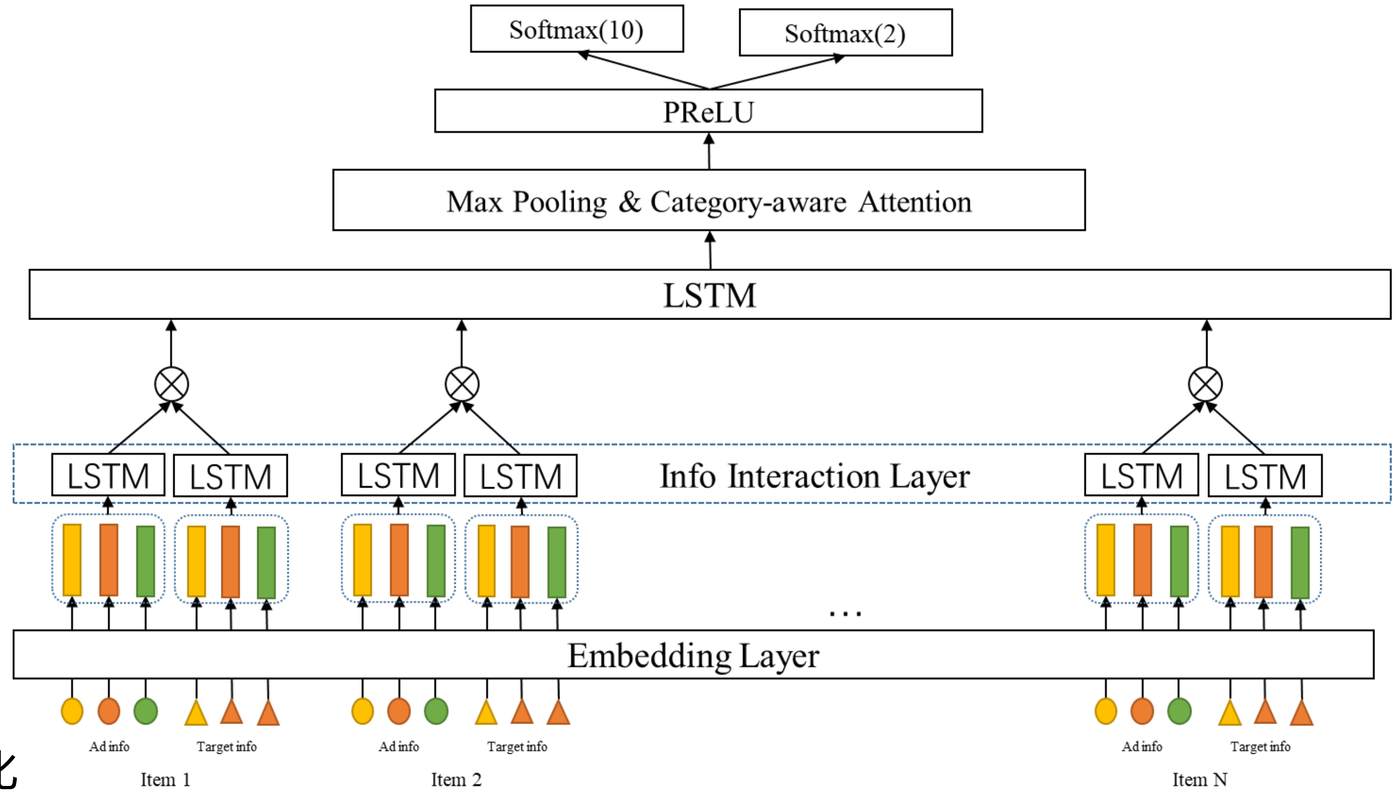
该模型的输入主要为用户的点击行为信息，点击序列中的每个广告item由两类信息组成:

- Ad info(广告属性)
  - creative_id
  - ad_id
  - advertiser_id
  - Industry

  利用word2vec和glove算法训练上述id类型信息，最终序列向量为两类词向量拼接后的词向量

- Target info（目标编码）

  将目标编码得到的特征，进行离散化（qcut）转化为类别id,然后随机初始化为一个稠密向量。



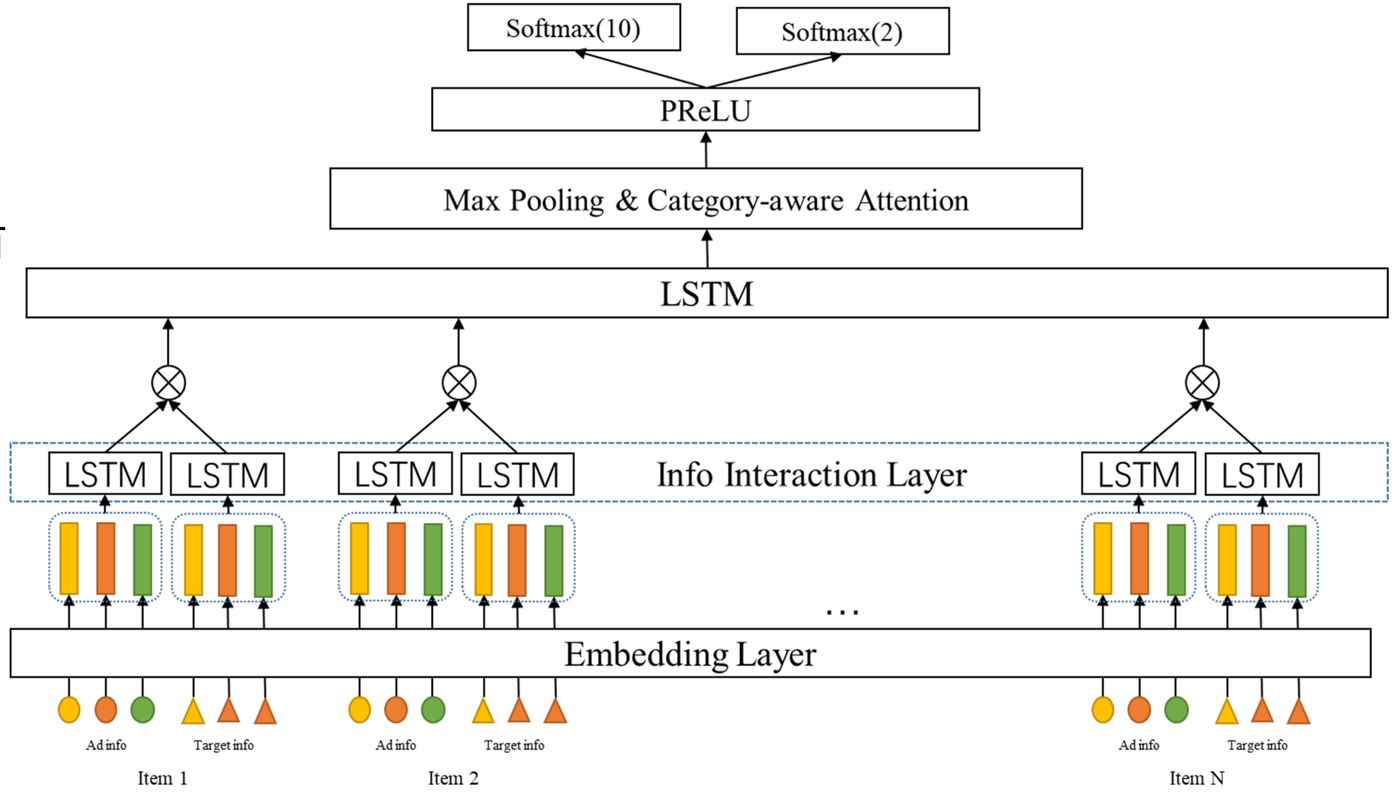
4.1 模型结构

- Info Interaction Layer

该层可以认为是对item的字编码。对于Ad info 和Target info所包含的各个id信息，我们认为其各id间具有序列关系，故选择BiLSTM进行短序列的交互。

对于Ad info中，creative_id, ad_id, advertiser_id, industry这些id具有一定的层级关系；

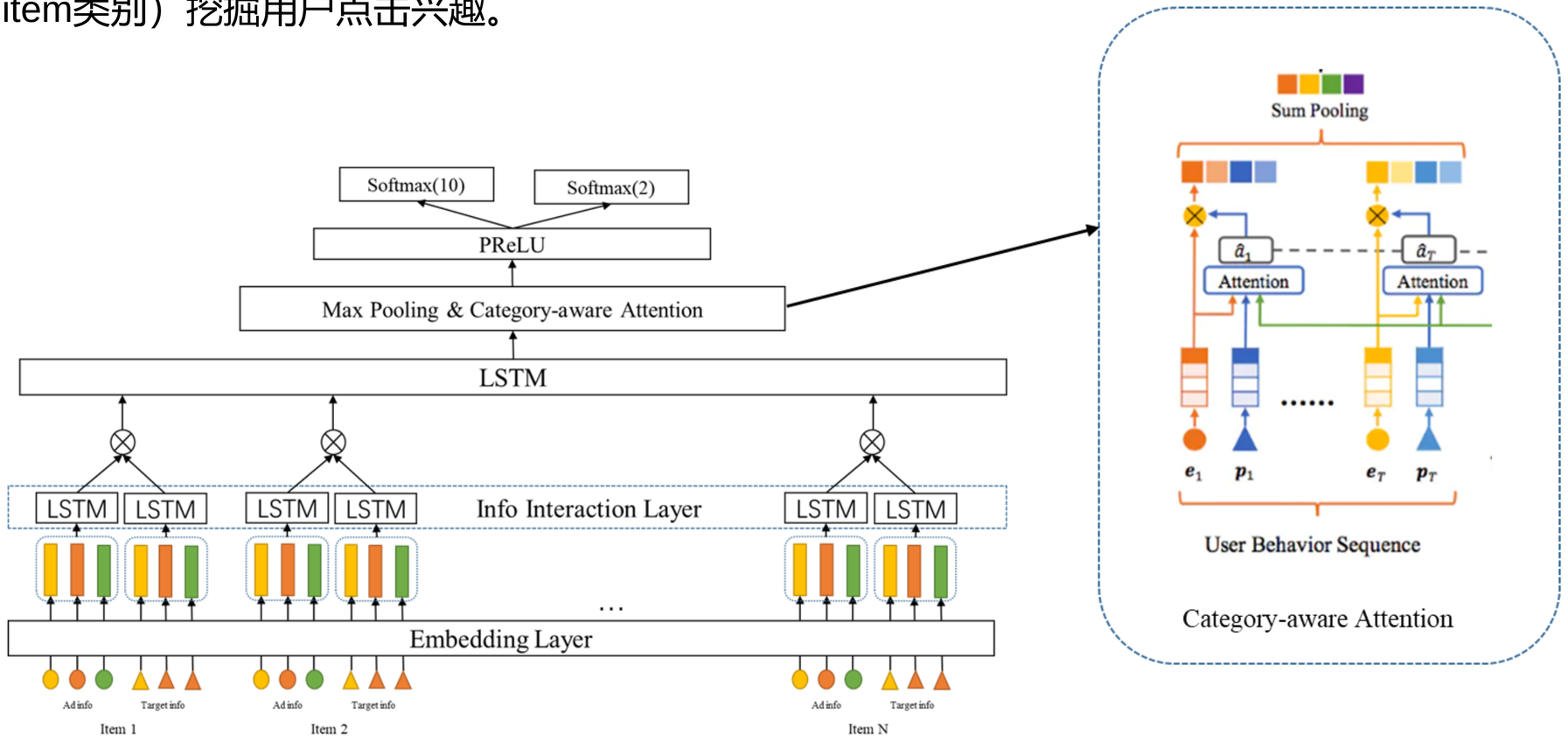
对于Target info， 我们认为年龄段从小到大，也是一个序列信息，故对各id的age的目标特征进行交互。



4.1 模型结构

- Pooling Layer

该层是对用户的点击兴趣进行聚合，采用Max Pooling和Category-aware Attention，其中Category-aware Attention见下图，e表示点击行为编码信息，p表示product_category随机初始化后的embedding。采用这两种聚合操作，是为了从不同层级（具体的item和item类别）挖掘用户点击兴趣。



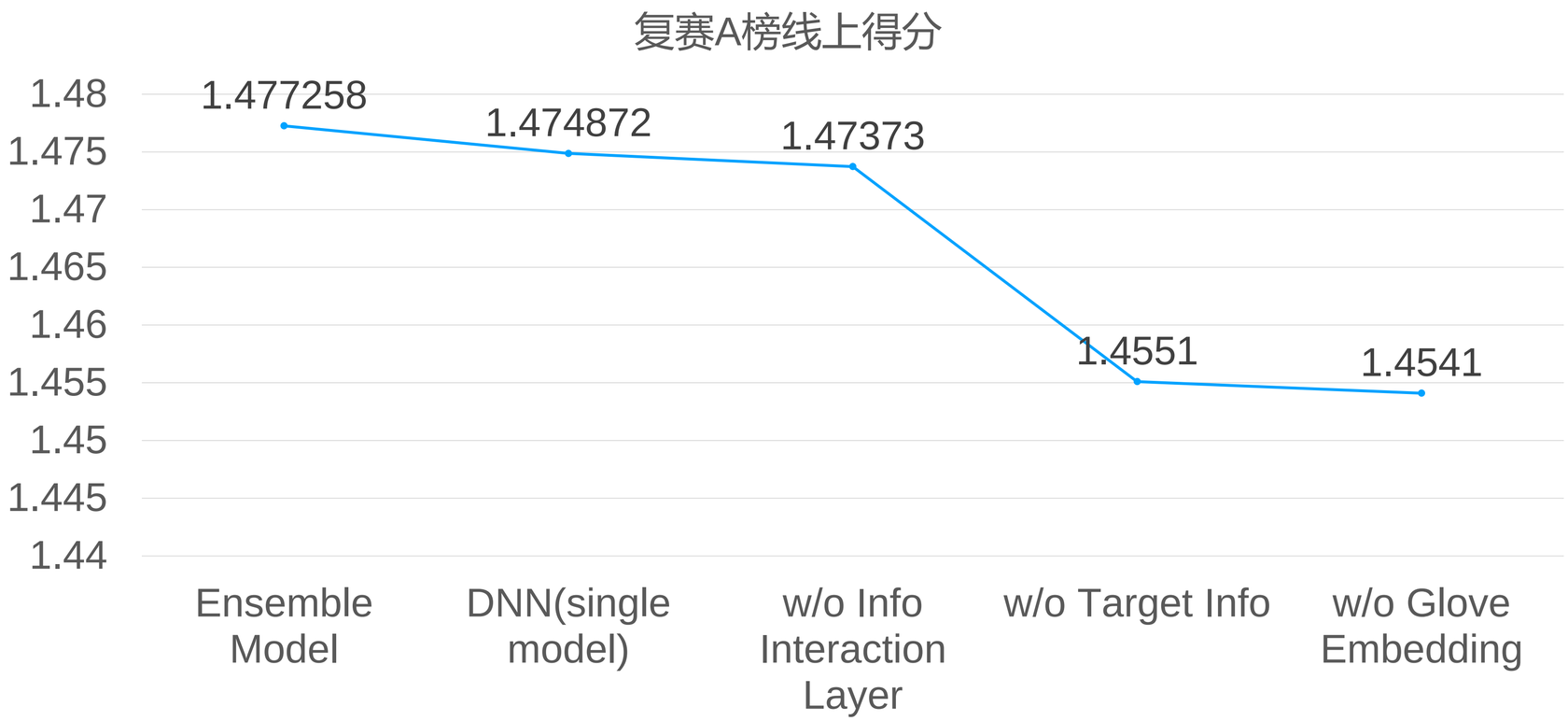
4.2 模型融合

- 数据多样性
  - 点击序列局部重排
    由于一天内的点击记录顺序未知，故对一天内的点击序列进行随机打乱，打乱前后的训练集约有80%不同；
  - 点击序列长度选择不同
    对一天内的点击记录，按照其点击次数进行筛选，筛选数量分别为10和66；
  - 数据划分选择不同
    采用5折、10折划分数据训练。
- 特征多样性
  - 预训练item向量差异
    对于广告id的预训练item向量选择不同的窗口大小进行训练；
  - 广告属性差异
    对于输入采用两类id序列特征：
    （1）creative_id, ad_id, advertiser_id, **industry**
    （2）creative_id, ad_id, advertiser_id, **industry&product_categoty**

总结与思考

DEMO/总结与思考

DEMO


复赛A榜线上得分
1.48
1.475
1.47
1.465
1.46
1.455
1.45
1.445
1.44
1.477258
1.474872
1.47373
1.4551
1.4541
Ensemble Model
DNN(single model)
w/o Info Interaction Layer
w/o Target Info
w/o Glove Embedding

总结与思考

总结

- 采用目标编码方式，有效引入标签信息，提升模型准确率；
- 采用“广告属性-广告点击行为-广告类别”的多层次特征提取结构，使模型能够关注不同层级的特征，增加特征多样性。

不足

- 构建的特征少，例如，仅围绕广告维度构建特征，没有构建用户维度的有效特征；
- 未有效解决测试集中出现大量未登录id所带来的特征编码损失问题。

THANKS