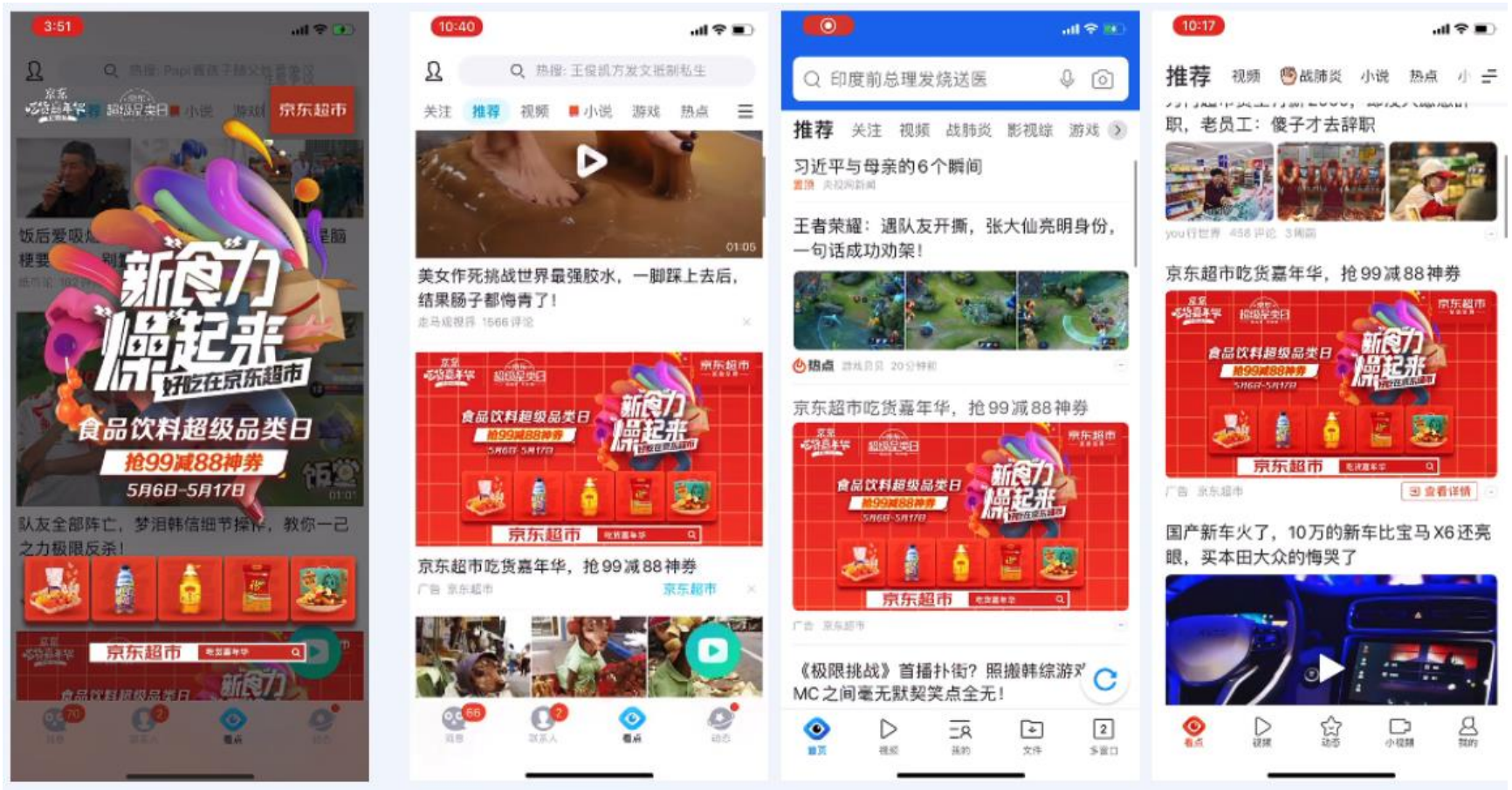
广告点击率预测

助教：许铮睿
程　磊

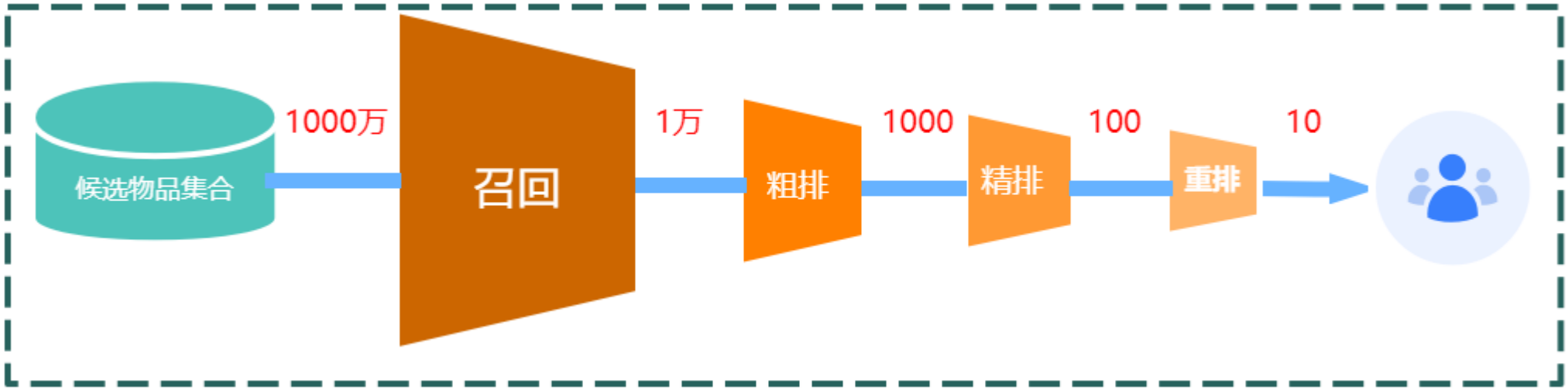
什么是点击率（CTR）预测？

CTR（Click-Through Rate）预估是搜索、推荐、广告等领域基础且重要的任务，主要目标是预测用户在当前上下文环境下对某一个候选物品（视频、商品、广告等） 发生点击的概率。例如：一个公司投放的移动广告推广活动产生了10000次展示量，并在App Store中产生了 500次点击量，则此次推广活动的CTR为5%。



为什么需要预测点击率？

CTR预估算法常用于推荐系统的“精排”阶段，可以根据预测出的CTR对物品进行排序，在某些领域（比如在线广告）也可以根据CTR*bid（出价）进行排序。但是CTR和推荐系统的目标又存在差异。假如给一个用户准备了A/B/C三个广告，那么无论预测出的CTR是0.9、0.8、0.6，还是0.5、0.4、0.3，都不影响三个广告的展现顺序，但是向客户的收费却有天壤之别。



任务简介

传统的点击率预测任务利用用户基本信息、广告日志数据来预测用户对某个广告可能的点击概率。但广告域（目标域）数据可能存在用户行为稀疏，行为类型单一的问题。引入同一媒体下的跨域数据可以获得同一用户在其它域（源域）的行为数据，丰富用户和广告特征，提升广告点击率预测的准确率。

本次任务提供6天的数据用于训练，1天的数据用于测试，希望同学们利用目标域数据（用户-广告交互记录、用户基本信息、广告素材信息）和源域数据（用户-文章交互记录、文章的基本信息）预测用户在目标域的点击率。

评价指标：统计测试集中的广告样本点击率预测值，计算AUC。

提交方式：上传csv文件，该csv文件包含两列，分别为id和pctr（点击率预测值），其中id为测试集中广告样本序号，顺序与测试集中顺序一致。提交示例如下：

```
id,pctr
0,0.007662377929002994
1,0.012495708357102096
2,0.012410172539432817
3,0.012463935534802377
4,0.010263819385989569
5,0.010263819385989569
```

数据介绍

由于数据维度较高，仅展示部分字段对数据分部分进行介绍，详细数据介绍会通过PDF的方式发给大家。

目标域数据

交互记录

字段名	字段含义	字段类型	取值样例
label	是否点击	int	0，1
ad_click_list_001	用户点击广告任务id列表	[string,]	[1^2…]
ad_close_list_001	用户关闭广告任务列表	[string,]	[1^2…]
……	……	……	……

用户信息

字段名	字段含义	字段类型	取值样例
user_id	用户id	int	1,2,3
age	年龄	int	1,2,3
gender	性别	int	1,2,3
……	……	……	……

广告信息

字段名	字段含义	字段类型	取值样例
log_id	广告id	int	1,2,3
adv_id	广告对应的素材id	int	1,2,3
slot_id	广告位id	int	1,2,3
……	……	……	……

源域数据

交互记录

字段名	字段含义	字段类型	取值样例
pro	文章浏览进度	int	1,2,3
label	是否点击	int	1,-1
cilLabel	是否点赞	int	1,-1
……	……	……	……

文章信息

字段名	字段含义	字段类型	取值样例
i_docId	文章docid	int	0001
i_cat	文章类别id	int	0001
i_upTimes	文章点赞量	int	10
……	……	……	……

关于数据集各字段的分布Kaggle上已有最初的分析

任务技术点

① 模型选择

eg：树模型？深度模型？孰优孰劣并不好说。

② 特征工程

eg：挖掘统计类特征（比如越流行的被点击率越高？）、进行特征减法（特征并不是越多越好，如何保留有效特征）。

③ 模型调参

eg：手动调参？自动调参（网格搜索）？

④ 交叉验证

eg：交叉验证在机器学习比赛中往往可以提升模型效果，但是如何设计一套交叉验证pipeline？最终结果能否保证线上线下同升同降？

⑤ 模型融合

eg：Bagging？Stacking？

⑥ 样本不均衡

eg：目前正负样本比例1：100+，建模时是否考虑该问题？

比赛限制及说明

① 数据方面

- 数据已做特殊处理，不包含冷启动用户；
- 可以使用穿越特征，本次比赛不对特征做限制；
- 不得使用所提供数据以外的任何数据。

② 比赛设置

- 本课程共54人，5-6人/组，分为10组，**每组注册一个team账号，按组提交比赛结果**，一天最多可提交3次；
- **队伍名称统一命名为：BJTU_2023ML1_组ID**；
- 支持高分队伍讨论区分享自己的trick；提供最基础的baseline；

评分规则及作业截止时间：

- **实验代码**：比赛排名、代码质量；
- **小组汇报**：汇报本组实验方案及结果4分钟，提问2分钟左右；
- **实验报告**：提供模板，**一组一份，由组长在课程平台提交word文档，命名为大作业报告-组ID**。注意实验报告最后需要写明每位成员的工作量及贡献，依此浮动给分。
- **Kaggle平台提交结果截止至12月27日（汇报前一天）；实验报告截止时间为2024年1月7日（17周周日）**。

参考资料

1. Catboost论文：https://arxiv.org/pdf/1706.09516.pdf
2. Catboost官方文档：https://catboost.ai/
3. 科大讯飞广告点击率预测冠军方案：https://www.bilibili.com/read/cv14089186/
4. DeepFM论文：https://arxiv.org/abs/1703.04247
5. xDeepFM论文：https://arxiv.org/abs/1803.05170
6. DeepCTR库：https://github.com/shenweichen/DeepCTR