2022全球算法精英大赛初赛参赛手册

题目：广告-信息流跨域ctr预估

一、简介

广告推荐主要基于用户对广告的历史曝光、点击等行为进行建模，如果只是使用广告域数据，用户行为数据稀疏，行为类型相对单一。而引入同一媒体的跨域数据，可以获得同一广告用户在其他域的行为数据，深度挖掘用户兴趣，丰富用户行为特征。引入其他媒体的广告用户行为数据，也能丰富用户和广告特征。

本赛题希望选手基于广告日志数据，用户基本信息和跨域数据优化广告 ctr 预估准确率。目标域为广告域，源域为信息流推荐域，通过获取用户在信息流域中曝光、点击信息流等行为数据，进行用户兴趣建模，帮助广告域 ctr 的精准预估。

二、赛题说明

本赛题提供 7 天数据用于训练，1 天数据用于测试，数据包括目标域（广告域）用户行为日志，用户基本信息，广告素材信息，源域（信息流域）用户行为数据，源域（信息流域）物品基本信息等。希望选手基于给出的数据，识别并生成源域能反映用户兴趣，并能应用于目标域的用户行为特征表示，基于用户行为序列信息，进行源域和目标域的联合建模，预测用户在广告域的点击率。所提供的数据经过脱敏处理，保证数据安全。

三、数据说明

提供的数据包括目标域用户行为数据，源域用户行为数据,以下按照这 2 个维度分别说明。

1、目标域用户行为数据

序号	字段名称	字段含义	是否可为空	字段类型	取值样例
1	label	是否点击，0：否，1：是	否	int	0，1
2	user_id	用户 id	否	String	1，2…
3	age	年龄	是	String	1，2,3…
4	gender	性别	是	String	1，2…
5	residence	常住地-省份	是	String	1，2…
6	city	常住地-市-编号	是	String	1，2…
7	city_rank	常住地-市-等级	是	String	1，2…
8	series_dev	设备系列	是	String	1，2…
9	series_group	设备系列分组	是	String	1，2…
10	emui_dev	emui 版本号	是	String	1，2…
11	device_name	用户使用的手机机型	是	String	1，2…
12	device_size	用户使用手机的尺寸	是	String	1，2…

13	net_type	行为发生的网络状态	是	String	1，2…
14	task_id	广告任务唯一标识	是	String	1，2…
15	adv_id	广告任务对应的素材 id	是	String	1，2…
16	creat_type_cd	素材的创意类型 id	是	String	1，2…
17	adv_prim_id	广告任务对应的广告主 id	是	String	1，2…
18	inter_type_cd	广告任务对应的素材的交互类型	是	String	1，2…
19	slot_id	广告位 id	是	String	1，2…
20	site_id	媒体 id	是	String	1，2…
21	spread_app_id	投放广告任务对应的应用 id	是	String	1，2…
22	Tags	广告任务对应的应用的标签	是	String	1，2…
23	app_second_class	广告任务对应的应用的二级分类	是	String	1，2…
24	app_score	app 得分	是	Int	4
25	ad_click_list_001	用户点击广告任务 id 列表	是	[string,]	[1^2…]
26	ad_click_list_002	用户点击广告对应广告主 id 列表	是	[string,]	[1^2…]
27	ad_click_list_003	用户点击广告推荐应用列表	是	[string,]	[1^2…]
28	ad_close_list_001	用户关闭广告任务列表	是	[string,]	[1^2…]
29	ad_close_list_002	用户关闭广告对应广告主列表	是	[string,]	[1^2…]
30	ad_close_list_003	用户关闭广告推荐应用列表	是	[string,]	[1^2…]
31	pt_d	时间戳	否	String	20220522 1430
32	log_id	样本 id	否	Int	12345678

2、源域用户行为数据

序号	字段名称	字段含义	是否可为空	字段类型	取值样例
1	u_userId	用户标识	否	String	0001
2	u_phonePrice	用户手机价格	是	String	13
3	u_browserLifeCycle	浏览器用户活跃度	是	String	10
4	u_browserMode	浏览器业务类型	是	String	11
5	u_feedLifeCycle	信息流用户活跃度	是	String	12
6	u_refreshTimes	信息流日均有效刷新次数	是	String	16

7	u_newsCatInterests	信息流图文 点击 分类偏好	是	[string,]	[1^2…]
8	u_newsCatDislike	信息流图文 负反馈 分类偏好	是	[string,]	[1^2…]
9	u_newsCatInterestsST	用户短时 兴趣 分类偏好	是	[string,]	[1^2…]
10	u_click_ca2_news	用户图文 类别 点击序列	是	[string,]	[1^2…]
11	i_docId	文章 docid	是	String	0001
12	i_s_sourceId	文章来源的 sourceid	是	String	0001
13	i_regionEntity	文章地域词 id	是	String	0001
14	i_cat	文章类别 id	是	String	0001
15	i_entities	文章实体词 id	是	[string,]	[1^2…]
16	i_dislikeTimes	文章负反馈量	是	String	60
17	i_upTimes	文章点赞量	是	String	22
18	I_dtype	文章展现形式	是	String	20
19	e_ch	频道	是	String	1,2…
20	e_m	事件来源设备机型	是	String	1,2…
21	e_po	第几位	是	String	9
22	e_pl	拜访地	是	String	1,2…
23	e_rn	第几刷	是	String	1
24	e_section	信息流场景类型	是	String	13
25	e_et	时间戳	否	String	20220522 1430
26	label	是否点击，-1：否，1：是	否	String	1
27	cilLabel	是否点赞，-1：否， 1：是	否	String	1
28	pro	文章浏览进度	否	String	1,2…

三、评估方式

评估方式：统计广告域的样本 ctr 预估值，计算 GAUC 和 AUC
评测指标：本次比赛使用 GAUC 和 AUC 的加权求和作为评估指标，具体公式如下：

$$xAUC = \alpha * GAUC + \beta * AUC$$

xAUC 越高，代表结果越优，排名越靠前。
其中，AUC 为全体样本的 AUC 统计，GAUC 为分组 AUC 的加权求和，以用户为维度分组，分组权值为分组内曝光量/总曝光）

$$GAUC = \frac{\sum_{k=i}^{n} AUC_i * Impression_i}{\sum_{k=i}^{n} Impression_i}$$

初赛：α 为 0.7，β为0.3

五、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：log_id,pctr。其中 log_id 为对应测试样本中的 log_id，pctr 对应测试样本经由模型计算出的预估 ctr 值，pctr 保留 6 位小数。

提交文件参考如下示例：

log_id,pctr
1, 0.002345
2, 0.010456
…