阿里巴巴点击与转化预估数据集说明

1. 简介

  本数据集采集自手机淘宝移动客户端的推荐系统日志，其中包含点击和与之关联的转化数据，二者的关系图1描述。淘宝平台作为全球最大的在线零售电子商务平台，为提升其用户体验，通过推荐系统提供商品推荐服务，用户可以在浏览（impression）推荐结果中点击（click）感兴趣的商品，或者进一步对商品进行购买（conversion）。因此用户的行为可以抽象为一个序列模式：浏览 -> 点击 -> 购买。



图 1: 手机淘宝上的单品广告

  本数据的整体描述为{(xi,yi 🡪 zi)} from 1🡪 n，样本的格式为(x, y 🡪z)，是来自定义域(x,y 🡪 z) 的分布D，其中X是特征空间，Y和Z是标签空间，N是样本总量。X代表观测样本的特征向量，通常是被划分为多个域的高维稀疏向量，例如，用户域、商品域等。y和z都是二值标签（0 or 1），其中y=1代表样本发生点击事件，而z=1代表样本发生转化事件，显然，由于业务归因分析的需要，转化事件被定义为由本次点击引导的转化事件，因此同一个样本的点击事件和转化事件不是独立的，转化事件的发生必然要求该样本存在先序点击事件，这种序列依赖关系我们用y 🡪 z来描述。

  根据上述说明，本数据集中各样本的标签y和z字段的值分布，服从如下约束：

表 1: 标签分布

|  |  |
| --- | --- |
| 标签值域 | 是否合法 |
| y=0 & z=0 | 合法 |
| y=0 & z=1 | 非法 |
| y=1 & z=0 | 合法 |
| y=1 & z=1 | 合法 |

2. 样本组织说明

  本数据集由两部分组成：第一部分为训练数据（sample\_train.tar.gz),第二部分为测试数据（sample\_test.tar.gz)，每个部分的数据还分别携带了一个MD5检校文件，详见图2描述。

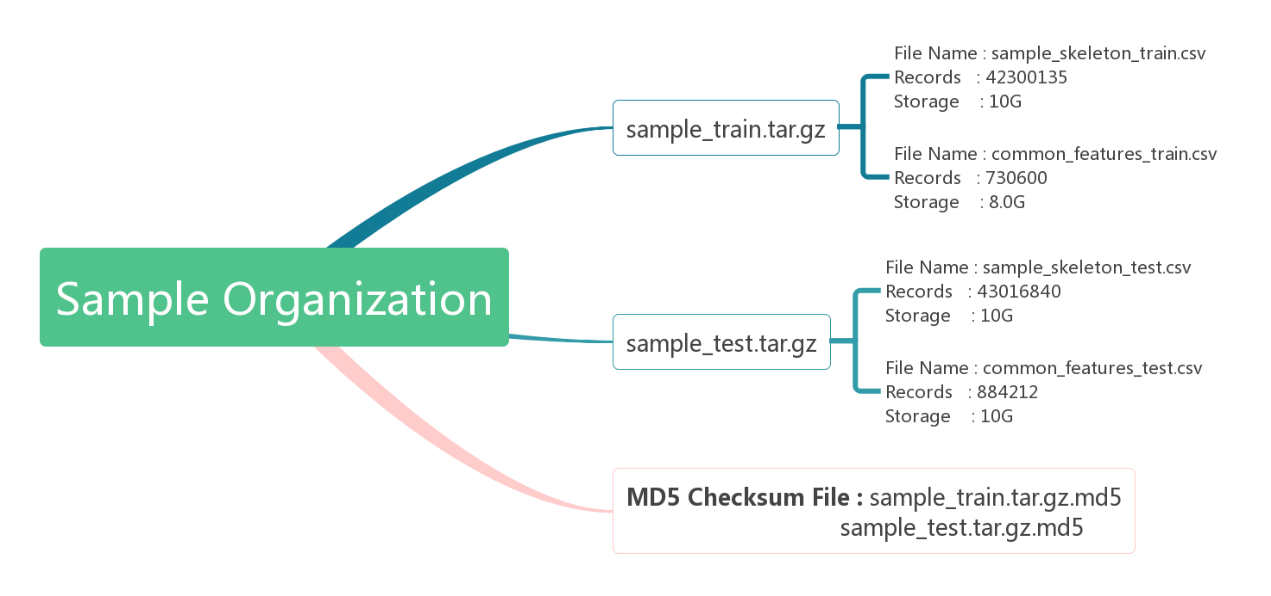
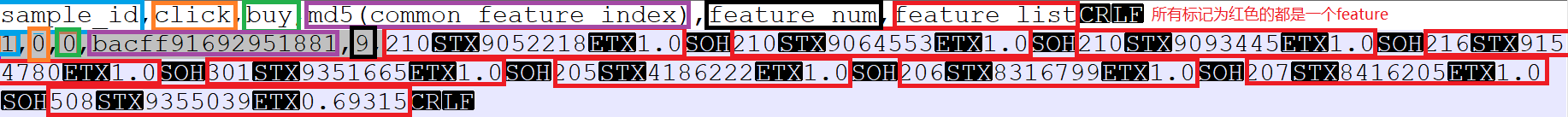


图 2: 样本组织说明

注意：训练和测试数据严格按照时间顺序划分——训练数据先于测试数据发生，这也遵循了传统工业数据集的实际应用场景。考虑到存储效率问题，每部分数据由两个CSV格式文件构成：一个是样本骨架文件，另一个是公共特征文件。使用本数据集合前，应首先确保样本骨架文件与公共特征文件已经进行正确地关联，具体关联方法详细参考本文第5部分样本关联方法。

3. 样本骨架说明

  样本骨架文件中的每一条记录代表一次用户浏览，并且由三个部分组成，图3详细描述了各部分属性与格式信息。



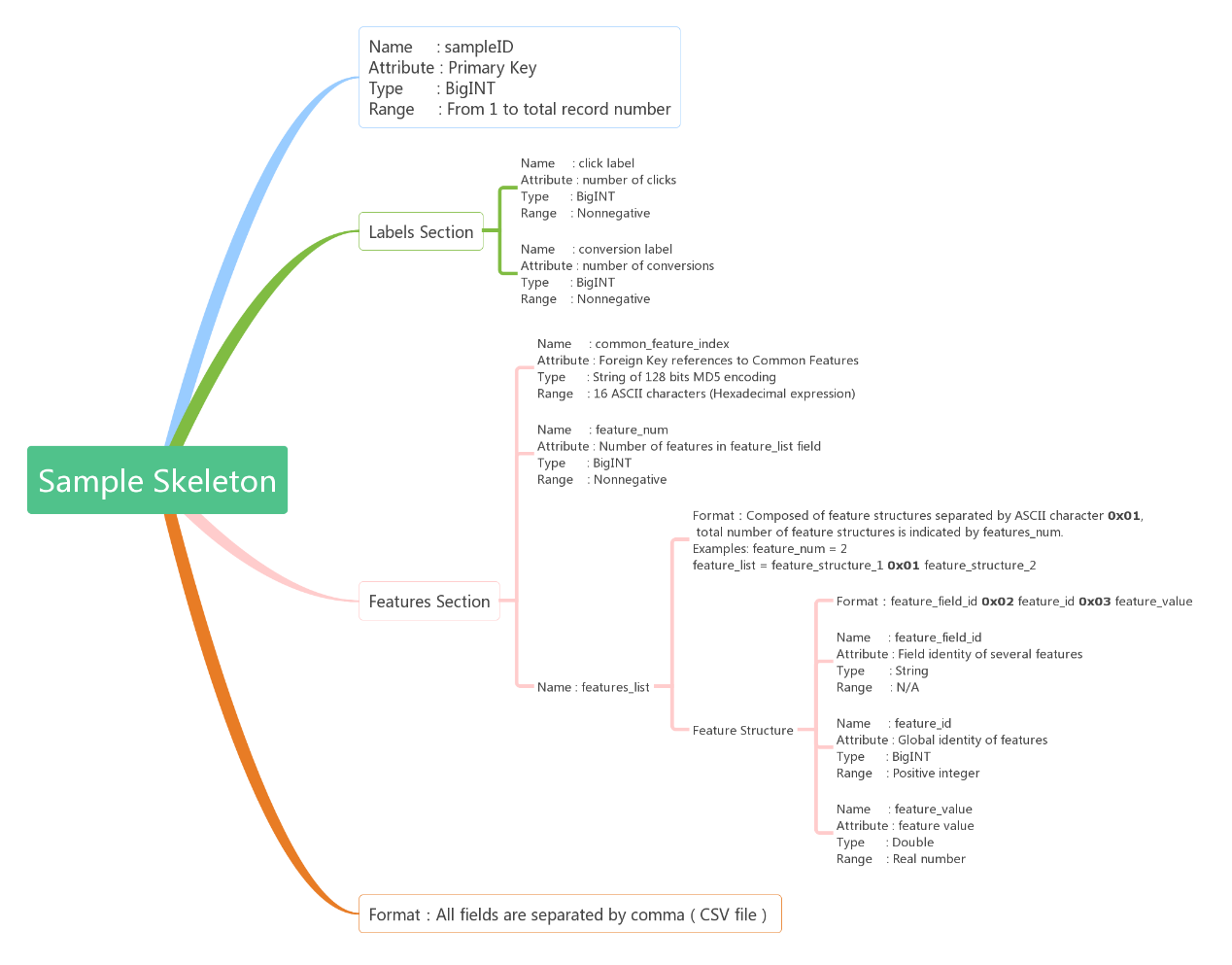


图 3: 样本骨架说明

Features\_list : feature\_feild\_id0x02feature\_id0x03feature\_value 0x01 feature\_feild\_id0x02feature\_id0x03feature\_value

样本ID部分sample\_id

唯一标识样本骨架中的一条记录，取值从1开始直到全部样本数，是样本骨架文件的主键。

标签部分 click buy

标签部分包含一次浏览记录上的两种类型的标签信息：点击和转化事件是否发生，

表1对取值范围进行了描述。

特征部分

特征部分包含三个域：

索引md5（common\_feature\_index）域：作为样本骨架文件的外键，用来关联公共特征文件中的信息，复原样本时使用。

特征数量feature\_num（feature\_num）域：指出本条记录中特征列表（feature\_list）域包含的特征总数，是一个非负整数。

特征列表（feature\_list）域：由ASCII字符0x01分割的若干特征组成，这些特征由特征数据结构（Feature Structure）描述。特征数据结构由ASCII字符0x02和0x03作为分隔符的字符串构成，例如：feature\_field\_id 0x02 feature\_id 0x03 feature\_value，其中feature\_field\_id字段代表特征域信息，在表2中进行了详细描述；feature\_id字段是被全局编码后的特征ID值；feature\_value字段是特征ID对应的特征值。

数据格式解析：

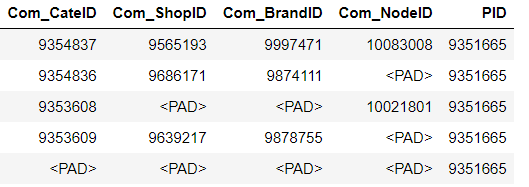
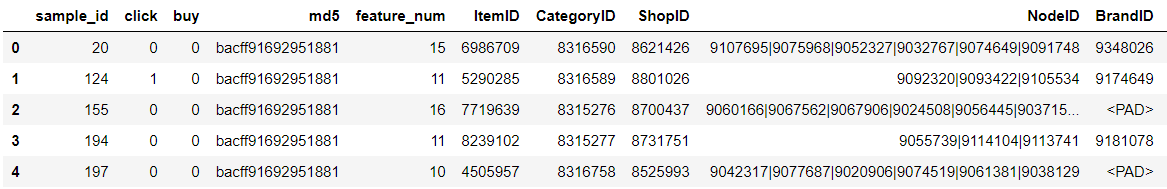


表 2: 特征描述

|  |  |  |
| --- | --- | --- |
| 特征域名称 | 特征域ID | 特征域说明 |
| 用户域 | 101 | 用户ID User ID. |
| 109\_14 | 商品类目ID以及用户在该类目上的历史行为累积数量。 |
| 110\_14 | 商品店铺ID以及用户在该店铺上的历史行为累积数量 |
| 127\_14 | 商品品牌ID以及用户在该品牌上的历史行为累积数量。 |
| 150\_14 | 用户意图ID以及用户在该意图上的历史行为累积数量 |
| 121 | 用户的一种分类ID。 |
| 122 | 用户的一种分类ID |
| 124 | 用户性别分类ID。 |
| 125 | 用户年龄分类ID。 |
| 126 | 用户消费水平分类I。 |
| 127 | 用户消费水平分类II。 |
| 128 | 用户是否就业。 |
| 129 | 用户地理信息分类ID。 |
| 商品域 | 205 | 商品ID |
| 206 | 商品所属类目ID |
| 207 | 商品所属店铺ID |
| 210 | 商品关联用户意图ID |
| 216 | 商品的品牌ID |
| 组合域 | 508 | 109\_14和206域的组合特征。 |
| 509 | 110\_14和207域的组合特征。 |
| 702 | 127\_14和216域的组合特征。 |
| 853 | 150\_14和210域的组合特征。 |
| 场景域 | 301 | 业务场景信息的一种分类表示。 |

∗ 用户历史行为信息来自过去两周。

|  |  |  |
| --- | --- | --- |
| **Feature Category** | **Feature Field ID** | **Feature Field Description** |
| User Features | 101 | User ID. |
| 109\_14 | User historical behaviors of category ID and count*.* |
| 110\_14 | User historical behaviors of shop ID and count |
| 127\_14 | User historical behaviors of brand ID and count*.* |
| 150\_14 | User historical behaviors of intention node ID and count |
| 121 | Categorical ID of User Profile. |
| 122 | Categorical group ID of User Profile. |
| 124 | Users Gender ID. |
| 125 | Users Age ID. |
| 126 | Users Consumption Level Type I. |
| 127 | Users Consumption Level Type II. |
| 128 | Users Occupation: whether or not to work. |
| 129 | Users Geography Informations. |
| Item Features | 205 | Item ID. |
| 206 | Category ID to which the item belongs to. |
| 207 | Shop ID to which item belongs to. |
| 210 | Intention node ID which the item belongs to. |
| 216 | Brand ID of the item. |
| Combination Features | 508 | The combination of features with 109\_14 and 206. |
| 509 | The combination of features with 110\_14 and 207. |
| 702 | The combination of features with 127\_14 and 216. |
| 853 | The combination of features with 150\_14 and 210. |
| Context Features | 301 | A categorical expression of position. |

∗ User historical behaviors collected within the past two weeks.

4. 公共特征说明

  公共特征文件中的一条记录代表一个特定的特征集合，这个特征集合被样本骨架文件中若干条记录共享，图4所描述。在公共特征文件中同样存在common\_feature\_index作为主键，与样本骨架文件中的同名字段沟通描述这种共享关系。

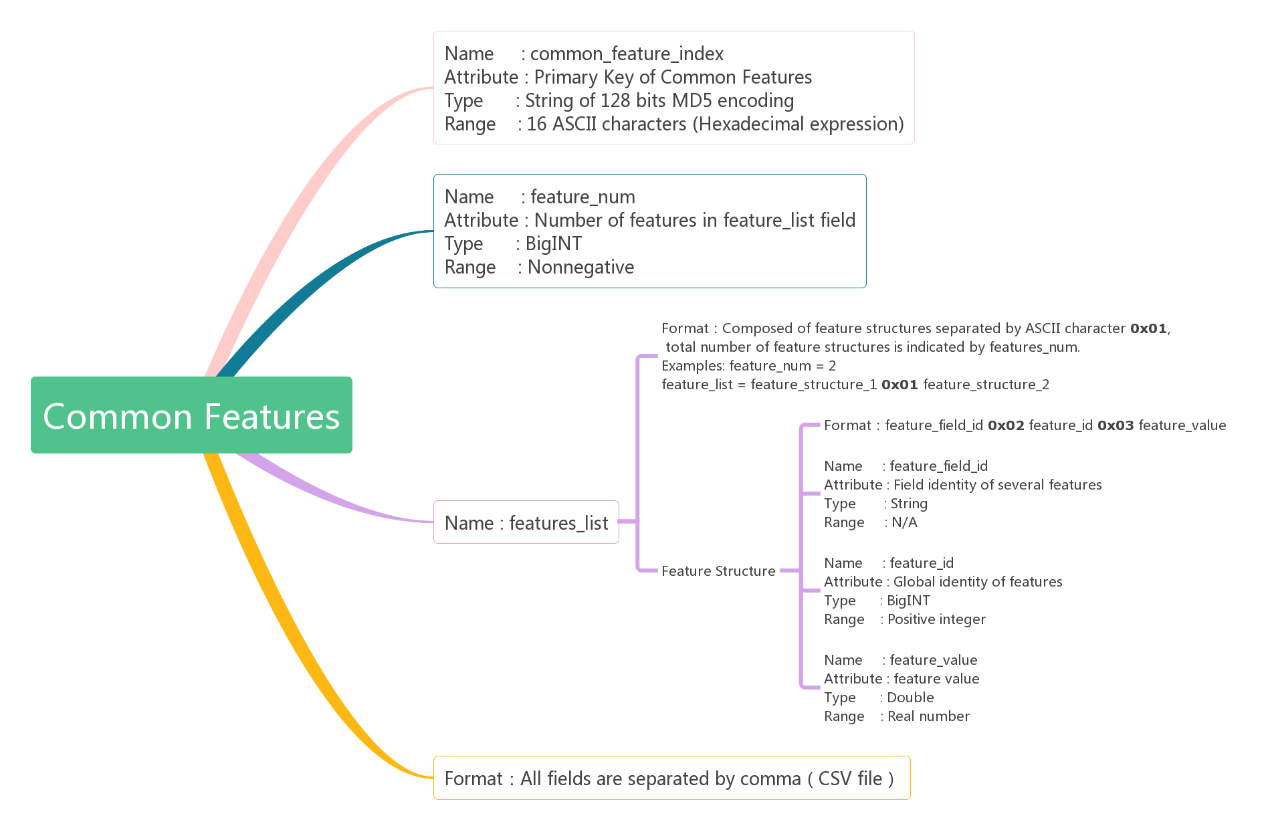
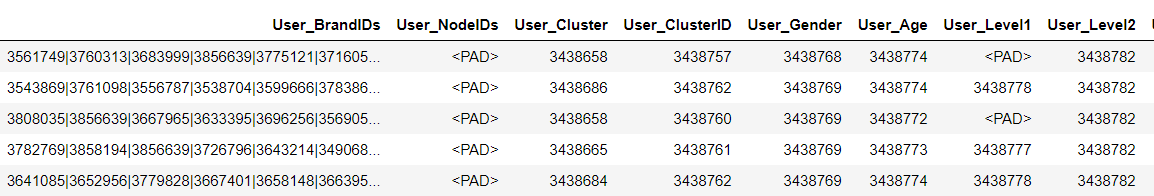
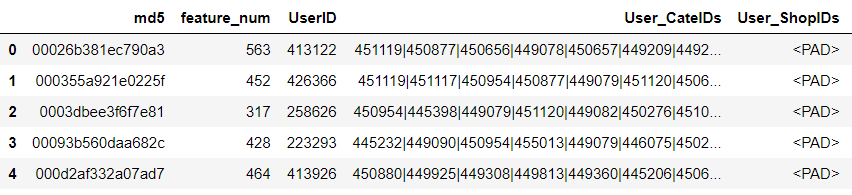
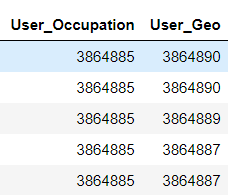
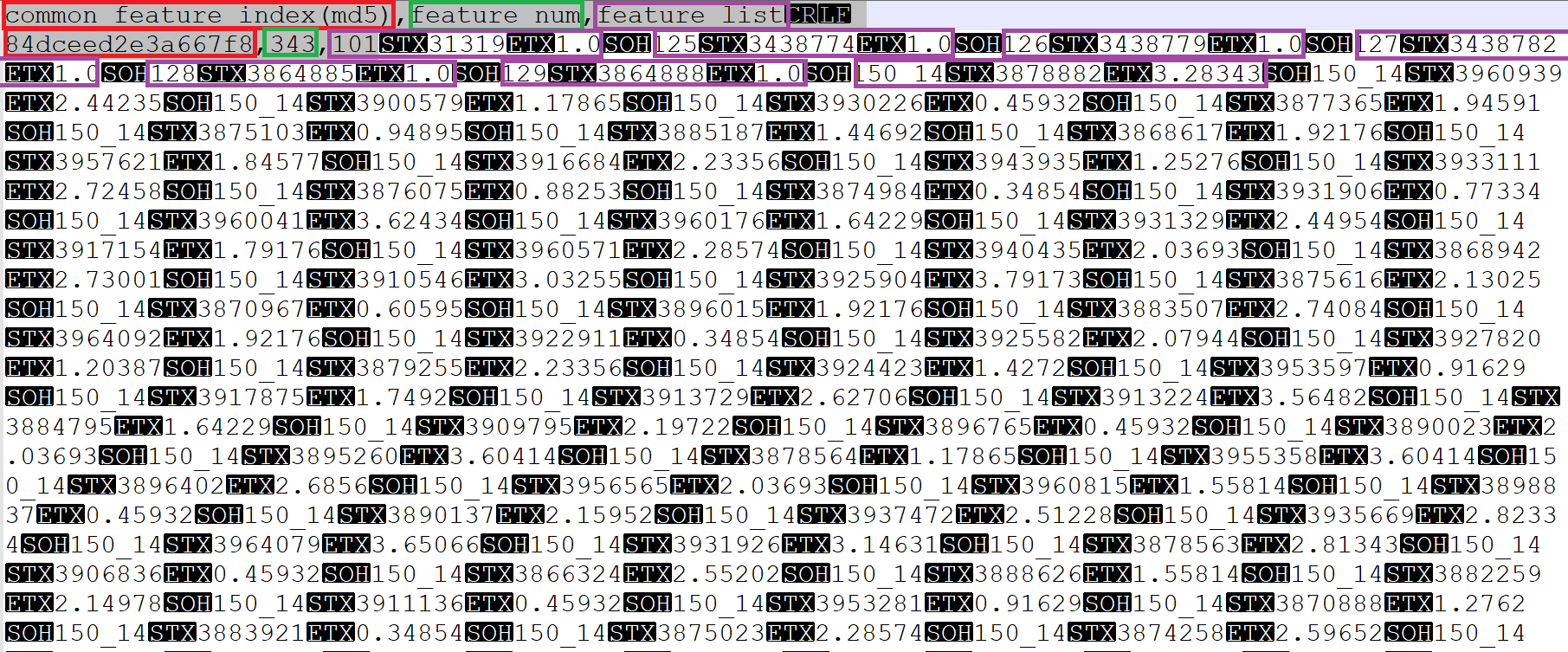


图 4: 公共特征说明



5. 样本关联方法

  注意：使用本数据集合前，应首先确保样本骨架文件与公共特征文件进行正确的关联，具体关联方法详细参考本图5中伪代码描述。

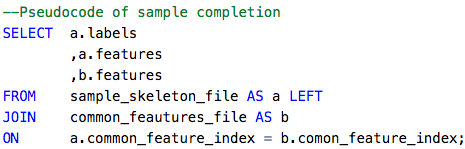
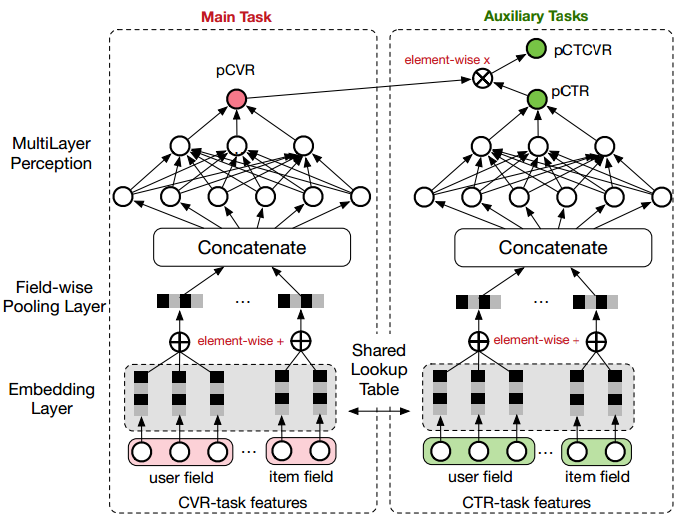


图 5: 样本关联伪代码



6. 引用说明

  在任何形式的出版物中声明使用本数据，应包含如下论文的引用信息：

Xiao Ma, Liqin Zhao, Guan Huang, Zhi Wang, Zelin Hu, Xiaoqiang Zhu, Kun Gai. Entire Space Multi-Task Model: An Effective Approach for Estimating Post-Click Conversion Rate. In SIGIR 2018-Proceedings of the 41th International ACM SIGIR Conference on Research and Development in Information Retrieval 2018 July 8-12. ACM.