t-17077185

首页 天池大赛 AI 学习 天池实验室 数据集 技术圈 其他

Ali-CCP：Alibaba Click and Conversion Prediction

天池小T 2018-04-19 10:44:09 992 18

内容 notebook 评论

描述

This data set is provided b

数据列表

数据名称	上传时间	大小	下载
sample_test.tar.gz	2018-11-07	4.68GB	
sample_test.tar.gz.md5	2018-04-21	53B	
sample_train.tar.gz	2018-04-21	4.10GB	
sample_train.tar.gz.md5	2018-04-21	53B	

文档

Data Set Description

1. Introduction

This is a dataset gathered from real-world traffic logs of the recommender system in Taobao. As the largest online retail platform in the world, Taobao provides item recommendation service for better user experience. Given recommended items(goods) when visiting Taobao.com, users might click interested ones and make a further purchase among them. In other words, user actions follow a sequential pattern of $impression \rightarrow click \rightarrow conversion$.

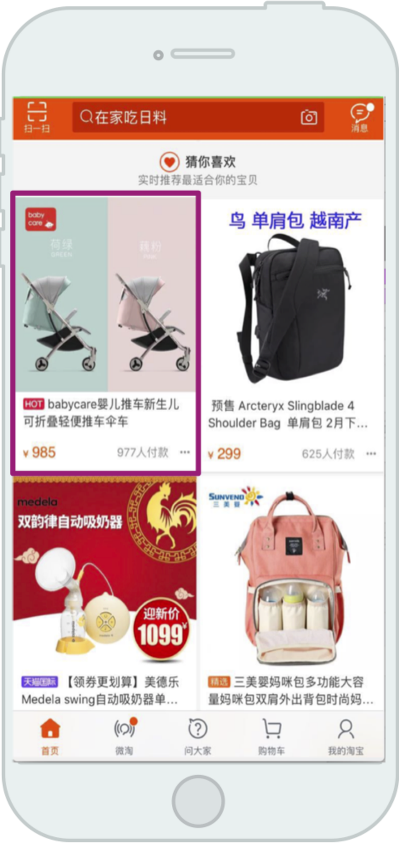

扫一扫
在家吃日料
消息
猜你喜欢
实时推荐最适合你的宝贝
baby care
荷绿
GREEN
藕粉
PINK
HOT babycare婴儿推车新生儿可折叠轻便推车伞车
¥985
977人付款
鸟 单肩包 越南产
预售 Arcteryx Slingblade 4 Shoulder Bag 单肩包 2月下...
¥299
625人付款
medela
双韵律自动吸奶器
迎新价
1099¥
天猫国际 【领券更划算】美德乐 Medela swing自动吸奶器单...
SUNVENO
三美婴
精选 三美婴妈咪包多功能大容量妈咪包双肩外出背包时尚妈...
首页
微淘
问大家
购物车
我的淘宝

Figure 1: Illustration of recommender system in Taobao.

In all, the observed dataset is in format of $\{(x_i, y_i \to z_i)\}|_{i=1}^{N}$, with sample $(x, y \to z)$drawn from a distribution D with domain $\mathcal{X} \times \mathcal{Y} \times \mathcal{Z}$, where $\mathcal{X}$ is feature space, $\mathcal{Y}$ and $\mathcal{Z}$ are label spaces, and N is the total number of impressions.$\boldsymbol{x}$ represents feature vector of observed impression, which is usually a high dimensional sparse vector with multi-fields, such as user field, item field etc.y and z are binary labels with $y = 1$ or $z = 1$ indicating whether click or conversion event occurs respectively.$y \to z$ reveals the sequential dependence of click and conversion labels that there is always a preceding click when conversion event occurs.

Values of y and z in the data can only belong to one of the following cases:

Table 1: Labels Distribution

Label Values	Is Validation
y=0 & z=0	Yes
y=0 & z=1	No
y=1 & z=0	Yes
y=1 & z=1	Yes

2. Sample Organization

The dataset consist of two parts : a train set (sample_train.tar.gz) and a test set (sample_test.tar.gz) with theirs own MD5 checksum files respectively, as shown below in Fig.2.

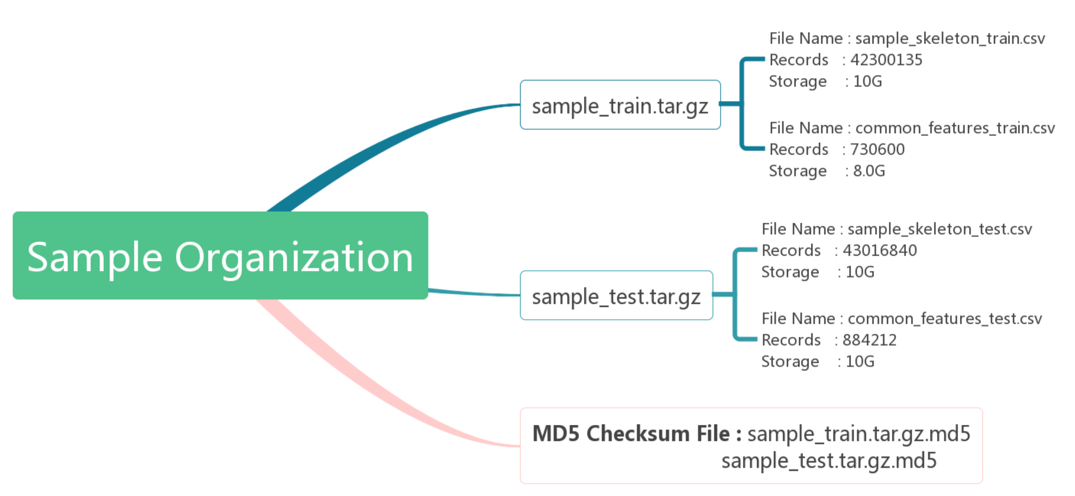


Figure 2: Illustration of sample organization.

Train and test set are split along the time sequence, which is a traditional industrial setting. Each set has two CSV format files: a sample skeleton file and a common features file. A complete sample record should contain not only the features from a record in sample skeleton file, but also the features from related record in common features file.

3. Sample Skeleton Description

Every row in sample skeleton file represents an impression and is composed of three sections. Fig.3 describes in detail the attributes and formats of different sections.

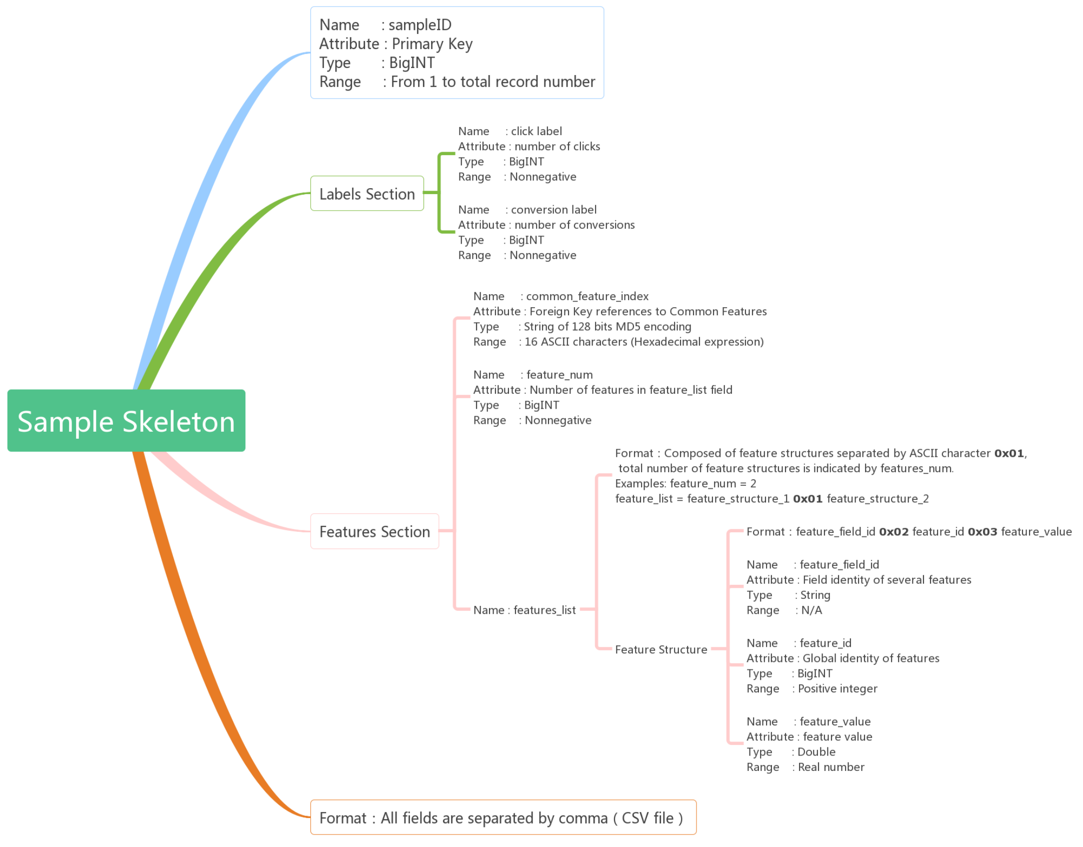

Sample Skeleton
Name : sampleID
Attribute : Primary Key
Type : BigINT
Range : From 1 to total record number
Labels Section
Name : click label
Attribute : number of clicks
Type : BigINT
Range : Nonnegative
Name : conversion label
Attribute : number of conversions
Type : BigINT
Range : Nonnegative
Features Section
Name : common_feature_index
Attribute : Foreign Key references to Common Features
Type : String of 128 bits MD5 encoding
Range : 16 ASCII characters (Hexadecimal expression)
Name : feature_num
Attribute : Number of features in feature_list field
Type : BigINT
Range : Nonnegative
Name : features_list
Format : Composed of feature structures separated by ASCII character 0x01,
total number of feature structures is indicated by features_num.
Examples: feature_num = 2
feature_list = feature_structure_1 0x01 feature_structure_2
Feature Structure
Format : feature_field_id 0x02 feature_id 0x03 feature_value
Name : feature_field_id
Attribute : Field identity of several features
Type : String
Range : N/A
Name : feature_id
Attribute : Global identity of features
Type : BigINT
Range : Positive integer
Name : feature_value
Attribute : feature value
Type : Double
Range : Real number
Format : All fields are separated by comma (CSV file)

Figure 3: Illustration of sample skeleton.

- **Sample ID Section.**

> The unique identity of a record in skeleton file, ranging from 1 to total record number. It is the primary key of the sample skeleton file.

- **Labels Section.**

> There are two types of labels, i.e., click and conversion, for every impression, which follows the sequential patten as described in Table 1.

- **Features Section.**

> There are three fields in this section:

1. an index field named common_feature_index which is foreign key of sample skeleton, references the field with the same name in common features file.
2. a nonnegative integer field named feature_num indicates the number of features in the field of features_list.
3. features_list which is composed of several features separated by the ASCII character 0x01. Each feature in the list is represented by Feature Structure, which is a three components data structure separated by the ASCII characters 0x02 and 0x03 , e.g. feature_field_id 0x02 feature_id 0x03 feature_value, where: i) feature_field_id represents the field of features, as shown below in Table 2, ii) all the features are encoded with global identities feature_id, iii) feature_value field gives a value of real number corresponding to feature_id.

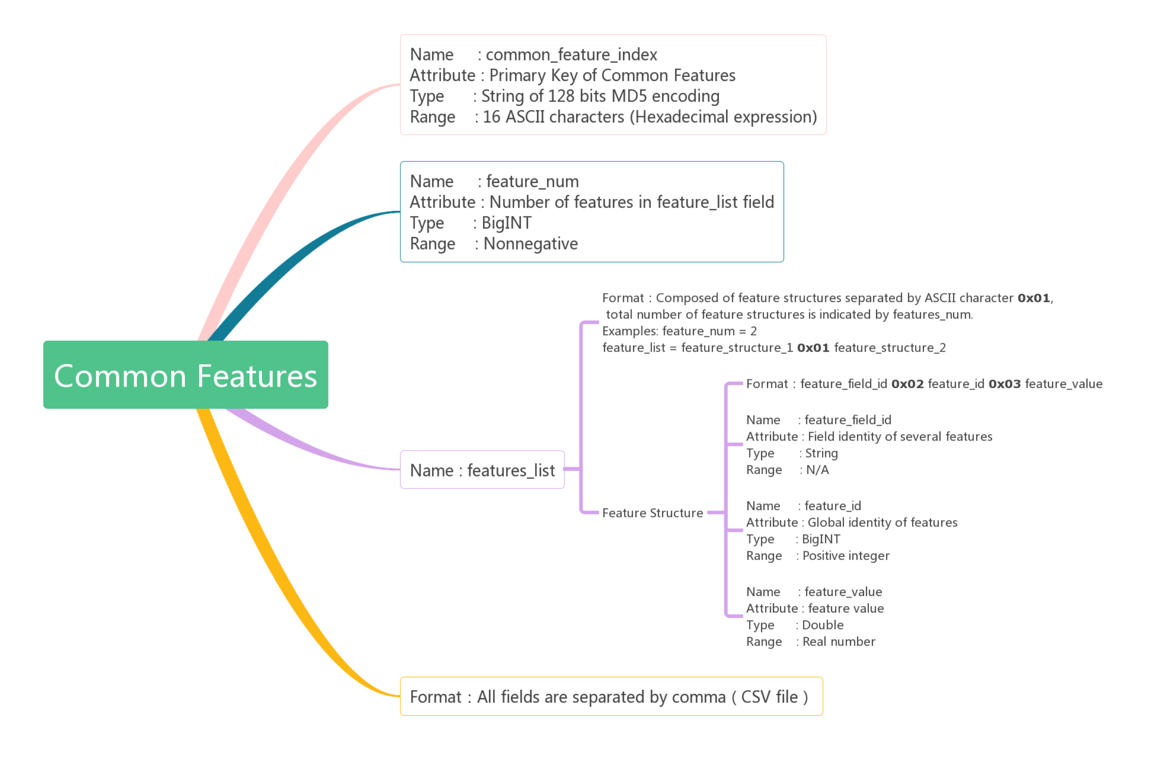
Table 2: Description of feature sets

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. Common Features Description

Every row in common feature file represents a collection of features shared by lots of impressions in sample skeleton file and has the same structure of Section 3 in Sample skeleton description, as shown below in Fig.4. The difference between them is that common_feature_index is the primary key of common features file while it is just a foreign key of sample skeleton file.



Common Features

Name : common_feature_index
Attribute : Primary Key of Common Features
Type : String of 128 bits MD5 encoding
Range : 16 ASCII characters (Hexadecimal expression)

Name : feature_num
Attribute : Number of features in feature_list field
Type : BigINT
Range : Nonnegative

Name : features_list

Format : Composed of feature structures separated by ASCII character **0x01**, total number of feature structures is indicated by features_num.
Examples: feature_num = 2
feature_list = feature_structure_1 **0x01** feature_structure_2

Feature Structure

Format : feature_field_id **0x02** feature_id **0x03** feature_value

Name : feature_field_id
Attribute : Field identity of several features
Type : String
Range : N/A

Name : feature_id
Attribute : Global identity of features
Type : BigINT
Range : Positive integer

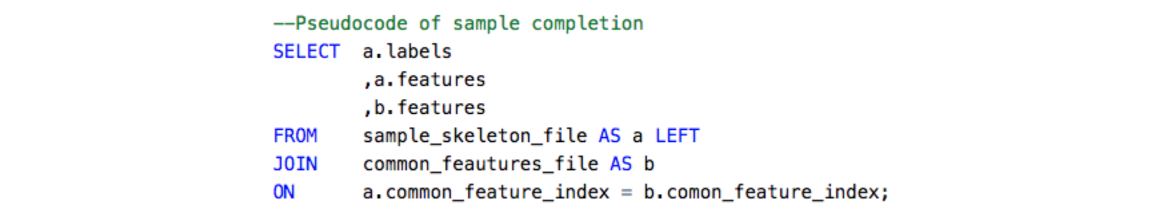
Name : feature_value
Attribute : feature value
Type : Double
Range : Real number

Format : All fields are separated by comma (CSV file)

Figure 4: Illustration of common features.

5. Pseudocode of Sample Completion

A complete sample record should contain not only the features from a record in sample skeleton file, but also the features from related record in common features file, according to the pseudocode shown in Fig.5.



```
--Pseudocode of sample completion
SELECT  a.labels
        ,a.features
        ,b.features
FROM    sample_skeleton_file AS a LEFT
JOIN    common_feautures_file AS b
ON      a.common_feature_index = b.comon_feature_index;
```

Figure 5: Pseudocode of sample completion.

6. Citation

To acknowledge use of the dataset in publications, please cite the following paper:

> ***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.

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。

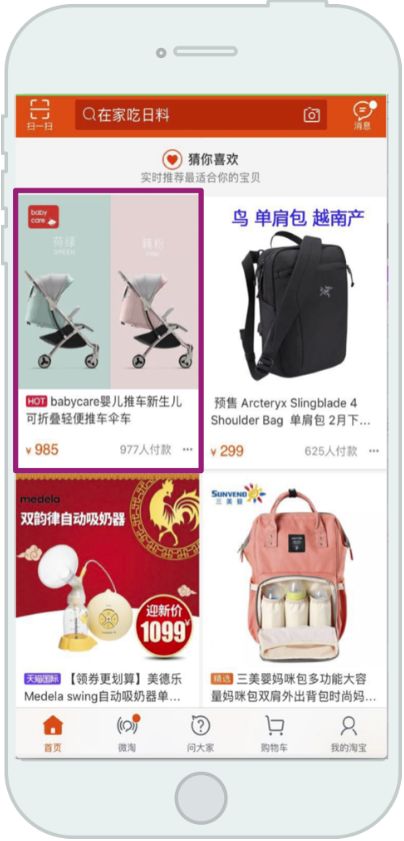

扫一扫
在家吃日料
消息
猜你喜欢
实时推荐最适合你的宝贝
HOT babycare婴儿推车新生儿可折叠轻便推车伞车
¥985 977人付款
鸟 单肩包 越南产
预售 Arcteryx Slingblade 4 Shoulder Bag 单肩包 2月下…
¥299 625人付款
medela
双韵律自动吸奶器
迎新价 1099
天猫国际 【领券更划算】美德乐 Medela swing自动吸奶器单…
精选 三美婴妈咪包多功能大容量妈咪包双肩外出背包时尚妈…
首页
微淘
问大家
购物车
我的淘宝

图 1: 手机淘宝上的单品广告

本数据的整体描述为$\{(x_i, y_i \rightarrow z_i)\}|_{i=1}^{N}$，样本的格式为$(\boldsymbol{x}, y \rightarrow z)$，是来自定义域$(x, y \rightarrow z)$的分布D，其中$\mathcal{X}$是特征空间，$\mathcal{Y}$和$\mathcal{Z}$是标签空间，$N$是样本总量。$\boldsymbol{x}$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$y \rightarrow z$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。

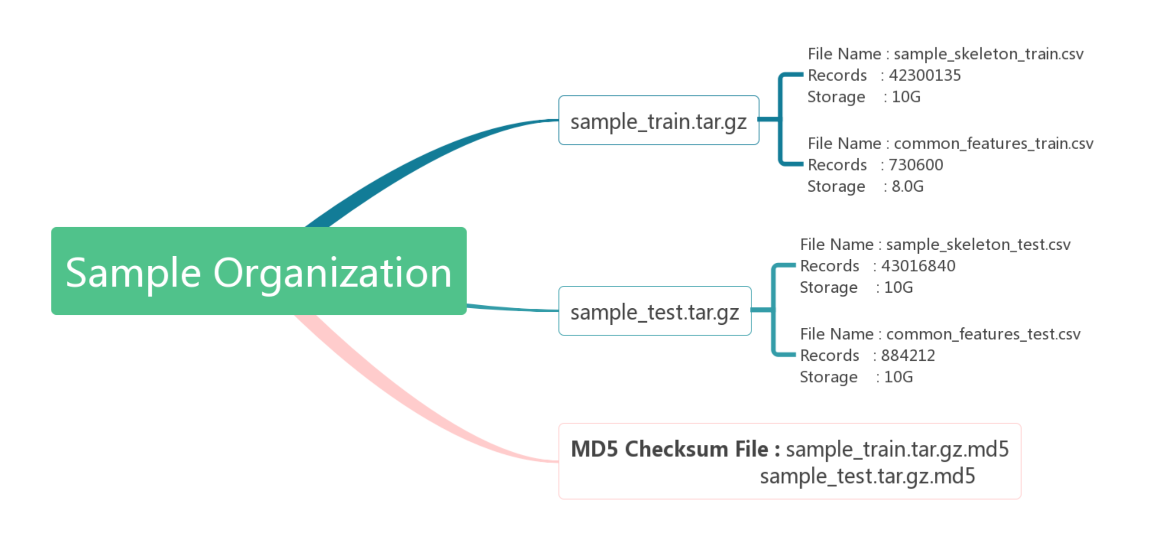


图 2: 样本组织说明

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。

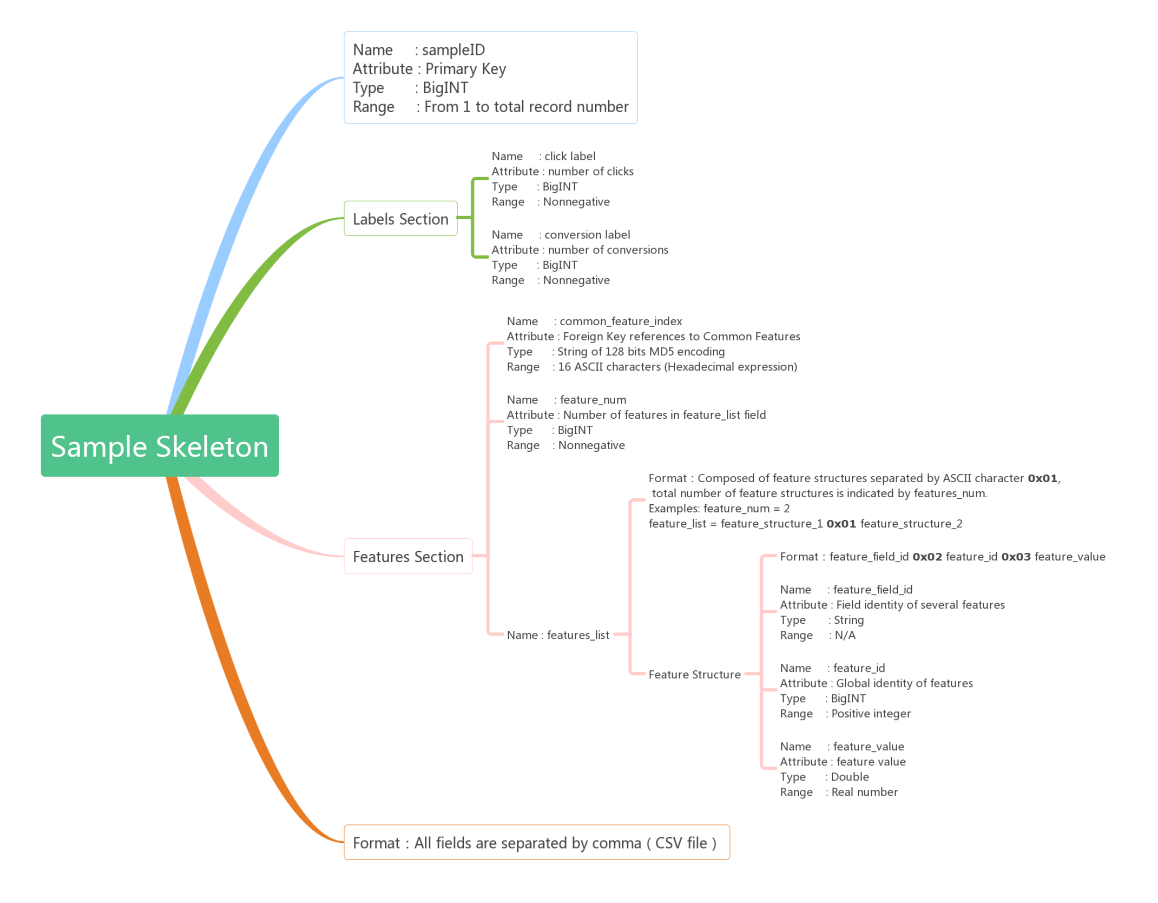

Sample Skeleton
Name : sampleID
Attribute : Primary Key
Type : BigINT
Range : From 1 to total record number
Labels Section
Name : click label
Attribute : number of clicks
Type : BigINT
Range : Nonnegative
Name : conversion label
Attribute : number of conversions
Type : BigINT
Range : Nonnegative
Features Section
Name : common_feature_index
Attribute : Foreign Key references to Common Features
Type : String of 128 bits MD5 encoding
Range : 16 ASCII characters (Hexadecimal expression)
Name : feature_num
Attribute : Number of features in feature_list field
Type : BigINT
Range : Nonnegative
Name : features_list
Format : Composed of feature structures separated by ASCII character **0x01**,
total number of feature structures is indicated by features_num.
Examples: feature_num = 2
feature_list = feature_structure_1 **0x01** feature_structure_2
Feature Structure
Format : feature_field_id **0x02** feature_id **0x03** feature_value
Name : feature_field_id
Attribute : Field identity of several features
Type : String
Range : N/A
Name : feature_id
Attribute : Global identity of features
Type : BigINT
Range : Positive integer
Name : feature_value
Attribute : feature value
Type : Double
Range : Real number
Format : All fields are separated by comma (CSV file)

图 3: 样本骨架说明

- 样本**ID**部分

> 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

> 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：

1. 索引（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。
2. 特征数量（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。
3. 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。

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。
	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。

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。

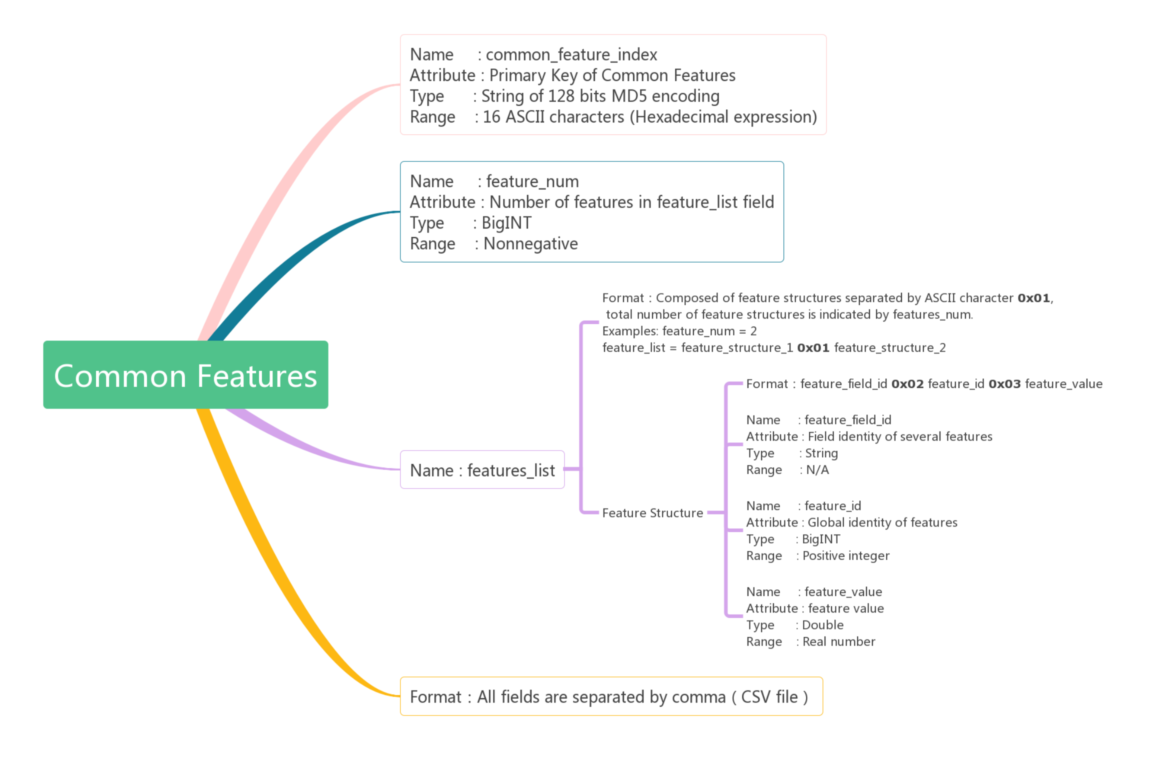
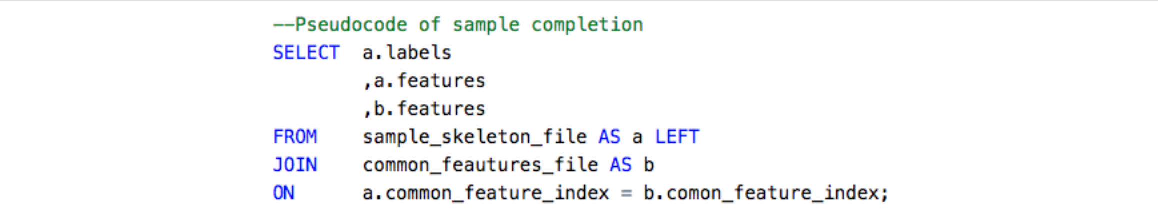

Common Features
Name : common_feature_index
Attribute : Primary Key of Common Features
Type : String of 128 bits MD5 encoding
Range : 16 ASCII characters (Hexadecimal expression)
Name : feature_num
Attribute : Number of features in feature_list field
Type : BigINT
Range : Nonnegative
Name : features_list
Format : Composed of feature structures separated by ASCII character **0x01**,
total number of feature structures is indicated by features_num.
Examples: feature_num = 2
feature_list = feature_structure_1 **0x01** feature_structure_2
Feature Structure
Format : feature_field_id **0x02** feature_id **0x03** feature_value
Name : feature_field_id
Attribute : Field identity of several features
Type : String
Range : N/A
Name : feature_id
Attribute : Global identity of features
Type : BigINT
Range : Positive integer
Name : feature_value
Attribute : feature value
Type : Double
Range : Real number
Format : All fields are separated by comma (CSV file)

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。



```
--Pseudocode of sample completion
SELECT  a.labels
        ,a.features
        ,b.features
FROM    sample_skeleton_file AS a LEFT
JOIN    common_feautures_file AS b
ON      a.common_feature_index = b.comon_feature_index;
```

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.

关于我们 法律协议 了解更多，请关注天池微信

解决方案

行业架构师咨询

成熟解决方案 | 专业架构师咨询 | 成功客户案例分享 | 专属定制服务

热门产品 云服务器ECS 云计算 云数据库RDS 云存储OSS NAT网关 负载均衡 域名注册 网站建设 大数据

用户热搜 网站备案 云安全 网安法 CDN加速 API网关 企业邮箱 whois查询 视频直播 视频转码

更多推荐 全民云计算 免费套餐 学生机 IT论坛 数据可视化 云虚机 com域名 cn域名

合规安全解决方案

关于我们　法律声明及隐私权政策　廉正举报　联系我们　加入阿里云

阿里巴巴集团　淘宝网　天猫　聚划算　全球速卖通　阿里巴巴国际交易市场　1688　阿里妈妈　飞猪　阿里云计算　YunOS　阿里通信　万网　高德　UC　友盟　虾米　阿里星球　钉钉　支付宝