Train & Test

We randomly sampled 100000 users from an on-line shopping website for 8 days (20170506-20170513) of ad display / click logs (more than 2 million records) to form the original sample skeleton. Field description is as follows:

(1) user: User ID(int);

(2) time_stamp: time stamp(Bigint, 1494032110 stands for 2017-05-06 08:55:10);

(3) adgroup_id: adgroup ID(int);

(4) pid: scenario, location of the AD on the screen

(5) noclk: 1 for not click, 0 for click;

(6) clk: 1 for click, 0 for not click;

NOTE: noclk and clk are somehow replicated information, you only have to predict the value of clk.

(The training set contains the data of the first 7days, and test set is the last day)

ad_feature

This data set covers the basic information of all ads in raw_sample. Field description is as follows:

(1) adgroup_id： Ad ID(int) ;

(2) cate_id： Ad category ID;

(3) campaign_id： campaign ID;

(4) brand： brand ID;

(5) customer_id: Advertiser ID;

(6) price: the price of item

Each ad ID corresponds to an item, an item belongs to a category, an item belongs to a brand.

user_profile

This data set covers the basic information of 1060000 users. All the users in the Train & Test are included in. Field description is as follows:

(1) userid: user ID;

(2) cms_segid: Micro group ID;

(3) cms_group_id: cms_group_id;

(4) final_gender_code: gender 1 for male , 2 for female

(5) age_level: age_level

(6) pvalue_level: Consumption level, 1: low, 2: mid, 3: high

(7) shopping_level: Shopping depth, 1: shallow user, 2: moderate user, 3: depth user

(8) occupation: Is the college student 1: yes, 0: no

(9) new_user_class_level: City level

behavior_log

This data set covers the shopping behavior in 22 days of all users in train and test data (totally sixty million records). Field description is as follows:

(1) nick: User ID(int);

(2) time_stamp: time stamp(Bigint, 1494032110 stands for 2017-05-06 08:55:10);

(3) btag: Types of behavior, include the following four:

type	explanation
ipv	browse
cart	add to the shopping cart
fav	favor
buy	buy

(4) cate: category ID of the item(int);

(5) brand: item brand ID(int);

Here if we use userID and timestamp as primary key, we will find a lot of duplicate records. This is because the behavior of different types of the data are collected from different departments and when packaged together, there are small deviations (i.e. the same two timestamps may be two different time with a relatively small difference).