构建训练集、预测集、离线测试集 II --- 特征抽取

> *训练集和预测集的构建包括了操作记录抽取、特征提取、正负样本标注、正负样本比例调整等步骤*
> *本文档主要包含 `特征抽取` 部分*

涉及表的含义

1.之前存在的表

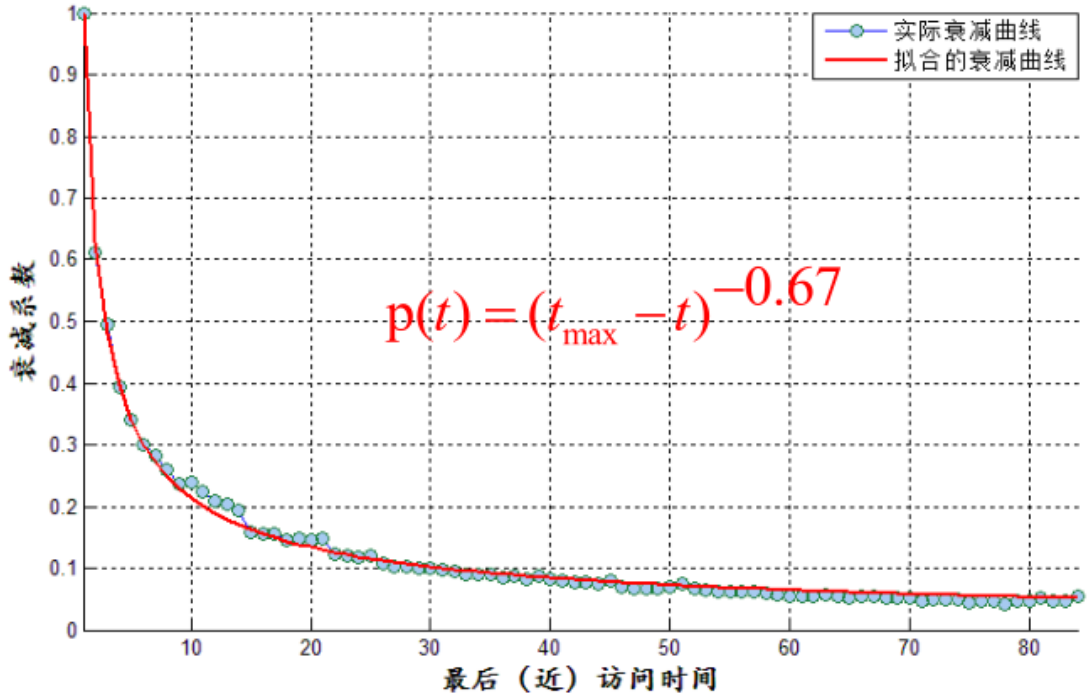
- `tianchi_p` 所有对商品子集P中商品的操作记录
- `tianchi_p_1_30` 11月18日到12月17日对商品子集的操作记录
- `tianchi_p_ten_1_30` 11月18日到12月17日间最后10天有操作的用户对商品子集的操作记录
- `tianchi_p_2_31` 11月19日到12月18日对商品子集的操作记录
- `tianchi_p_ten_2_31` 11月19日到12月18日间最后10天有操作的用户对商品子集的操作记录
- `tianchi_p_31_buy` 12月18日当天的购买情况
- `tianchi_p_ten_buy_1_30` 11月18日到12月17日间有购买过且最后10天有操作的用户对商品子集的操作记录
- `tianchi_p_ten_buy_2_31` 11月19日到12月18日间有购买过且最后10天有操作的用户对商品子集的操作记录
- `tianchi_p_ten_nobuy_1_30` 11月18日到12月17日间没有购买过、最后10天有操作且最后5天操作次数大于10小于500（有待商榷）的用户对商品子集的操作记录
- `tianchi_p_ten_nobuy_2_31` 11月19日到12月18日间没有购买过、最后10天有操作且最后5天操作次数大于10小于500（有待商榷）的用户对商品子集的操作记录

2.新创建的表

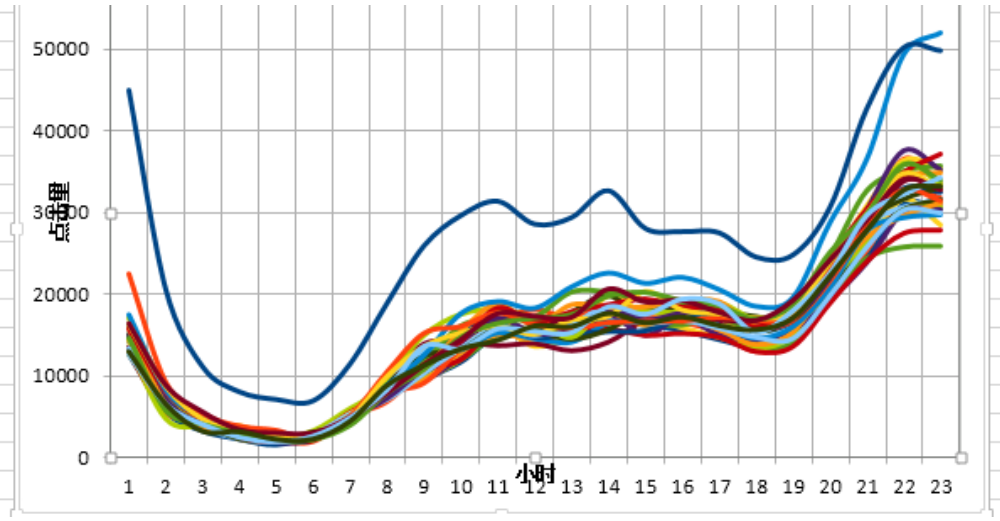
- `train_Nf_buy` 对购买过用户构建的训练集，包含了user_id,item_id以及从`tianchi_p_ten_buy_1_30`抽取的N个特征，但未加入正负标注
- `pre_Nf_buy` 对购买过用户构建的预测集，包含了user_id,item_id以及从`tianchi_p_ten_buy_2_31`抽取的N个特征

时间段划分

> - *由于双十二操作量激增，故将双十二（12月11日:Day 24 和12月12日:Day 25）作为一个时间段来单独考虑；同时将12月10日（Day 23）和12月13日(Day 26)当做连续的两天。*



- *按照购物衰减曲线选取距离预测日最近N天内的操作特征，离预测日越近分的越细，如最近1、2、3、5、7、10、14、20、28天的操作特征*
- *按照固定时间间隔来统计操作特征，如每一天、每三天、每一周的操作特征*
- *对时间间隔操作特征抽取统计特征，如最近N天内有多少个单天、三天、一周操作过某商品*
- *关键时间点：第一次交互、最后一次交互的时间*
- *根据小时操作分布曲线，将小时取值分成三类：白天Daytime(8-18)、晚上Evening(19-23 & 0-1)、深夜Night(2-7)，将小时和日期联合起来抽取特征*



行为划分

“

- `交互行为`*是购买、点击、收藏、购物车4种行为的总和*

- *留待行为是收藏、购物车2种行为的总和，这是因为2者皆是先留下再等待后续操作的一种行为*

特征键命名规则

“

- *对于 训练集，命名为 train_...*
- *对于 预测集，命名为 pre_...*
- *对于 最后N天，命名形式为 train_l7_...(最后7天)*
- *对于 时间段，双12命名为 train_d12... ; 其余以起止时间命名，如 train_1_7_...*
- *对于 时间点，第一次命名为 train_f... , 最后一次命名为 train_l_...*
- *对于 小时，白天命名为 train_l7_day , 晚上命名为 train_l7_eve , 深夜命名为 train_l7_nt_...*
- *对于 行为，购买命名为 train_l7_buy_... ,点击、购物车、收藏、交互、留待为clk、cart、favor、act、later*
- *对于 最近N天有多少个单天、三天、一周 这种统计特征，命名为 train_l28_h7_... (意思为how many 7days)*
- *命名顺序 训练集/预测集——最后N天/时间段/时间点——（多少个M天/小时分类，如果有）——行为类别*

特征种类

“

- *I.用户特征：描述用户属性，表明用户的 购买偏好*
- *II.商品特征：描述单个商品属性，表明商品的 购买价值*
- *III.类别特征：描述商品类的属性,表明类别的 大致性质，如大件还是小件、是否会被周期购买*
- *IV.用户-商品特征：描述用户-商品对的属性，它包含两类特殊的特征：*
  - *商品竞争特征：描述不同商品之间的属性*
  - *类别竞争特征：描述商品类之间的属性*

I. 用户特征

1. 最近N天特征

- 用户在最近的1、3、7、10、14、20、28天内的购买、点击、交互、留待所有商品的总次数
- 用户在最近的1、3、7、10、14、20、28天内购买、点击、交互、留待了多少个不同的商品

2. 时间段特征

- 用户在双12（Day24、25）期间购买、点击、交互所有商品的总次数
- 用户在每一周（对于训练集为：Day1-7、Day8-14、Day15-21、Day22-23&26-30）购买、点击、交互、留待所有商品的总次数 （可能和最近N天特征有重复）

3. 时间点特征

- 用户第一次、最后一次购买、交互任意商品的日期

4. 时间段统计特征

- 用户在最近7天、14天、28天有多少单天购买过、交互过任意商品

5. 日期加小时特征

- 用户在最近1、3、5、7天的白天、晚上、深夜购买、点击、交互所有商品的总次数（有待商榷）

6. 转化率特征

- 用户对所有商品的 点击to购买转化率、点击to留待转化率、留待to购买转化率（如果购买或留待为0，则设为0，因为并没有转化过）（还可采用Laplace平滑 $\frac{x}{y} = \frac{x+ab}{y+b}$）

II. 商品特征

1. 最近N天特征

- 商品在最近的1、3、7、14、28天内的日均购买量、日均交互量
- 商品在最近的1、3、7、14、28天内被多少不同用户购买、点击、交互
- 商品在最近的1、3、7、14、28天内的人均购买量、人均点击量、人均交互量

2. 时间段特征

- 商品在双12（Day24、25）期间的日均购买量、日均交互量、日均点击量、人均购买量、人均交互量、人均点击量
- 商品在每一周（对于训练集为：Day1-7、Day8-14、Day15-21、Day22-23&26-30）的日均购买量、日均交互量、人均购买量、人均交互量（可能和最近N天特征有重复）

3. 时间点特征

- 商品第一次、最后一次被购买、交互的日期

4. 重复购买特征

- 商品总共被多少人在不含双12的不同天（把TA称为回头客）、不同周购买过
- 商品的回头客率：回头客数/总的购买人数

5. 日期加小时特征

- 商品在所有天（含双12）的白天、晚上、深夜的人均购买量、人均点击量、人均交互量 （有待商榷）

III. 类别特征

1. 最近N天特征

- 类别在最近的7、14、28天内的日均购买量、人均购买量

2. 时间段特征

- 类别在双12（Day24、25）期间的日均购买量、日均交互量、人均购买量、人均交互量
- 类别在每一周（对于训练集为：Day1-7、Day8-14、Day15-21、Day22-23&26-30）的日均购买量、人均购买量（可能和最近N天特征有重复）

3. 重复购买特征

- 类别总共被多少人在不含双12的不同天、不同周购买过 （不同天意味着应是可重复购买的类别、不同周意味着有可能是周期性类别）

4. 日期加小时特征

- 类别在所有天（含双12）的白天、晚上、深夜的人均购买量、人均交互量 （有待商榷）

IV. 用户-商品特征

1. 最近N天特征

- 用户在最近的1、2、3、5、7、10、12、15、20、28天内的购买、点击、交互、留待此商品的总次数

2. 时间段特征

- 用户在双12（Day24、25）期间购买、点击、交互此商品的次数
- 用户在每一周（对于训练集为：Day1-7、Day8-14、Day15-21、Day22-23&26-30）购买、点击、交互、留待此商品的次数 （可能和最近N天特征有重复）
- 用户在最近15天的每3天购买、点击、交互此商品的次数 （可能和最近N天特征有重复）

3. 时间点特征

- 用户第一次、最后一次购买、交互此商品的日期
- 用户第一次和最后一次购买、交互此商品的日期相隔多少天
- 用户对此商品的第一个交互天的点击数、收藏数、购买数
- 用户对此商品的最后一个交互天的点击数、收藏数、购买数

4. 时间段统计特征

- 用户在最近3天、5天、7天、14天，最初14天有多少单天购买过、交互过此商品

5. 日期加小时特征

- 用户在最近1、3、5、7天的白天、晚上、深夜购买、点击、交互此商品的总次数（有待商榷）

6. 转化率特征

- 用户对此商品的 点击to购买转化率、点击to留待转化率、留待to购买转化率（如果购买或留待为0，则设为0，因为并没有转化过）（还可采用Laplace平滑 $\frac{x}{y} = \frac{x+ab}{y+b}$）

7. 行为间特征

- 用户在购买此商品之前的点击次数、收藏次数、加购物车次数 （抽取时，若在同一天，则由小时定义是否在之前！！！）

8. *商品竞争特征*

- 用户最后一次交互此商品与最后一次交互任意商品相隔多少天
- 用户最后一次交互此商品的当天、后一天分别交互了多少个其它商品
- 用户最后一次交互此商品的当天、后一天分别点击、购买、交互了多少次其它商品
- 用户在最近1、3、5、7、10、14、28天内有多少个单天只交互了此商品
- 用户在四周（Week 1-4）中有几周只交互了此商品

9. *类别竞争特征*

- 用户最后一次交互此类别与最后一次交互行为相隔多少天

创建训练集和预测集

> *对表`tianchi_p_ten_buy_1_30`抽取的N维特征放入表`train_Nf_buy`中作为`训练集`，它包含了user_id,item_id及抽取的N个特征，但未加入正负标注（result=1/0，即在Day31买或者没买）*
> *对表`tianchi_p_ten_buy_2_31`抽取的N维特征放入表`pre_Nf_buy`中作为`预测集`，它包含了user_id,item_id及抽取的N个特征*