构建训练集、预测集、离线测试集 III --- 正负样本

> *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`正负样本`部分*

涉及表的含义

- `tianchi_p_31_buy` 12月18日当天的购买情况
- `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

正负样本标注

- 根据表`tianchi_p_31_buy`中的*用户-商品对*为`train_Nf_buy`的条目加入`result`列。若*用户-商品对*恰好吻合，则result为1，即正样本，反之result为0，为负样本。

正负样本比例调整

- 由于正样本要`远少于`负样本，所以要对比例进行调整
- 对正样本进行复制，同时对负样本进行抽样
- 最后正负样本比例达到`1：5~1：10`左右