阿里移动推荐算法大赛答辩

2015天池大数据竞赛

TIANCHI天池

CHLL

陈靖

中国科学技术大学

提纲

- 解题思路
- 算法介绍
  - 样本
  - 特征
  - 模型
- 总结回顾
  - 涨分技巧
  - 参赛收获

解题思路

问题描述：

给出一定量用户在11.18~12.18之内的移动端行为数据（D），需要预测12.19对商品子集（P）的购买数据。行为分为四种：1(浏览)，2(收藏)，3(加购物车)，4(购买)

user_id	item_id	behavior_type	user_geohash	item_category	time
99512554	37320317	3	94gn6nd	9232	2014-11-26 20
9909811	266982489	1		3475	2014-12-02 23
98692568	27121464	1	94h63np	5201	2014-11-19 13
96089426	114407102	1	949g5i3	836	2014-11-26 07
90795949	402391768	1	94h6dlp	3046	2014-12-09 21
96363456	379126815	1		10732	2014-12-07 23
95993830	78579528	1		5027	2014-12-10 10
95591350	58429334	1	95ipq3o	4190	2014-12-11 13
96927552	101192540	2	94oid72	4280	2014-12-05 19

解题思路

- 问题

分类问题→二分类，1:买，0:不买

- 样本

样本选取→有交互P子集样本

在哪一天交互的样本可能会在19号购买？

16号、17号、18号？

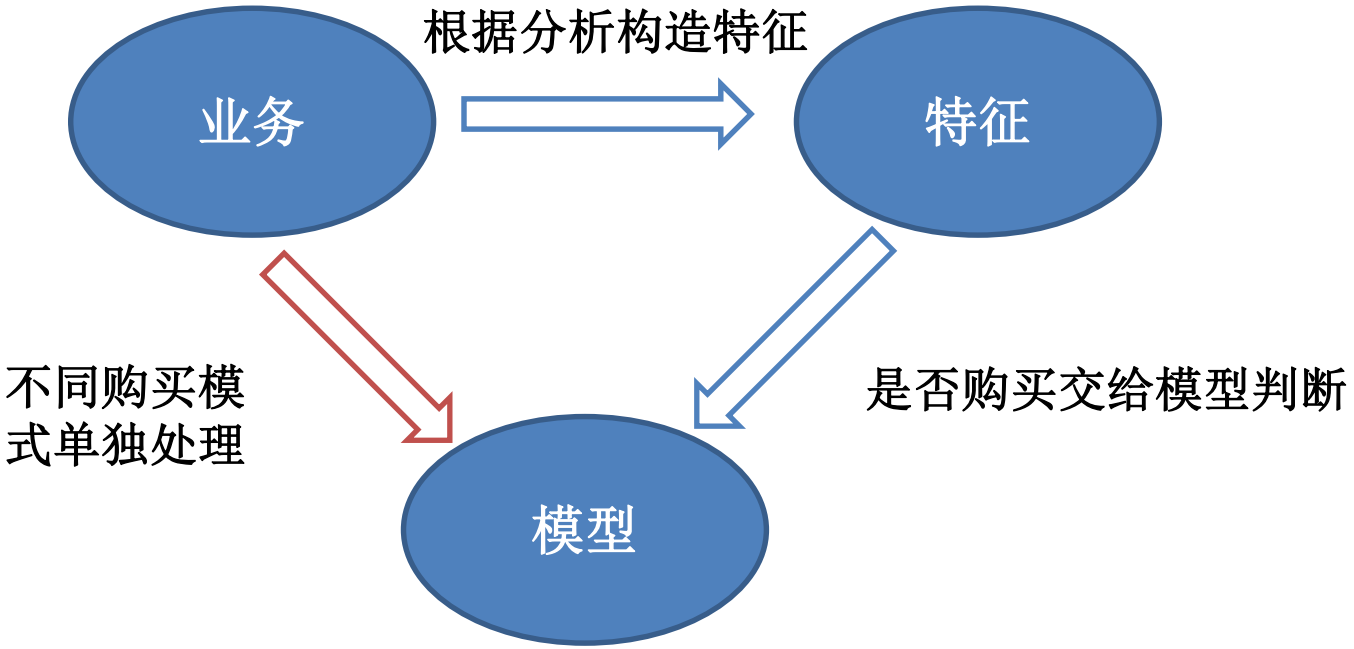
- 特征

什么样的用户在什么时候对什么样的商品有过什么样的操作之后可能会在19号购买？是否买过？商品销量？加购物车？

- 模型

分类、回归。（LR、RF、GBDT）

解题思路



解题思路

第一赛季

数据量小，更换数据前后模型效果波动大

规则过滤：过滤18号的交互样本

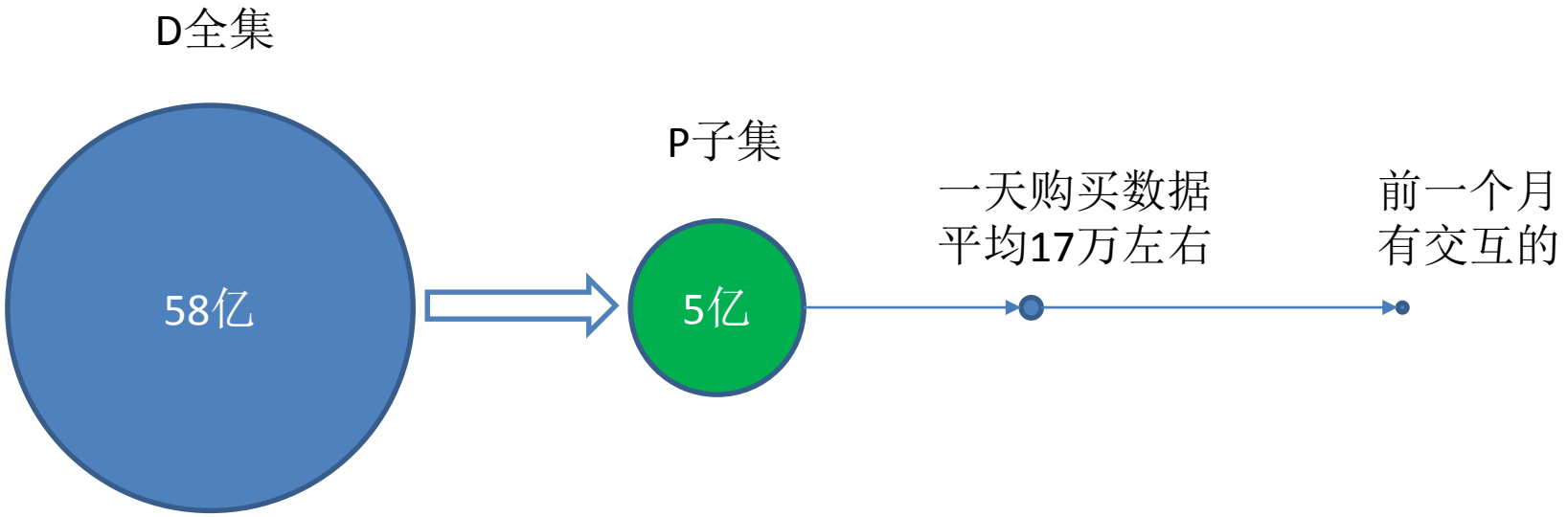
cart(u, i) && cart_time(u, i) > 13 && buy_time(u) < cart_time(u, i)　　F1=9.9

转化成特征

是否加购物车、加购物车时间、用户发生购买行为的时间（小时）

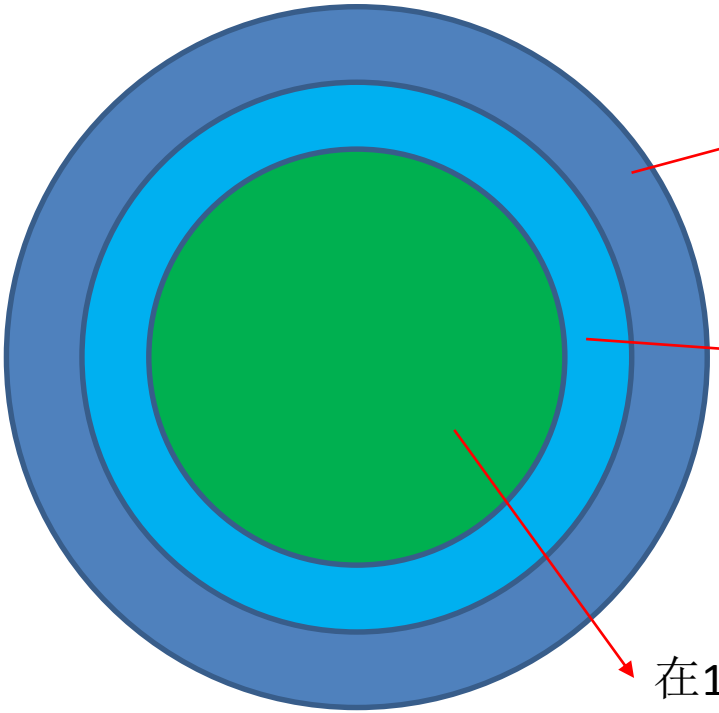
解题思路

第二赛季



解题思路

- 购买数据分布



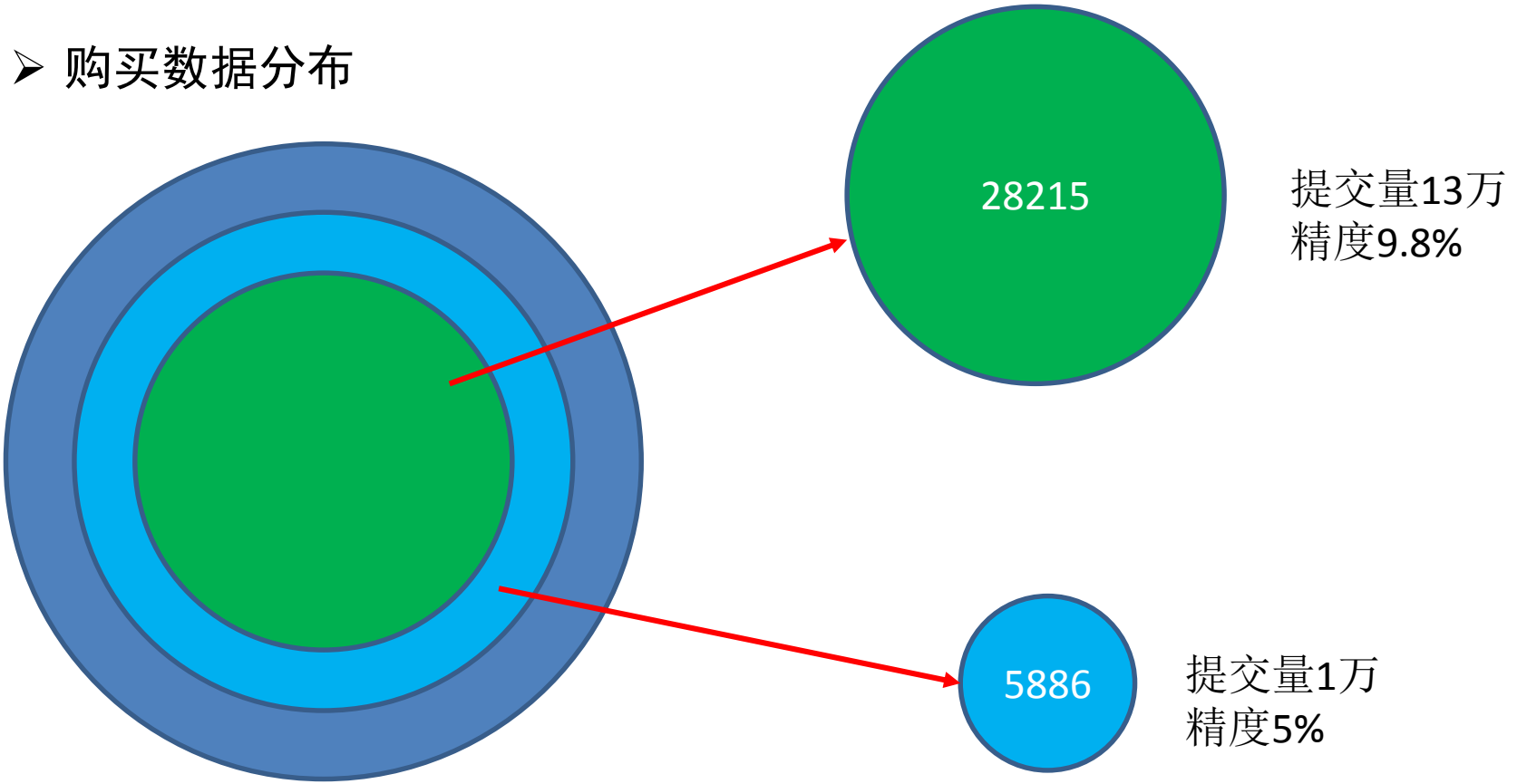
在12.18号购买且前一个月交互过

在12.18号购买且前两天交互过

在12.18号购买且前一天交互过

解题思路

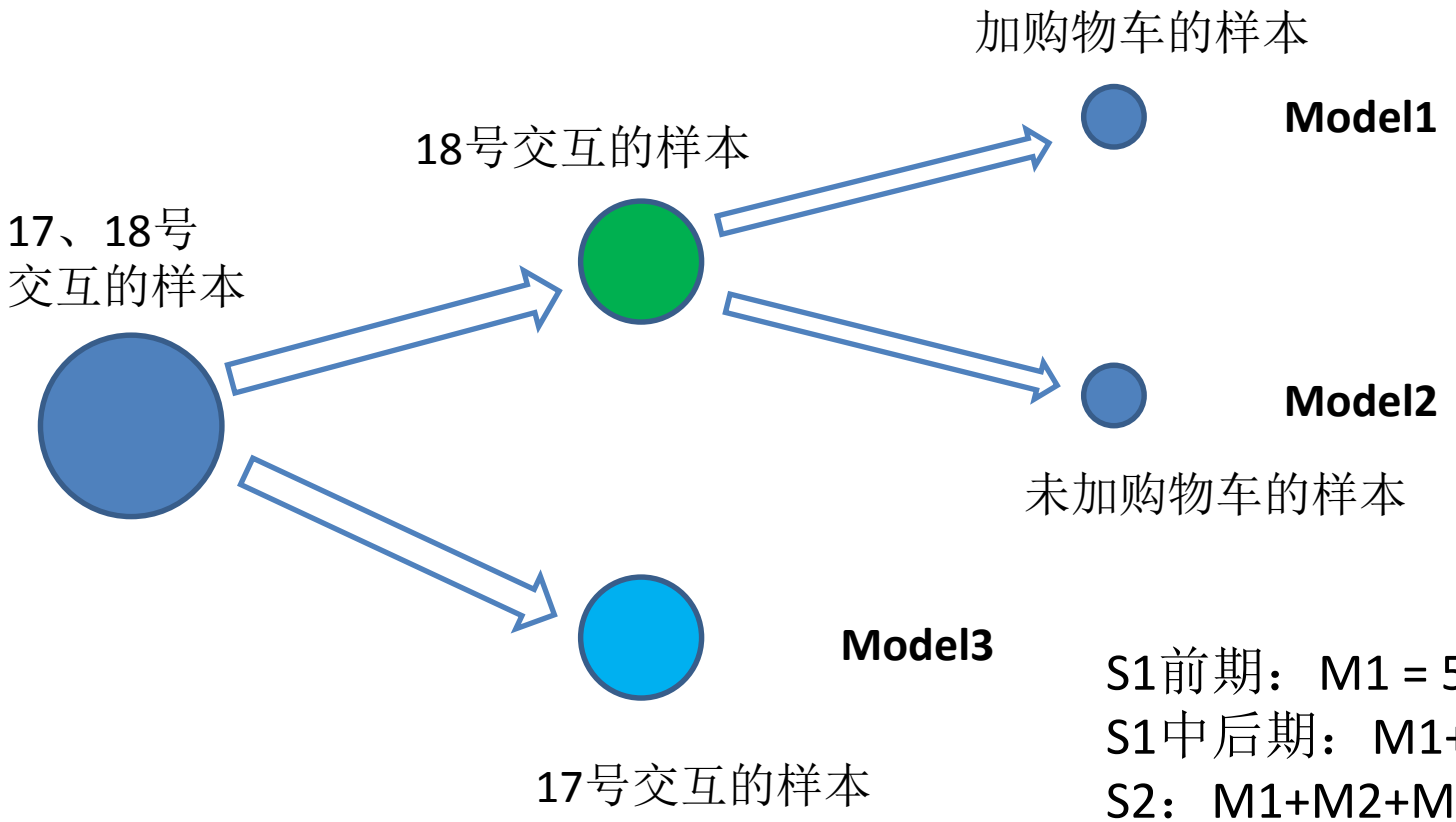
- 购买数据分布



越往前交互越难预测，购买可能性越低

算法介绍

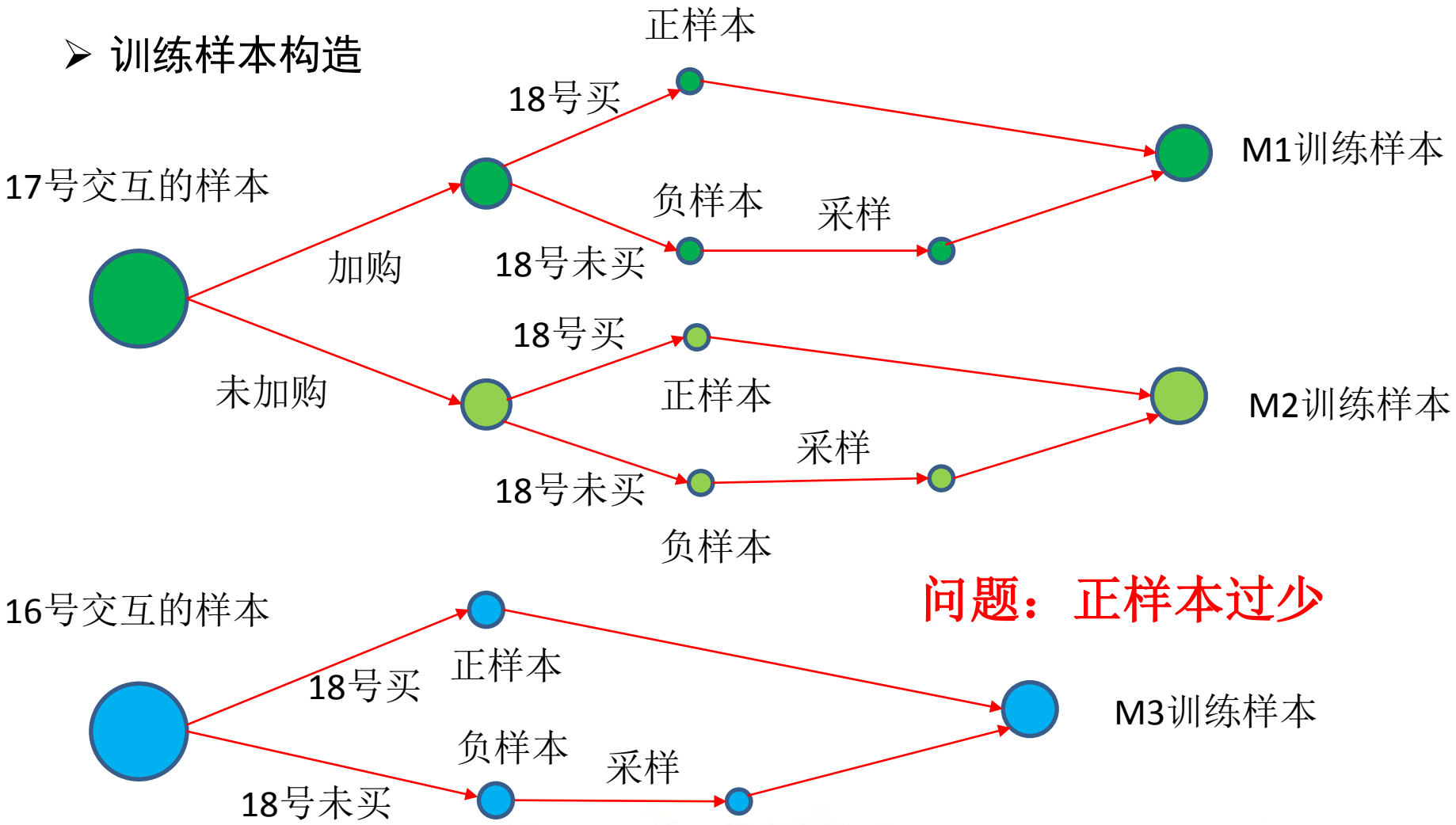
- 待预测样本选取



S1前期：M1 = 5.3+
S1中后期：M1+M2 = 6.2+
S2：M1+M2+M3 = 8.63

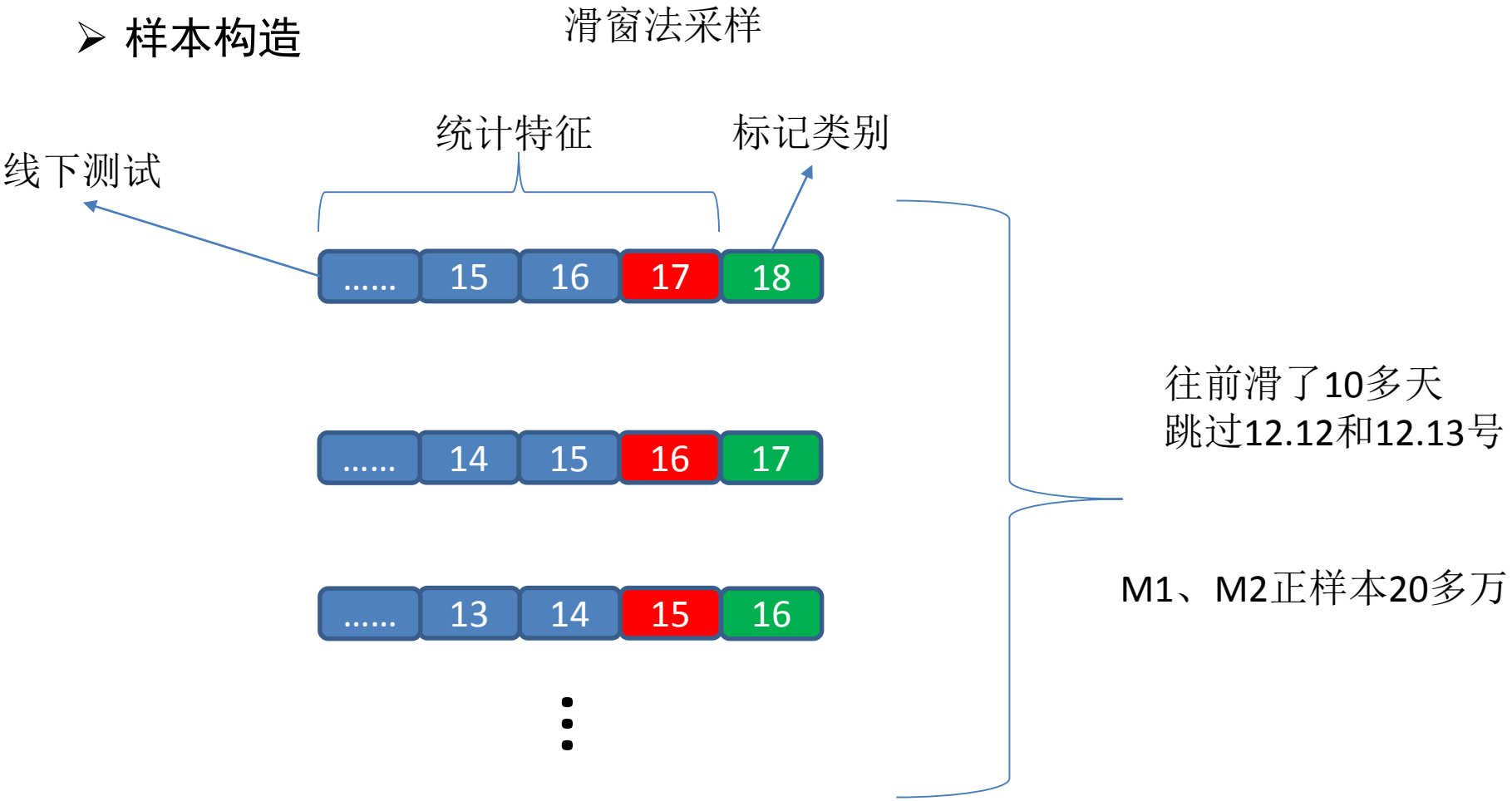
算法介绍

- 训练样本构造


正样本
18号买
17号交互的样本
M1训练样本
负样本
采样
加购
18号未买
18号买
未加购
正样本
M2训练样本
采样
18号未买
负样本
问题：正样本过少
16号交互的样本
正样本
18号买
M3训练样本
负样本
采样
18号未买

算法介绍

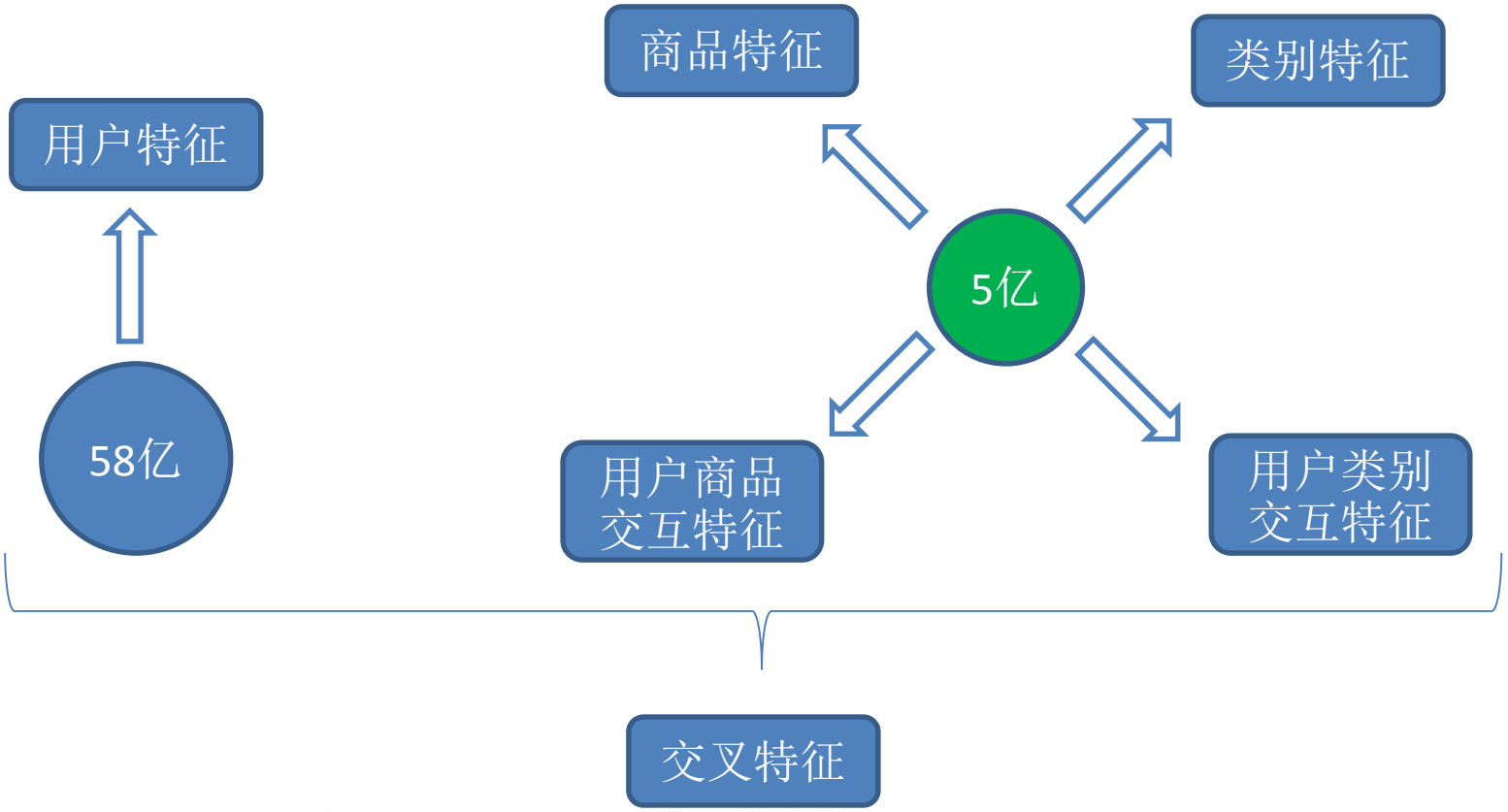
- 样本构造


滑窗法采样
统计特征
标记类别
线下测试
..... 15 16 17 18
..... 14 15 16 17
..... 13 14 15 16
往前滑了10多天
跳过12.12和12.13号
M1、M2正样本20多万

算法介绍

- 特征

六大类特征



算法介绍

- 特征

统计型特征、比值类特征、时间特征

可能影响用户购买的因素：
加购物车？
访问该商品后是否买了其他东西？
交互当天访问了多少商品？
购买点击比？
是否买过？
……

uif_add_cart
uif_visit_time
uf_buy_time
uf_visit_count
uf_buy/uf_click
uf_buy_all

侧重点：用户特征—用户在交互当天的行为特征
是否会购买主动权在用户，把用户的特征做细

算法介绍

- 特征

交互当天	用户在当天交互了多少类别、商品，在线时长，离线时间，当天是否购买，购买时间，交互商品时间，交互商品之后是否发生购买行为。商品、类别在当天的销售情况
3/5/7/10/all天内	用户的购买习惯、购买天数，点击、收藏、购买量，购买量/点击量 商品、类别销售情况，商品购买量占所在类别购买量的比例 用户访问商品次数占访问该类别的比例

特征处理：缺省或出现分母为0时用-1填充

最终使用特征维数：610

算法介绍

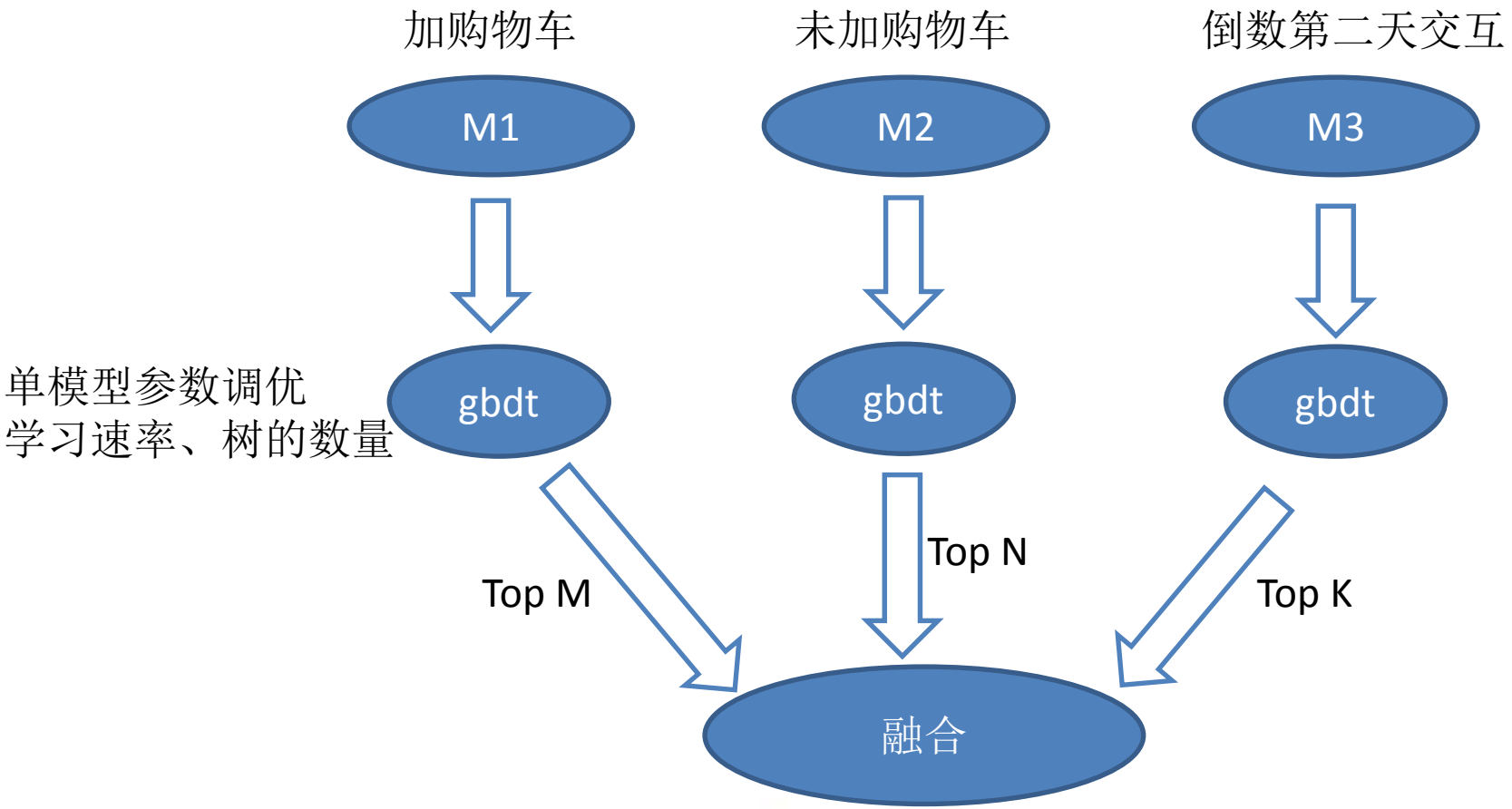
- 模型

测试效果

模型选择

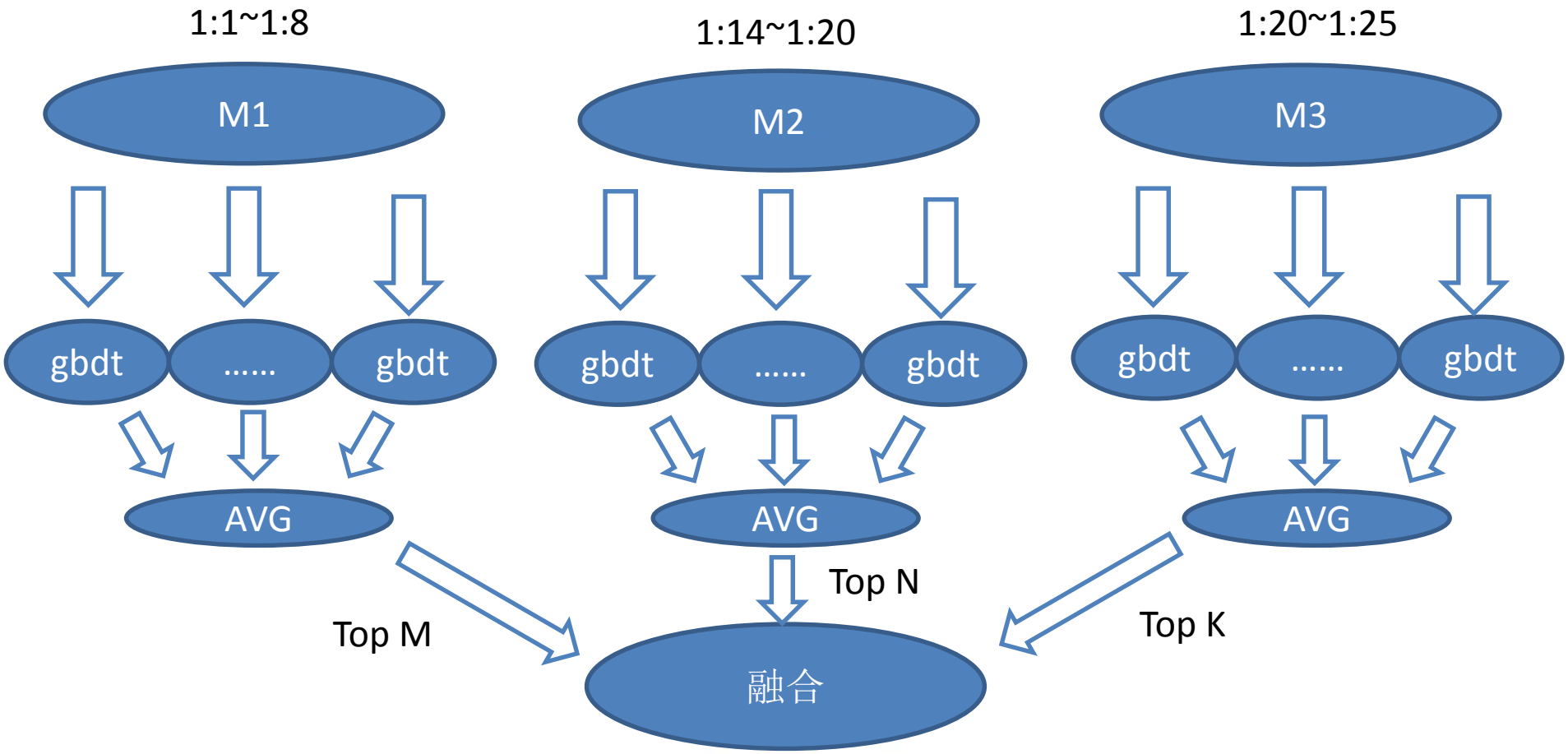
算法介绍

- 模型


加购物车
未加购物车
倒数第二天交互
M1
M2
M3
gbdt
gbdt
gbdt
单模型参数调优
学习速率、树的数量
Top M
Top N
Top K
融合

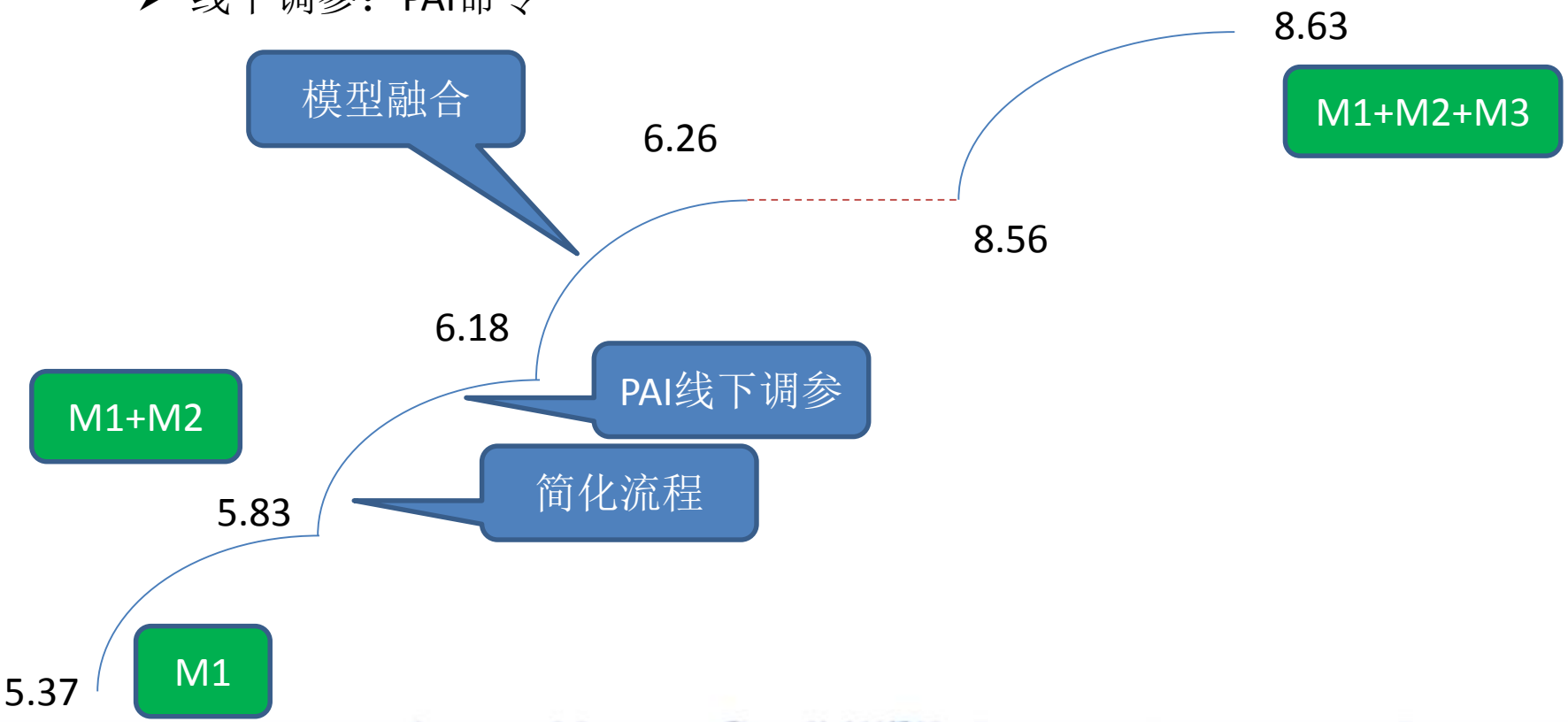
算法介绍

- 模型融合


1:1~1:8
1:14~1:20
1:20~1:25
M1
M2
M3
gbdt
……
gbdt
gbdt
……
gbdt
gbdt
……
gbdt
AVG
AVG
AVG
Top N
Top M
Top K
融合

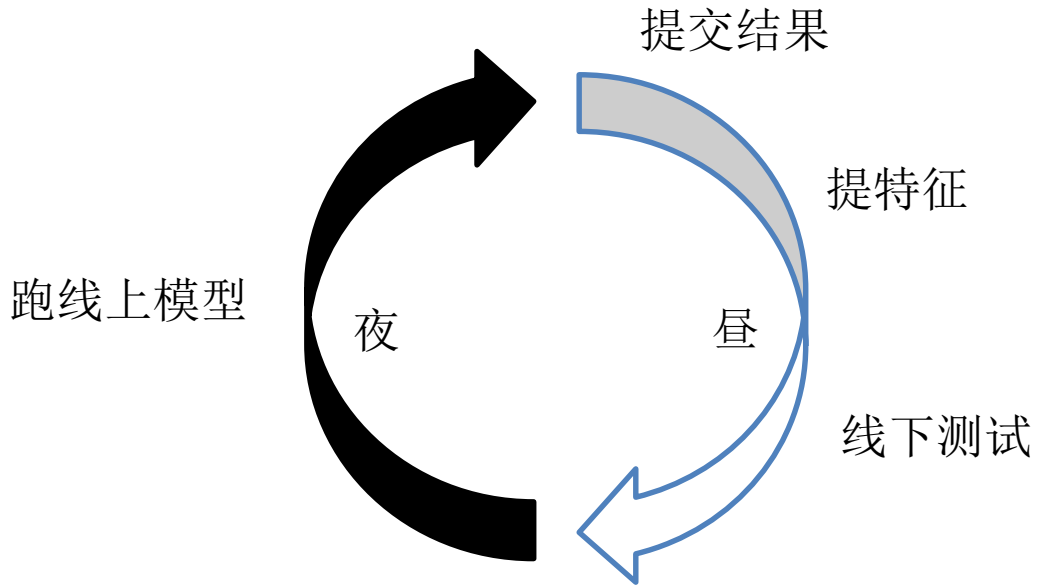
总结回顾

- 涨分技巧
  - 流程尽量简化，从新增特征到模型输出步骤精简（4步）
  - 线下调参：PAI命令



总结回顾

- 如何高效打比赛



总结回顾

- 参赛收获
  - 真正接触大数据，特征提取技能getV
  - 学习了数据挖掘相关算法
  - 认识了很多志同道合的人

谢谢