
阿里移动推荐算法大赛答辩
2015天池大数据竞赛
TIANCHI天池

北京仰望星空大学第一Carry

阿里移动推荐算法大赛分享

中科院计算技术研究所
顾茂杰 李强

目录

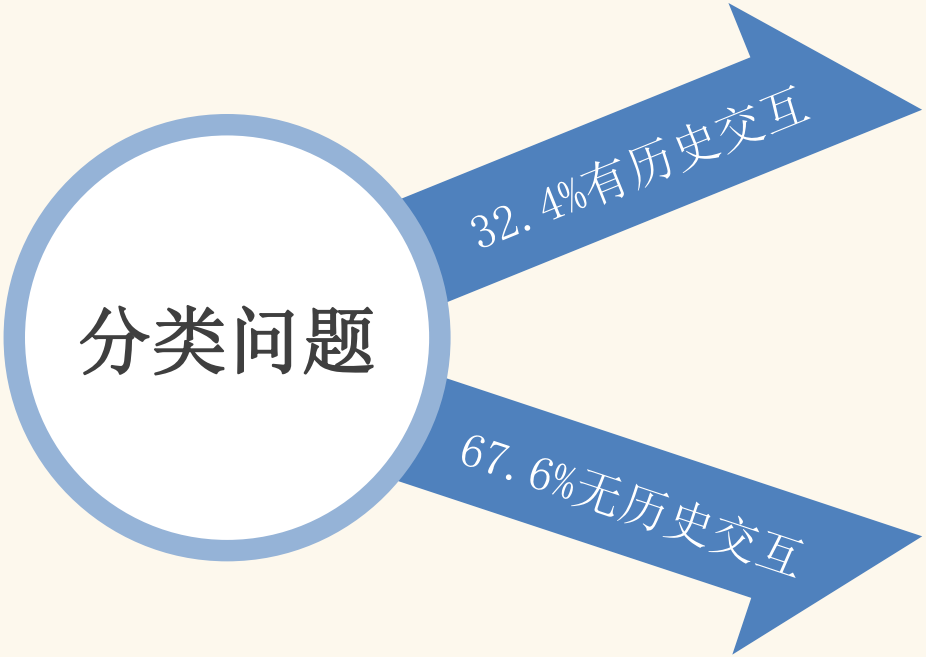
第1部分

赛题建模

问题描述


500万用户
一个月
58亿次交互
1
2
1.5亿商品
浏览 收藏
购物车
购买
使用
F1值
评估
4
3
预测12月19日
用户是否会购
买这个商品

问题建模



分析

根据历史交互行为预测
用户偏好
商品价值

可用模型

神经网络、逻辑斯蒂回归、GBDT、SVM……

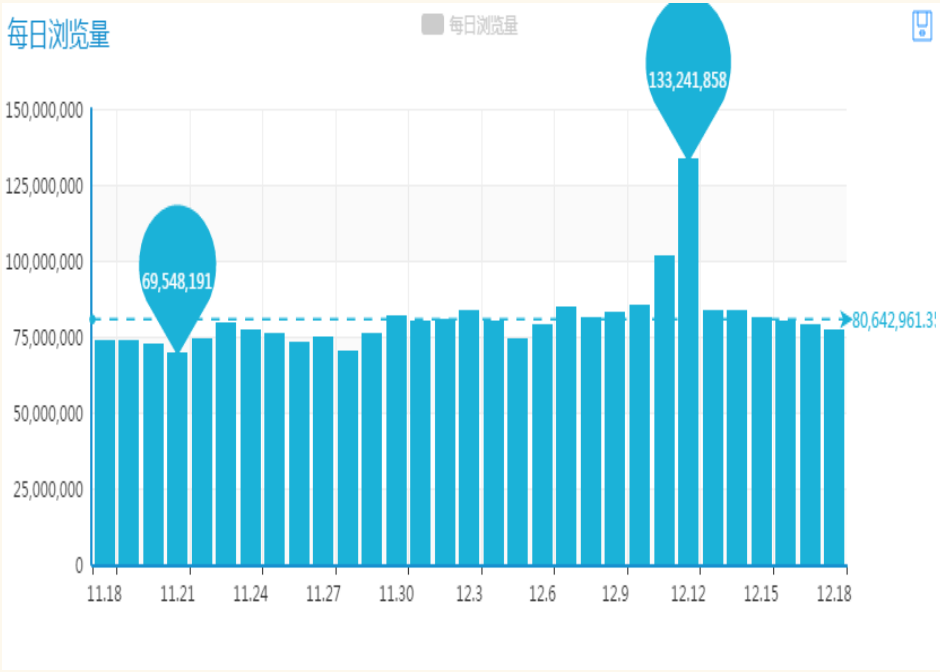
分析

数据稀疏
预测一天购买随机性强
海里捞针

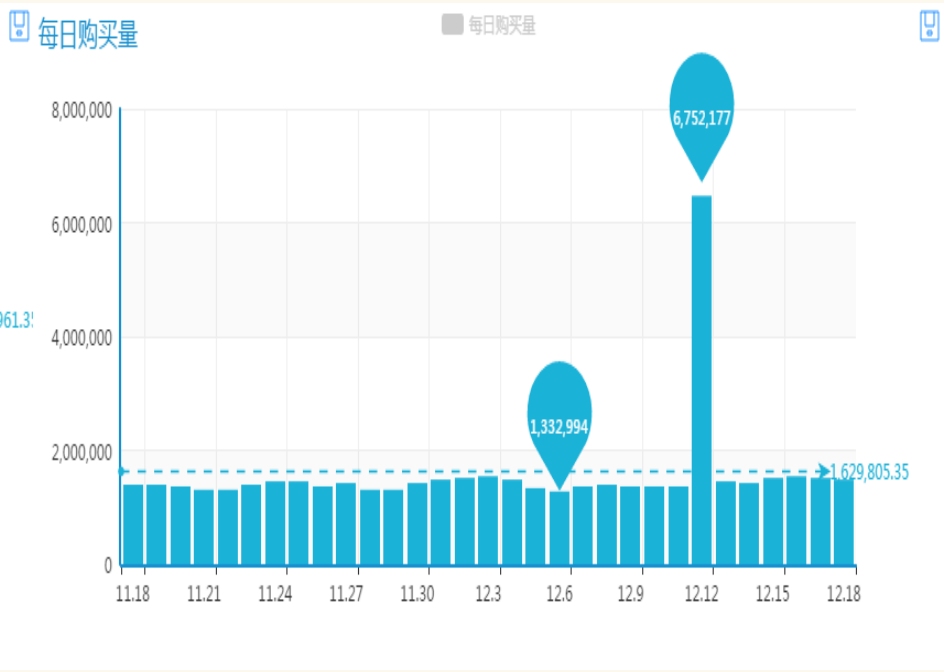
可用模型

协同过滤、SVD分解、关联规则
……

宏观分析——每日行为统计



千万量级



百万量级

微观分析——用户行为观察

用户	浏览数	收藏数	购物车	购买数	分类
A	1441872	0	0	0	X
B	100722	0	17	2	X
C	70111	996	1877	9	X
D	66694	202	0	0	X
E	46578	532	412	177	√
F	5998	92	107	43	√
G	5990	0	0	0	X
H	3485	326	10	0	X

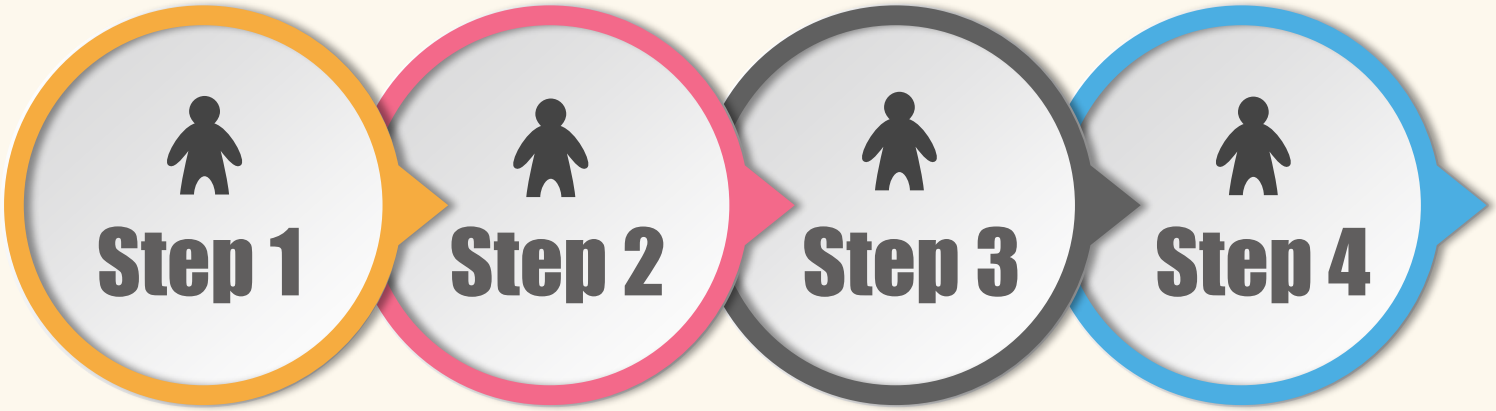
潜在问题

1. 双12影响数据分布 → 数据平滑
2. 存在爬虫作弊用户 → 用户清洗
3. 正负样本比例不均 → 滑动窗口

第2部分

数据处理

用户清洗



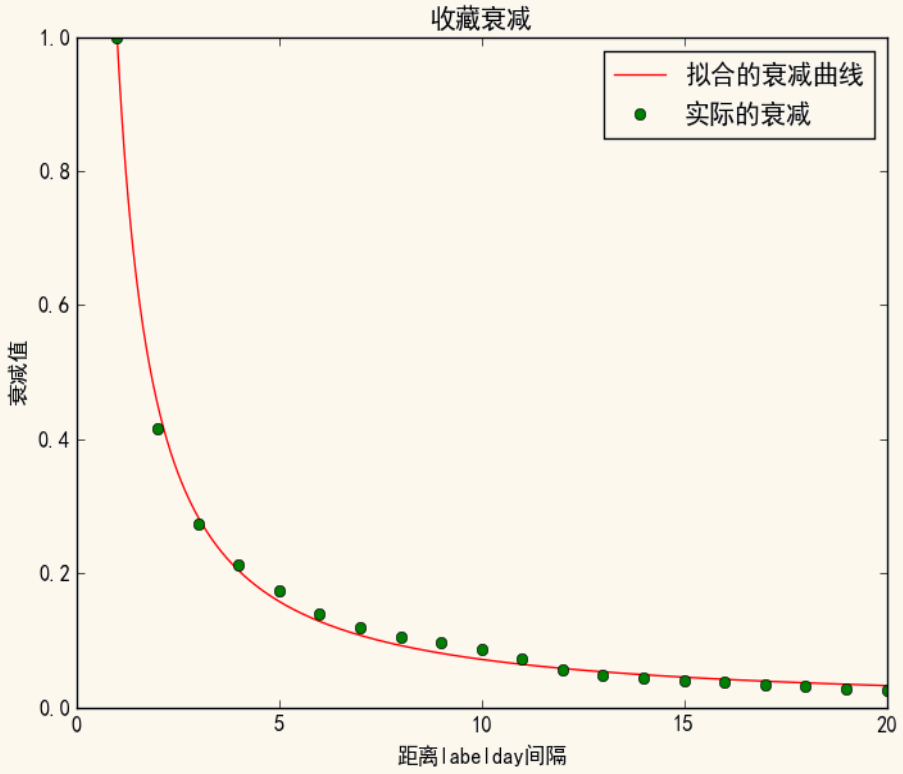
无收藏、购物车、购买行为

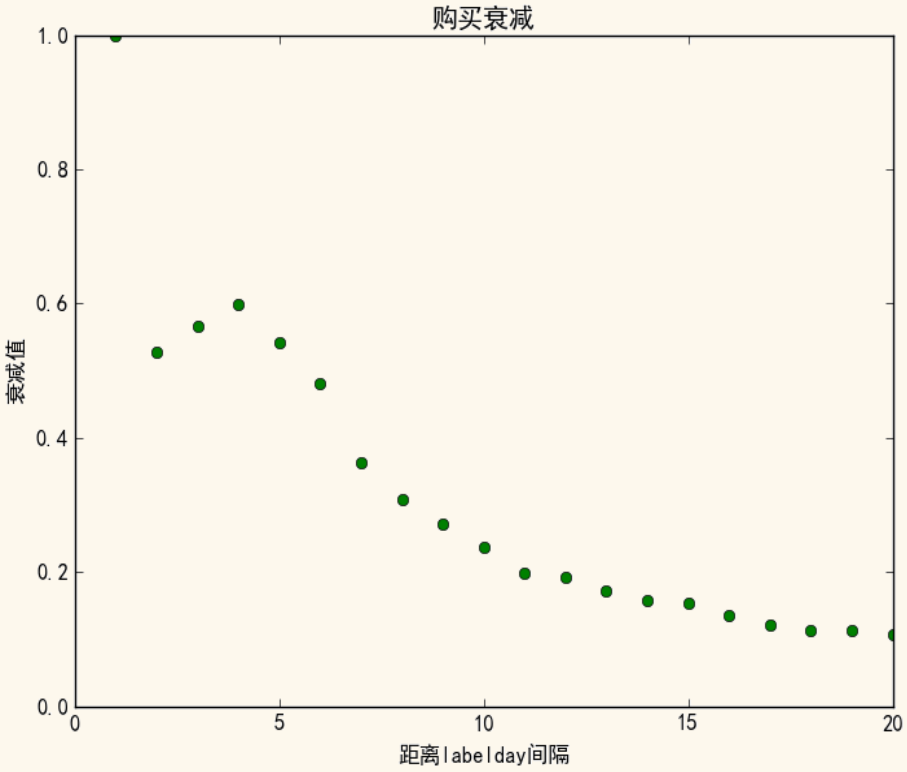
浏览数过多从没购买

对商品子集无收藏、购物车、购买

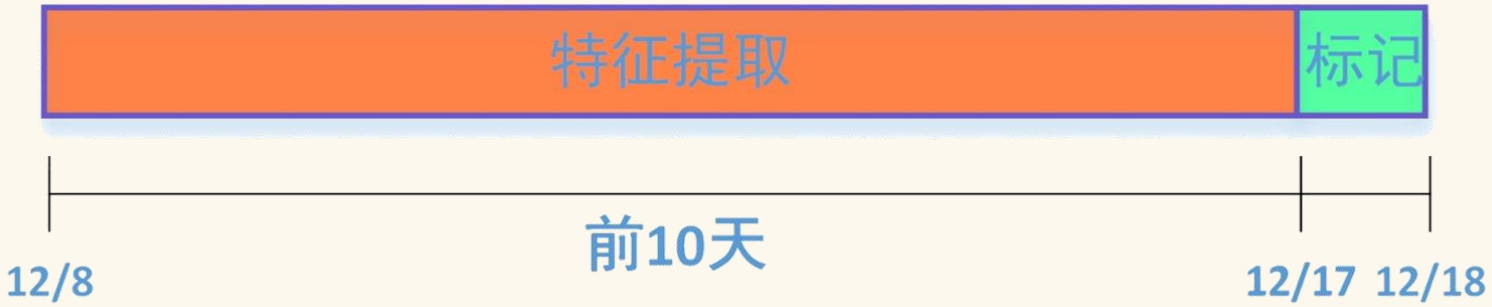
浏览数/购买数比例过大

拟合用户行为


收藏衰减
拟合的衰减曲线
实际的衰减
衰减值
距离label day间隔
1.0
0.8
0.6
0.4
0.2
0.0
0
5
10
15
20

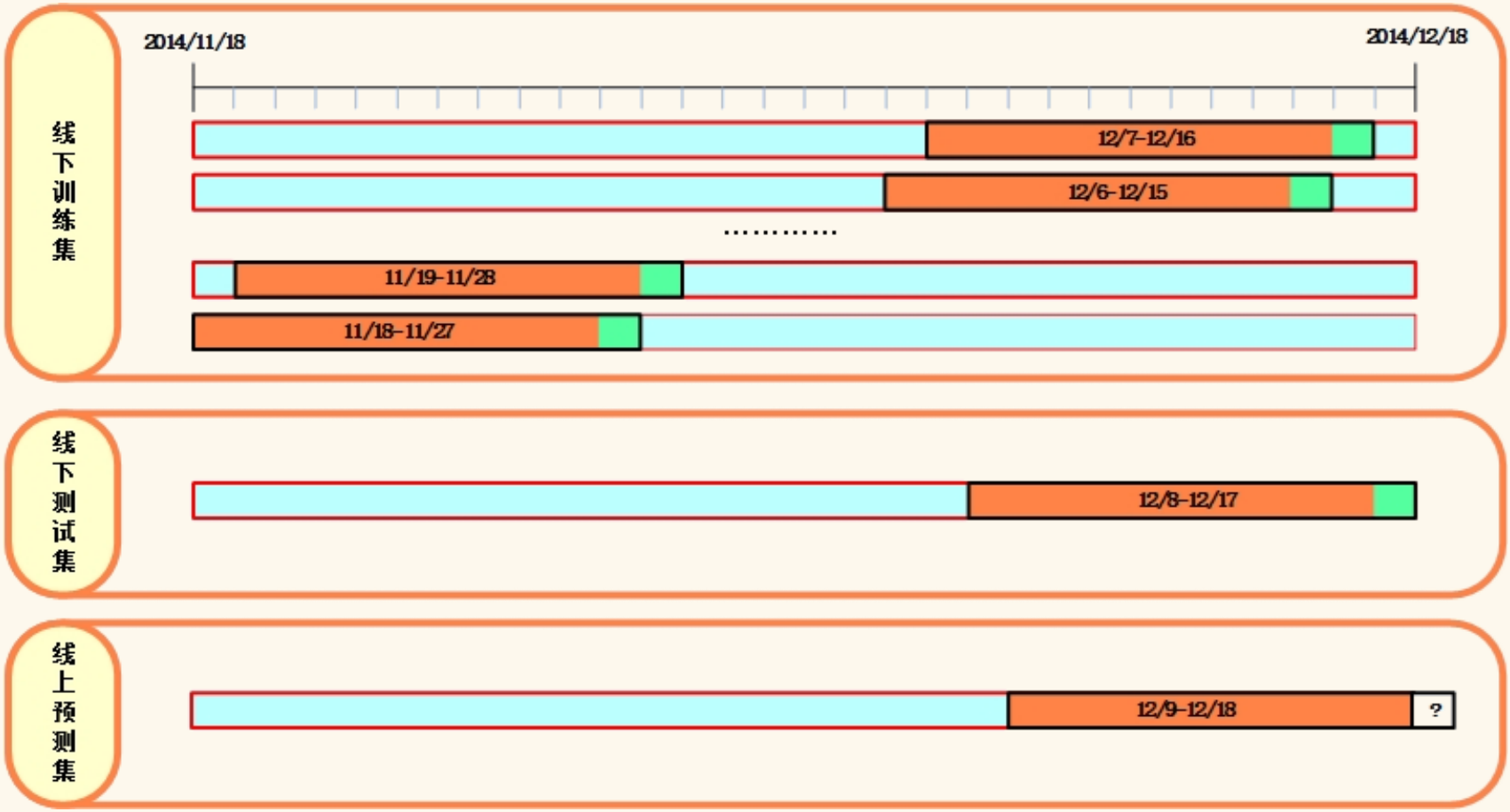

购买衰减
衰减值
距离label day间隔
1.0
0.8
0.6
0.4
0.2
0.0
0
5
10
15
20

滑动窗口



- 时间区间内有交互的user-item对
- 根据接下来一天是否购买给样本标记0或1
- 结合效率和性能，只用商品子集覆盖的行为记录来构造训练集（54E->5E）

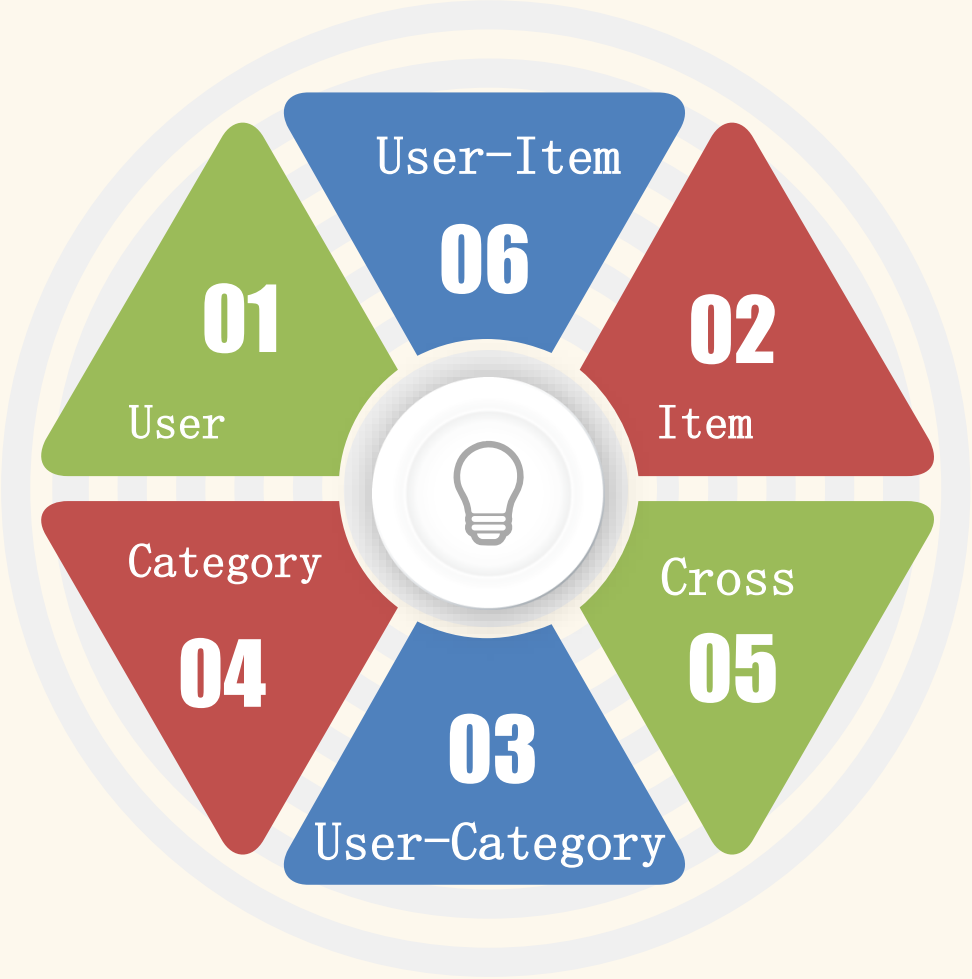
训练集构造


线下训练集
2014/11/18
2014/12/18
12/7-12/16
12/6-12/15
.....
11/19-11/28
11/18-11/27
线下测试集
12/8-12/17
线上预测集
12/9-12/18
?

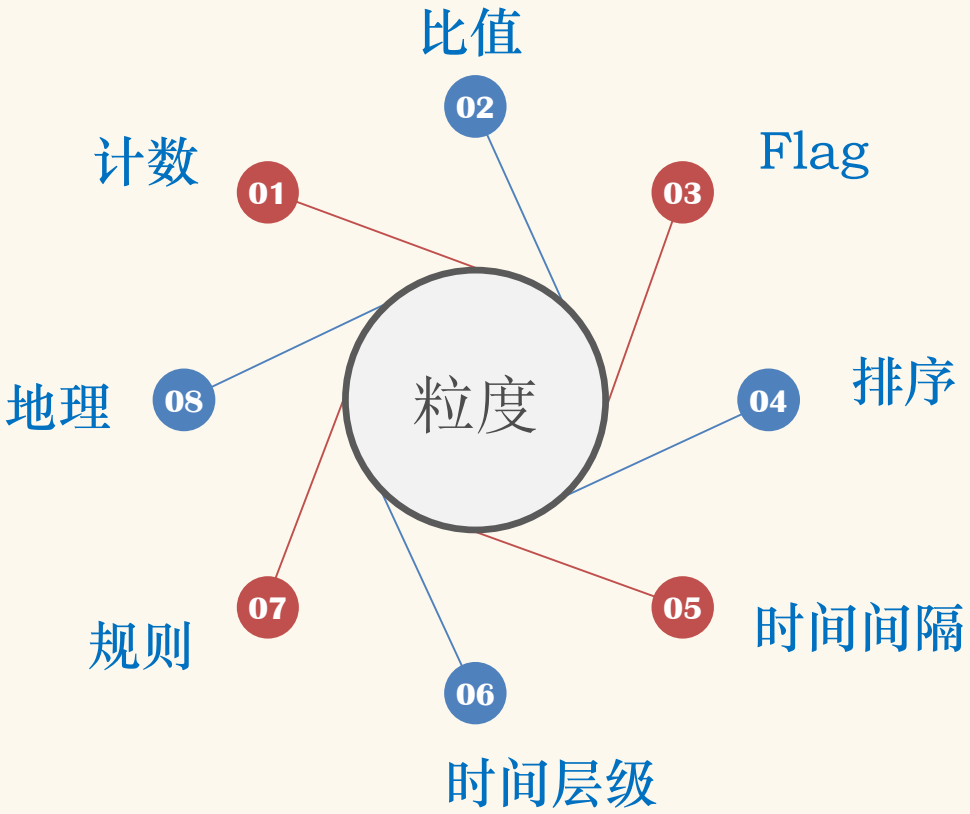
第3部分

特征工程

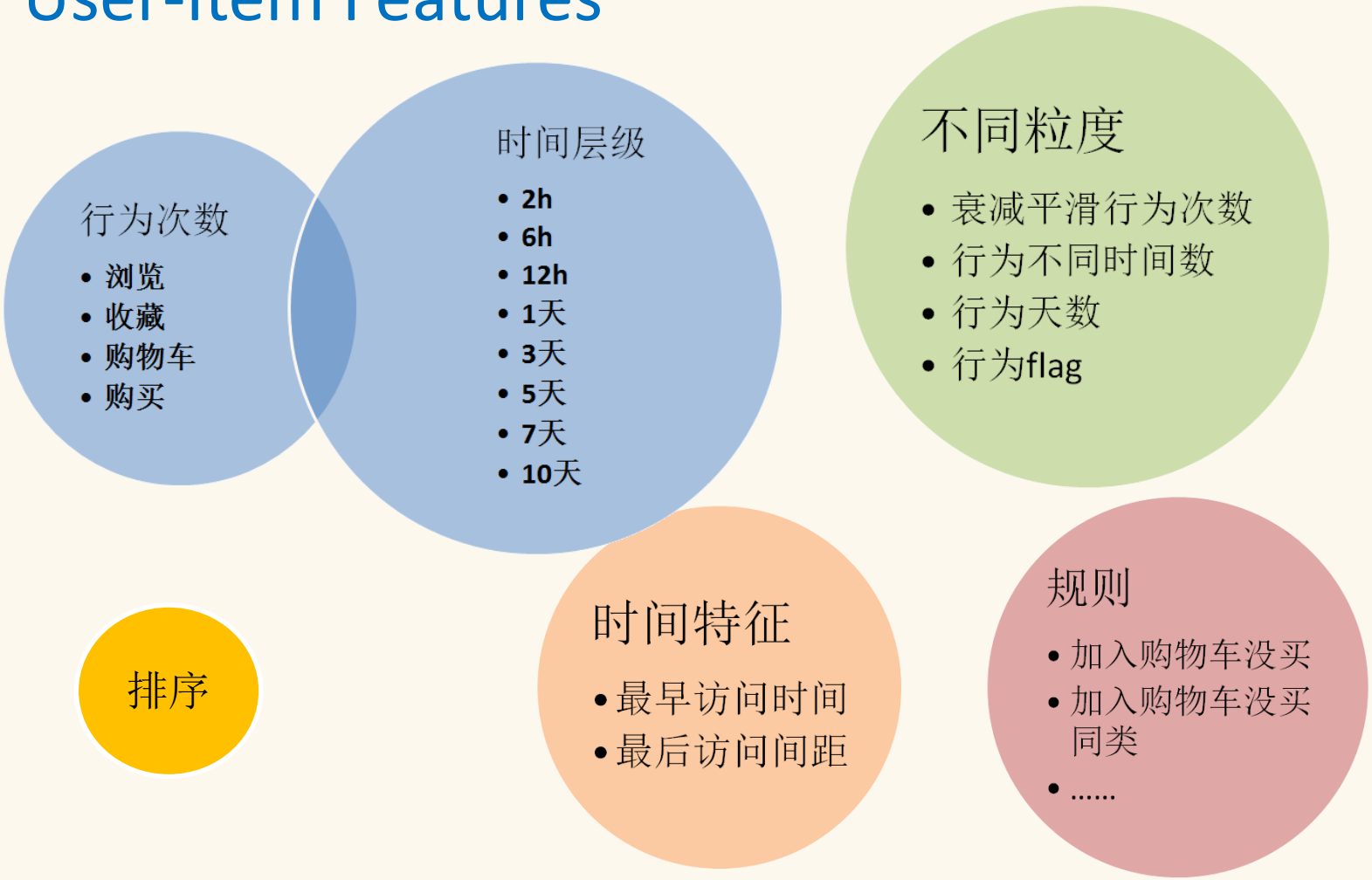
六大类特征


User-Item
06
01
User
02
Item
Category
04
Cross
05
03
User-Category

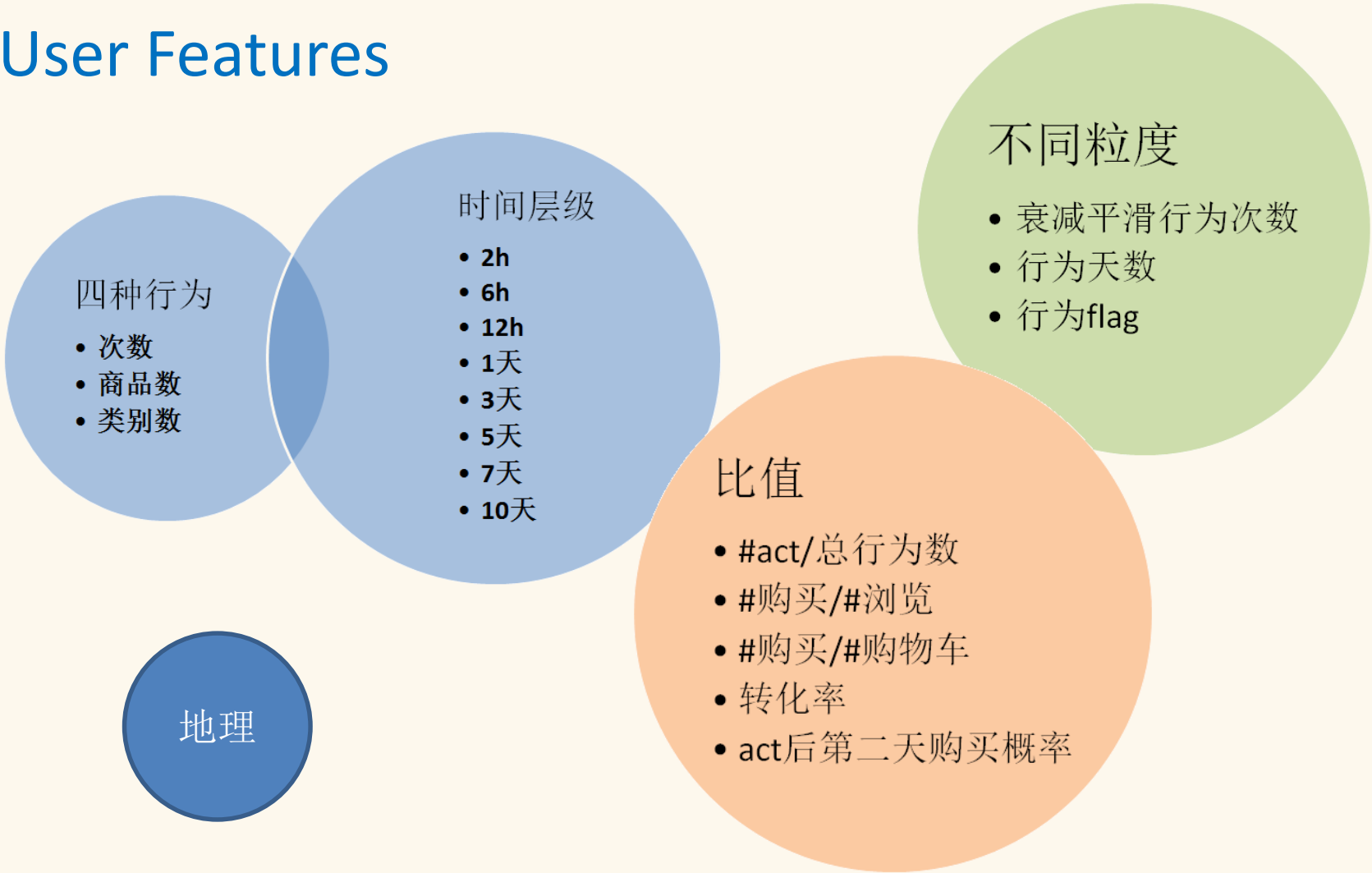
八种角度



- User-Item Features



- User Features



- Item Features

四种行为
- 次数
- 用户数

时间层级
- 1天
- 3天
- 5天
- 7天
- 10天

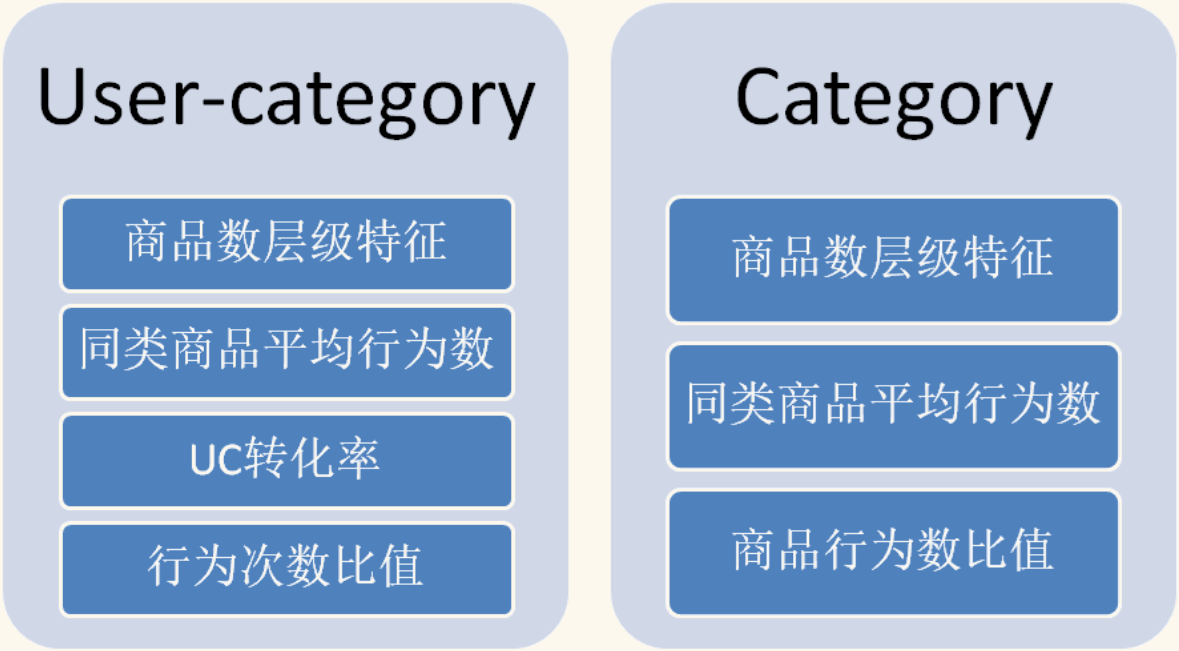
不同粒度
- 衰减平滑行为次数
- 行为天数
- 行为flag

比值
- #act/总行为数
- #购买用户/#浏览用户
- #购买用户/#购物车用户
- #购买量/#浏览量
- #购买量/#购物车量
- 用户转化率

比值
- 老客户率
- 跳出率
- 用户平均行为数

排序

- User-category & Category



- Cross Features

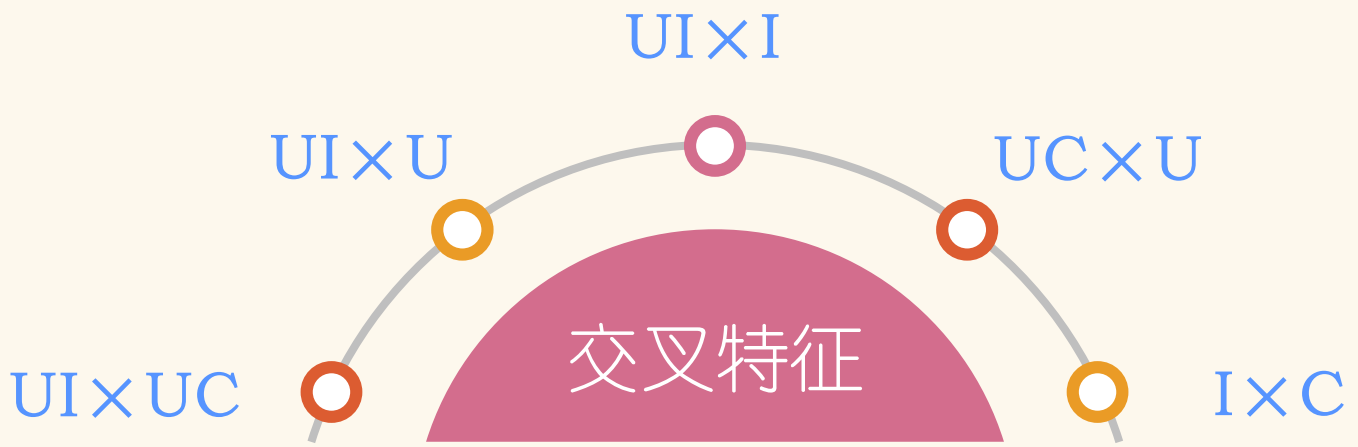
UI行为数*U购买比例

UI行为/UC行为

UI行为数*I转化率

UI行为/U行为

Flag特征叉乘



- **动机**
  - 用户在购物前会不断对比同类商品
  - 对于用户来说，购买就是在心中对商品排序
  - 通过历史数据，这种排序可以捕捉
- **排序特征**
  - 将原有的特征值按照大小排序，得到的排名作为排序特征

用户A	商品	浏览次数
	W	3
	X	10
	Y	25
	Z	1

用户A	商品	浏览次数	排序特征
	W	3	3rd
	X	10	2nd
	Y	25	1st
	Z	1	4th

- User-Item Rank

对数值排序

- 同uc中ui浏览次数的排名
- 同uc中ui购物车次数的排名
- 同uc中ui购买次数的排名
- 同uc中ui所花时间的排名
- 同一用户中ui行为次数的排名

排序->flag

- 同uc中该ui是6小时内最先访问的
- 同uc中该ui是6小时内最先加购物车的
- 同uc中该ui是点击次数最多的
- 同uc中该ui是花费时间最多的

对时间排序

- 同uc中ui的浏览次序
- 同uc中ui的收藏次序
- 同uc中ui的购物车次序

- Item Rank

对数值排序

- 同类商品中item行为次数的排名
- 同类商品中item购买人数的排名
- 同类商品中item老客户率的排名
- 同类商品中item转化率的排名
- 同类商品中item人均行为的排名

排序->flag

- 同类商品中该item销量最多
- 同类商品中该item点击最多
- 同类商品中该item购买用户最多
- 同类商品中该item人均行为最多

地理特征——不同地区用户购买力不同

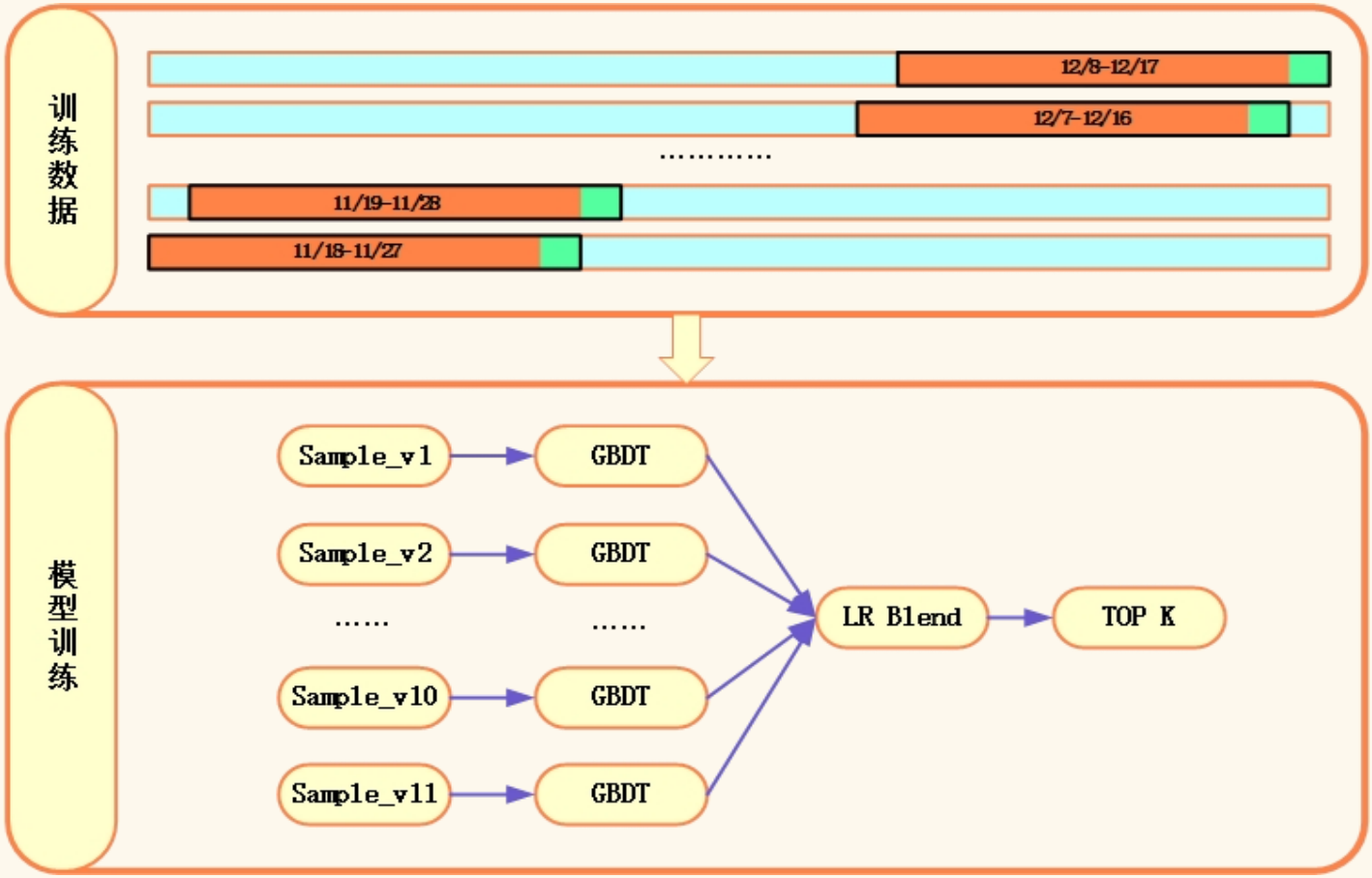
- 对用户按照地区分类
  - 统计地区内行为数以及转化率等
- 如何确定地区？
  - 三点定位
  - Kmeans聚类
- 如何统计？
  - 依据行为记录->所属地区
  - 依据用户->所属地区

- **地理特征**
  - 统计地区内行为：人均和总量
  - 实践效果：性能提升较少（F值提升0.02%）
- **原因分析**
  - 带有地理信息的记录较少，许多用户没有地理信息，数据覆盖率太低
  - 地区行为统计粒度较大，特征强度较弱

第4部分

算法实现

算法框架


训练数据
12/8-12/17
12/7-12/16
.....
11/19-11/28
11/18-11/27
模型训练
Sample_v1
Sample_v2
.....
Sample_v10
Sample_v11
GBDT
GBDT
.....
GBDT
GBDT
LR Blend
TOP K

模型选择

单模型效果最好

GBDT > RF > LR

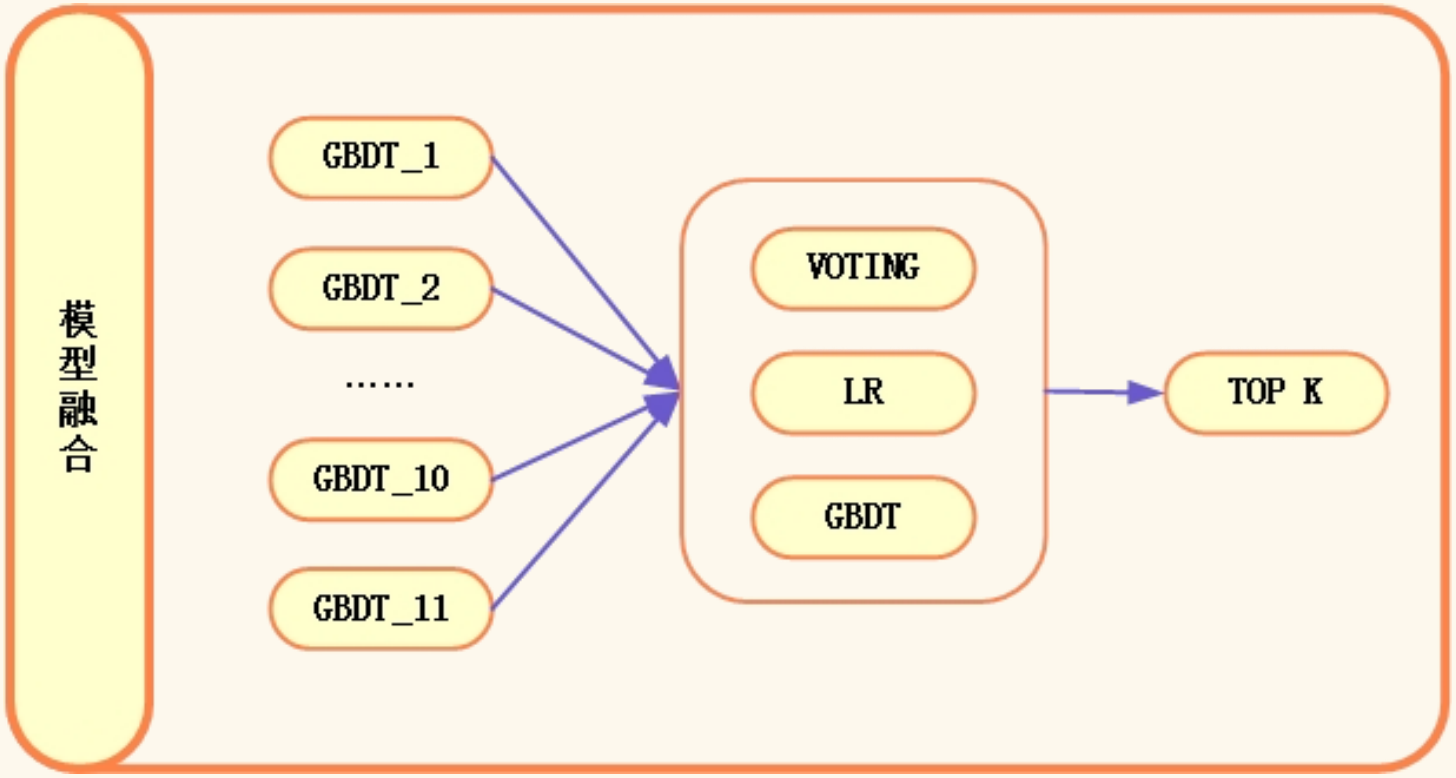
不易过拟合

泛化能力强

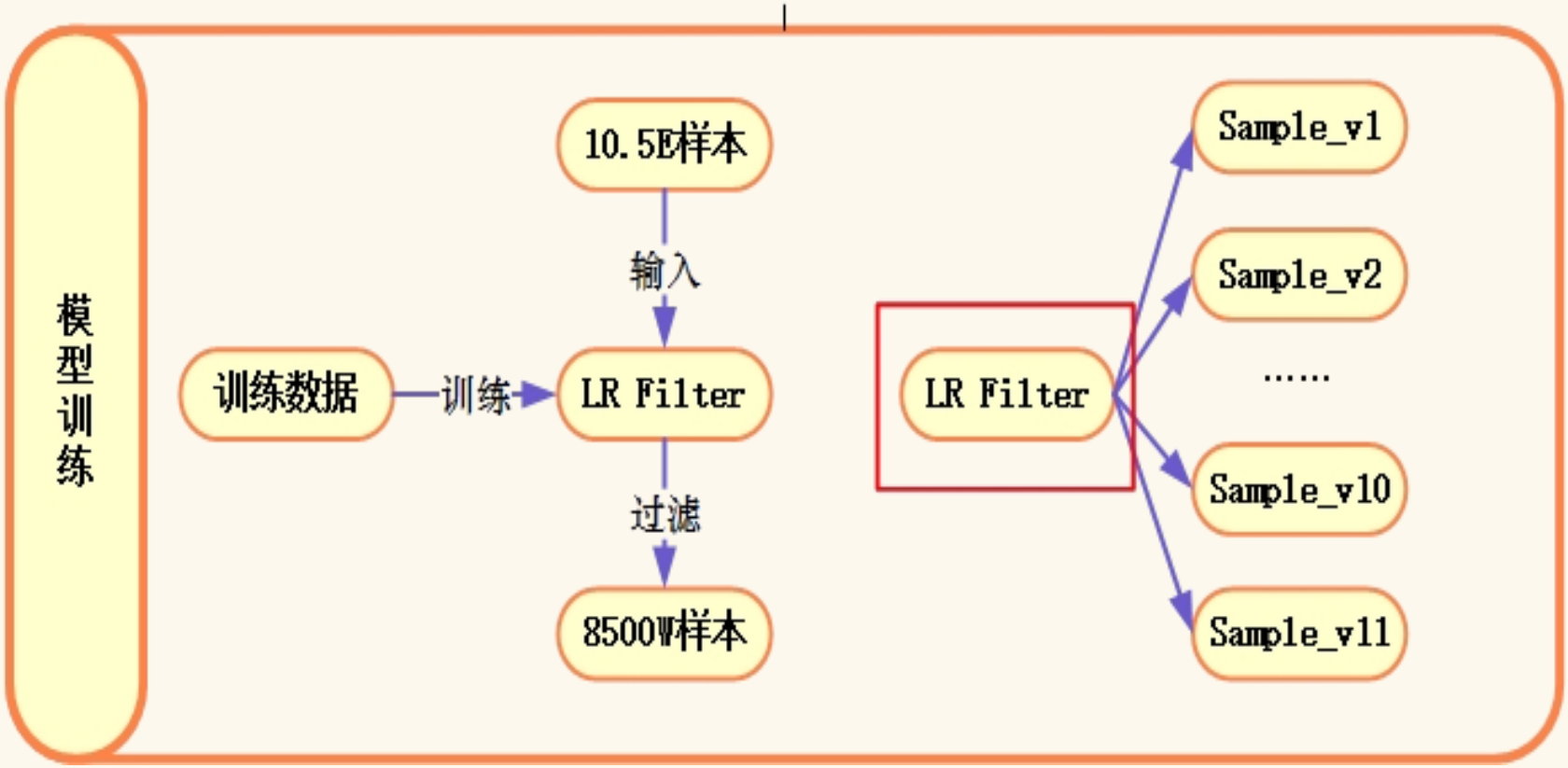
预测速度较快

LR > GBDT > RF

模型融合



样本筛选


模型训练
10.5E样本
输入
训练数据
训练
LR Filter
过滤
8500W样本
LR Filter
Sample_v1
Sample_v2
.....
Sample_v10
Sample_v11

未实现的Idea

A 不同大小的窗口训练得到的模型进行融合

B 商品全集的影响：譬如用户购买了同类别的商品对购买同类别商品子集的影响

C 使用GBDT等非线性算法做样本过滤

参赛心得

A 队友之间code review非常重要

B 使用OneNote记录整理，细节决定成败

C 不要轻言放弃，火箭随时可能出现

平台使用心得

A

尽量使用工作流和手动任务，不会丢失日志且比算法平台效率更高

B

希望平台可以加上GBDT和RF的可视化训练过程中训练集和验证集的正确率曲线

C

平台卡时，稍安勿躁，泡杯茶等等就行。

致谢

感谢阿里巴巴集团提供数据和平台

感谢天池团队的完美组织

感谢在比赛中互相成长的小伙伴

感谢所有坚持走完比赛旅程的选手们

Contact us:
gmj_py@163.com