分析

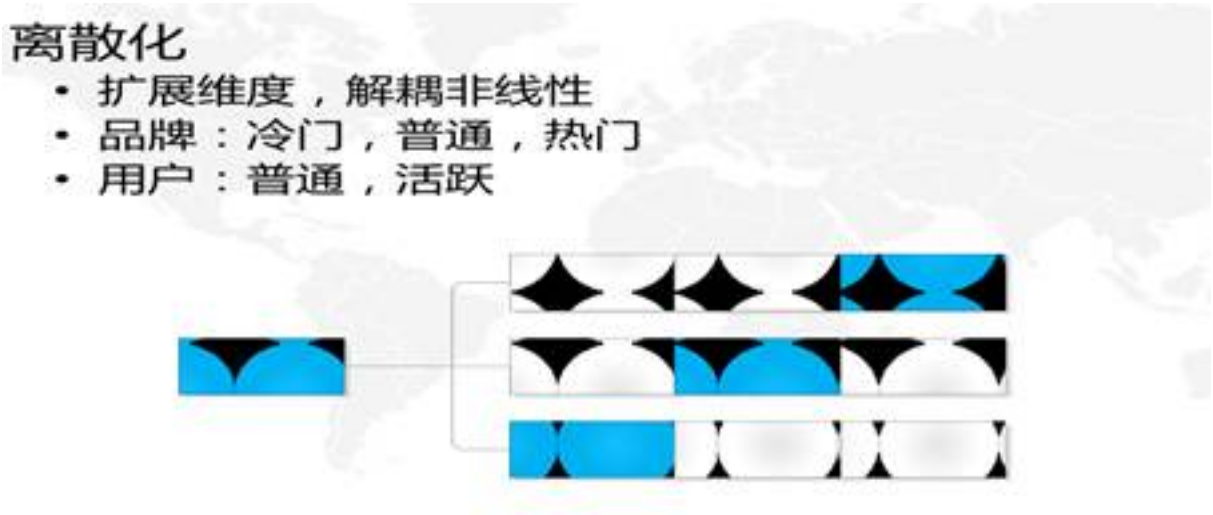
1.数据的过滤

a。在历史信息中没有任何交互的样本线上和线下不能丢弃（ 可以利用训练模型让他进行训练，后面的操作对买和不买影响较大）
b。 点击大于500(经验值)，且没有购买过
c.没有点击，但有其他操作（我们没有对这一部分数据进行处理，我们可以统计一下这样的行为有多少，对我们的结果的影响度大不大）

2.特征的提取

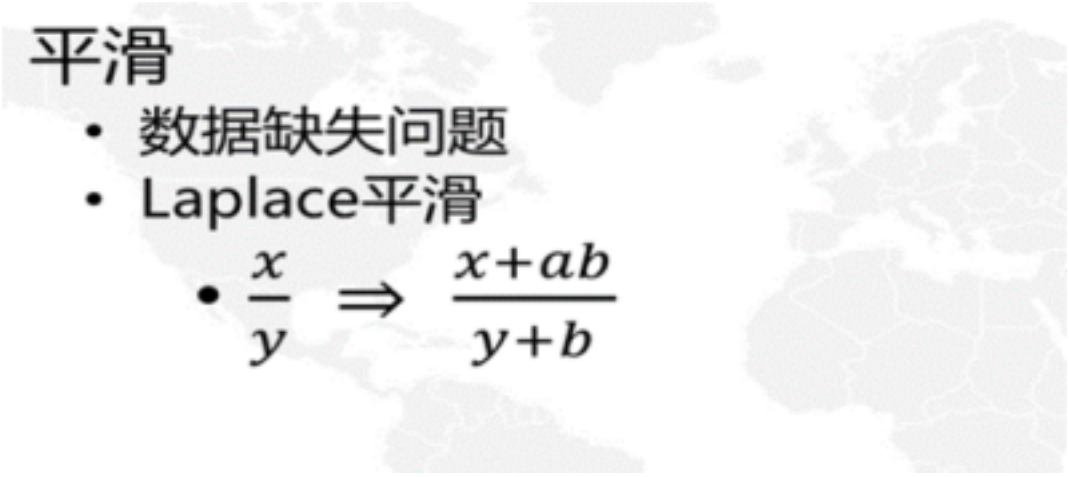
click me	click me
特征类别	特征描述
当前用户总体特征	用户最近1/3/5天有多少天有交互/购买任意商品
	用户最近一次/第一次交互/购买任意商品到现在的时间距离
	用户最近1/3/5天购买/点击所有商品的总次数
	用户最近1/3、5天购买/点击过的不同商品的个数
	用户的属性（普通，活跃）
	用户的购买周期
	回头的次数
当前商品总体特征	商品最近第1/2/3天以及最近7/15天被多少不同用户交互/购买
	商品总共有多少人购买（不去重）
	商品总共有多少人在不同天重复购买
	商品总共有多少人在不同周重复购买
	商品第一次/最后一次被任意用户交互/购买/加购的时间
	商品的人均销量
	商品的日均销量
	增长率
	回头率（回头客的数量）
	商品的属性（冷门，普通，热门）
	商品的购买周期
商品竞争	当前用户与当前商品的最后一个交互天的前一天,当天和后一天这3天中一共交互了多少个不同其他商品
	用户最近1/2/3/4/5/6/7/8/9/10/12/14/16/18/20/25/30天有多少单天/三天/单周/十天/双周对当前商品存在纯交互日（只与这一个商品进行交互）
	用户最后一次交互当前商品距离用户最后一次交互任意商品相隔多少天

click me	click me
	当前用户与当前商品的最后一个交互天的前一天,当天和后一天这3天中一共点击/购买/收藏了多少次其他商品
当前用户当前商品特征	用户交互当前商品的最近/第2近/第3近/第4近/第5近/第6近/第7近/第8近的交互天的点击数/购买数/收藏数以及时间
	用户交互当前商品的最远一个交互天的点击数/购买数/收藏数以及时间
	用户第一次交互/购买当前商品和最后一次交互当/购买前商品相隔多少天
	用户最近1/2/3/4/5/6/7/8/9/10/12/14/16/18/20/25/30天有多少单天/三天/单周/十天/双周有交互/购买当前商品以及交互/购买/收藏的总次数
	用户30天/60天以前有多少单天/三天/单周/十天/双周有交互/购买当前商品以及交互/购买/收藏的总次数



特征提取之后要进行归一化的处理

数据的平滑处理

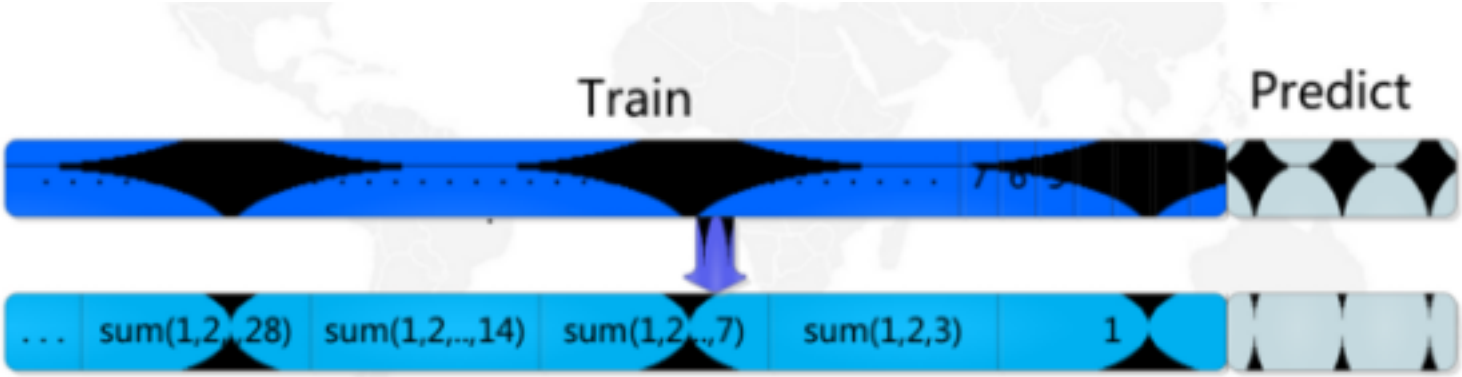


3.训练集测试集的划分

将序列信息分段压缩

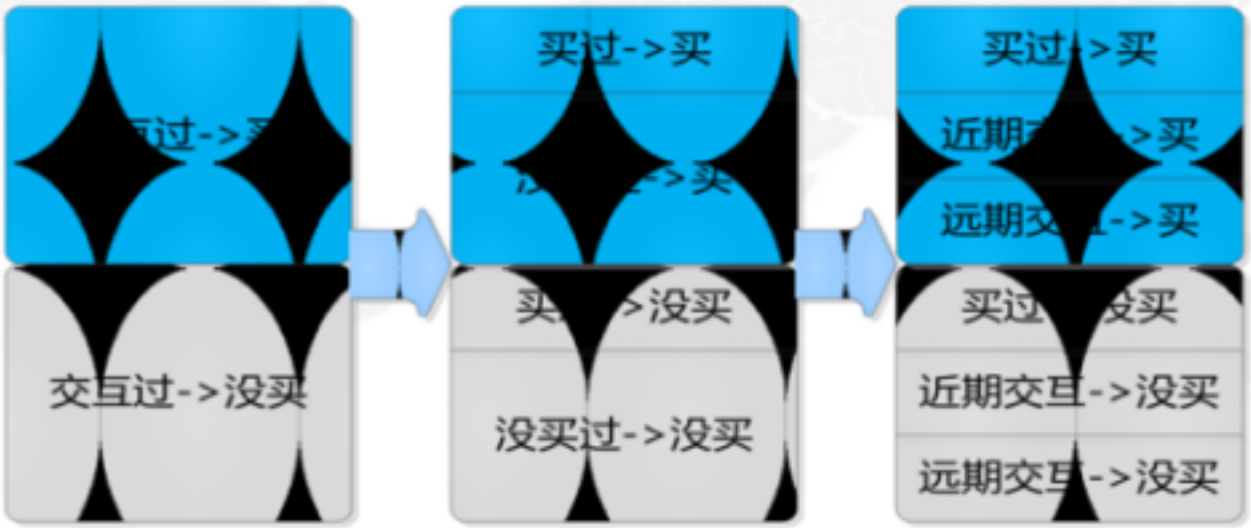
• 近期密，远期粗，重叠

• 1,3,7,14,28,56,max



已交互推荐

- 二分类问题->多分类问题



4.模型的训练

LR特殊处理

- 所有特征都进行Dummy Coding

基本思想: 将有多个取值的1个特征转换成多个取值为0,1的特征

举例: 一个用户对一个商品有交互的天数nday(nday<8)为例,假设一条样本n_day=5,那么dummy之后变为8个特征,其含义和取值如下: (他这样做相当于将特征的数量增加了)

nday=1	nday=2	nday=3	nday=4	nday=5	nday=6	nday=7	nday=8
0	0	0	0	1	0	0	0

变种: 对数值型特征,基于大于等于或者小于等于的包含式dummy

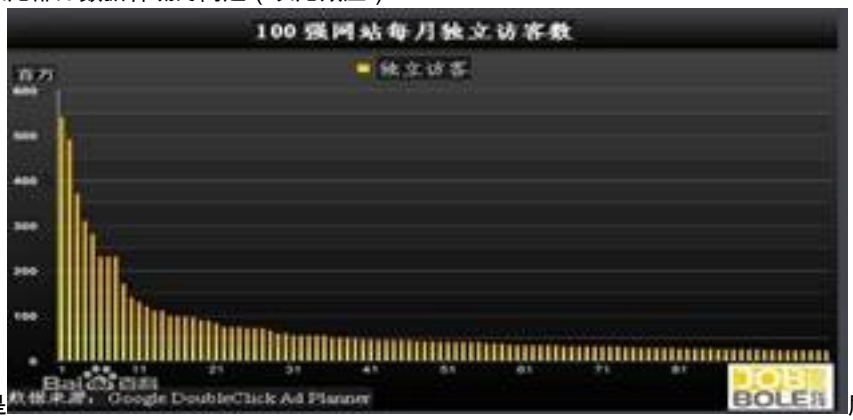
nday>=1	nday>=2	nday>=3	nday>=4	nday>=5	nday>=6	nday>=7	nday>=8
0	0	0	0	1	1	1	1

LR为什么要用dummy coding？

实现单特征的非线性权重, 以及为了使用非数值类特征（比如商品id）

LR为什么要用包含式的dummy coding？（在总决赛答辩（4）中他画的那个直方图，显示就和下面这个图很类似，就是长尾效应，所以我觉得我们是不是也可以画一下，看看是不是长尾效应，是不是长尾效应运用这个原理就会方便很多）

解决特征值长尾部分数据稀疏的问题（长尾效应）

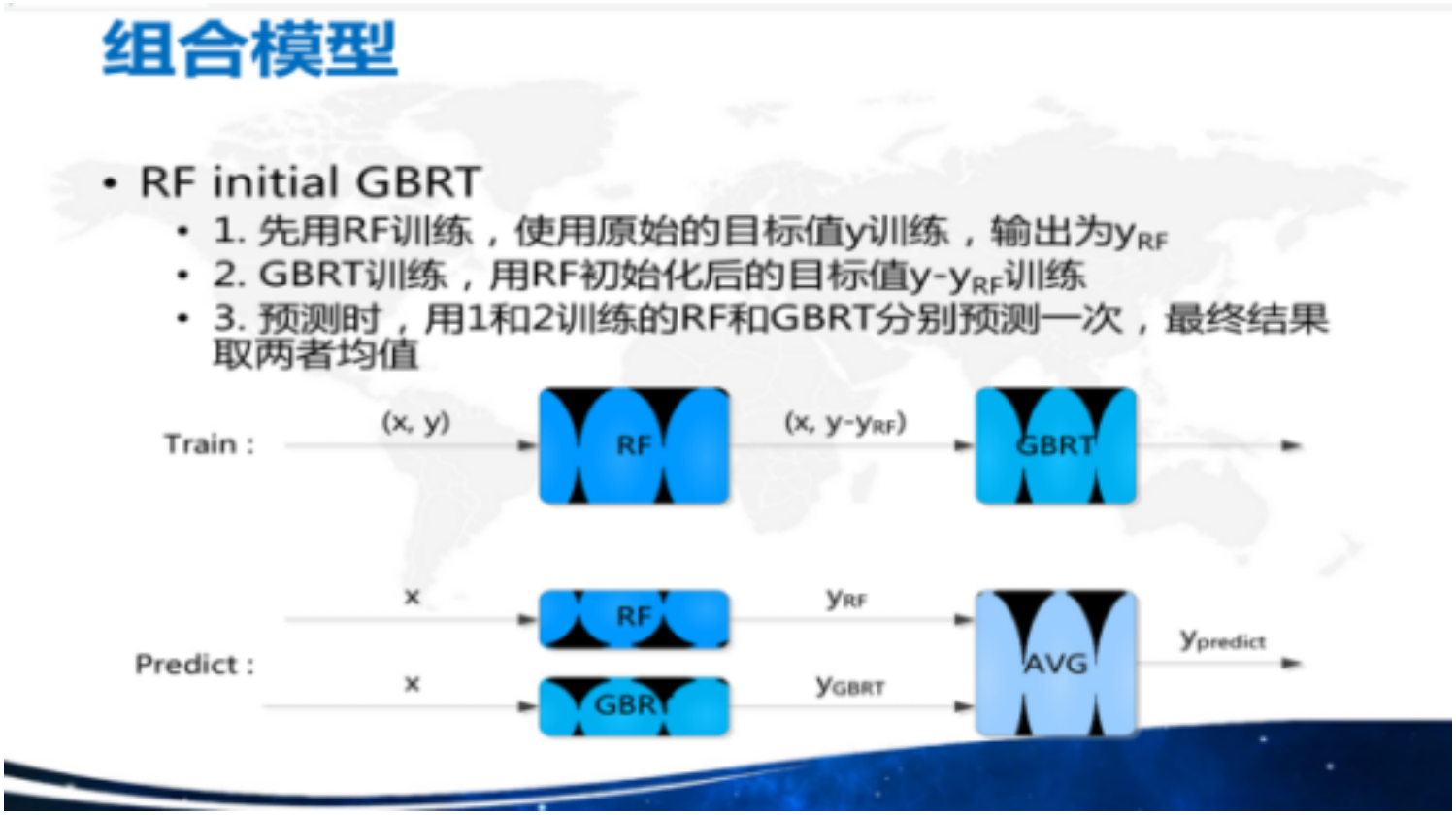


长尾效应就是尾巴部分累加和不一定比流行部分少

模型的融合问题（这是我一直考虑的问题，模型的融合不单纯的是求交集）

我是这样理解的（根据不同的模型求并出来的结果进行加权和）

另外一种组合的方法



这里还有一个模型(我不太懂)

其中各模型的特点


逻辑回归
线性模型
正则化项
逐步逻辑回归
训练预测效率高
大数据量下精度下降


随机森林
averaging方法
降低方差
基于分类决策树
抗噪能力强
训练快、 预测慢


GBRT
boosting方法
降低偏差
回归树
训练慢、 预测快

5.前车之鉴

个人经验总结

- 线上调参
  - 不要过早陷入根据线上结果调整算法,要多依赖线下结果
- 特征选择
  - 从追求分数的角度来说,不要花太多时间反复尝试人工删除特征并验证效果
- 人工规则
  - 不要过早陷入人工规则的调节
- 数据分布
  - 多关注线下和线上数据的不一致性
- 不断尝试
  - 不能光从原理上分析就拍板结论,也不能不懂原理盲目不断尝试