
新浪微博互动预测大赛答辩
2015天池大数据竞赛
TIANCHI天池

Jokeren说我们水

新浪微博互动预测算法大赛分享

目录

第1部分

赛题建模

问题描述

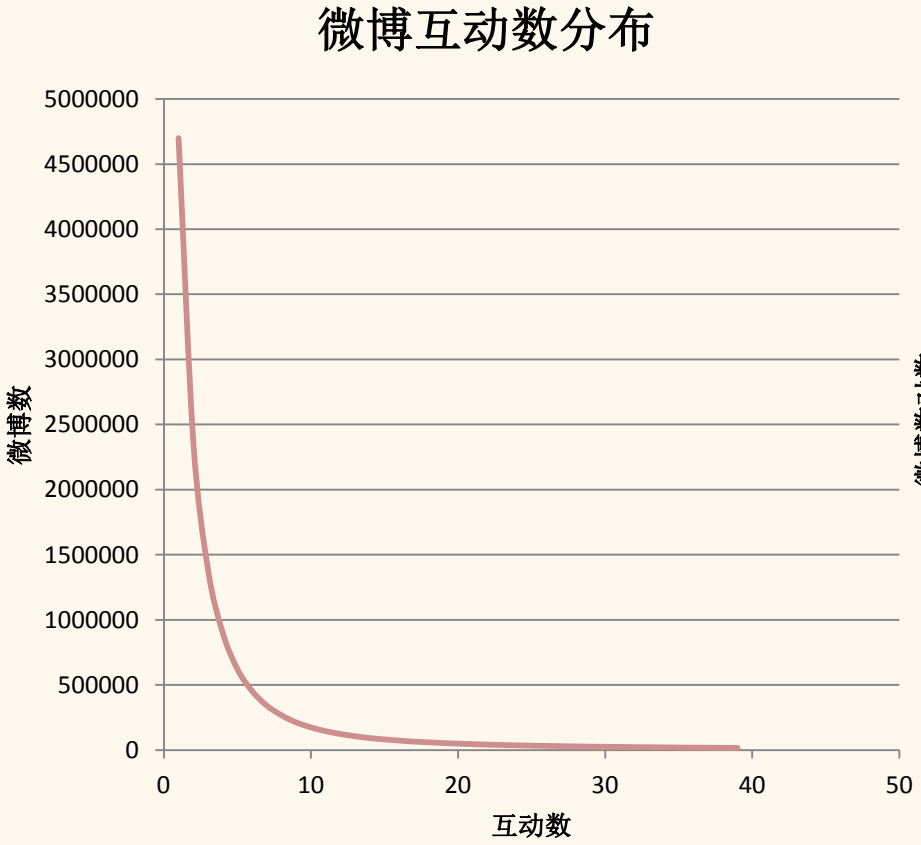

196W用户
5个月
1.1E微博
1
2
1779W用户
对1407W微
博的2.39E
次转评点
预测4月份
每条微博
转评点数
的等级
4
3
4100W用户对
122W用户的
1.8E条关注

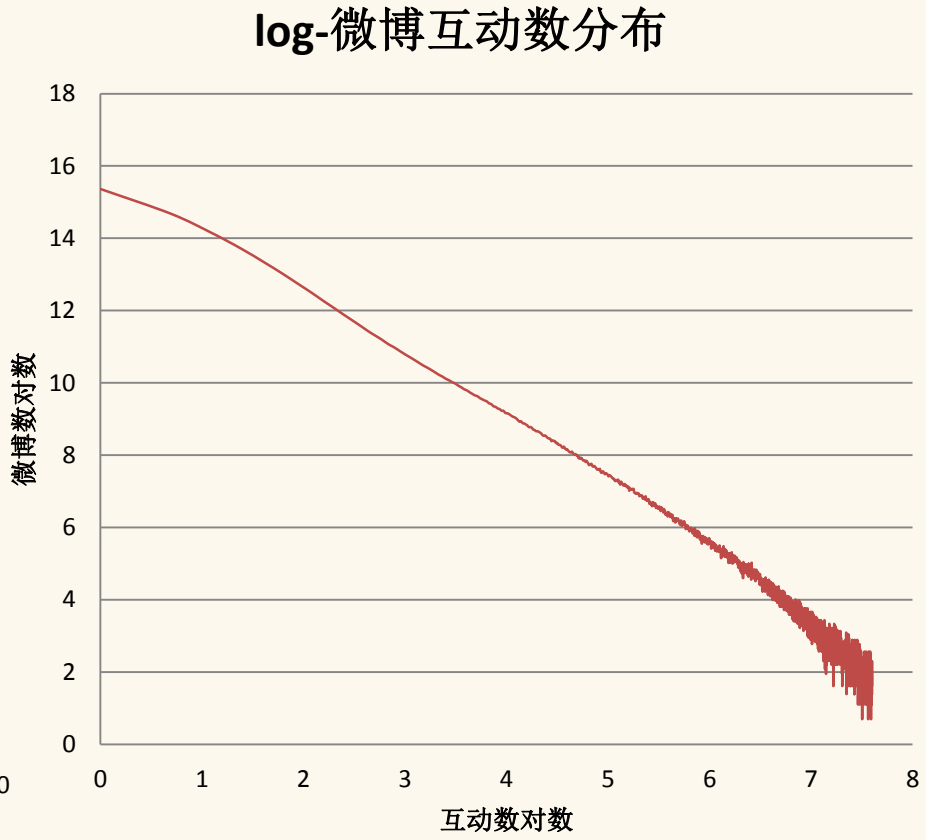
初步建模

- 带权分类问题

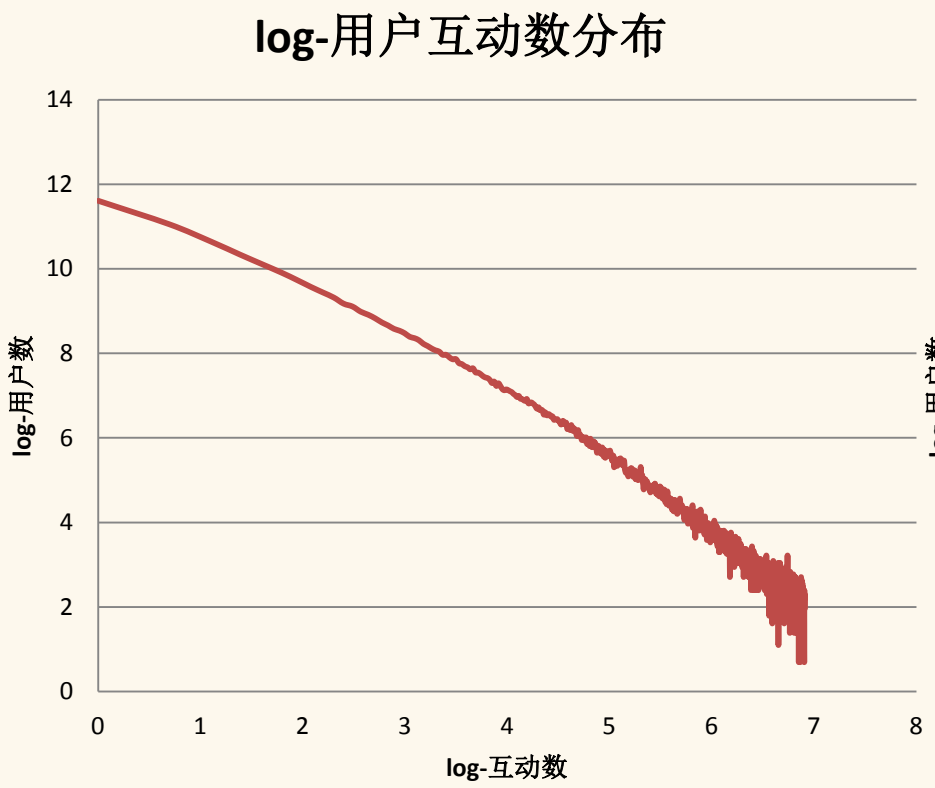
- 结合用户历史行为和微博特征去分类
- 预测目标不是用户，而是每条微博
- 可用模型：RF、LR、GBDT

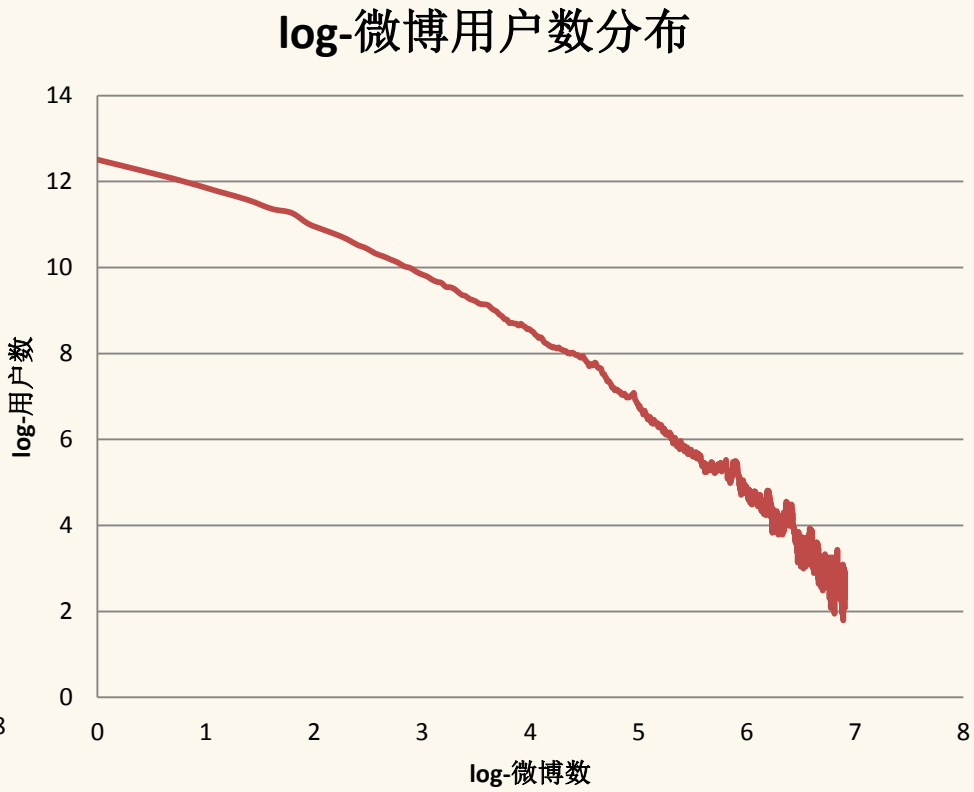
微博转评点数分析


微博互动数分布
5000000
4500000
4000000
3500000
3000000
2500000
2000000
1500000
1000000
500000
0
微博数
0
10
20
30
40
50
互动数


log-微博互动数分布
18
16
14
12
10
8
6
4
2
0
微博数对数
0
1
2
3
4
5
6
7
8
互动数对数

用户分析





Top 1k的用户发了156W微博，收获1.78E/2.39E转点评。

uid	weibo_count
090626b134b357fe5515e5db4fae6869	7073252

blog_time	blog
2014-11-17 16:17:22	REPORT1_PUBLISH_PIC180.149.135.2301416212241.378059
2015-02-23 21:52:02	UPDATE_10.73.32.189_1424699521.979176
2015-03-05 16:11:01	UPDATE_10.73.32.189_1425543061.337783

但是仍然存在垃圾用户，僵尸粉！

文本统计

word	count	df
(+86)186205190...	4	4
(010)62074712	1	1
(010)62178811	1	1

word	count	df
----dliaobp517----chenyuxua...	1	1
----hheft173----goujiangtao08...	1	1
----ixiangu628----yucocojy@...	1	1

word	count	df
0.01999265	1	1
0.01万	36	36
0.01亿	2	1

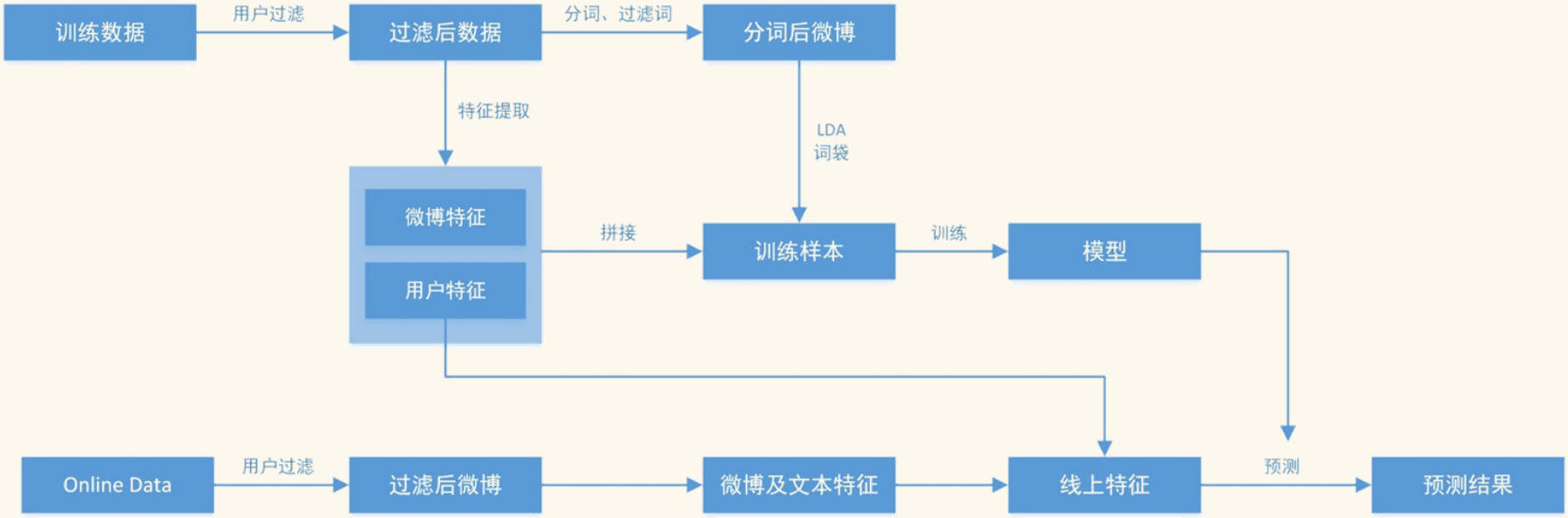
word	count	df
00年度	1	1
01日	36	33
01月	121	108
01月01日	124	123

-平均微博长度：35个单词
-词库大小：277W单词/6428W微博（过滤微博后的数据）
-杂乱单词很多

赛题挑战

1. 微博转评点分布不均衡
2. 存在大量垃圾僵尸用户
3. 如何利用文本及社交信息
4. 如何做带权多分类

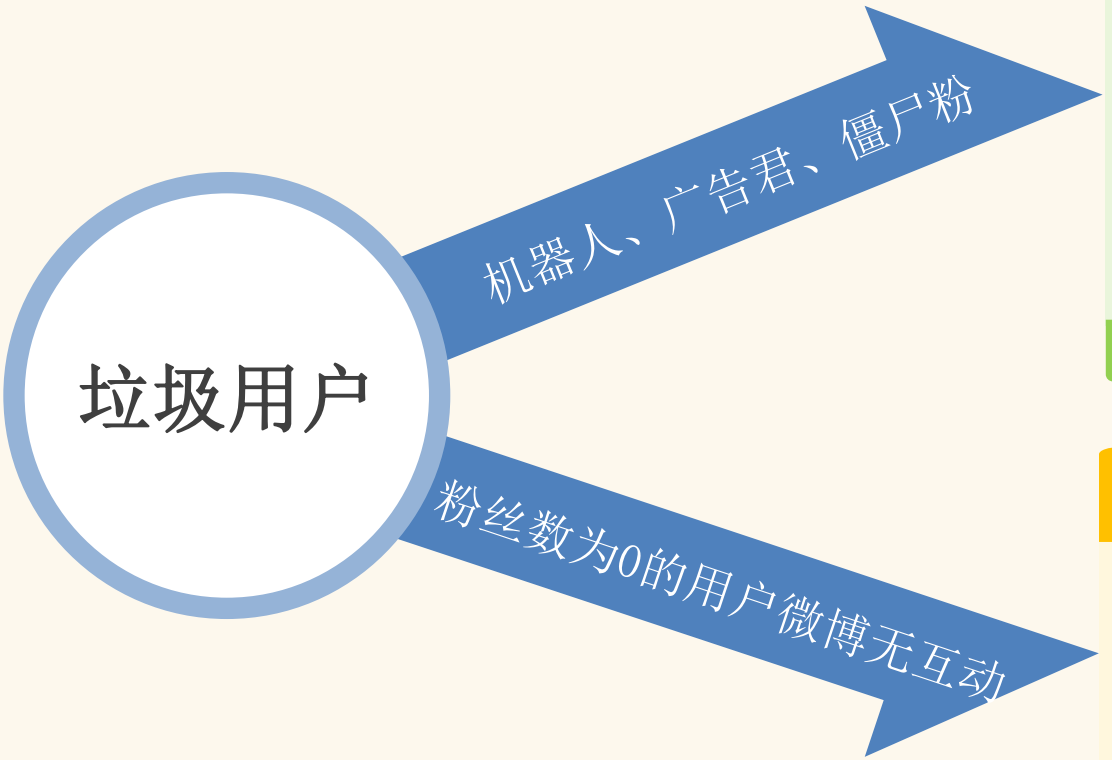
整体流程


训练数据
用户过滤
过滤后数据
分词、过滤词
分词后微博
特征提取
LDA
词袋
微博特征
用户特征
拼接
训练样本
训练
模型
Online Data
用户过滤
过滤后微博
微博及文本特征
线上特征
预测
预测结果

第2部分

数据处理

用户清洗



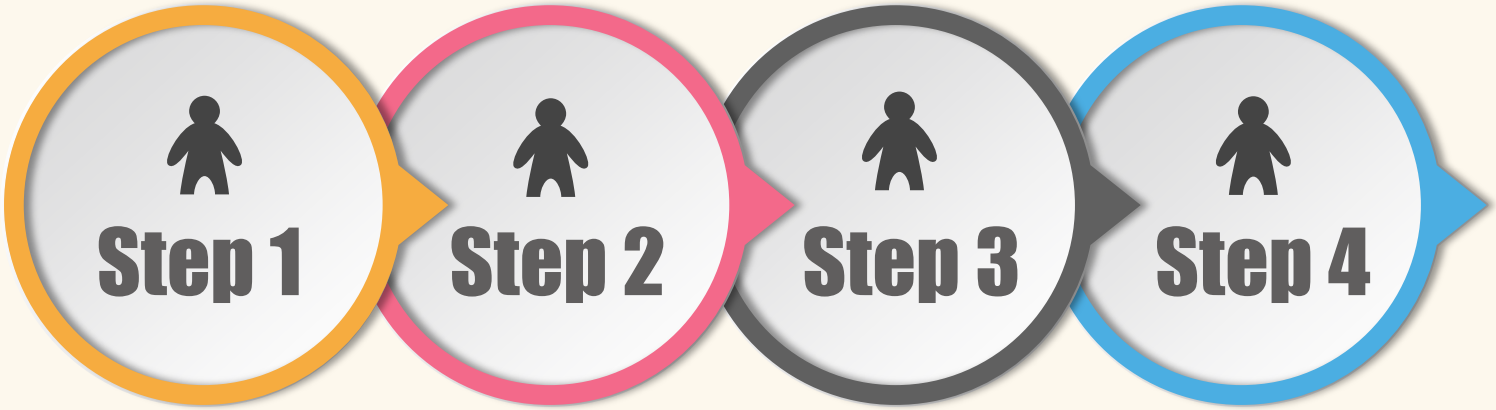
结果

方法：规则过滤，在测试集验证效果
共性：微博发的多 and 无人互动
规则：weibo_count > 2000 and max(sum)=0
验证：测试集中只有36条微博行为数大于0，最大仅有3。

结果

结果：5503W微博/77W用户
效果：极大地减少了样本数量
降低了噪声
比赛初期可带来1%的提升

文本预处理



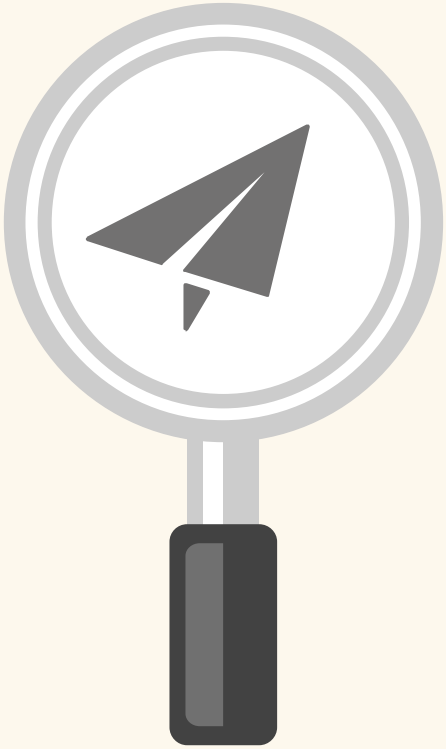
停用词
标点符号

数字、邮箱
电话号码、
日期、URL

频率为1
的单词

过滤后词袋大小
111W，LDA主题
分布更明显

数据划分

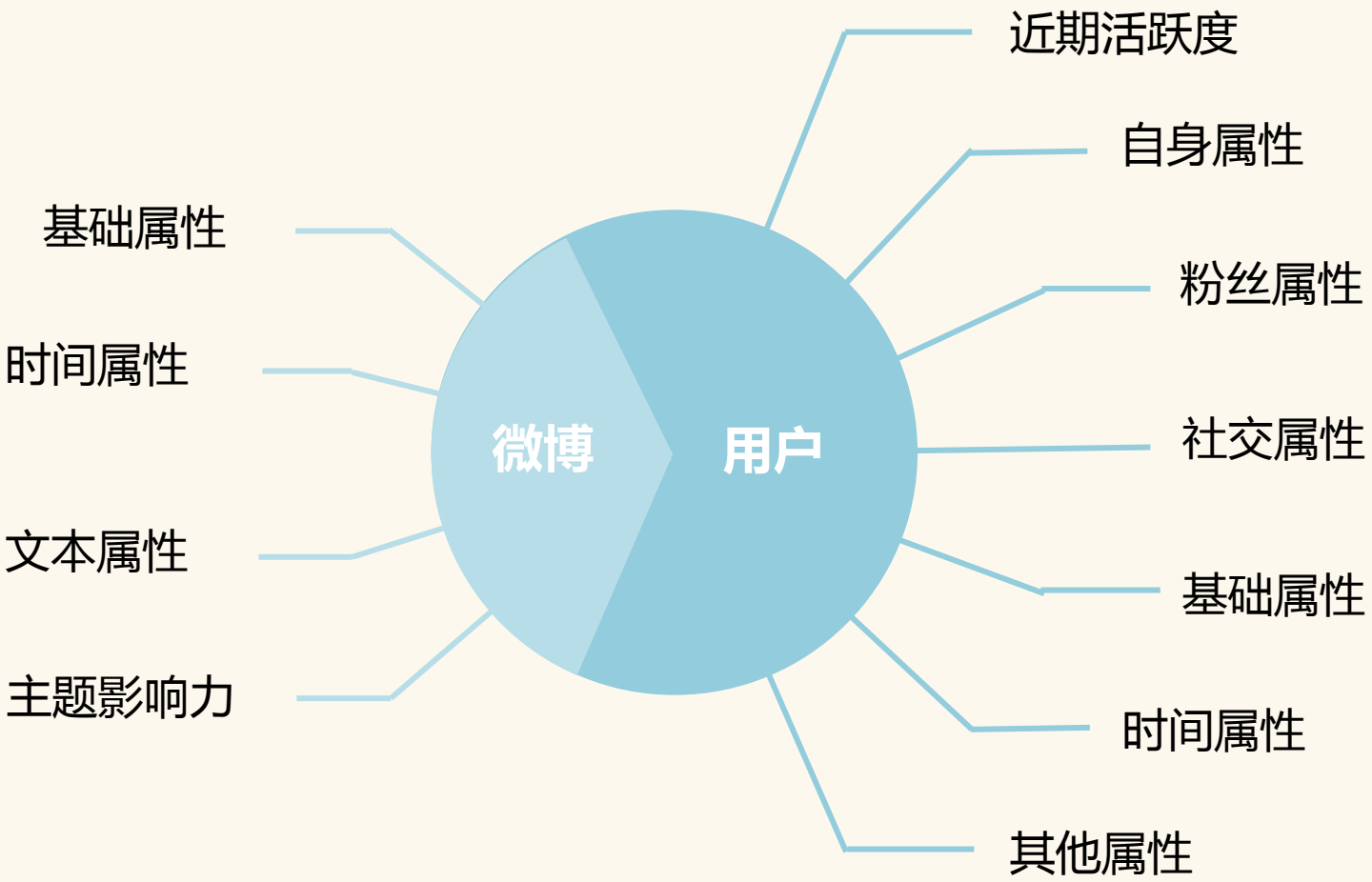


- 滑动窗口
  - 前三个月抽取用户特征，结合下一月微博内容抽取微博特征
- 数据预测
  - 未过滤的用户的微博用模型预测结果
  - 过滤掉的用户微博直接判为1档

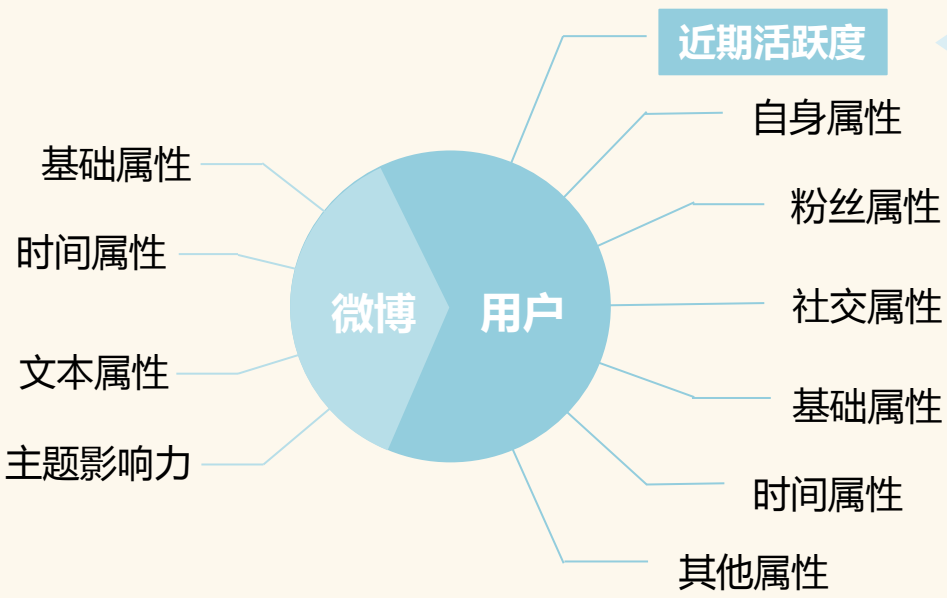
第3部分

特征工程

特征工程

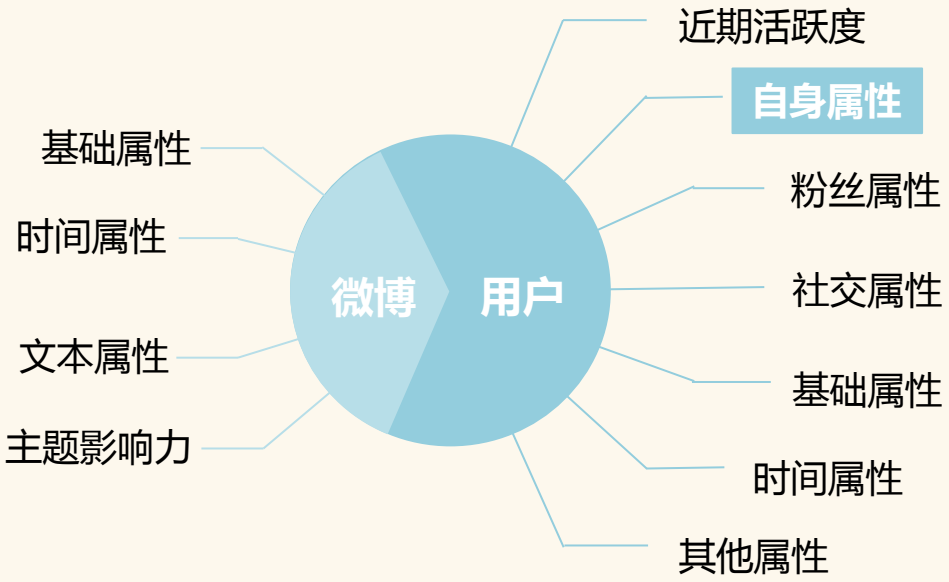

近期活跃度
自身属性
粉丝属性
社交属性
基础属性
时间属性
其他属性
基础属性
时间属性
文本属性
主题影响力
微博
用户

特征工程



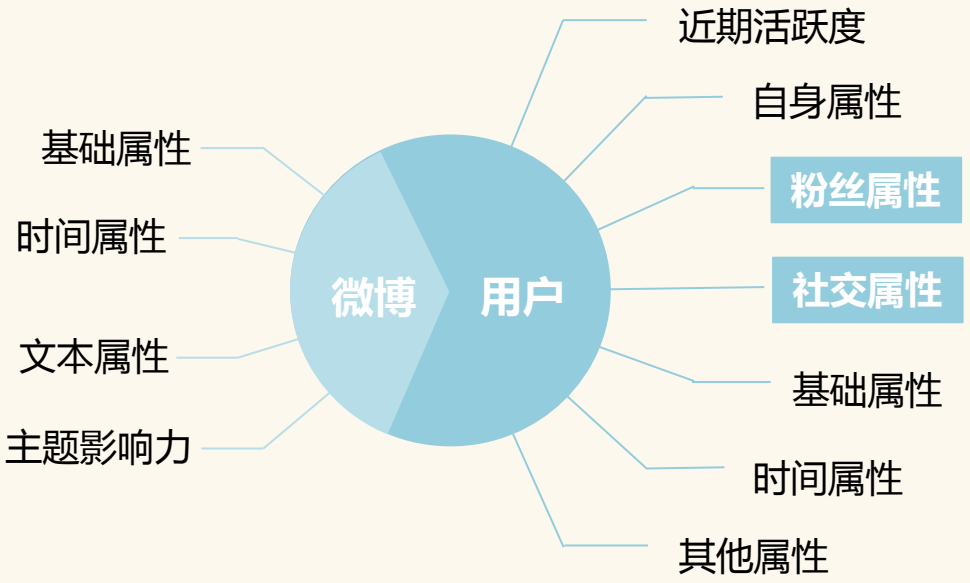
- 近1，3，7天发微博的条数
- 近7，15，30，90总发微博天数
- 连续发微博天数
- 连续不发微博天数
- 近7天平均每天发微博数
- 近1，3，5，10条微博时间间隔
- 总action数，平均action数
- 近7，15，30，90是否每天都发微博
- 当天发了几条微博

特征工程



- 所有微博中3无微博的条数和比例
- 发出微博后，收到前3个action的平均时间间隔
- 上个月最后三条微博的action sum
- 连续发了多少条小于等于level几的微博数以及比例
- 窗口内第一条和最后一条距离窗口最后一天的时间间隔

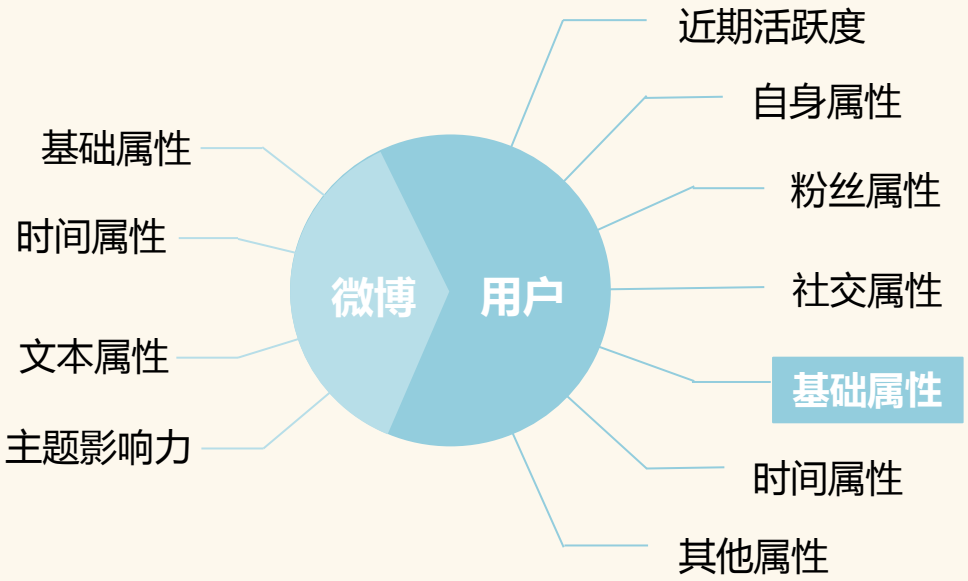
特征工程



- 粉丝level的mid，avg以及偏度
- 粉丝活跃程度特征
- 行为数大于2，5，10的粉丝数

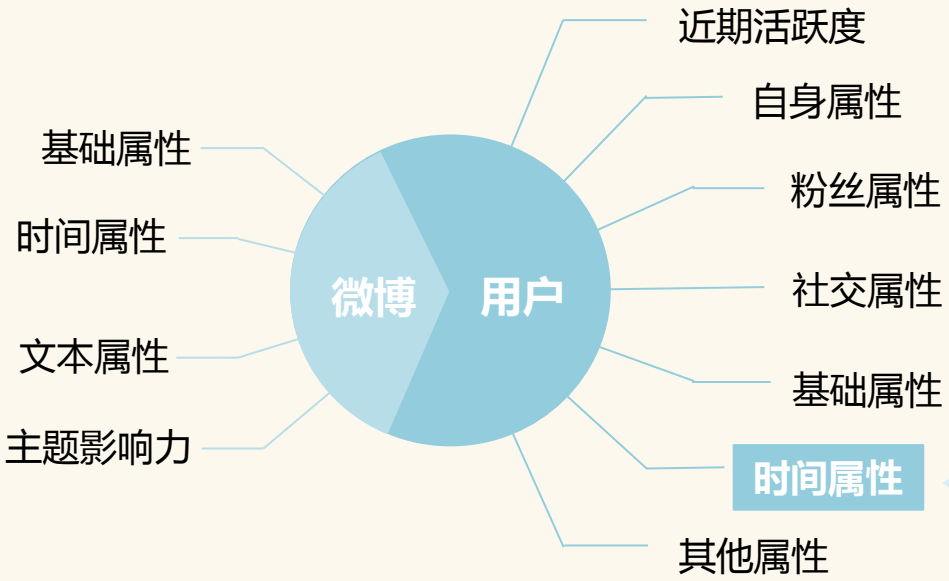
- 互动粉丝数
- 关注的人数
- 粉丝数
- 用户发出的fw cm lk all的数量及天数

特征工程



- 收到的fw,cm,lk,all的sum,mid,max,avg
- 收到的fw,cm,lk占all的比例
- 不同level的微博数，以及占总微博数的比例
- 加权后每个level微博分值比例
- level众数，加权level众数以及得分

特征工程

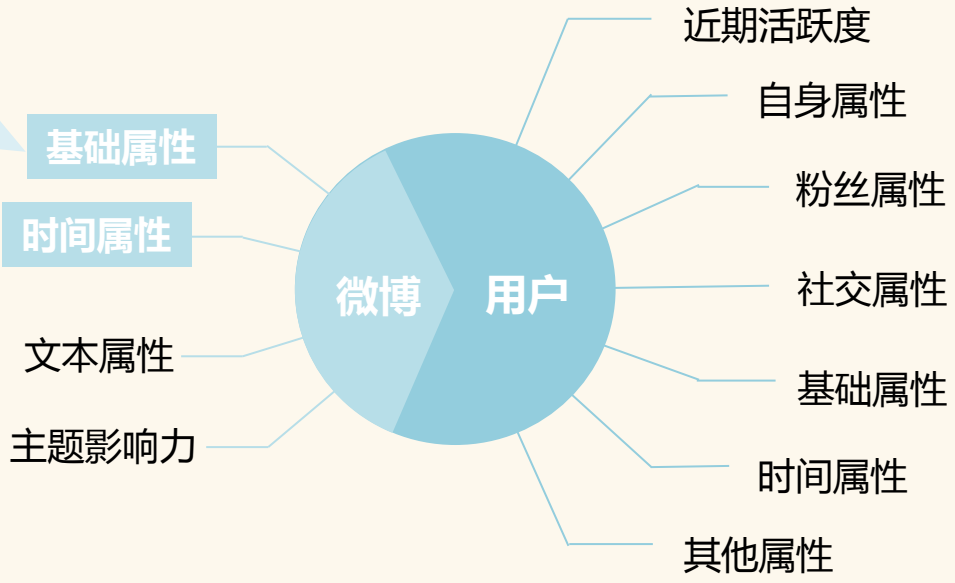


- 在时间段range(2,22,4)内收到的action sum的mid,max,avg
- 用户在当前时间段历史微博的sum的mid,max,avg,std
- 微博发出后的行为趋势：1,2,3,4,6,8,12,24小时内的action的max,avg,sum
- 微博发出后的用户趋势：1,2,3,4,6,8,12,24小时内不同行为独立用户数的max,avg,sum
- 历史六周活动趋势
- 用户在星期几的微博action sum的mid,max,avg,min

特征工程

- 微博长度
- @数量
- http数量
- topic数量
- 标点符号*25

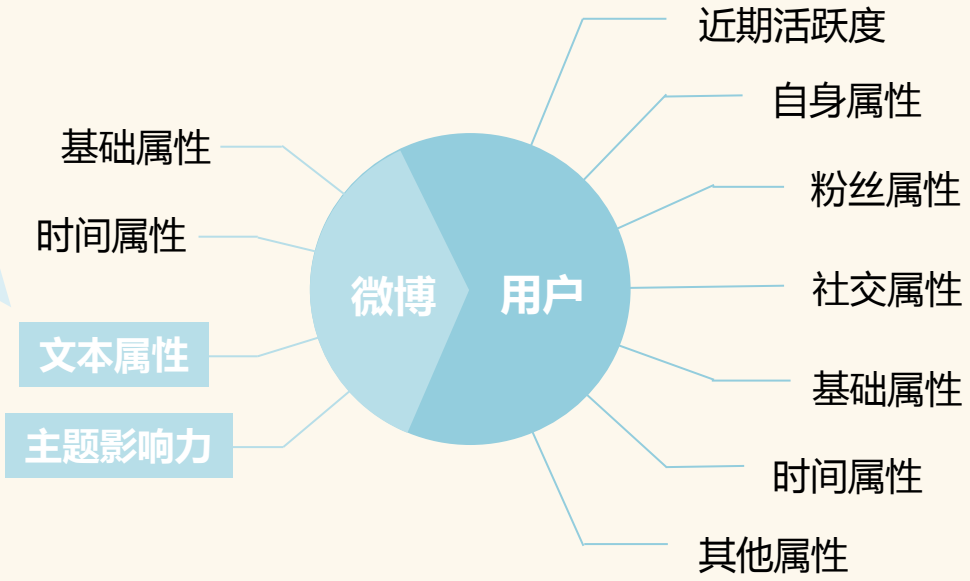
- 周几特征
- 微博所在时间段
- 今明天是否约会日，节假日，调休日，休息日



特征工程

- LDA topic分布*20
- 词袋模型*24
- 微博发出后前后1min，15min，1hour，3hour内的本人微博相似度

- 粉丝fw cm lk微博内容的偏好
- 粉丝活跃度的sum和平均值
- 活跃粉丝对微博内容的偏好
- 粉丝的活跃带权偏好



LDA微博主题分布

- 20个主题的分布作为微博的20维特征
- 通过KL散度，可计算不同博文的相似度
  - $\mathrm{D_{KL}(P||Q)} = \sum_i p_i \log(\frac{p_i}{q_i})$
  - $\mathrm{D_{JS}(P||Q)} = \frac{1}{2}\mathrm{D_{KL}(P||M)} + \frac{1}{2}\mathrm{D_{KL}(Q||M)}$
  - $\mathrm{Sim_{JS}(P||Q)} = 1 - \sqrt{\mathrm{D_{JS}(P||Q)}}$
- 计算微博与前后1min,15min,1hour,3hour内博主其他微博的相似度

词“袋”模型

- 词袋模型太稀疏，改为一个袋子多个词
- 寻找具有区分性的词
  - 对每个词，统计包含它的所有微博
    - 不同档次数量及百分比
    - 不同档次用户数
- 构建了24个词袋，大约900个词

word	label_1	label_2	label_3
帮忙	0.8114921855618953	0.04825105432895063	0.1100688414785413
爆料	0.7789415974145891	0.054103185595567864	0.1093028624192059
预计	0.7955833807626637	0.05857712009106431	0.105475241889584...
透露	0.7841352190866433	0.05627519661621836	0.104652699755468...

影响力模型

- 影响力模型
  - 一次互动行为可以看做博主成功影响了粉丝
  - 计算用户u对粉丝fan_i的影响力$f_{u \to fan_i}$
    - 难点：需要计算下个月的影响力
- 解决方案
  - 训练GBRT预测下月影响力
  - 训练目标：下个月博主的微博被该粉丝互动的比例 [0,1]
  - 样本：前三个月有交互的user-fan对
  - 特征：前三个月的博主的行为，粉丝的行为，交互特征（34维）
  - 效果：训练集测试集的RMSE都在0.048左右

影响力模型and用户偏好特征

- 用户偏好
  - 用户的转评点是有偏好的
  - 分别统计用户转评点微博主题分布的期望
  - 只保留至少有十次互动行为的粉丝
- 设置阈值：影响力大于0.1的为该博主的“铁粉”
- 博主u的某条微博
  - Sum(粉丝的偏好与微博的相似度)
  - Sum(对粉丝的影响力), Avg(对粉丝的影响力)
  - Sum(“铁粉”的偏好与微博的相似度)
  - Sum(粉丝的偏好与微博的相似度*对该粉丝影响力)
  - 比赛后期提升0.1~0.2

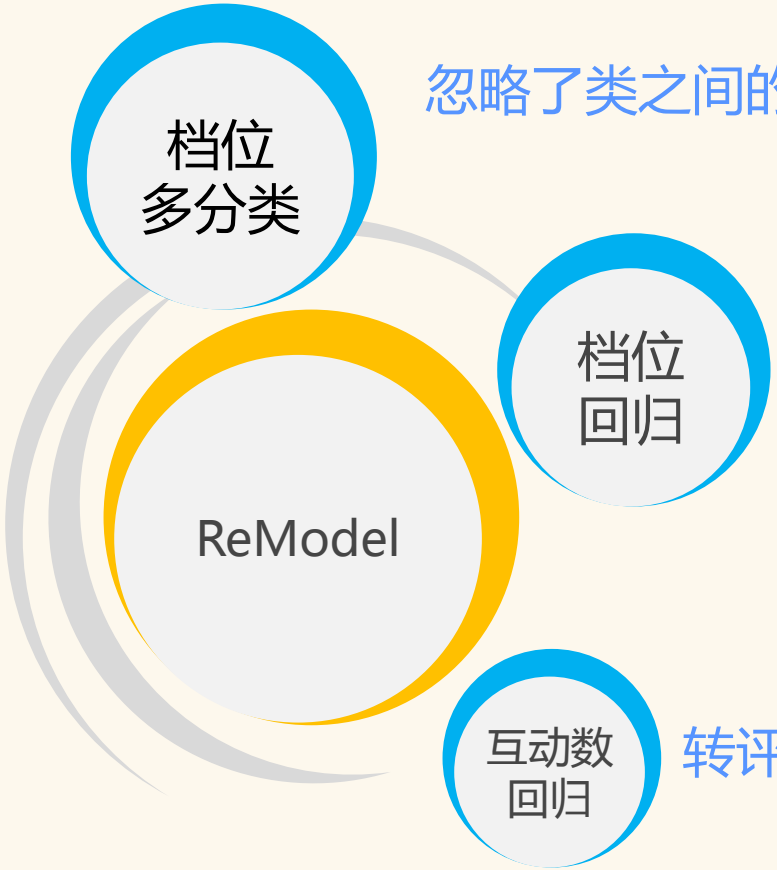
第4部分

算法实现

模型衍变历程

1. GBRT回归action_sum
2. GBRT回归档位
3. 加权采样-GBRT回归档位、log(action_sum+1)
4. One-vs-other 二分类GBDT*5
5. 带权多分类RF
6. 加权采样-多分类LR
7. 我的心好累……

二次建模



忽略了类之间的次序信息

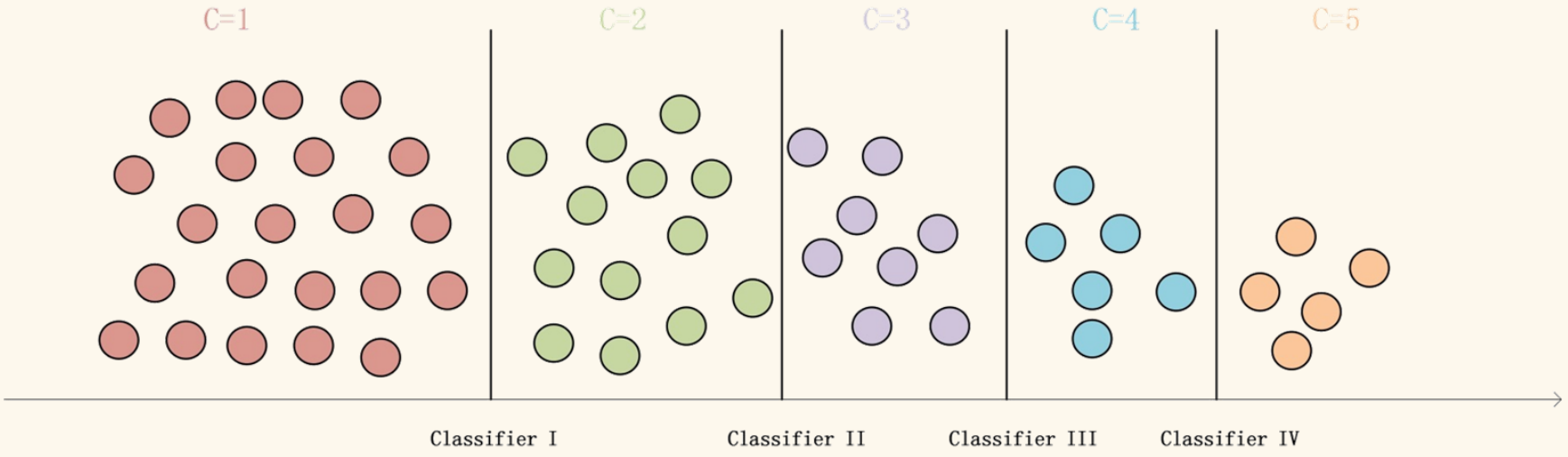
档位不是标称属性

转评点数差异较大，回归误差较大

序数分类 -> 二分类

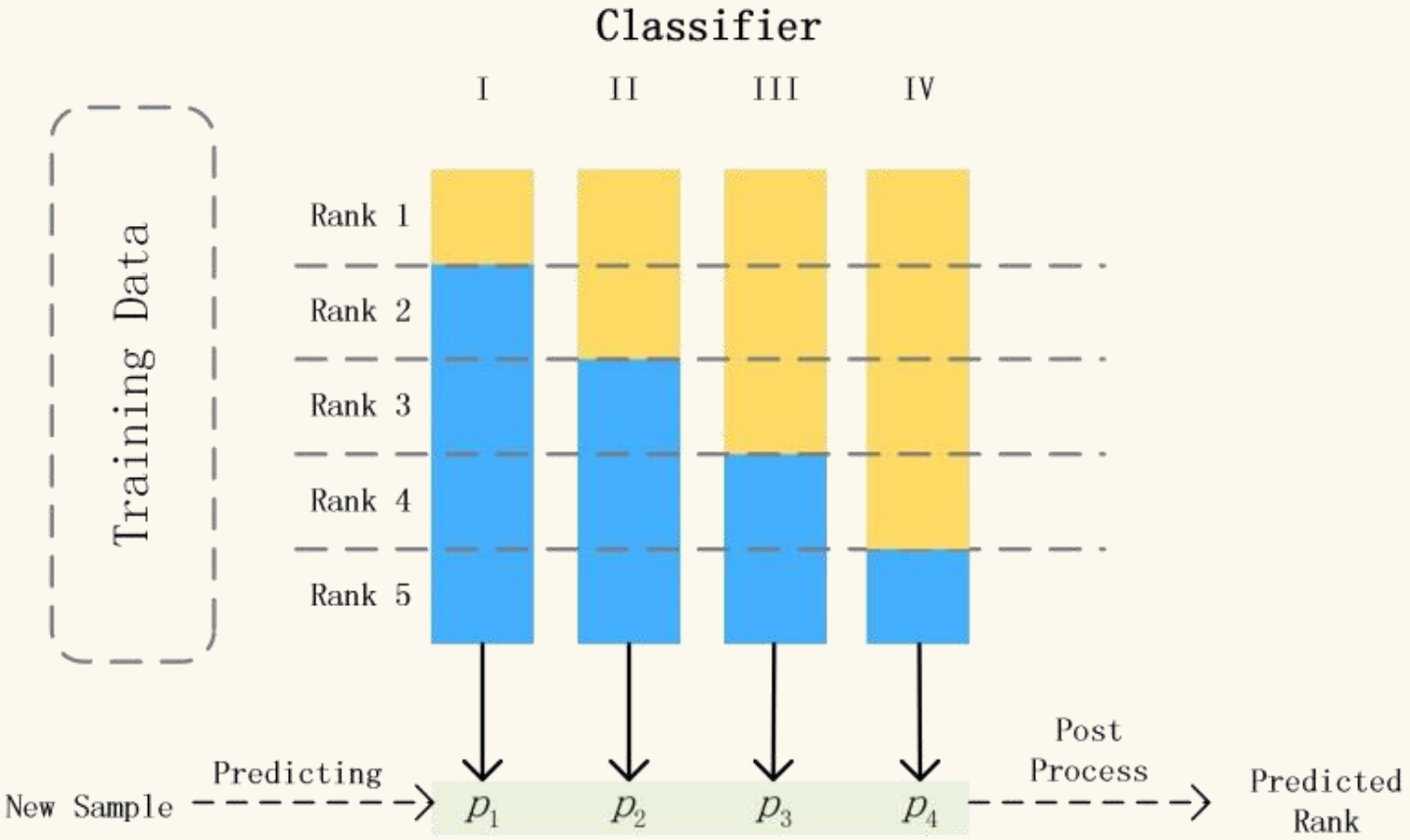
重新定义问题：序数分类(Ordinal Classication)

- 既不是multi-class classification也不是metric regression
- 转化为多个子问题：是否档次K -> 4个子问题，是否大于i



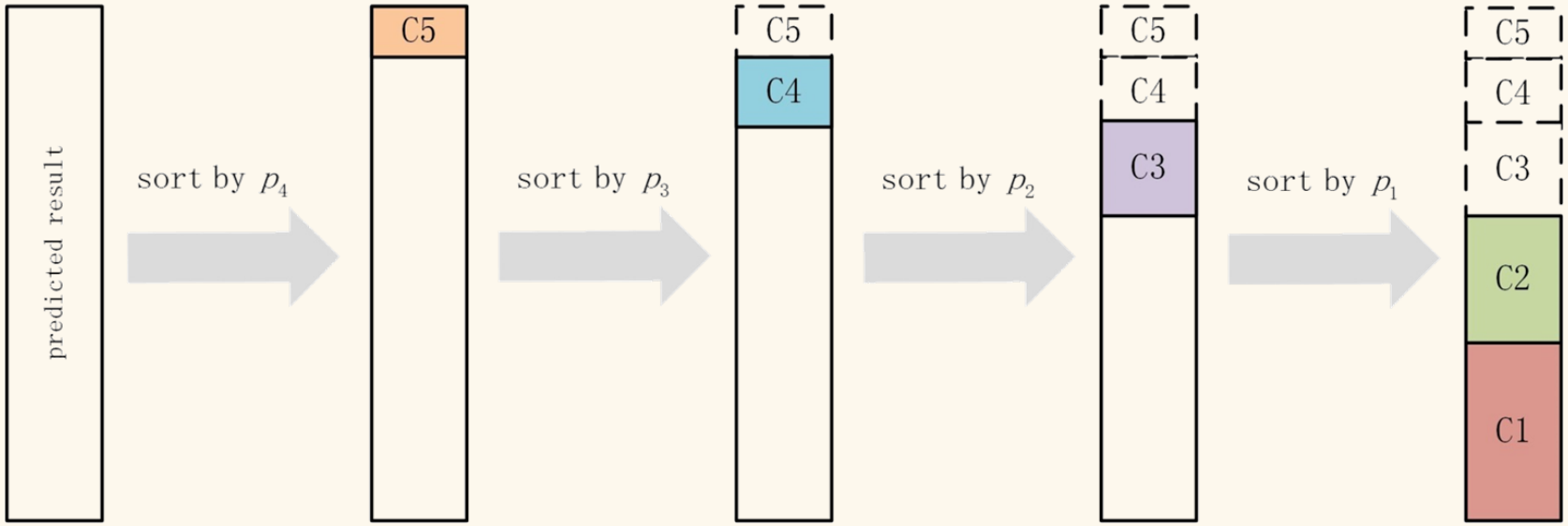
Frank, Eibe, and Mark Hall. A simple approach to ordinal classification. ECML'2001.
Li, Ling, and Hsuan-Tien Lin. Ordinal regression by extended binary classification. NIPS'2006.

训练和预测


Classifier
I
II
III
IV
Training Data
Rank 1
Rank 2
Rank 3
Rank 4
Rank 5
Predicting
New Sample
p_1
p_2
p_3
p_4
Post
Process
Predicted
Rank

预测结果后处理

训练集样本比例和权重比差异很大，没有按照样本权重去采样，
不能用$\Pr(\mathrm{c}=\mathrm{i})=\Pr(c>i-1)\prod_{i\le k<5}(1-\Pr(c>k))$



预测结果后处理

Algorithm 1 Post-processing the predicted score

Require: Samples X, score P where $\vec{p} = (p_1, p_2, p_3, p_4)$

function POST-PROCESSING(X, P, T)
    $V[] \leftarrow$ list of sets;
    $S \leftarrow X$;
    $i \leftarrow 5$;
    while $i > 1$ **do**
        sort(S) according to p_{i-1};
        $V_i \leftarrow$ top T_i elements in S;
        label(V_i) $\leftarrow$ **rank$_i$**;
        $S \leftarrow S \setminus V_i$;
        $i \leftarrow i - 1$;
    end while
    $V_1 \leftarrow S$;
    label(V_1) $\leftarrow$ **rank$_1$**;
    return $V_5 \cup V_4 \cup V_3 \cup V_2 \cup V_1$;
end function

TODO

A 大规模稀疏化LR

B 模型融合

C 结合用户在线情况实时预测

团队亮点

A 代码自动化生成，系统化数据清理

B LDA和影响力模型的结合

C 序数排序 -> 二分类

比赛心得

A 代码自动化生成，快速迭代

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

致谢

感谢新浪和阿里巴巴提供数据和平台

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


THANK
YOU