数据挖掘杂谈
机器学习应用视角

@龙星镖局

2016/07/15

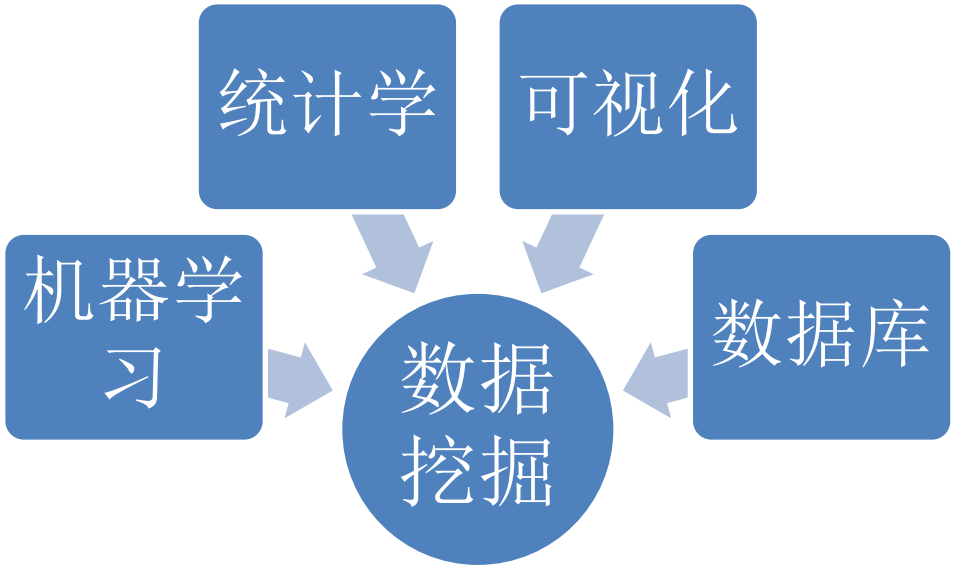
深圳大学

目标

- 数据挖掘是做什么的
- 在工业界是怎么用的
- 学术界和工业界的差别
- 微信购物推荐实例
- 数据挖掘的若干准则
- 工程师的日常

什么是数据挖掘

- 数据挖掘一般是指从数据中发掘有价值信息的过程。是一个交叉学科。

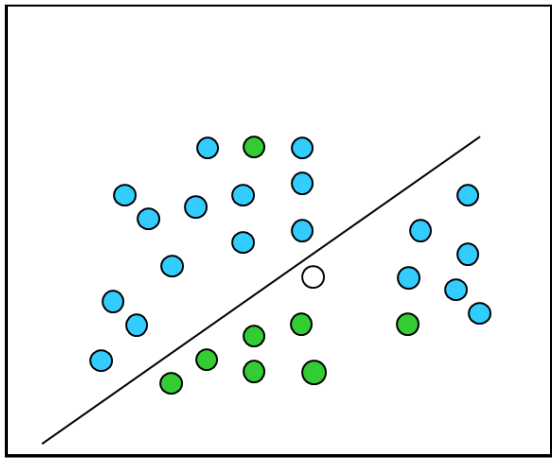


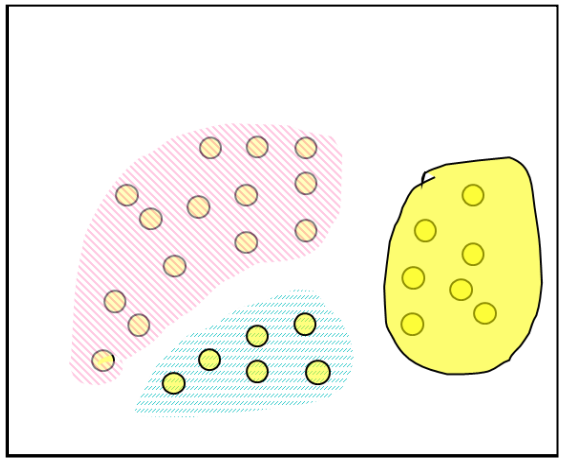
数据挖掘构成

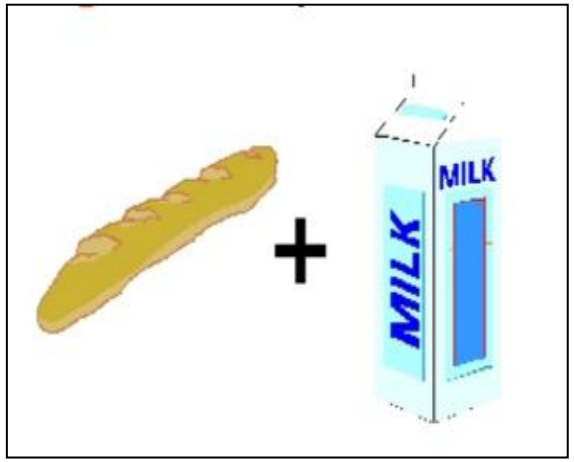
- 数据挖掘三要素
  - 数据
  - 工具
  - 人
- 数据挖掘两个点
  - 挖什么
  - 怎么挖

数据挖掘技术

- 分类&回归
- 聚类&降维
- 关联分析







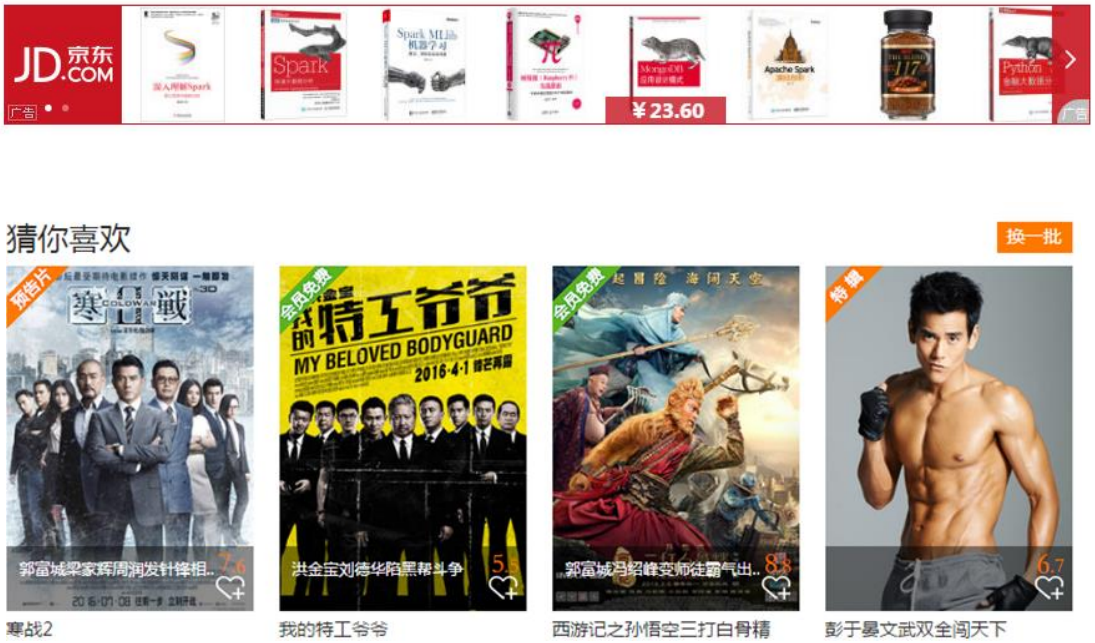
在工业界

- Netflix 《纸牌屋》

- 网易花田

在腾讯

- 腾讯视频



- 微信购物

更多应用

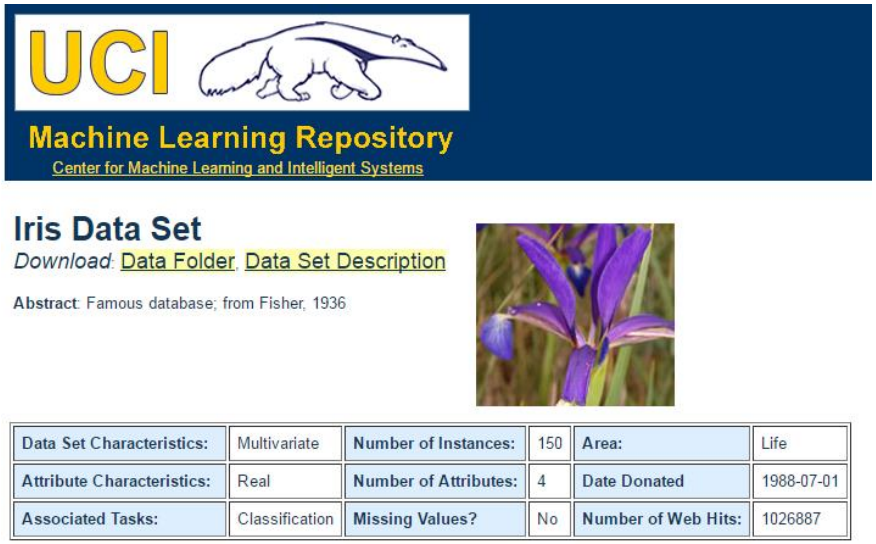
- 个性化推荐
- 电商购物
- 在线广告
- 互联网金融
- 欺诈检测
- 客户关系管理
- 精准营销
- 游戏

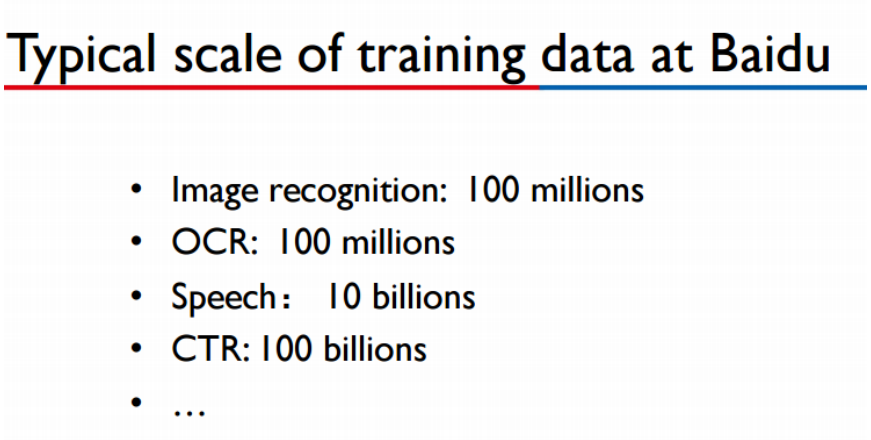
接下来

- 数据挖掘是做什么的
- 在工业界是怎么用的
- 学术界和工业界的差别
- 微信购物推荐实例
- 挖掘的一些基本准则
- 工程师的日常

举个例子

- 学术界
  - 发明一个算法提升分类准确率
    - 理论分析
    - 实验分析
- 工业界
  - 组一个团队提升产品的变现能力
    - 拆解目标，技术团队优化投放/CTR预估/
    - 离线/上线实验CTR/CPM





相关数据参考：yukai, http://www.cikm2013.org/slides/kai.pdf, CIKM 2013

关注点不同

数据收集 → 数据清洗 → 特征工程 → 数据建模







学术界 vs 工业界

对比项	学术界	工业界
数据	小清新 比较规范 一定的预处理	大流氓 比较原始 需要特别处理
工具	原型 比较随意 比较专用	成品 相对固定 相对通用
人	研究者 少 单打独斗 一条龙	工程师 多 团结互助 分工明确
挖什么	不那么明确	一般比较明确
怎么挖	一般要炫酷	较常规接地气

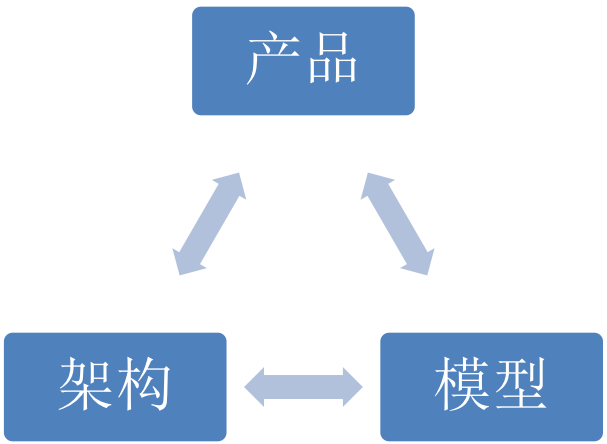
学术界 vs 工业界

- 挖掘结果的影响
  - 一般般 vs 影响说多大就有可能有多大
- 数据挖掘过程
  - 一锤子买卖 vs 多次迭代
- 评测指标不同
  - 从准确率/召回率/AUC到点击率/CPM等
- 模型稳定健壮性要求
  - 样本/特征监控
  - 系统服务监控

工业界独有的特点

- KPI导向的挖掘
  - 老板拍目标，小老板拆分目标，工程师干活
  - 分阶段KPI，从糙快猛到精细化
- 持续的挖掘
  - 一拨人持续优化
  - 人不在，业务还在
- 受业务形态影响较大
  - 模型场景多，且在变化，追求尽量通用
  - 模型只是一个因素，且潜力发挥依赖作用形式

模型非独立存在



- 好的产品设计降低模型的难度
  - 微信语音信息自然开始和结束
  - 某语音产品需要算法识别
- 好的架构设计提升模型的效率
  - 清晰的数据流，数据采集方便
  - 完善的属性，丰富的特征

殊途同归

- 学术界想向工业界靠拢，而工业界也在等着拥抱学术界。
  - 利用工业界数据做研究/竞赛
    - KDD CUP，大数据竞赛
  - 工业界借鉴学界的研究成果，实用化
    - 大公司挖学术界人才组实验室
- 无论学界还是工业界，终极目标是用技术改变人们的生活。

接下来

- 数据挖掘是做什么的
- 在工业界是怎么用的
- 学术界和工业界的差别
- 微信购物推荐实例
- 挖掘的一些基本准则
- 工程师的日常

微信购物业务


中国移动 16:48
发现
扫一扫
摇一摇
附近的人
漂流瓶
购物
游戏
微信
通讯录
发现
我


中国移动 16:49
返回
京东购物
频道
JD
带着宝宝探索世界
换旧版
立即抢购
中兴威武3 青春版 ¥999
首发
威武3S新品首发
新品
品牌
香奈儿满600减60
特价
内增高休闲女鞋
¥138
发现好店
心跳秒杀
吃货场
酒类秒杀专场
¥69
购物
搜索
购物圈
108
购物车
个人中心


中国移动 16:49
返回
品牌
上新
热卖
限时
预告
分类
每天 9:00 准时上新
剩2天
梦幻七夕
Tanabata Festival
奢品珠宝人气推荐
每满600减60
PANDORA
京东自营
奢品珠宝七夕专场
4折起
POLO
剩2天
购物
搜索
购物圈
购物车
个人中心

微信购物的需求

- 给定一个用户和待展现的候选item list，返回最佳排序
  - 点击率/转化率导向
  - 业务规则
    - 类目差异性
    - 人工置顶部分item
    - 弱可解释性

- 排序问题/pctr问题

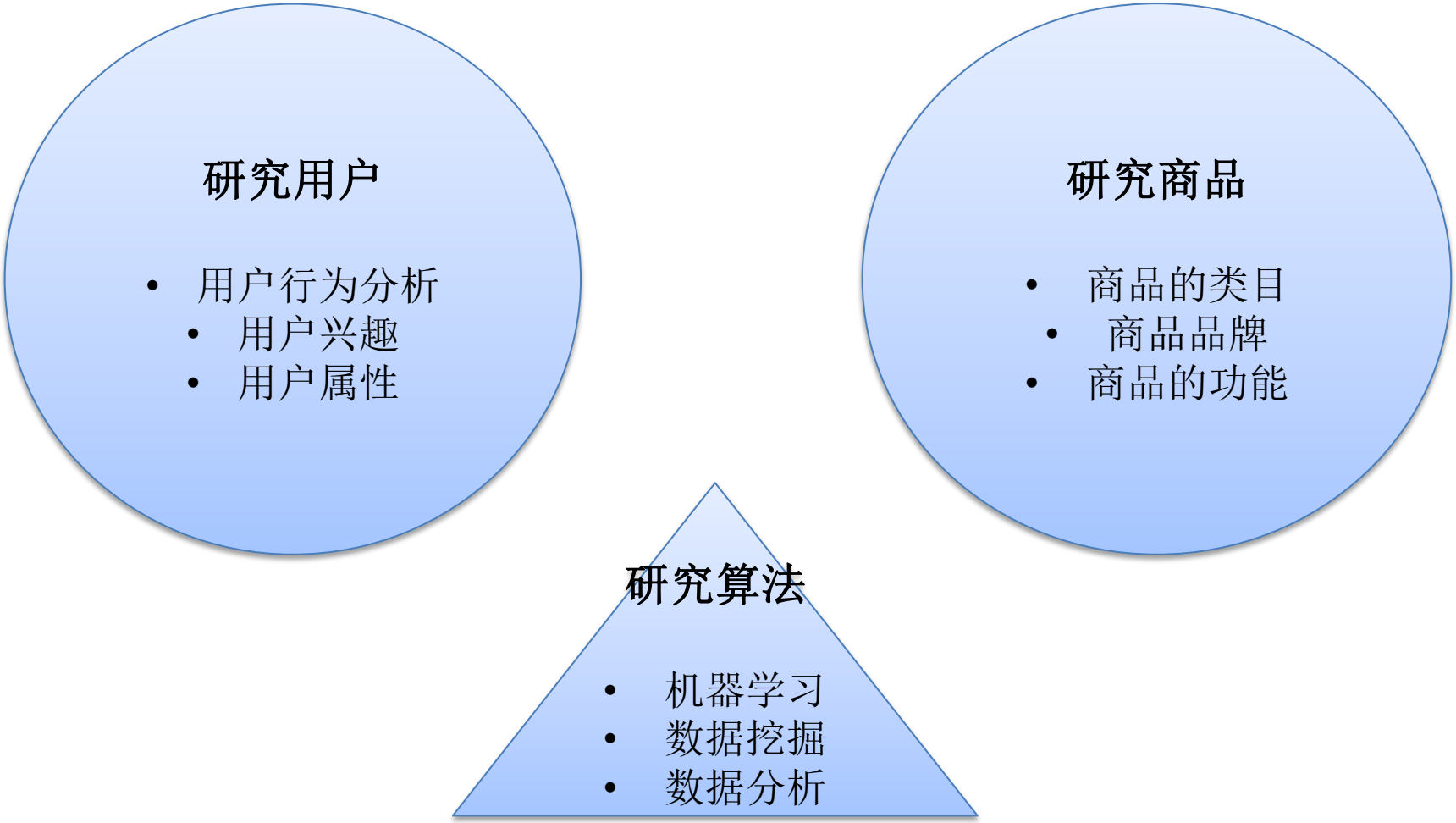
问题的规模

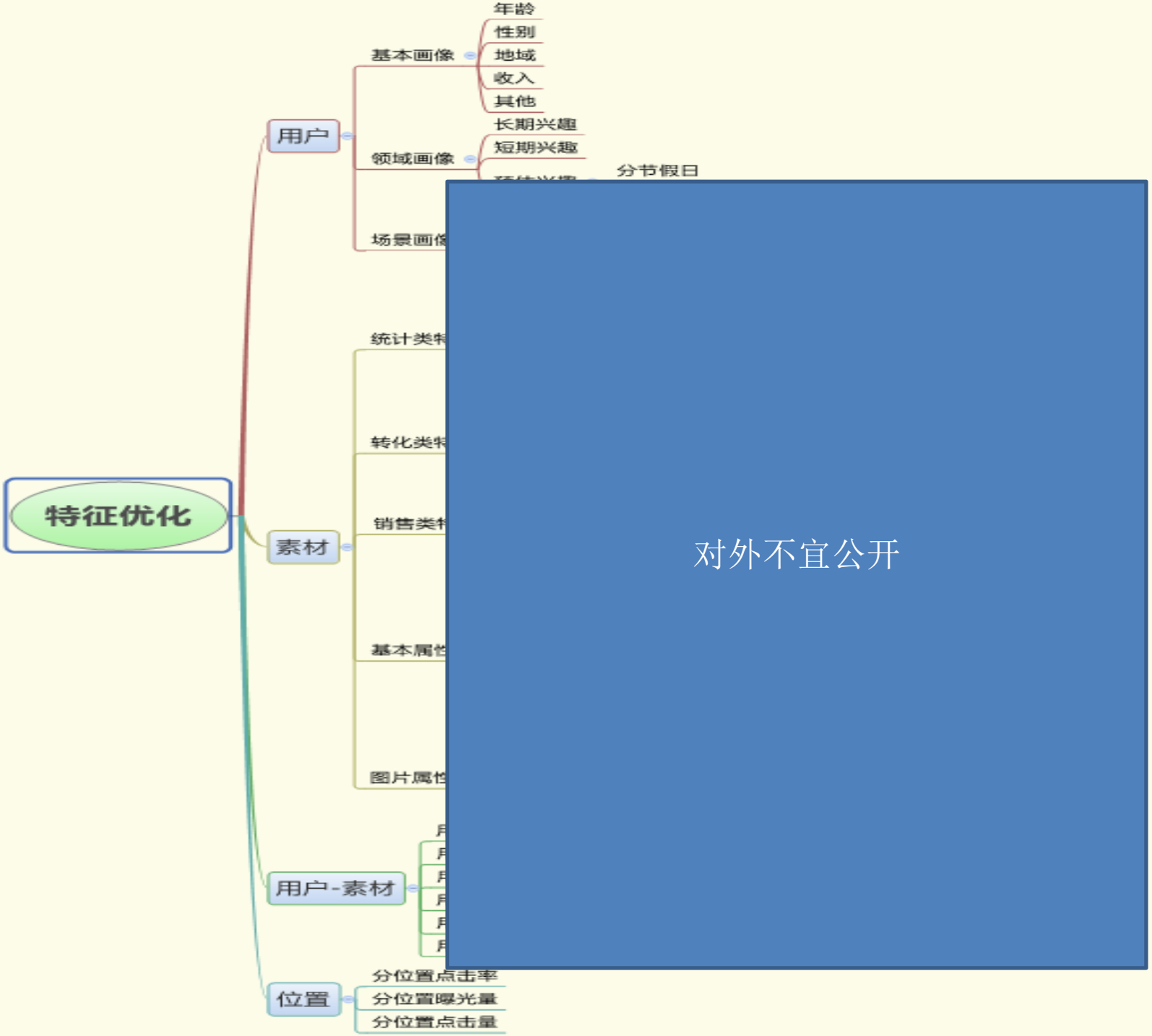
- 每天用户数 千万
- 每天请求数 亿
- 每天素材 百
- 商品池 十万
- 场景 几十
- 平台：微信/手Q/APP

推荐问题

- Find the best match between a given user **u**, in a given context **c**, and a suitable item list **l**
- 什么是 the best match？
  - Max $\sum_i^T ROI(l_i,u_i,c_i)$
- 预测单次$ROI(l_i,u_i,c_i)$
  - 单次CTR / CVR预估

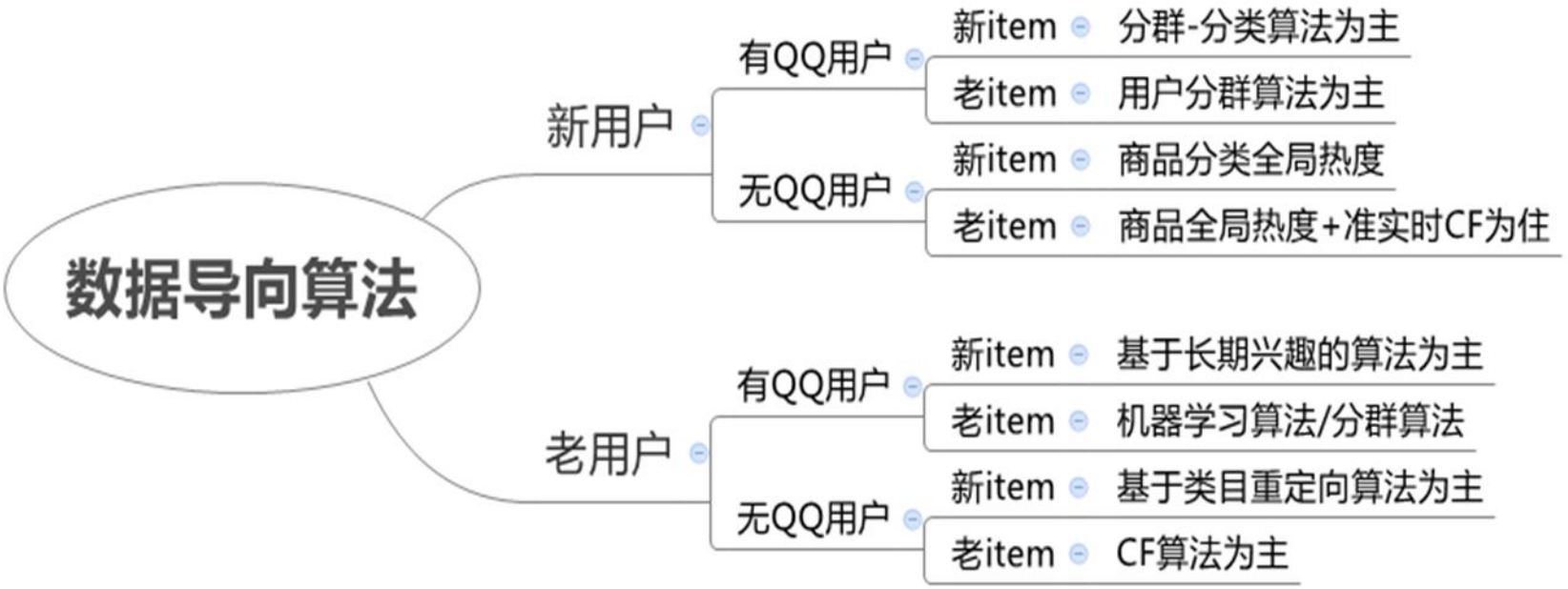
推荐套路


研究用户
• 用户行为分析
• 用户兴趣
• 用户属性
研究商品
• 商品的类目
• 商品品牌
• 商品的功能
研究算法
• 机器学习
• 数据挖掘
• 数据分析

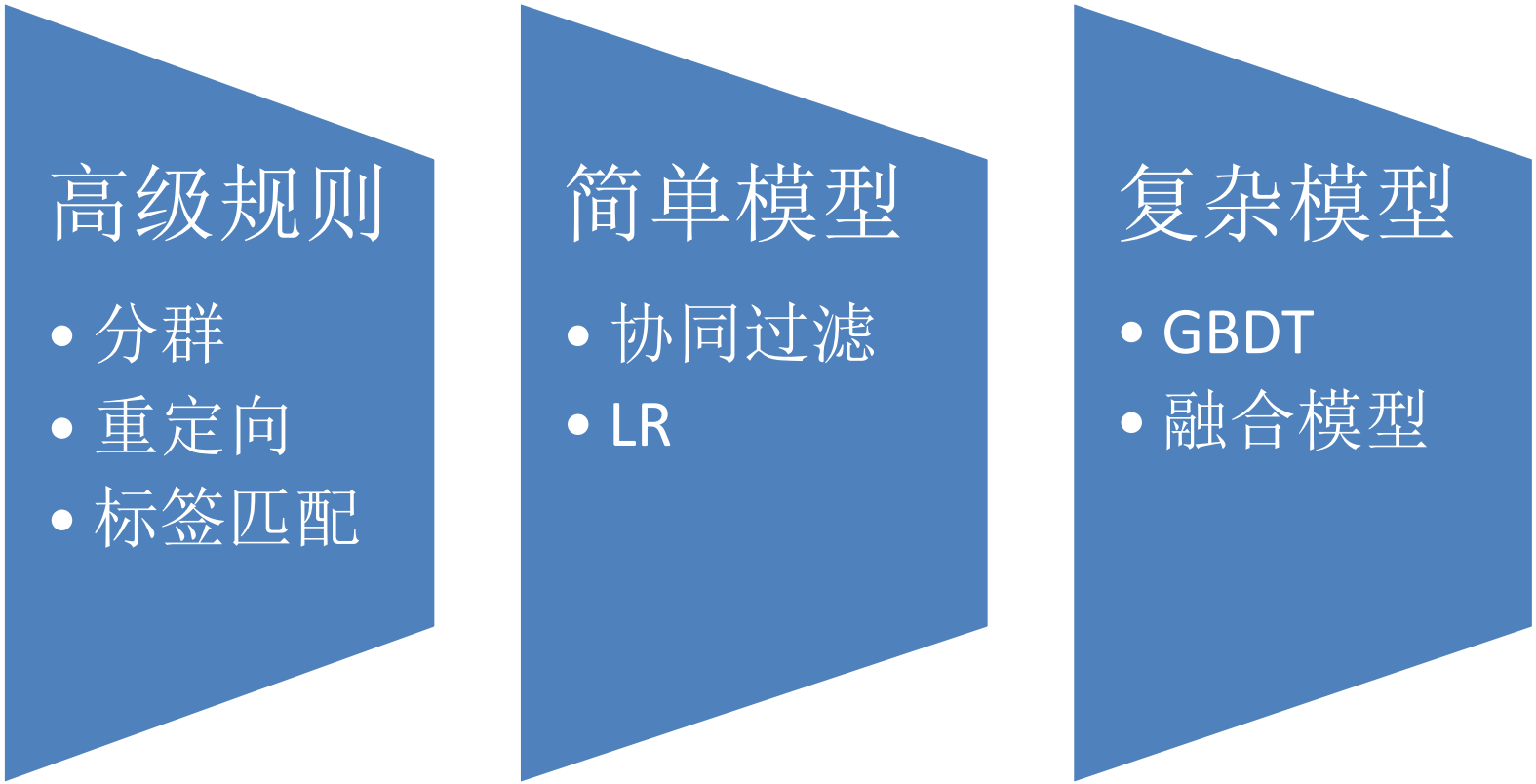

特征优化
用户
基本画像
年龄
性别
地域
收入
其他
领域画像
长期兴趣
短期兴趣
分节假日
场景画像
素材
统计类特
转化类特
销售类特
基本属性
图片属性
用户-素材
位置
分位置点击率
分位置曝光量
分位置点击量
对外不宜公开

打法套路

- 面向数据特性的算法布局



推荐算法进化



模型虽然越来越复杂，但提升幅度却是越来有限，为什么？

推荐技术1.0

- 全局热度/分群热度
  - 合理的分群很关键，考虑画像和行为
- 重定向逻辑
  - 基于用户行为进行重定向，看神马就推浮云
  - 不同粒度的重定向
    - 二级类目>三级类目>品牌>一级类目>商品

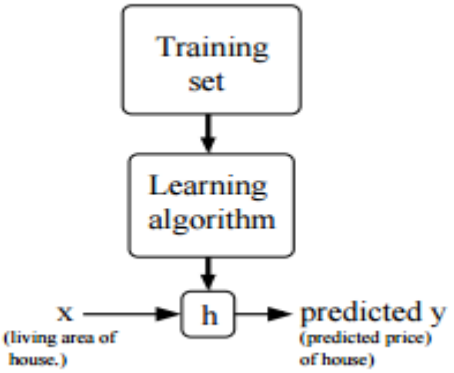
注意：去除position bias/ 不同行为权重分配 / 时间影响

推荐技术2.0

- 协同过滤算法
  - 更高级点/更细的分群
  - 使用Item-Based CF
    - 优势：商品量级<<用户量级
    - 劣势：素材/商品生命周期不长
- 基于用户/商品倾向性推荐
  - 年龄/购买力/性别等
  - 统计的置信性问题

推荐技术2.0

- 机器学习技术
  - 希望自动分配各因素的权重
  - 设定一个优化目标，求解参数



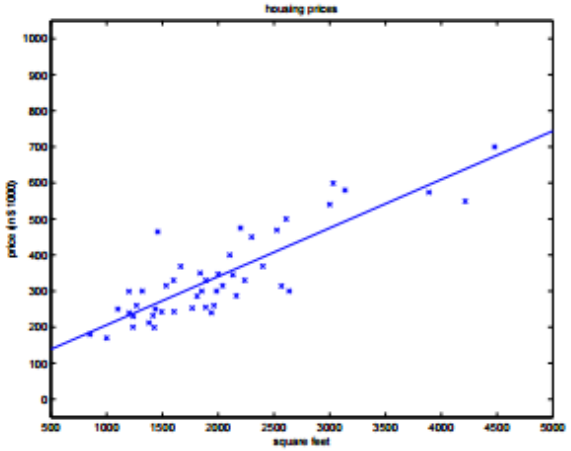
$$Obj(\Theta) = L(\Theta) + \Omega(\Theta)$$

- Logistic Regression

$$L = \sum_{i=1}^{n} l(y_i, \hat{y}_i)$$

$$l(y_i, \hat{y}_i) = y_i \ln(1 + e^{-\hat{y}_i}) + (1 - y_i) \ln(1 + e^{\hat{y}_i})$$

$$\Omega(w) = \lambda \|w\|^2$$



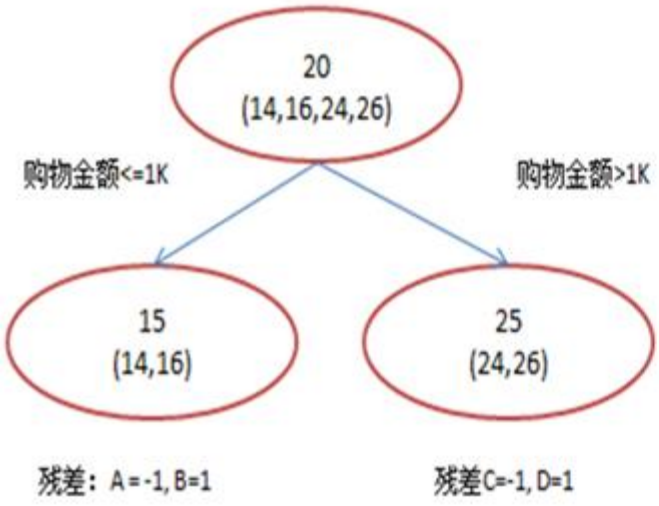
Why LR？

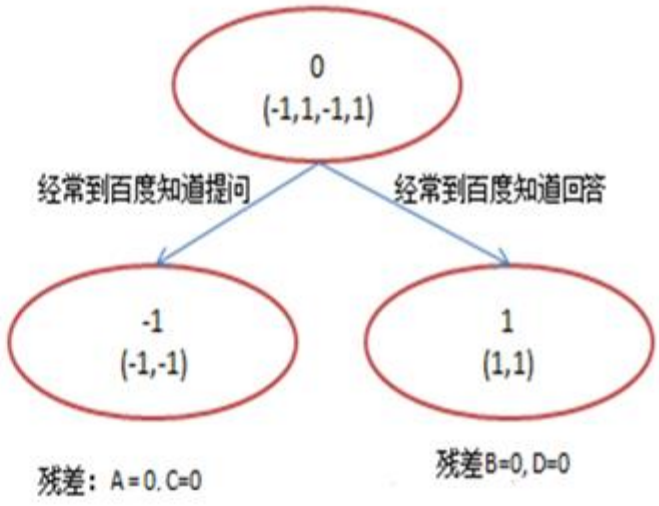
- 大部分人说
  - 好解释
  - 易并行
  - 有概率意义
  - 好实现
  - 理论严谨
  - …
- 实际情况是
  - 大家都在用
  - 没有其他工具用
  - **大规模**
  - **易于Debug**
  - **减少特征并发症**
  - **阈值选择**

Why not LR？

- 预测目标和因素（特征）线性关系
  - 精细离散化
  - 特征交叉
    - 男->喜欢电子产品 0.6
    - 男 & 25岁 ->喜欢电子产品 0.7
    - 男 & 25岁 & 码农 -> 喜欢电子产品 0.8
    - ….
- 线性模型需要太多的（人工）预处理，有没方法解放人？

GBDT算法





- 决策树
  - 自动地特征预处理
  - 过拟合

- GBDT
  - 决策树
  - Boosting

如何学习？

- Boosting or additive training

$$
\begin{aligned}
\hat{y}_i^{(0)} &= 0 \\
\hat{y}_i^{(1)} &= f_1(x_i) = \hat{y}_i^{(0)} + f_1(x_i) \\
\hat{y}_i^{(2)} &= f_1(x_i) + f_2(x_i) = \hat{y}_i^{(1)} + f_2(x_i) \\
&\cdots \\
\hat{y}_i^{(t)} &= \sum_{k=1}^{t} f_k(x_i) = \hat{y}_i^{(t-1)} + f_t(x_i)
\end{aligned}
$$

Tree model

加入一个新的函数

第t轮的模型预测

保留前面t-1轮的模型预测

- 如何选择f?

$$
\begin{aligned}
Obj^{(t)} &= \sum_{i=1}^{n} l(y_i, \hat{y}_i^{(t)}) + \sum_{i=1}^{t} \Omega(f_i) \\
&= \sum_{i=1}^{n} l\left(y_i, \hat{y}_i^{(t-1)} + f_t(x_i)\right) + \Omega(f_t) + constant
\end{aligned}
$$

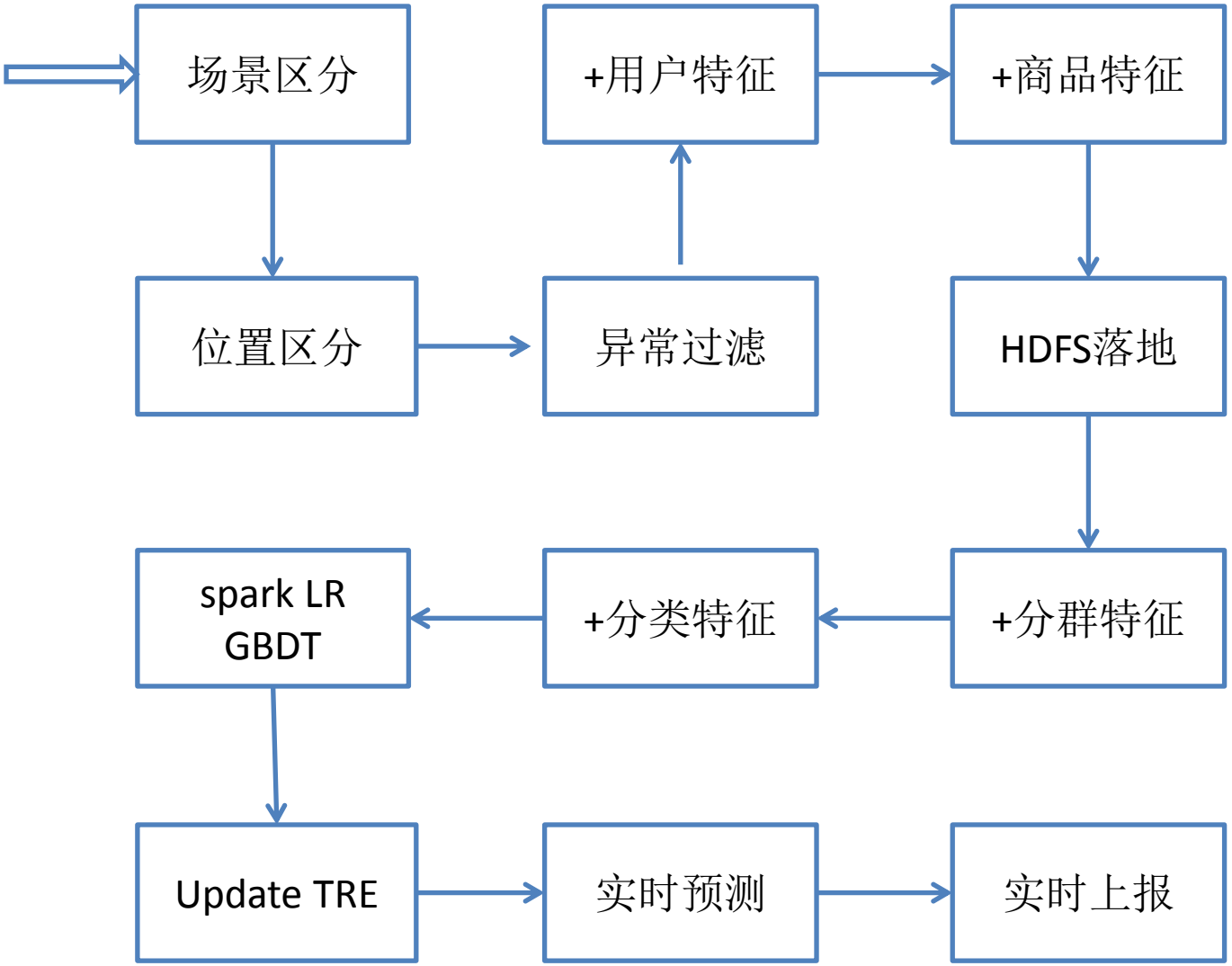
目标：找到 f_t 来优化这一目标

- 如何预测

$$\hat{y}_i = \sum_{k=1}^{K} f_k(x_i)$$

参考：陈天奇《boosted tree》，
http://www.52cs.org/?p=429

完整的流程

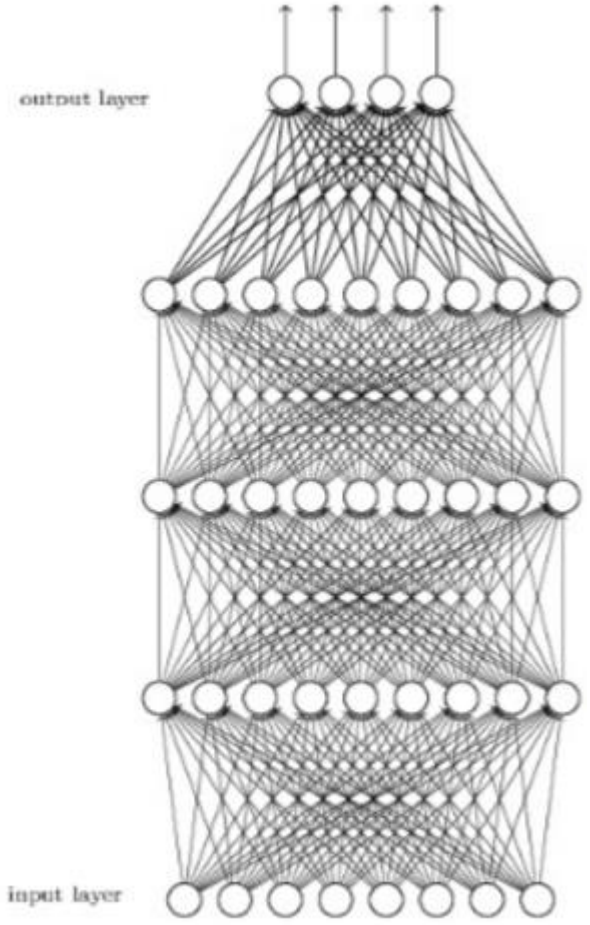

场景区分
+用户特征
+商品特征
位置区分
异常过滤
HDFS落地
spark LR
GBDT
+分类特征
+分群特征
Update TRE
实时预测
实时上报

接下来

- 数据挖掘是做什么的
- 在工业界是怎么用的
- 学术界和工业界的差别
- 微信购物推荐实例
- 挖掘的一些基本准则
- 工程师的日常

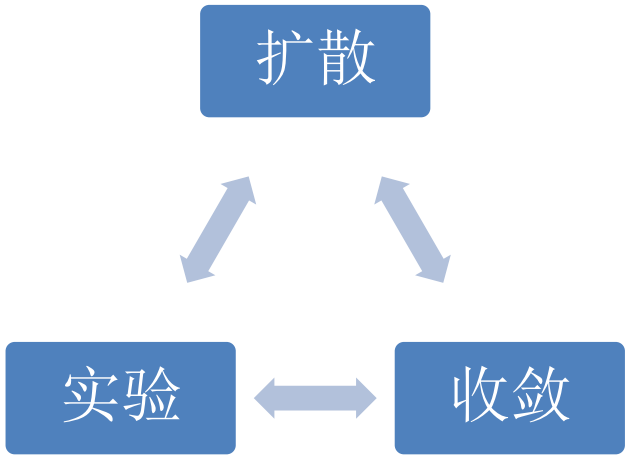
准则一：由浅入深

- 浅模型对外接的依赖较少
  - 更高的性能
  - 更易于维护
- 深模型依赖较多
  - 更高效的架构支持
  - 更复杂的调参和debug
  - 更高的硬件支持，如GPU



准则二：特征构造有迹可循

- 特征分类
  - 低级 vs 高级
  - 单维 vs 多维 vs 降维
  - 非实时 vs 实时
  - 直接来源 vs 间接来源
  - Bias特征 vs 有效特征
- 从简单到复杂构造特征
- 特征构造是个持续循环的过程



准则三：重视特征预处理

来自：https://www.zhihu.com/question/29316149/answer/110159647

准则四：越简单越美丽

- 简单的模型泛化能力强，稳定
  - 正则项
  - 决策树减枝
  - SVM最大间隔
  - 特征选择
  - 特征降维
- Why？Y = w1*x1 + w2*x2 + w3*x3 +…
  - 最小化|w|的p次方
  - 减少w个数
  - 最小化||w||
  - 减少w个数
  - 减少w个数

准则五：目标导向的评测

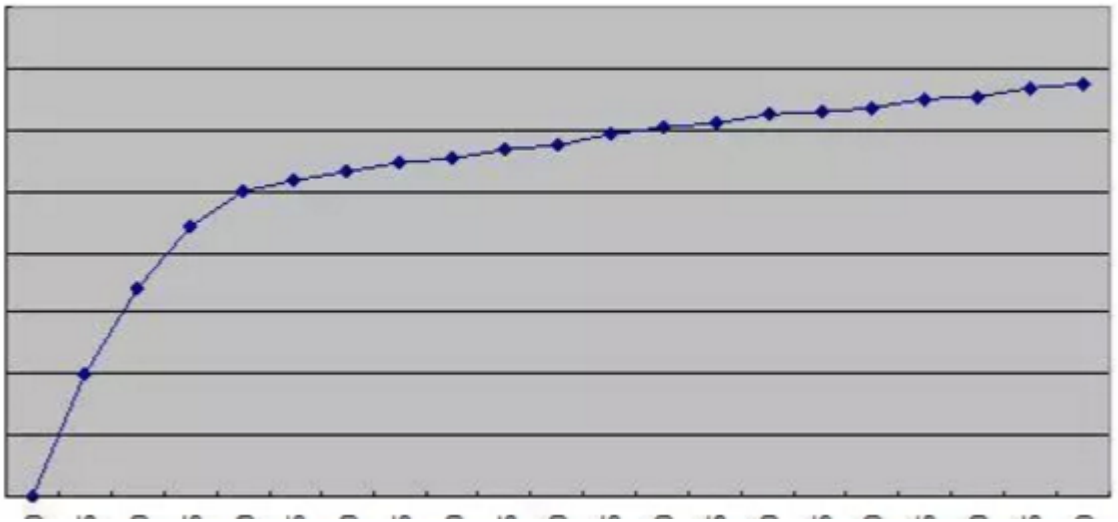
- 评测挖掘成功的标准要和应用场景一致
  - 购物场景中购物性别和真实性别
  - 主题模型好不好要看用在什么地方

准则六：算法差异没有那么大

- 每个算法都可以做到极致，极限情况下大家的表现趋于近似
  - 设计好非线性特征，线性模型媲美非线性
  - 理解每个算法的区别，可以“曲线救国”
    - 决策树构造特征给LR用
    - 非线性算法学习模型，线性模型解释模型

准则七：数据够用就好

- 持续增加训练数据的收益有瓶颈点，够用就好，不宜追求数据的大
  - 太少的数据，不足以支撑真实的样本的空间
  - 太多的数据，影响学习的效率。
    - FB负采样10%，效果基本相当。



准则八：融合模型

- 三个臭皮匠顶一个诸葛亮
- 多模型融合效果已经被多次证明有效，基本成为数据竞赛必备
  - 模型简单组合投票
  - 模型嫁接融合
  - 模型级联融合
  - 模型输出为融合模型能够输入

准则九：拆分目标

- 将优化目标拆分，控制算法能控制的
  - 某公司要优化某场景的广告收入，就做了一个回归模型拟合，效果很差
    - CPM = CTR * ACP，CTR预估有意义，ACP人为因素
  - 大目标制约因素太多，拆分为小目标各个击破
    - 提升挖掘效率和团队作战能力

准则十：重视特征选择

- 特征选择比想象的重要
  - 无用的特征干扰了模型训练
- 特征选择的不那么容易
  - 特征之间关系的捕获
  - 正则项只是最naïve的方式
  - 单个特征重要性和一类特征重要性

终极准则：“人工”智能

- 有多少人工，就有多少智能。人能做的不要留给机器。
  - 数据清洗
  - 特征构造
  - 调参
  - 调网络结构

接下来

- 数据挖掘是做什么的
- 在工业界是怎么用的
- 学术界和工业界的差别
- 微信购物推荐实例
- 挖掘的一些基本准则
- 工程师的日常

挖掘工程师的logo

- **二十一世纪最性感的职业**
  - 数据科学家
  - 人傻钱多
  - 很高深，很神秘
  - 做的事情很高大上
  - 很人工智能
  - 敲几行代码，搞定一切。
- 够了，这些都是不对的。

工程师的日常

- 大部分时间都在准备数据，做自己不那么喜欢的事情

不喜欢的 vs 耗时间的

参考：http://www.forbes.com/sites/gilpress/2016/03/23/data-preparation-most-time-consuming-least-enjoyable-data-science-task-survey-says/

主要内容

- 数据挖掘是做什么的
- 在工业界是怎么用的
- 学术界和工业界的差别
- 微信购物推荐实例
- 挖掘的一些基本准则
- 工程师的日常

参考资料

- http://www.52cs.org/?p=429
- https://research.facebook.com/publications/practical-lessons-from-predicting-clicks-on-ads-at-facebook/
- http://www.zhihu.com/question/28641663/answer/41653367
- https://github.com/dmlc/xgboost
- http://www.cikm2013.org/slides/kai.pdf
- http://google.com
- http://www.forbes.com/sites/gilpress/2016/03/23/data-preparation-most-time-consuming-least-enjoyable-data-science-task-survey-says/