"千人千面" —— 机器学习

一点就通

liangdong@smzdm.com

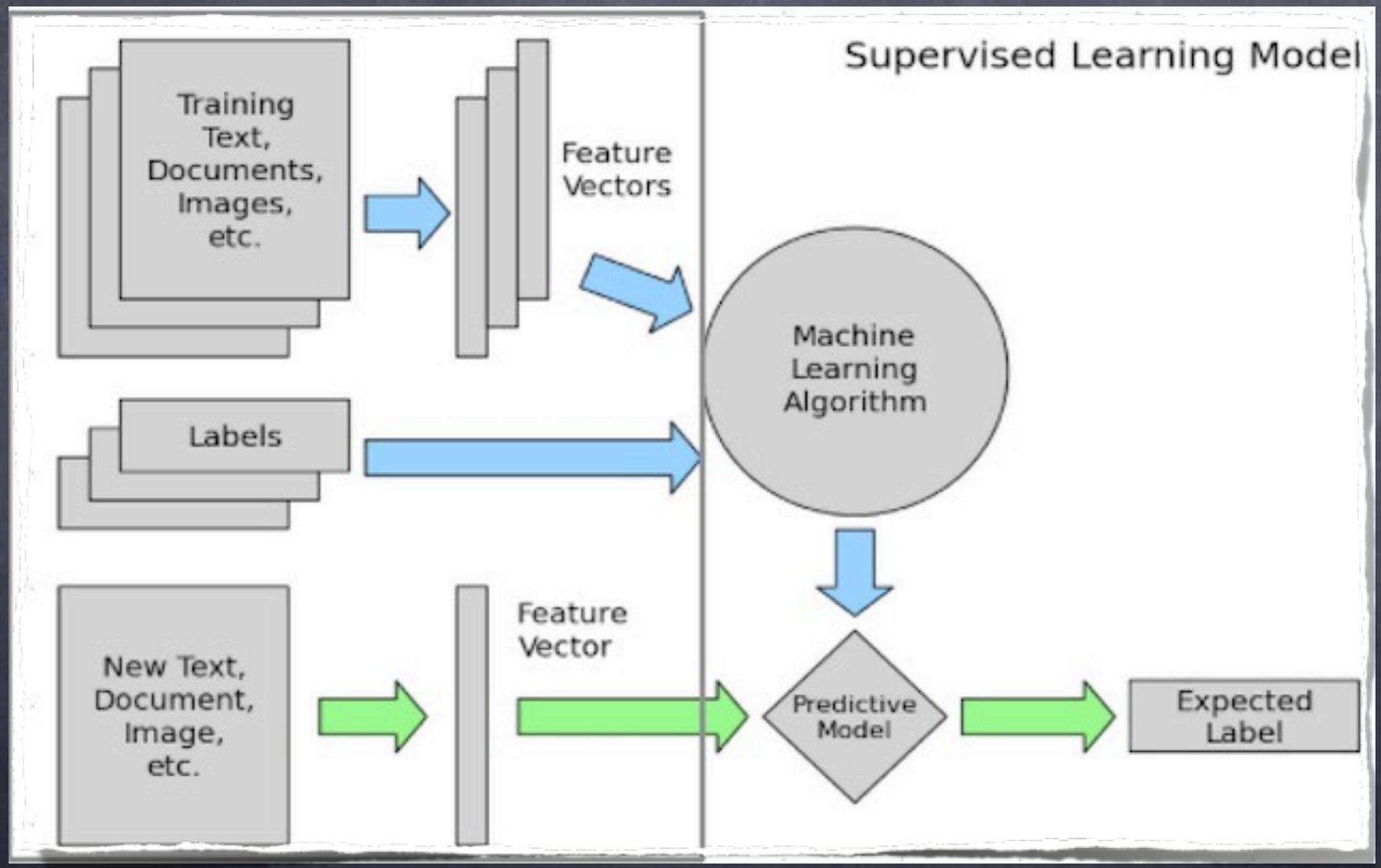
目录

- 看懂机器学习
- 用户画像 -> 用户特征
- 商品属性 -> 商品特征
- [用户特征，商品特征] -> 千人千面

案例分析

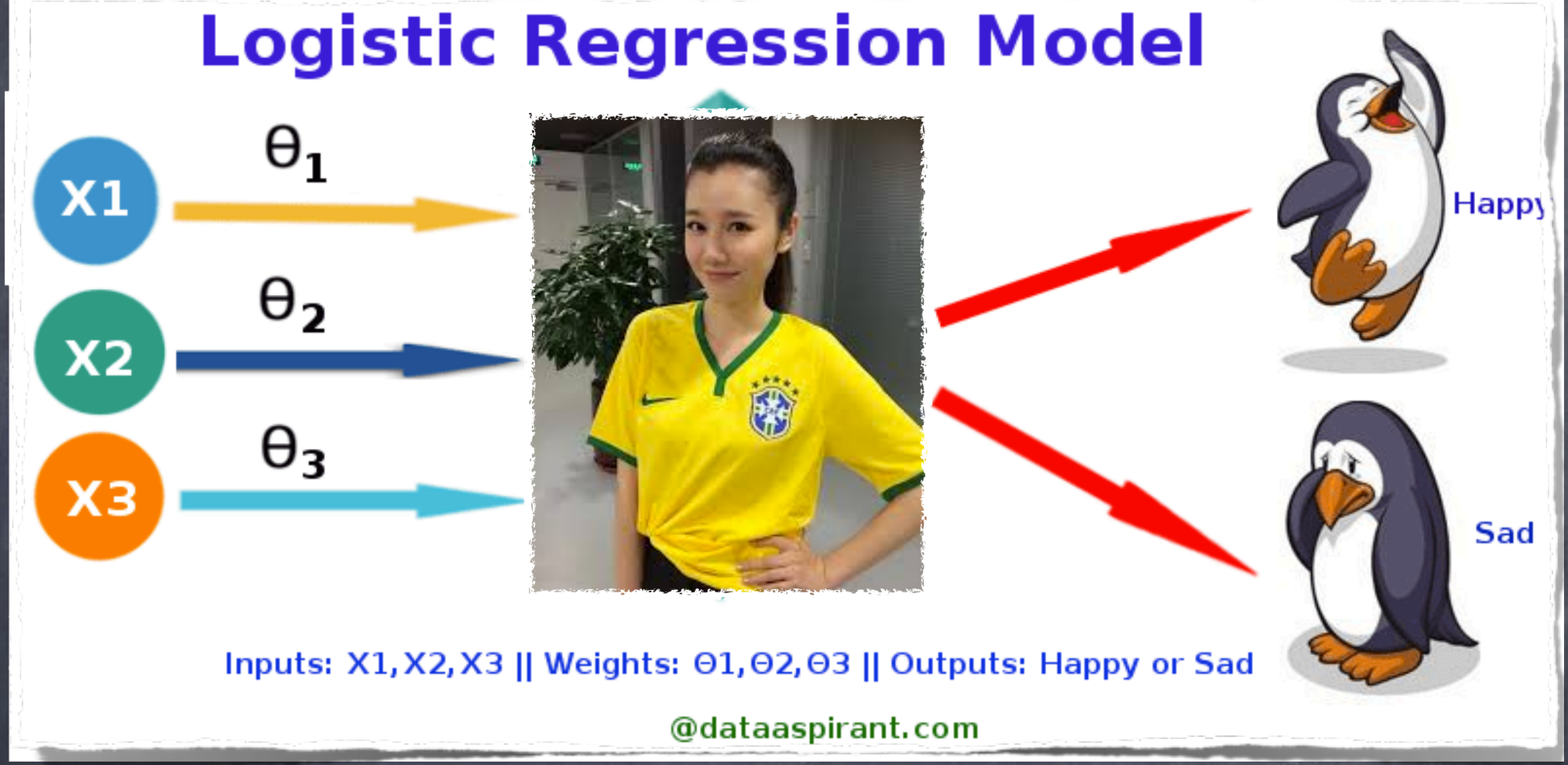
- 先发后审，为了实现机器自动审核，开发了大量策略
- 最终判定是否自动发，需要大量人工规则if else if ... else ...
- 规则是人设计的，代码是人写的... 陷入僵局

下图描述了几个流程?


Supervised Learning Model
Training Text, Documents, Images, etc.
Feature Vectors
Machine Learning Algorithm
Labels
New Text, Document, Image, etc.
Feature Vector
Predictive Model
Expected Label

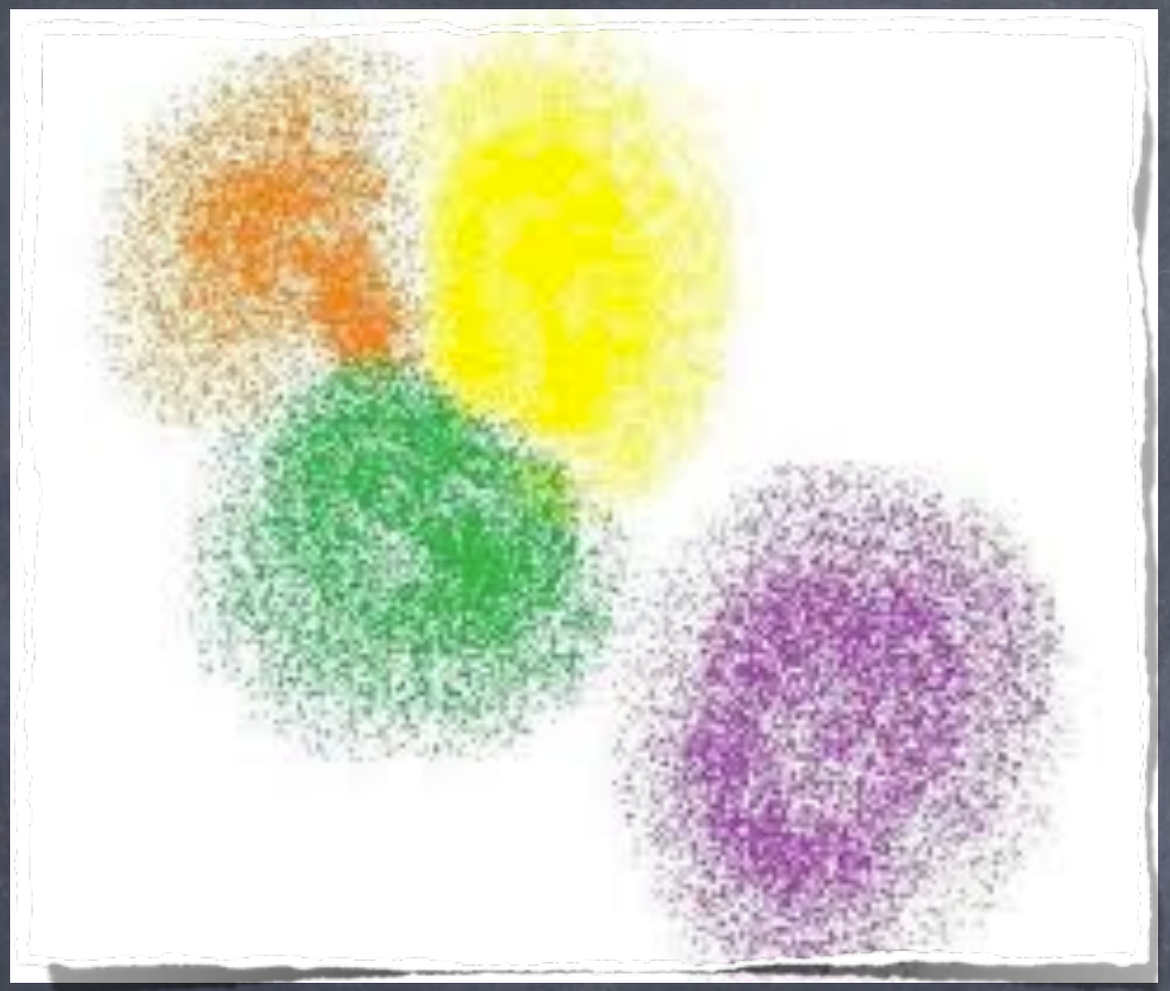
监督机器学习

- 训练模型
  - 特征+标签 -> 算法 -> 预测模型
- 预测标签：
  - 特征 -> 模型预测 -> 标签


Logistic Regression Model
X1
X2
X3
θ1
θ2
θ3
Happy
Sad
Inputs: X1, X2, X3 || Weights: Θ1, Θ2, Θ3 || Outputs: Happy or Sad
@dataaspirant.com

预测

监督机器学习



聚类

非监督机器学习

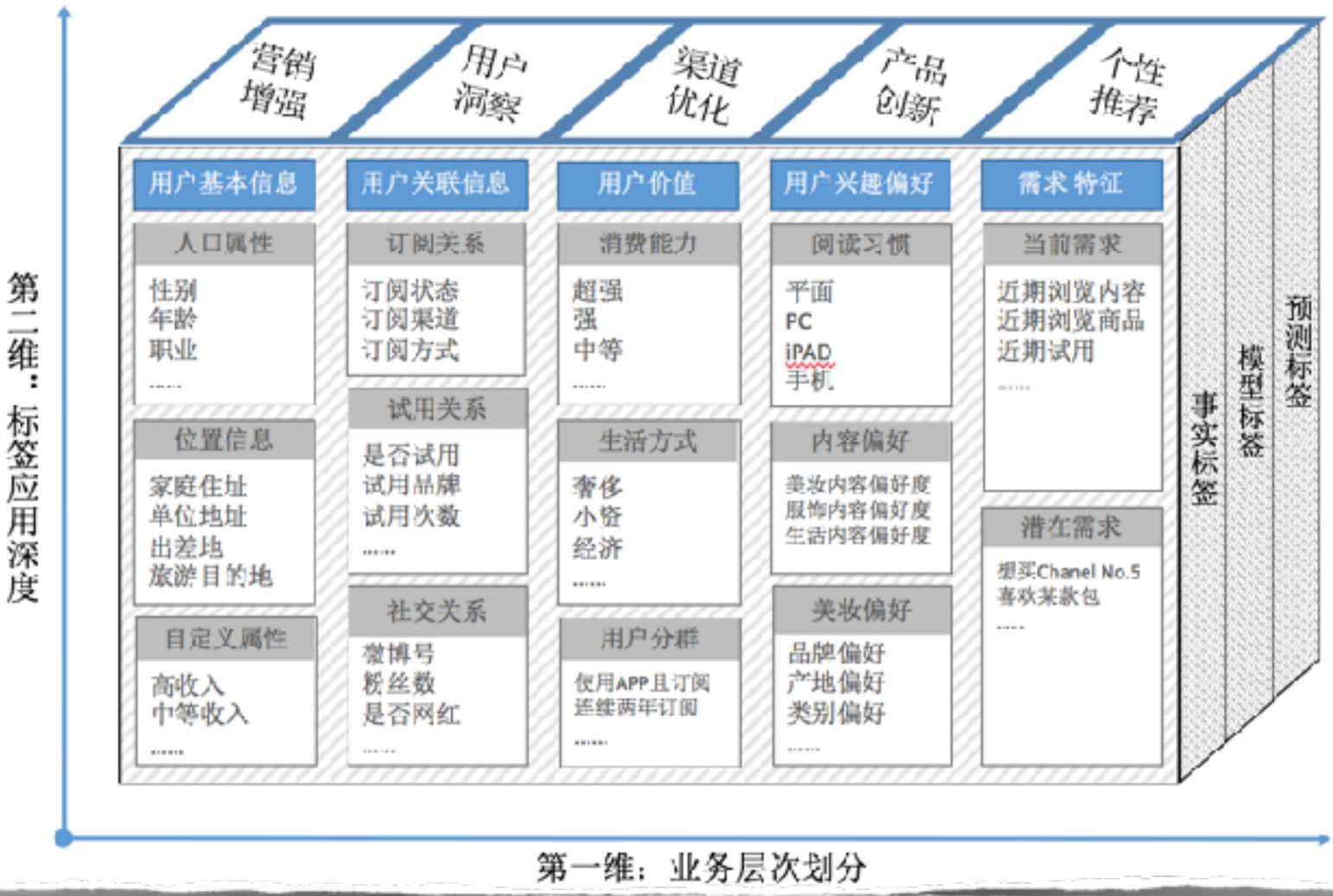
AI是什么?

人工智障

学习误区！

- 学习各种算法的实现原理：
  - 数学不行，看不懂本质。
  - 看懂了，也没有深度。
- 我们是应用技术，而不是科研。

用户画像体系



用户画像

做这个对公司有什么用?

用户画像 -> “用户特征向量”

[性别，收入，年龄，活跃，偏好电脑，偏好运动....]

->

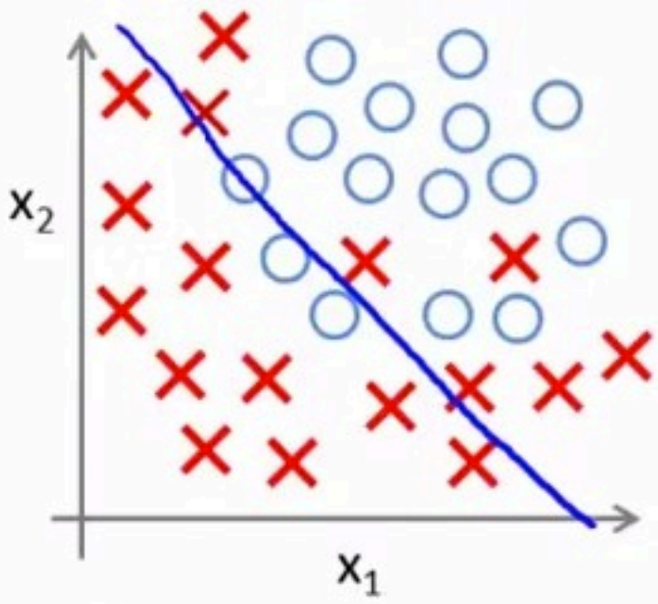
[0, 4000, 23, 3, 1, 0, ...]

特征黑魔法

- 性别（离散）
  - 男、女 == [1], [0]
    - 是男、是女 == [1, 0], [0, 1]
- 收入（连续）
  - [1000], [1000000]
  - 屌丝、土豪 == [0], [1]

结论：
好特征让算法更容易发现规律

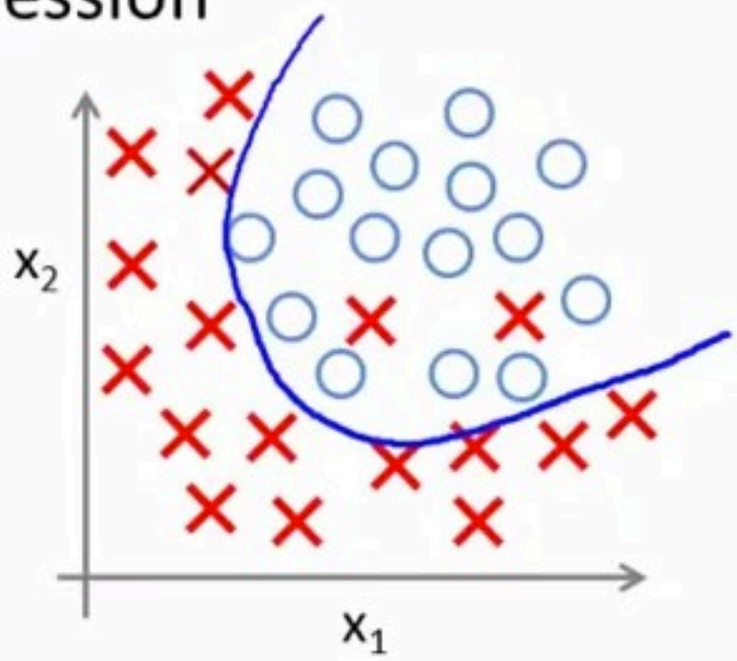
Example: Logistic regression



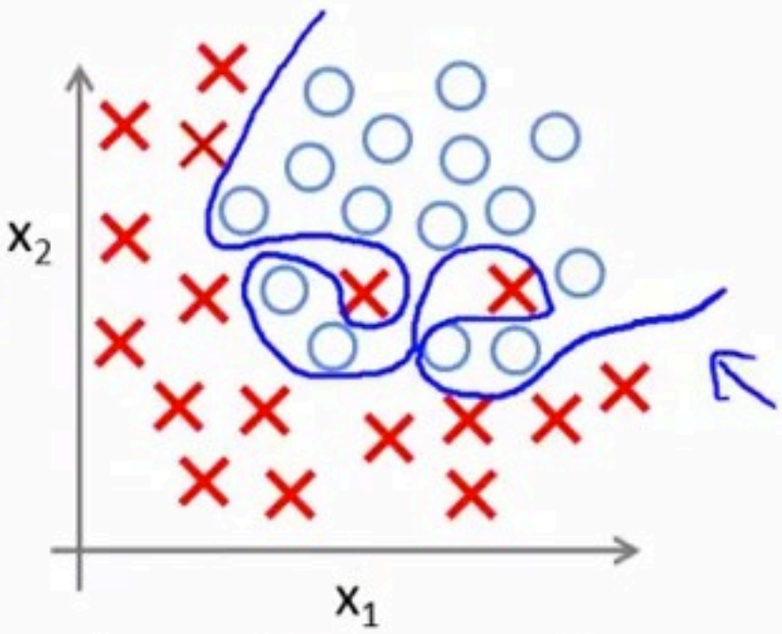
$h_\theta(x) = g(\theta_0 + \theta_1 x_1 + \theta_2 x_2)$

(g = sigmoid function)

"Underfit"



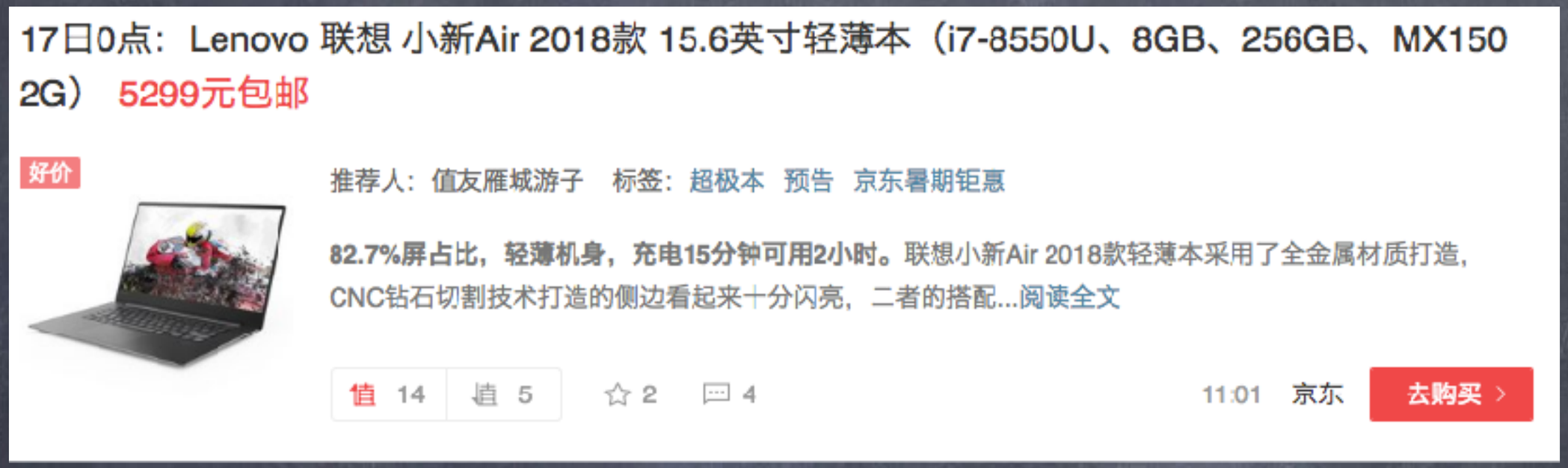
$g(\theta_0 + \theta_1 x_1 + \theta_2 x_2 + \theta_3 x_1^2 + \theta_4 x_2^2 + \theta_5 x_1 x_2)$



$g(\theta_0 + \theta_1 x_1 + \theta_2 x_1^2 + \theta_3 x_1^2 x_2 + \theta_4 x_1^2 x_2^2 + \theta_5 x_1^2 x_2^3 + \theta_6 x_1^3 x_2 + \dots)$

"Overfit"

商品有哪些属性?



- 品牌、分类、商城
- 值/不值、收藏、评论
- 价格
- ...

练习：
商品特征向量

[品牌ID，商城ID，1级分类ID，2级分类ID，价格，访问次数，点值率，收藏数，评论数]

更好的特征向量

- 离散特征one-hot:
  - [是京东吗?，是成人用品吗?]
- 连续特征:
  - 归一化: [log(收藏数)]
  - 分段: [0], [1], [2]分别表示价格<1000, <10000, >= 10000

思考:
smzdm推荐的2个关键要素

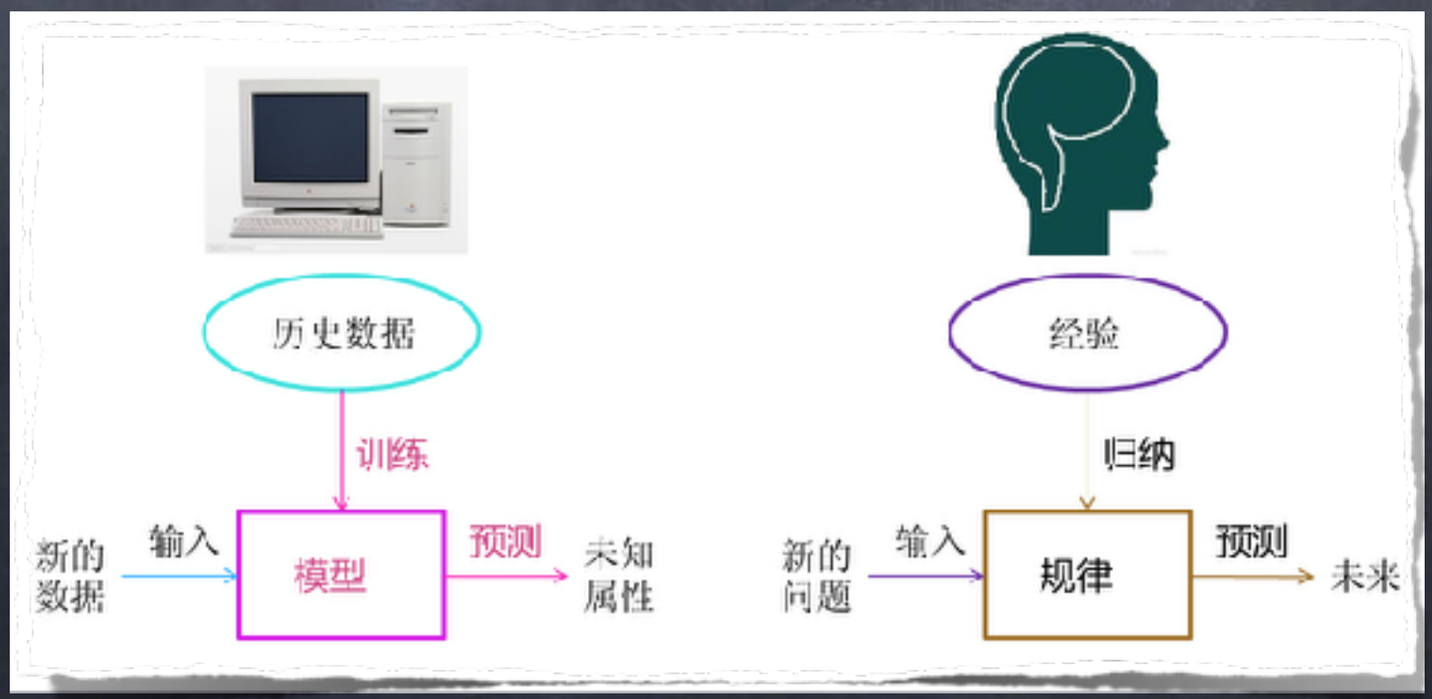
- 用户
  - 用户特征向量
- 文章
  - 文章特征向量

准备训练数据

- 文章曝光/点击/不感兴趣 等行为日志：
  - 哪个用户，哪篇文章，什么行为
- 离线处理行为日志，得到训练数据：
  - [用户特征向量，文章特征向量] -> 是否感兴趣(0/1)

训练模型

- 选择LR、GBDT、XGBOOST等机器学习算法完成模型训练
- 生成的模型保存到文件中，供线上推荐服务预测使用



首页 - 个性化推荐

- 粗排+召回：取最近24小时的文章，运营位的文章，用户感兴趣的分类...
- 精排：对召回的1000篇文章，遍历完成模型打分(感兴趣的可能性)，并结合业务策略决定最终排序.
- 缓存：用户刷新时完成1000篇排序计算，文章ID缓存到redis中，翻页请求取redis.（我是有底线的）

体验 - 机器学习

- sklearn: python，机器学习算法单机实现
- spark：大数据挖掘、分布式机器学习
- tensorflow：机器学习分布式框架、深度学习
- PMML：预测模型标记语言，跨语言兼容

sklearn, spark, pmml等用法demo~

https://github.com/owenliang/machine-learning

thanks