大数据后的用户画像

主讲人：周涛

有米科技

CATALOG 目录

有米科技

What

Why

How

What

用户画像

01 What

有米科技

- Alan Cooper （交互设计之父）最早提出了 persona 的概念:“Personas are a concrete representation of target users.”

- Persona 是真实用户的虚拟代表,是 建立在一系列真实数据(Marketing data,Usability data)而抽象出的一个标签化的用户模型。

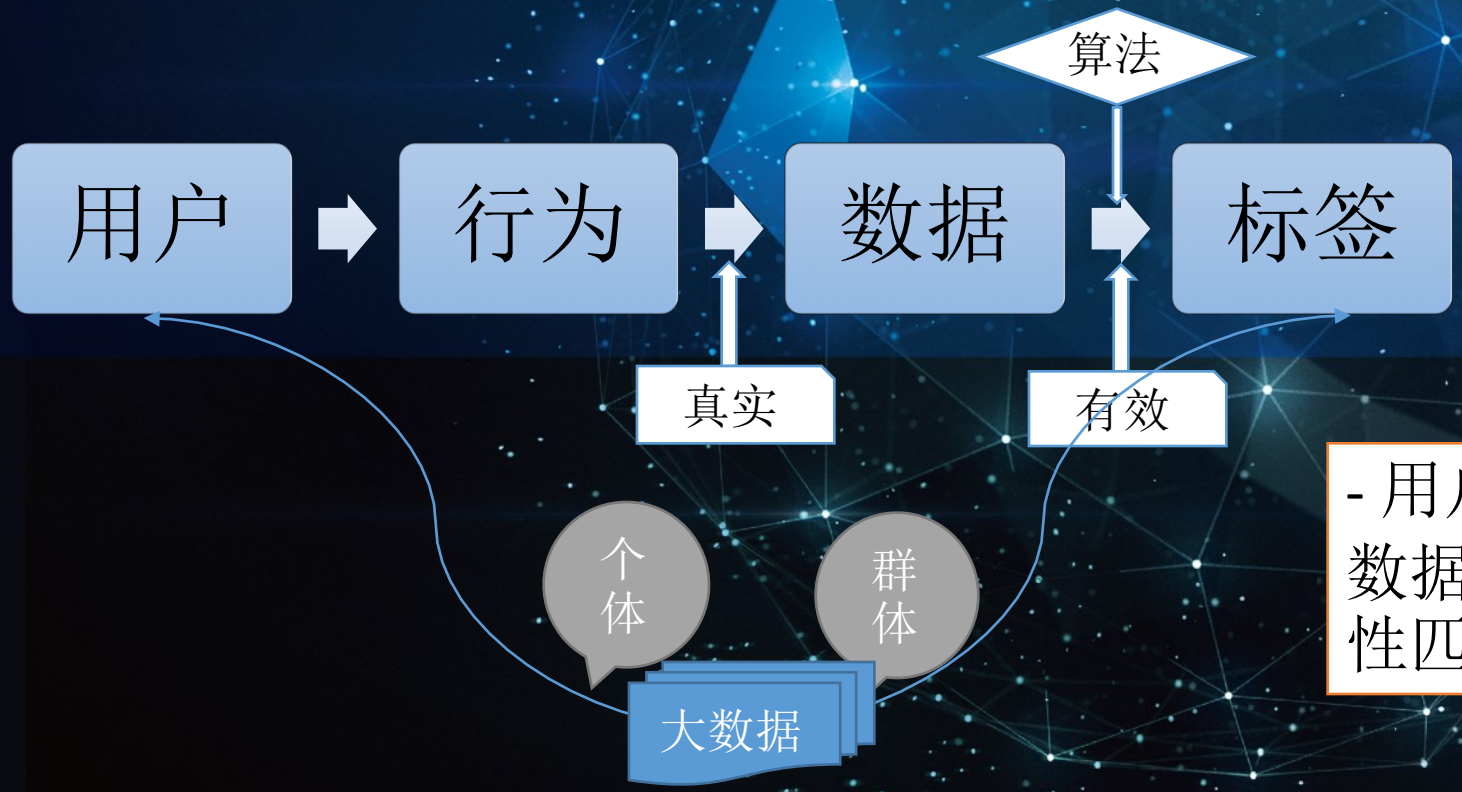
- 构建用户画像的核心工作即是给用户打“标签”，而标签是通过对用户信息分析挖掘而来的高度精炼的特征标识。

Why

画像浅析

02 Why

有米科技



- 用户的属性和行为的可逆性越强，数据的特征越明显，标签和用户属性匹配的准确率越高。

How

性别模型

03 How

有米科技

性别模型：基于安装包列表推测用户的性别

1. 数据分析

- 行为 → 安装列表：验证数据的真实性 ← 先验知识
- 安装列表 → 标签：验证数据的有效性 ← 后验&先验

2. 特征选取

2.1 one-hot特征：0-1向量

- 若有APP集合中有5个，分别为0, 1, 2, 3, 4
  - 用户A安装了0,1,2，其one-hot特征为[1, 1, 1, 0, 0]
  - 用户B安装了0,4,5，其one-hot特征为[1, 0, 0, 1, 1]

\+ 编码简洁，基准特征

\- 维度过高

2. 特征选取

2.2 类别特征：app映射成类别

- 每个APP映射成类别集合中的一个或多个类型

+ 加入先验知识，高度归纳package属性；降低维度
- 严重依赖先验知识，类别的选取较依赖标签

有米科技

2. 特征选取

2.3 几率特征

*

$$odd = \frac{p(gender = male|package = i)}{p(gender = female|package = i)}$$

+ 后验；降低维度
- support低的package计算较为敏感

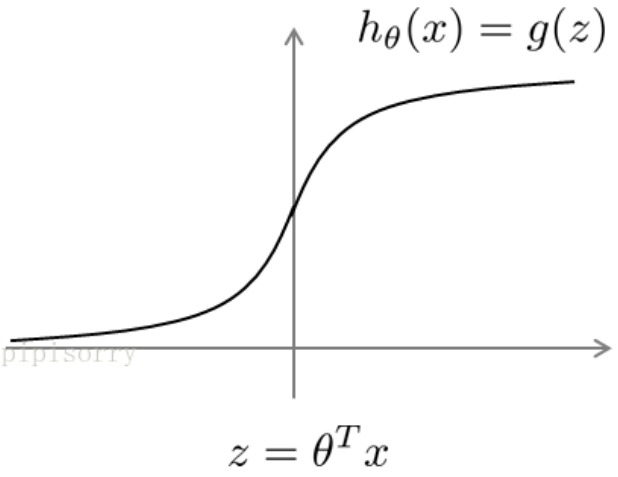
有米科技

3. 模型构建

3.1 Logistic Regression

*

$$h_\theta(x) = \frac{1}{1 + e^{-\theta^T x}}$$



http://blog.csdn.net/pipisorry

If $y = 1$, we want $h_\theta(x) \approx 1$, $\theta^T x \gg 0$

If $y = 0$, we want $h_\theta(x) \approx 0$, $\theta^T x \ll 0$

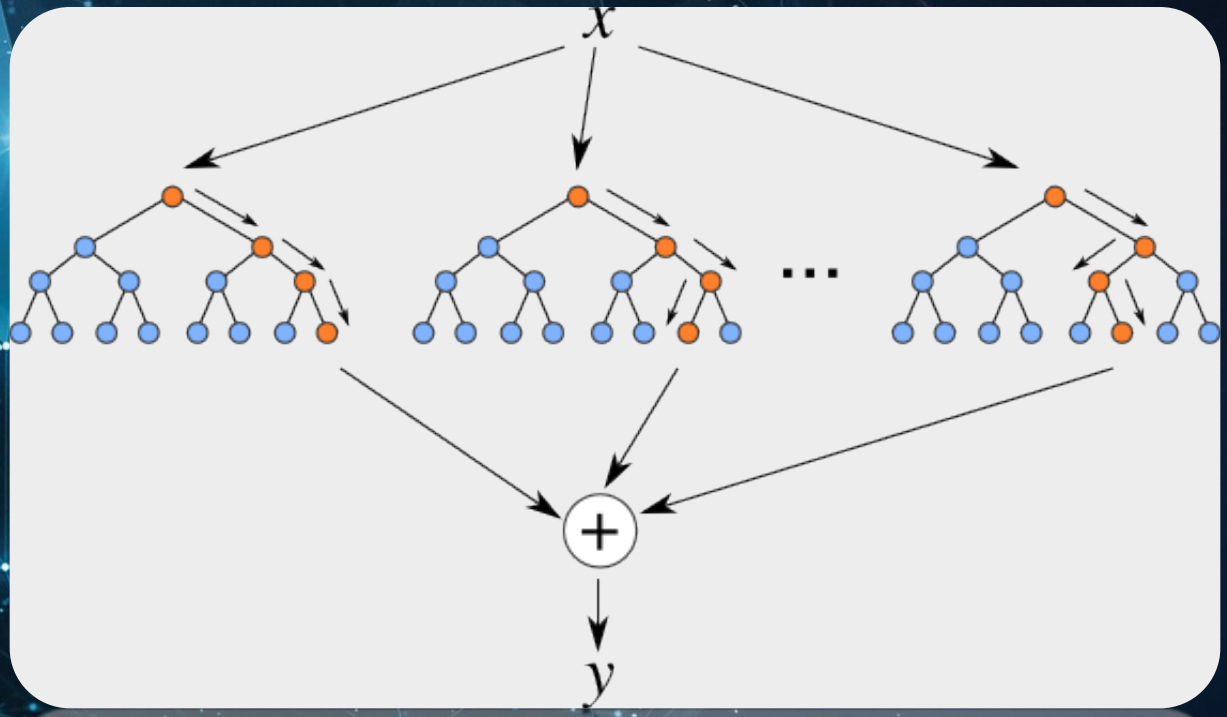
\+ 模型简单，计算量小，可解释性强；可作为基准模型

\- 容易欠拟合，准确率低

3. 模型构建

3.2 Random Forest

*



+ 训练速度较快；准确率高；特征排序
- 在噪声较大的数据中过拟合

有米科技

3. 模型构建

3.3 GBDT

*

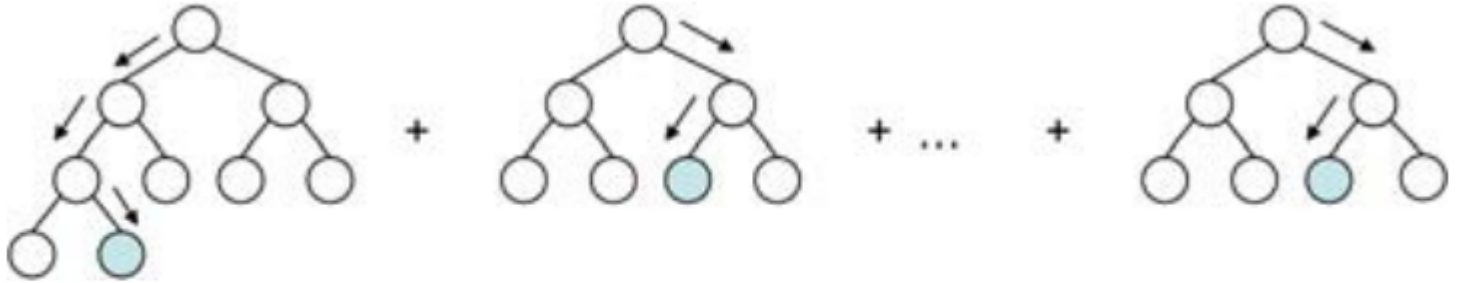


Figure 1: **A gradient boosted decision tree ensemble.**

+ 准确率高，适合低维稠密数据
- 训练速度慢

4. 模型评估

- 测试集准确率
- DSP用户定向

有米数据

有米科技

- DMP数据覆盖约4亿设备
- 标签覆盖上千属性
- 高准确率的标签算法

基础数据能力	数据标签能力	数据营销能力
ID Mapping	标签查询	种子用户放大
设备信息查询	标签计算	垂直领域分析

THANKS

有米科技 | 全球领先的综合性移动互联网企业