第四课时：算法工具使用

TIANCHI天池

主讲人：李强

- 就读学校：中科院计算所
- 研究领域：社会网络，计算复杂性，数据挖掘
- 实习经历：MSRA Theory Group
- 比赛经历：阿里移动推荐算法 季军

  新浪微博互动预测 亚军

提纲

1. 数据到样本（特征，标志）
2. 牛刀小试 — 逻辑回归
3. 进 阶 版 — 归一化、样本均衡、模型融合
4. 一些小tips

数据到样本（特征，标志）

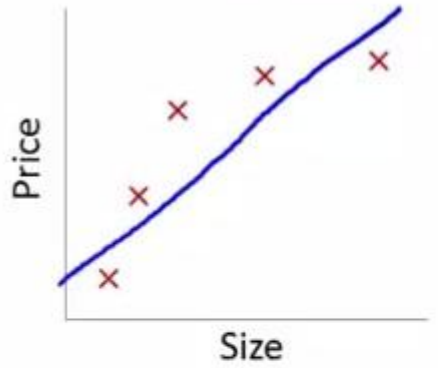
- 确定样本
  - 问题建模：二分类问题，UI对是否被购买
  - 样本选择：10天内有过交互的UI对
- 从数据到特征
  - 针对每个(user_id, item_id)统计一些属性
  - 基本特征：浏览、收藏、购物车、购买量
  - 基于规则：头天是否加入购物车没买
- 从数据到label
  - 为样本添加标志

牛刀小试 — 逻辑回归

进阶版 — 归一化、样本均衡、模型融合

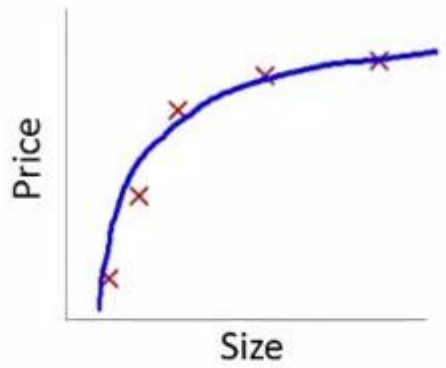
- 样本归一化
- 正负样本均衡
- 多个模型结果融合

Example: Linear regression (housing prices)



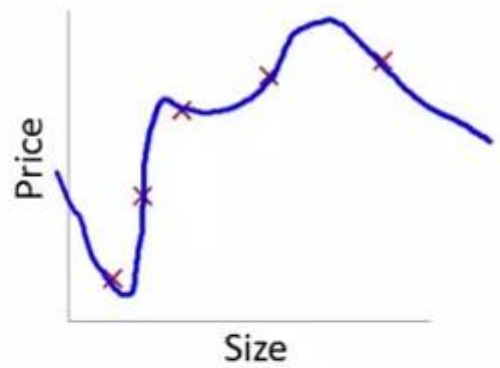
$\rightarrow \theta_0 + \theta_1 x$

"Underfit" "High bias"



$\rightarrow \theta_0 + \theta_1 x + \theta_2 x^2$

"Just right"



$\rightarrow \theta_0 + \theta_1 x + \theta_2 x^2 + \theta_3 x^3 + \theta_4 x^4$

"Overfit" "High variance"

Overfitting: If we have too many features, the learned hypothesis may fit the training set very well ($J(\theta) = \frac{1}{2m} \sum_{i=1}^{m} (h_\theta(x^{(i)}) - y^{(i)})^2 \approx 0$), but fail to generalize to new examples (predict prices on new examples).