广告数据上的大规模机器学习

@夏粉_百度

目录

- 背景
- 问题
- 技术
- 小结

计算广告学

计算广告学与CTR预估

- 计算广告学的核心问题:
  - 给定的环境下，用户与广告的最佳匹配

- 流量变现:

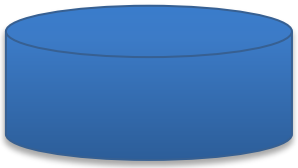
$$profit = PV * CTR * ACP$$

- 方法: 依赖机器学习和历史数据，做精准CTR预估

CTR预估问题

广告库

$D = \{ad_1, ad_2, \cdots, ad_l\}$



广告请求

$q^{(i)}$

CTR 预估系统

广告的预估CTR

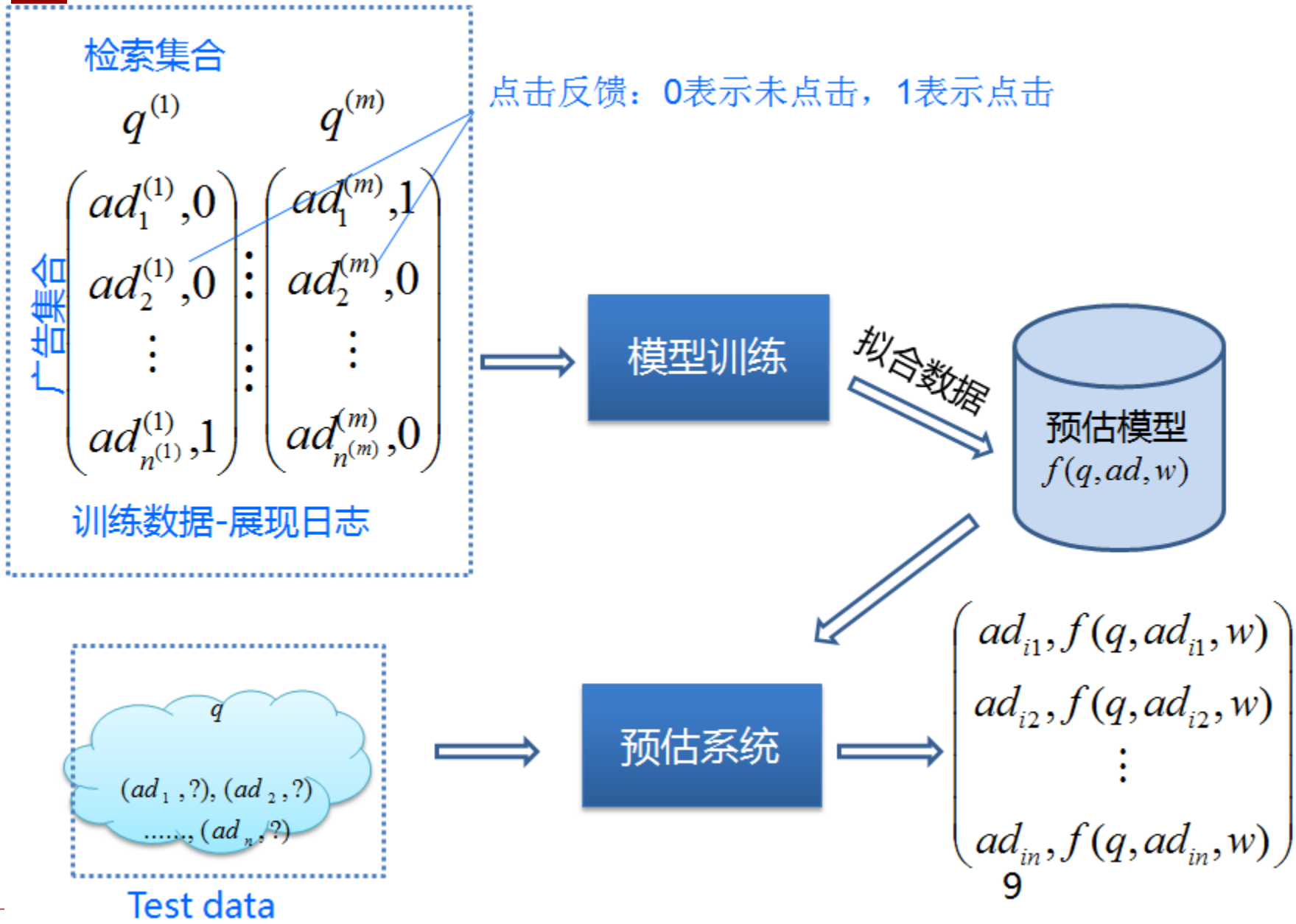
$Pctr\_ad_1^{(i)}$

$Pctr\_ad_2^{(i)}$

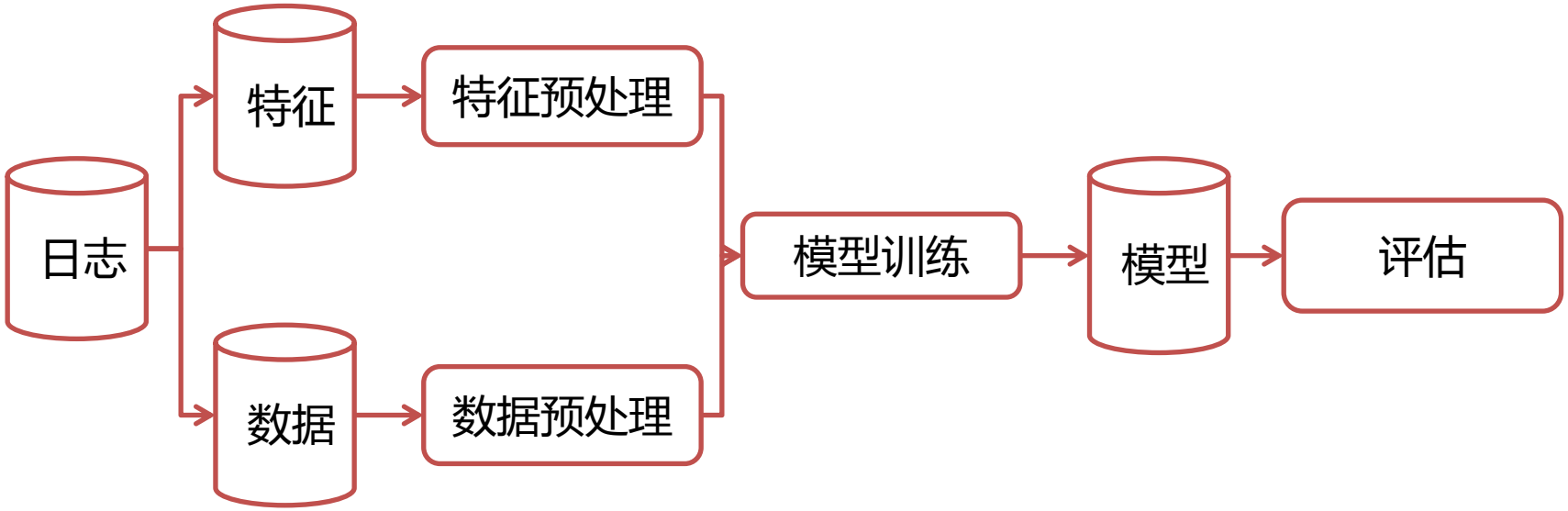
$\vdots$

$Pctr\_ad_{n_i}^{(i)}$

点击率预估-机器学习模型



数据处理流程



大规模机器学习问题

大规模机器学习问题

数据特征规模大	•每天百亿广告展现，十亿特征 •类别不平衡、噪音大
特征复杂度高	•特征之间存在高度非线性关系
数据时效性高	• 点击率随时间变动， e.g., 兴趣变化 • 新广告和流量上线，旧广告和流量下线
数据训练频繁	• 模型更新 • 策略调研

大规模机器学习技术

数据处理技术

- 解决方法：
  - 不可见和不完整样本过滤
  - 样本采样
  - 异常样本检测

数据采样:

- Google:
  - 采样：
    - Any query for which at least one of the ads was clicked.
    - A fraction $r \in (0, 1]$ of the queries where none of the ads were clicked.
  - 矫正：

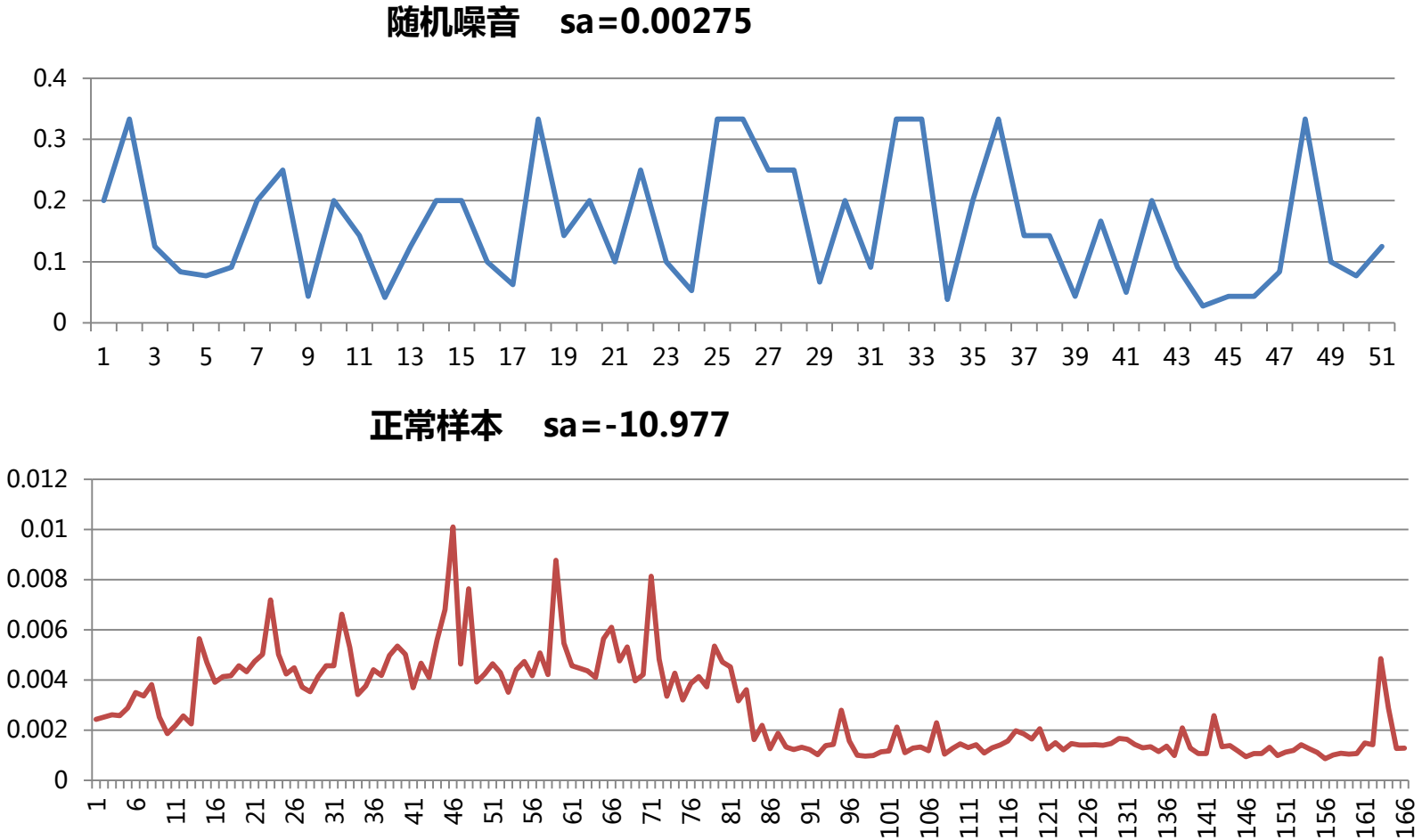
$$\omega_t = \begin{cases} 1 & \text{event } t \text{ is in a clicked query} \\ \frac{1}{r} & \text{event } t \text{ is in a query with no clicks.} \end{cases}$$

  - 原理： 采样后的期望损失等于原损失

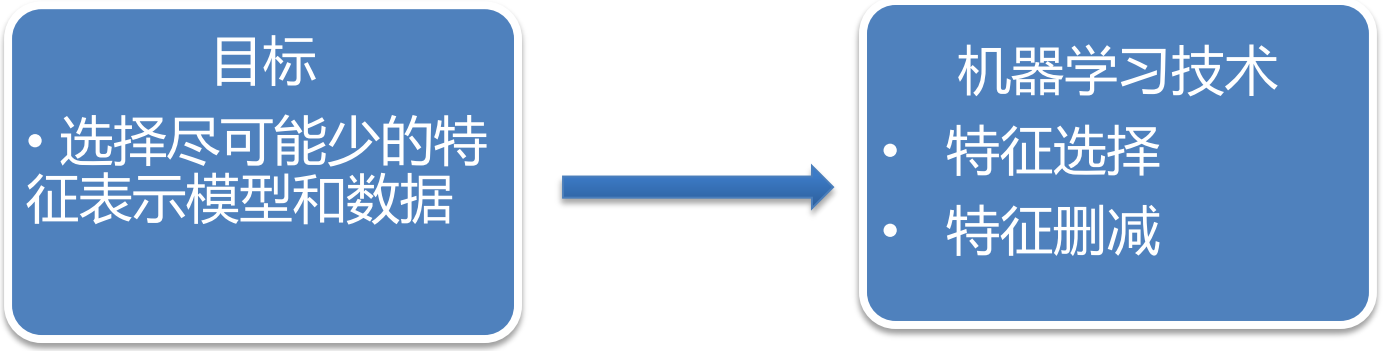
$$\mathbb{E}[\ell_t(\mathbf{w}_t)] = s_t \omega_t \ell_t(\mathbf{w}_t) + (1 - s_t)0 = s_t \frac{1}{s_t} \ell_t(\mathbf{w}_t) = \ell_t(\mathbf{w}_t).$$

噪音检测

- 计算点击率随时间变化趋势 –百度首创：SA算法

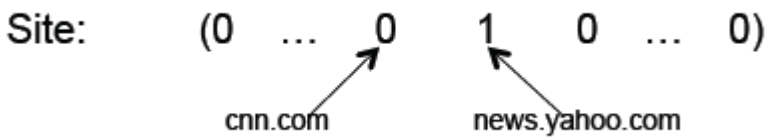


特征处理技术



- 特征类型：
  - ✓ 类别型特征(categorical features)、连续值特征
- 特征表示：
  - ✓ 使用one-hot编码类别型特征

特征海量！

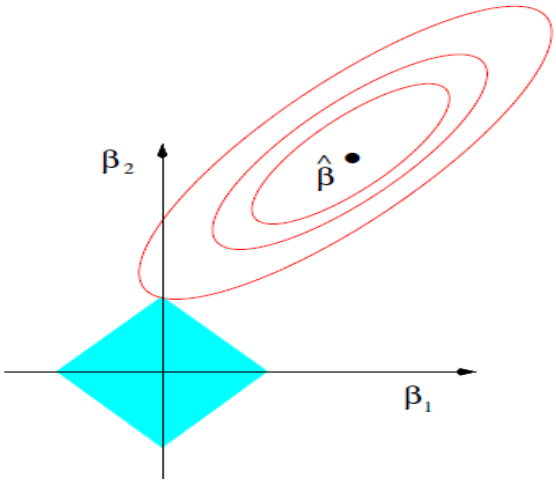


特征选择

$$\min_{\mathbf{w}} \quad f(\mathbf{w}) \equiv \|\mathbf{w}\|_1 + C\sum_{i=1}^{l}\xi(\mathbf{w};\mathbf{x}_i,y_i)$$

Regularization term，特征选择 ，降低模型复杂度

数据拟合项：拟合训练数据，使得预估CTR尽可能靠近经验CTR。



其中损失取似然损失

$$\xi_{\log}(\mathbf{w};\mathbf{x},y) = \log(1+e^{-y\mathbf{w}^T\mathbf{x}})$$

特征删减

- 背景：
  - ✓ 模型大小占特征大小比例极低
- 技术挑战：
  - ✓ 训练前，判断哪些特征权值为0

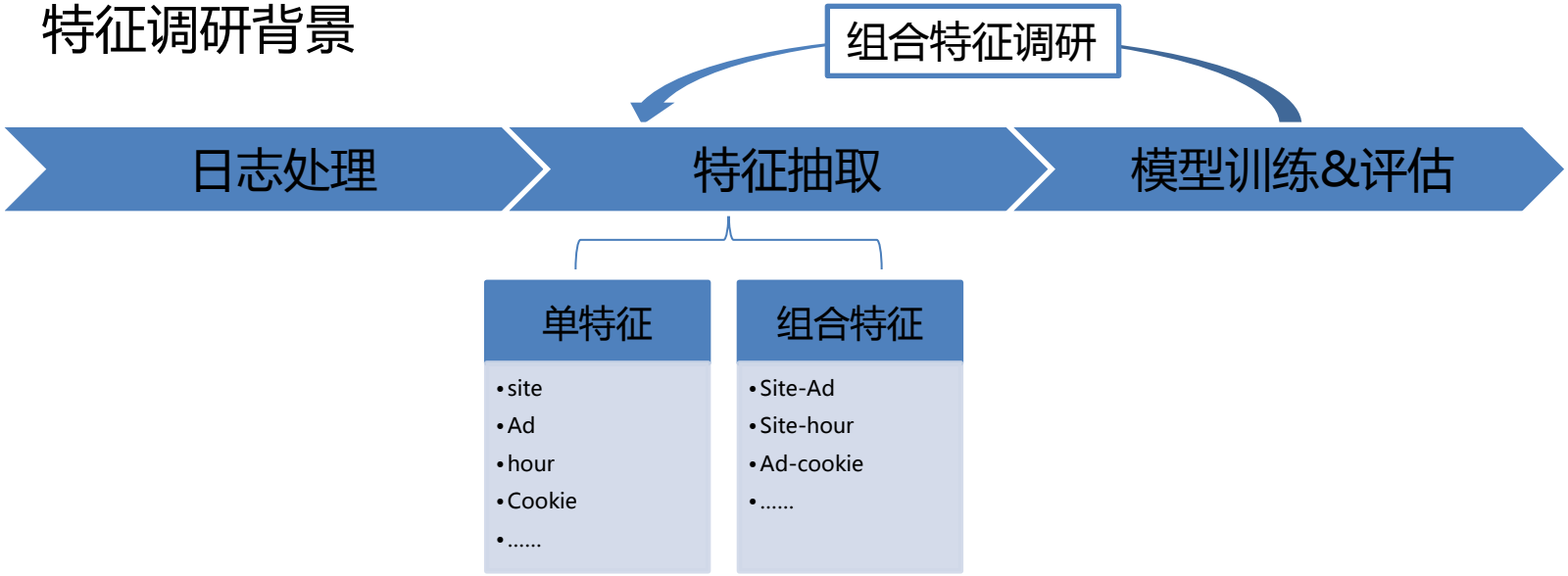
谷歌：
新特征按概率p加入
Bloom Filter+次数超过n

百度首创：
Fea-G算法: 理论保证效果无损

方法	内存节省	AucLoss升高
Bloom(n=2)	66%	0.008%
Bloom(n=1)	55%	0.003%
Poisson(p=0.003)	60%	0.020%
Poisson(p=0.1)	40%	0.006%
Fea-G	97%	0%

深度特征学习技术

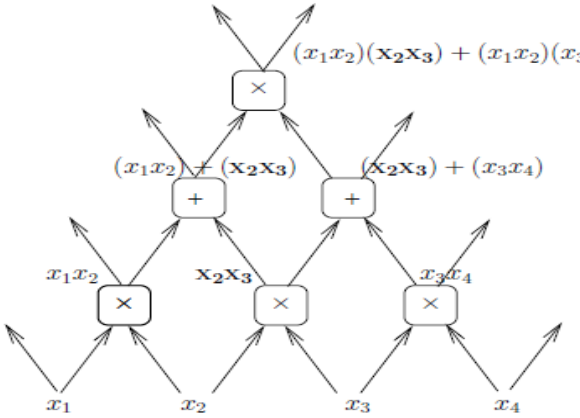
- 特征调研背景



- 构造高阶组合特征，描述特征之间非线性关系
- 人工挖掘，耗时！耗力！依赖先验，无推广性！
  - ✓ 假设有N个单特征类，组合特征候选类： $2^N \approx C_N^1 + C_N^2 + ... + C_N^{N-1} + C_N^N$
  - ✓ 选最优特征类，需要时间： 2^N

深度特征学习算法

- 特征学习
  - ✓ 深度学习在语音、图像上取得突破性进展
  - ✓ 广告数据特征维数非常高（单特征百亿），尚无大规模稀疏特征学习算法
- DANOVA: 首个直接应用于大规模稀疏特征的深度特征学习算法



高阶组合：Site-Ad-Cookie，...

二阶组合：Site-Ad, Site-Hour, Ad-Cookie, ...

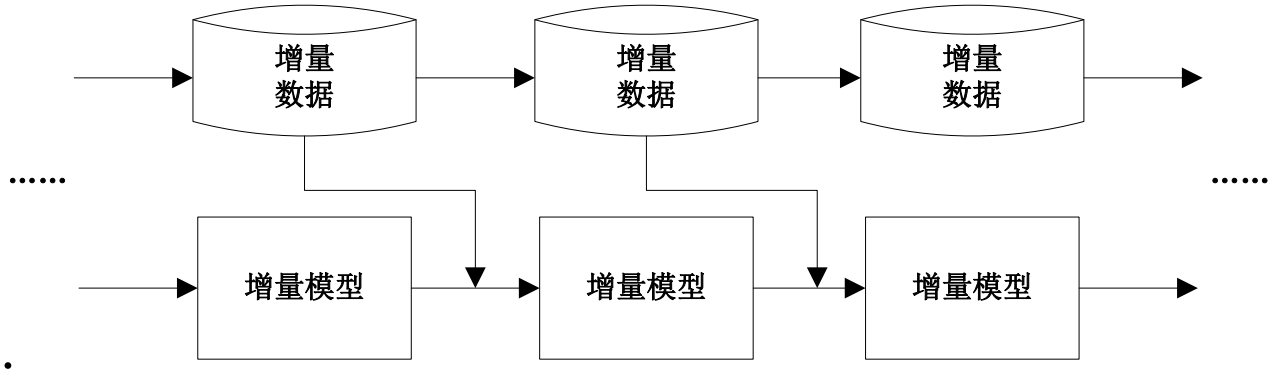
单特征：Site, Ad, Hour, Cookie, ...

逐层贪婪学习

- 上线效果
  - ✓ 特征挖掘效率提升上千倍
  - ✓ CTR，CPM显著增长

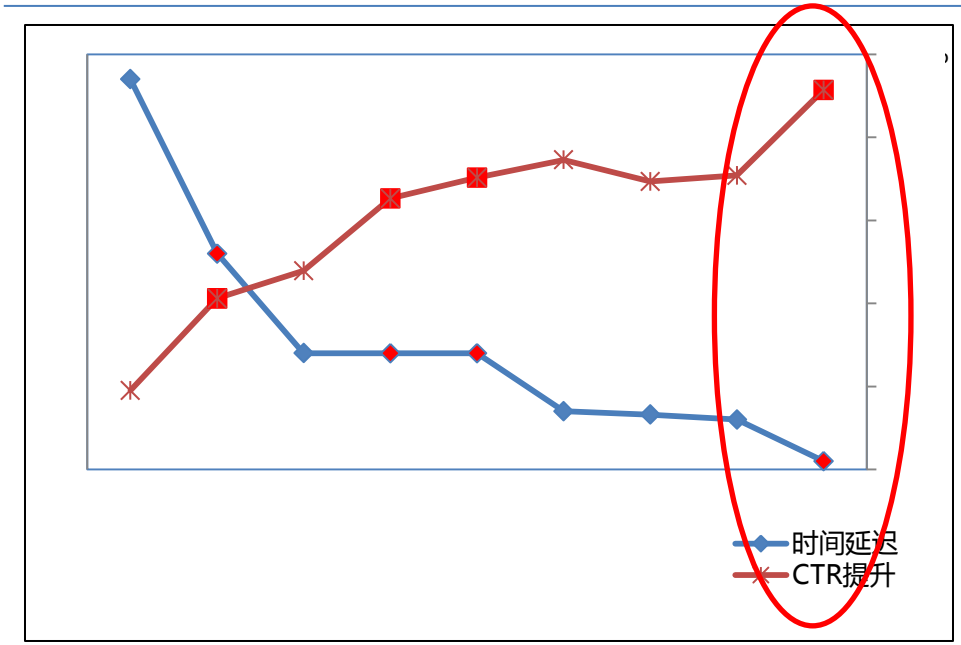
模型时效性

- 背景：
  - ✓ 模型更新时，训练数据尽可能少
- 技术挑战：
  - ✓ 稀疏性、时效性、稳定性
- 方法：稀疏在线算法



- 现状：
  - ✓ 大部分在线算法非稀疏
  - ✓ Google保留前N次模型梯度方法，不够稳

增量效果汇总



- 时效性从20-30小时降到分钟
- Ctr累积大幅提升

- 在线学习
  - ✓ 时效性为分钟
  - ✓ Ctr显著提升
  - ✓ 资源节省50%

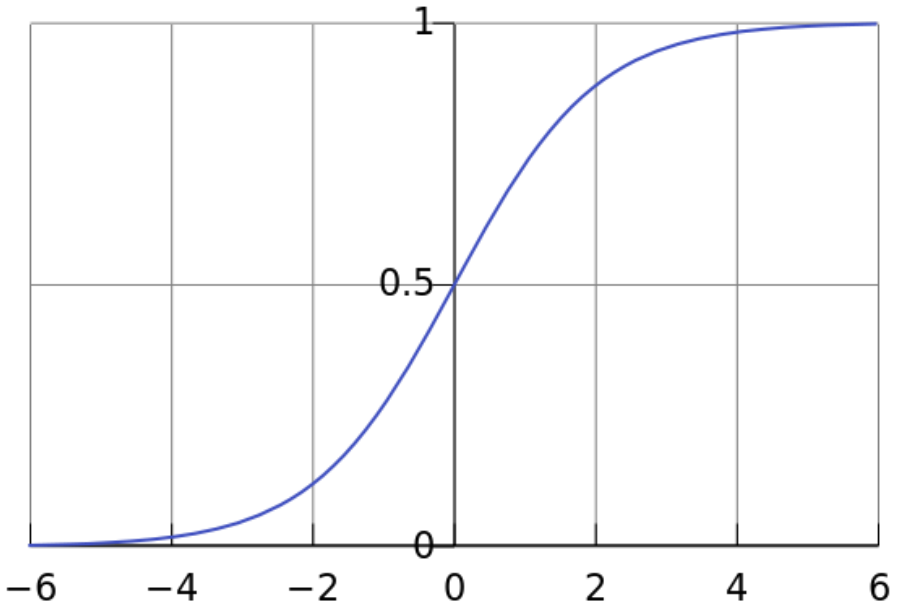
- 技术创新点：
  - ✓ 训练算法：首创SOA算法，使模型稳定性更好
  - ✓ 训练架构：批处理改为在线，节省资源80%以上
  - ✓ 在线学习平台：在大数据上实现分钟级别的在线学习

模型训练

- 线性逻辑回归模型

$$P(Y=1|X)=\frac{1}{1+e^{-(\beta_0+\sum\beta_i*x_i)}}$$

$$\ln(\frac{p_1}{p_0})=\beta_0+\sum\beta_i*x_i$$

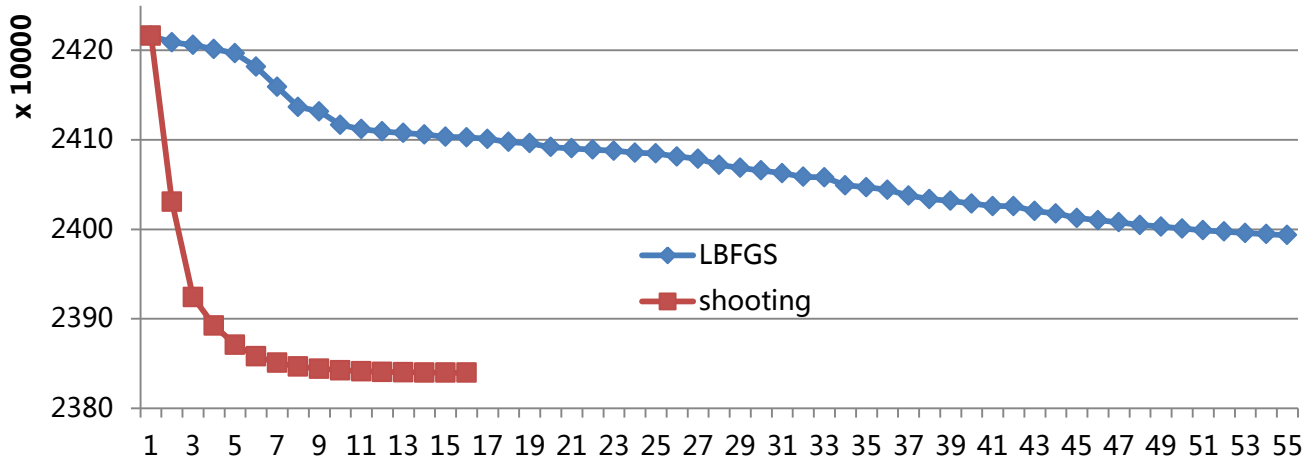


- 目标函数

$$\arg\min_w L(w)=\frac{1}{n}\sum_{i=1}^{n}\log(1+e^{-y_i x_i^T w})+c\|w\|_1$$

训练算法优化

- 背景：
  - ✓ 寻找更好优化方向，减少迭代轮数
- 技术方案：
  - ✓ 算法创新：Shooting算法，更准的方向
  - ✓ 性能变化：相比于LBFGS训练轮数从平均50轮下降到5轮，训练更充分



单元小结

小结

- 以CTR预估为例，大数据学习技术应用计算广告学，尽可能少的资源尽可能多的提升CTR准确率

- 大数据学习技术：
  - ✓ 数据和特征过滤算法，容纳百亿数据特征
  - ✓ 深度特征学习算法，学习效率提升千倍
  - ✓ 稀疏在线算法，模型分钟更新
  - ✓ 模型训练算法，速度提升十倍

与Google Seti对比

- **网盟CTR预估模型：**
  - ✓ 数据和特征过滤算法，容纳百亿数据特征
  - ✓ 深度特征学习算法，学习效率提升千倍
  - ✓ 稀疏在线算法，模型分钟更新
  - ✓ 模型训练算法，速度提升十倍

- **Google Seti:** (4/06/2010 08:00:00 AM Posted by Simon Tong, Google Research)
  - ✓ Binary classification (produces a probability estimate of the class label)
  - ✓ Parallelized
  - ✓ Scales to process hundreds of billions of instances and beyond
  - ✓ Scales to billions of features and beyond
  - ✓ Automatically identifies useful combinations of features
  - ✓ Accuracy is competitive with state-of-the-art classifiers
  - ✓ Reacts to new data within minutes

Thanks

百度技术沙龙

关注我们：t.baidu-tech.com

资料下载和详细介绍：infoq.com/cn/zones/baidu-salon

“畅想•交流•争鸣•聚会”是百度技术沙龙的宗旨。 百度技术沙龙是由百度与InfoQ中文站定期组织的线下技术交流活动。目的是让中高端技术人员有一个相对自由的思想交流和交友沟通的的平台。主要分讲师分享和OpenSpace两个关键环节，每期只关注一个焦点话题。

讲师分享和现场Q&A让大家了解百度和其他知名网站技术支持的先进实践经验，OpenSpace环节是百度技术沙龙主题的升华和展开，提供一个自由交流的平台。针对当期主题，参与者人人都可以发起话题，展开讨论。

InfoQ 策划·组织·实施

关注我们：weibo.com/infoqchina