大规模机器学习

夏粉

2016年7月25日

Big Data Lab, Baidu Research

- Big Data Lab
  - Led by ***Tong Zhang***
  - Focusing on **Machine Learning & Data Mining** for
    - **Ads/Search** Engine, **Recommendation** System, **NLP,**
    - **Spatial Temporal** Analysis,
    - **Heath Care AI** (e.g. DuNurse),
    - **Financial AI**, **Sales AI**, **Knowledge AI**, **Security AI ……**
  - Research on Large-scale Algorithms, Architecture and Applications.
  - Locations:
    - Beijing, Silicon Valley, Shenzhen

Outline

- 广告背景
- 大规模机器学习
- 实战技术
- 总结展望

搜索广告: Search Ads

新闻 网页 贴吧 知道 音乐 图片 视频 地图 文库 更多»

北京美食 百度一下

推荐：用手机随时随地上百度

北京美食 首选国内领先的吃喝玩乐信息平台-易吃易乐 bj.echiele.com 推广链接
北京美食首选国内领先的吃喝玩乐信息平台-易吃易乐,每天有上百万网
易吃易乐 餐饮美食 休闲娱乐 美容美发

找北京美食?来DaoDao.com www.daodao.com
找北京美食?DaoDao.com为您提供210000条北京市旅游点评/攻略.

北京美食-大众点评网
根据合理的商区、地标和美食商户分类系统,为你提供北京83892家美食商户,并通过海量亲身消费者的点评聚合,以各种评分、星级的标准让你选择。
www.dianping.com/beijing/f... 2013-7-5 - 百度快照

北京美食攻略_北京美食推荐_美食街,小吃,指南-驴妈妈旅游网
驴妈妈旅游网关于北京美食攻略,包含更多北京特色美食小吃(美食,餐饮,娱乐),【旅游预订】打折门票,周边酒店,自由行及跟团游信息,就在(www.lvmama.com)
www.lvmama.com/travel/zhongguo_beiji... 2013-6-29 - 百度快照

北京有什么特色美食?_百度知道
13个回答 - 提问时间: 2011年12月25日
最佳答案: 1.烤鸭:在北京您要是想吃到便宜实惠的烤鸭,您可以去便宜坊、大鸭梨、安贞烤鸭店。 当然您要是想吃最地道的烤鸭那就去和平门的全聚德。 2.涮羊肉:地...
zhidao.baidu.com/question/3585625... 2013-1-27 - 百度快照

北京美食_百度百科
北京美食guide是一款让你随时随地掌握北京美食信息的手机软件。北京美食拥有详尽的地图，十多种美食分类。
基本信息 - 软件介绍 - 安装指南 - 软件截图
baike.baidu.com/ 2013-07-03

在北京市搜索北京美食_百度地图
A. 辣尚瘾(人大店) - (010)82650566
1229条评论

北京美食

外文名: Guide
版本: V1.8.0
软件大小: 3941KB
来自百度百科>>

相关食物

北京小吃 北京烤鸭 门钉肉饼 褡裢火烧 护国寺小吃

推广链接

北京美食餐厅预定有折扣
上咕嘟妈咪,方便轻松享优惠北京美食;
咕嘟妈咪,不让亲朋排队等.
www.gudumami.cn

北京美食,金鼎鱼香渔村欢迎..
金鼎鱼香生态渔村,特色全鱼宴,灶台柴锅水库鱼,柴锅柴鸡,特色烧鸽
www.myjdyx.com

北京美食 刷雅酷卡 乐享无限..
找北京美食,精选北京美食折扣优惠!吃喝玩乐尽在雅酷卡网!
www.vacol.com

展示广告：Display Ads

1 **玻璃家具的挑选。**

玻璃家居，一定要注意玻璃的厚度，可以的情况下，最好选择是钢化玻璃的，可以安全很多，坚硬很多。因为钢化后，耐热，即使是破碎后，碎片也不易伤人。

家中有孩子有老人的特别要注意。

2 **塑料家具的选择。**

塑料家居，一般为椅子，桌子，层架等小物品。塑料制品，需要对材料进行分析，最好有说明材料的等级。

注意好塑料的软硬度，承载力，还有塑料制品需要避免风化，暴晒，否则容易缩短寿命。

快快选择自己喜爱的颜色吧。

如何让工作更具效率 50
春季菠萝滋润吃法 25
肠道排毒按摩手法示范 69
刚毕业大学生租房攻略 35
春日简单的盘发教程 73

3月30日北京家居团购博览会!..
中国超大规模一站式家居建材装修采购展家居团购博览会,鼎级家装设计,家居团购一线品..
jctg.com

北京 家居网,装房子,买家具,...
居然之家明码实价直降10%,“以旧换新”补贴5%!购物满10000还赠油卡灯具券,更多精彩..
www.juran.cn

4月12日中国网众大型家具家..
涵盖建材团购,家具团购,家具团购,地板团购,瓷砖团购,木门团购,橱柜团购,卫浴团购,洁..
bj.wzcbd.com

板式家具设备选山东明美==山..
选板式家具设备就到山东明美.板式家具设备一站式销售服务,全程专人负责调试,安排,培.

广告与点击率预估

- 广告的核心问题:
  - 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: 依赖机器学习和大数据，做精准CTR预估

Why "Computational" Advertising?

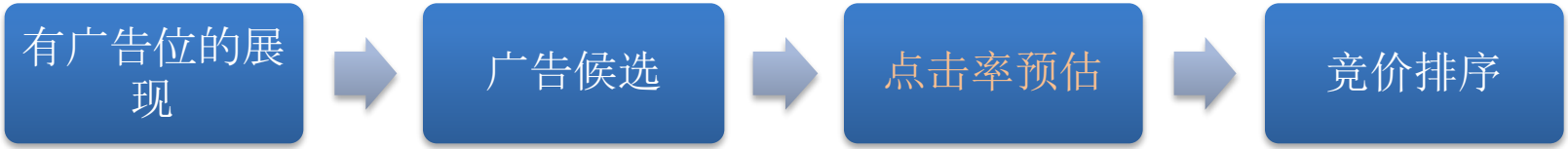
- "Half the money I spend on advertising is wasted; the trouble is I don't know which half"

John Wanamaker, ~ 1875

"Computational Advertising" attempts to solve it in a data-driven and AI way.

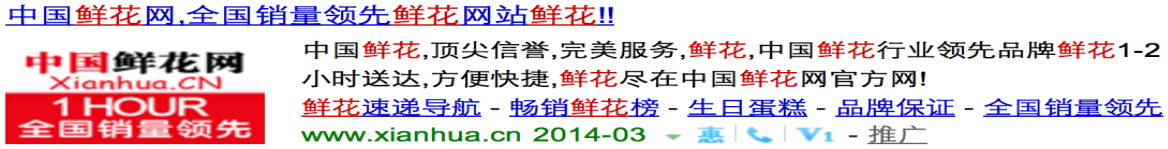
广告系统介绍

- 搜索广告系统:
  - Google AdWords、百度凤巢
- 展现广告系统:
  - Google AdSense、百度网盟
  - RTB (Real Time Bidding)
- 大致流程



点击率（CTR）预估问题

- Query: 鲜花
- 用户　：cookie\历史搜索
- 广告商：鲜花网
- 展现样式：



- 点击率预估：
  - 把广告放在第一位
  - 广告被用户点击的概率

Outline

- 广告背景
- 大规模机器学习
- 实战技术
- 总结展望

大规模机器学习

- 点击率（CTR）预估
- 问题规模：
  - 数据存储和管理：上万台机器
  - 数据量：百亿到千亿级
  - 特征数：百亿到千亿级 （稀疏离散值特征）

- 现状：大规模线性Logistic Regression模型
- 未来：大规模深度学习模型
- 核心技术：大型分布式模型训练

CTR预估的机器学习流程

特征生成

- 把广告展现变成一个向量

概率模型

- 把向量变成点击率

模型训练

- 从历史数据学习模型参数

线上预测

- 把模型用到新的广告展现上

离散特征生成

- 假设　：10000 查询; 1000 用户; 100 广告
- 查询(q)：1 , 2 , … , 10000
- 用户(u)：1 , 2 , … , 1000
- 广告(a)：1 , 2, … , 100
- 原始特征向量: q=1,u=2,ad=3

$$\left[[\underbrace{1, 0, \ldots, 0}_{\text{10000-dim q-vector}}][\underbrace{0, 1, 0, \ldots, 0}_{\text{1000-dim u-vector}}][\underbrace{0, 0, 1, 0, \ldots, 0}_{\text{100-dim ad-vector}}] \right]$$

高阶特征生成

- 1st 阶：3 种单维度特征 q,u,ad

$$\left[[\underbrace{1,0,...,0}_{\text{10000-dim q-vector}}][\underbrace{0,1,0,...,0}_{\text{1000-dim u-vector}}][\underbrace{0,0,1,0,...,0}_{\text{100-dim ad-vector}}] \right]$$

- 2nd 阶： q*u 查询和用户特征组合

$$[\underbrace{0,0,...,0,1,0,...,0}_{10000\times 1000 \text{ dim } q \times u \text{ vecotr}}]$$

离散特征影响

超大维度

- 百亿到千亿量级

稀疏特征

- 每个样本只有百个非零元

特征类型

- 百个特征类型

维数约简

- 离散到离散: Hashing

$$[\ \underbrace{0,0,\ldots,0,1,0,\ldots,0}_{10000\times1000\ \text{dim}\ q\times u\ \text{vecotr}}\ ] \rightarrow \left[[\ \underbrace{0,\ldots,1,\ldots}_{1000\text{dim hash table}}\ ][\ \underbrace{0,\ldots,1,\ldots}_{1000\ \text{dim hash table}}\ ]\right]$$

- 离散到统计：statistics

$$[\ \underbrace{0,0,\ldots,0,1,0,\ldots,0}_{10000\times1000\ \text{dim}\ q\times u\ \text{vecotr}}\ ] \rightarrow \left[\underbrace{0.4}_{\text{historic CTR}} \quad \underbrace{103}_{\text{historic show}} \quad \cdots\right]$$

- 更多先进技术?

模型: Logistic Regression

- 输入向量 x
- 输出点击率 ctr
- 模型参数 w
- 模型假设 ： $\mathrm{ctr} = \dfrac{1}{1 + \exp(-w^\top x)}$

模型训练

- 训练数据 $(x_1, y_1), \ldots, (x_n, y_n)$
  - X: 特征向量
  - Y: {-1,+1} ， -1: 未点检，+1：点检

- 求解如下优化问题

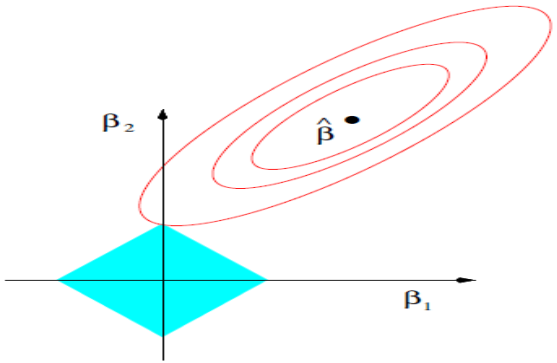
$$\min_{w} \sum_{i=1}^{n} \ln(1 + \exp(-w^\top x_i y_i))$$

正则化

- 稀疏性
  - 减少模型大小

$$\min_{w} \sum_{i=1}^{n} \ln(1+\exp(-w^\top x_i y_i)) + C\|w\|_1$$

$$\|w\|_1 = \sum_{j=1}^{d} |w_j|$$



求解算法

- 迭代算法:

$$w_t \leftarrow w_{t-1} - \eta_t d_t$$

d_t : 梯度方向(梯度或者牛顿方向)

- LBFGS: 使用 1st 阶梯度近似Hessian矩阵
- 坐标梯度下降: 使用单维特征梯度
- 随机梯度下降(SGD): 使用单个样本梯度

分布式计算架构

数据并行

- 每台机器存储所有参数
- 每台数据存储部分数据

模型并行

- 每台机器存储所有数据
- 每台机器存储部分参数

数据&模型并行

- 每台机器存储部分数据
- 每台机器存储部分参数

Outline

- 广告背景
- 大规模机器学习
- 实战技术
- 总结展望

实战技术

（从数据处理到算法优化再到架构）

数据处理	特征处理	组合特征学习	模型算法	时效性
• 样本采样至1% • 异常样本检测	• Fea-G算法：效果无损、内存节省97%	• DANOVA：首个直接应用于大规模稀疏特征、特征挖掘效率提升上千倍	• SOA算法：Online update • Shooting算法：收敛速度提升60%	• 增量架构：从20小时降到分钟级

数据处理技术

目标

- 获取主要信息
- 去除异常噪音

→

机器学习技术点

- 选择对点击概率分布预估足够多样本

- 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

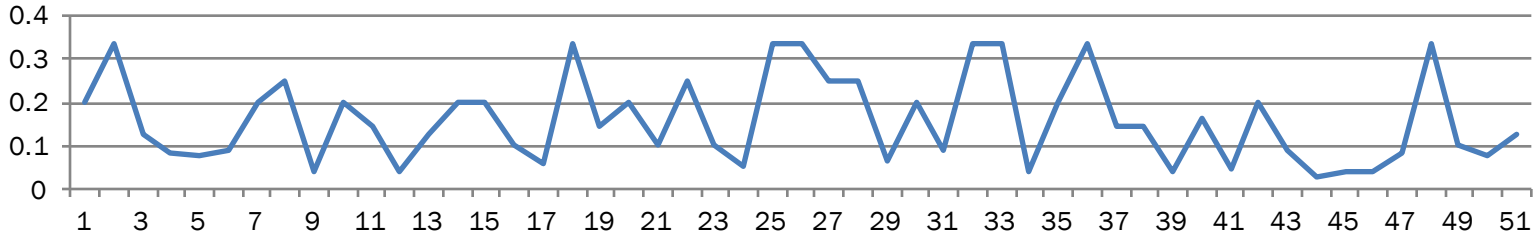
$$\mathbb{E}[\ell_t(\mathbf{w}_t)] = s_t \omega_t \ell_t(\mathbf{w}_t) + (1 - s_t)0 = s_t \frac{1}{s_t} \ell_t(\mathbf{w}_t) = \ell_t(\mathbf{w}_t).$$

- 百度首创: 1%样本的效果如90%+的随机采样效果

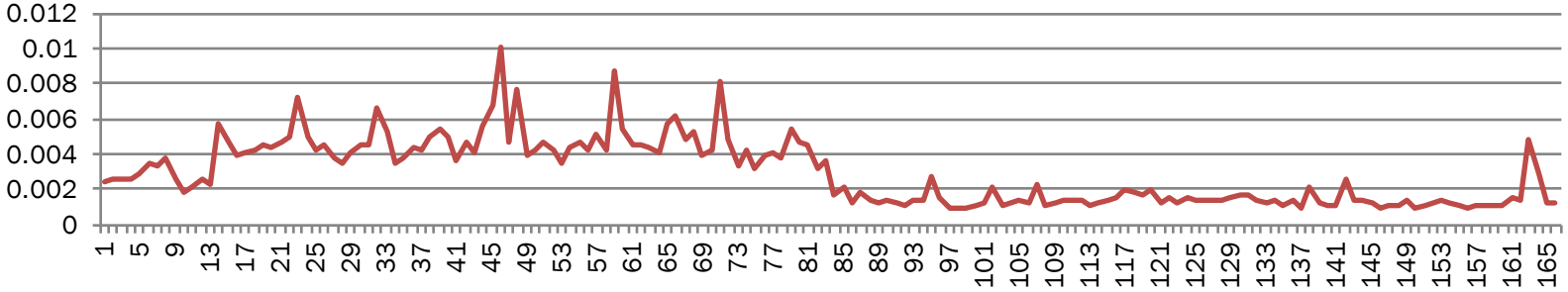
噪音检测

- 计算点击率随时间变化趋势 –百度首创：SA算法

随机噪音　sa=0.00275



正常样本　sa=-10.977



特征处理技术

目标

- 选择尽可能少的特征表示模型和数据

→

机器学习技术

- 特征选择
- 特征删减

- 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！

Site: (0 … 0 1 0 … 0)

cnn.com news.yahoo.com

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

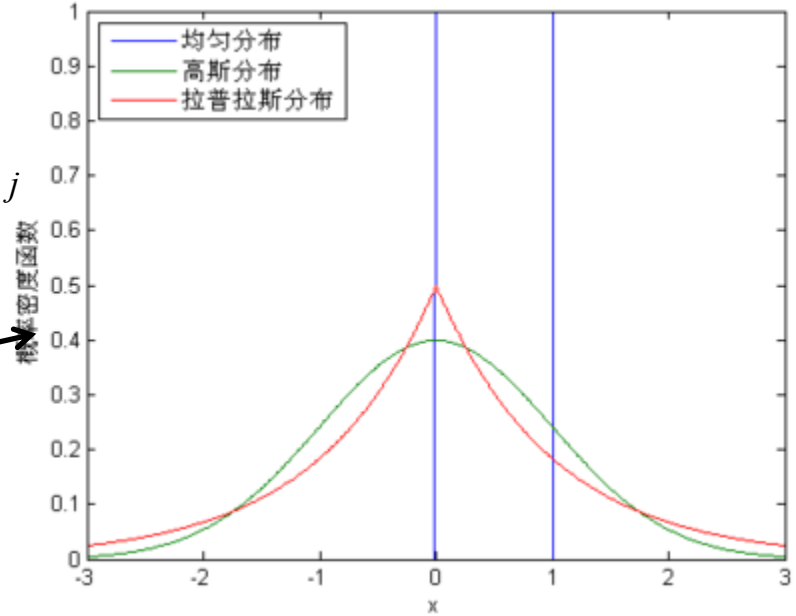
方法	内存节省	AucLoss升高
Bloom(n=2)	66%	0.008%
Bloom(n=1)	55%	0.003%
Poisson(p=0.003)	60%	0.020%
Poisson(p=0.1)	40%	0.006%
Fea-G	97%	0%

自动调参

- 背景

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

$$\arg\max_w L(w) = \sqrt[n]{\prod_{i=1}^{n}(1+\mathrm{e}^{-y_i x_i^T w})}\prod_{j=1}^{d} e^{\boxed{\frac{|w|_j}{1/c_j}}}$$



- 技术挑战： 人工选择超参数，费力
- 百度AdaLasso算法，智能选择超参数c_j

特征学习技术

- 特征调研背景

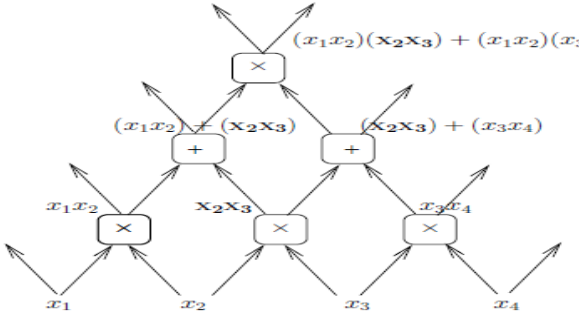
- 构造高阶组合特征，描述特征之间非线性关系
- 人工挖掘，耗时！耗力！依赖先验，无推广性！
  - ✓ 假设有N个单特征类，组合特征候选类：

$$2^N \approx C_N^1 + C_N^2 + \ldots + C_N^{N-1} + C_N^N$$

特征学习技术

- 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数据的特征学习算法



高阶组合：Site-Ad-Cookie，...

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

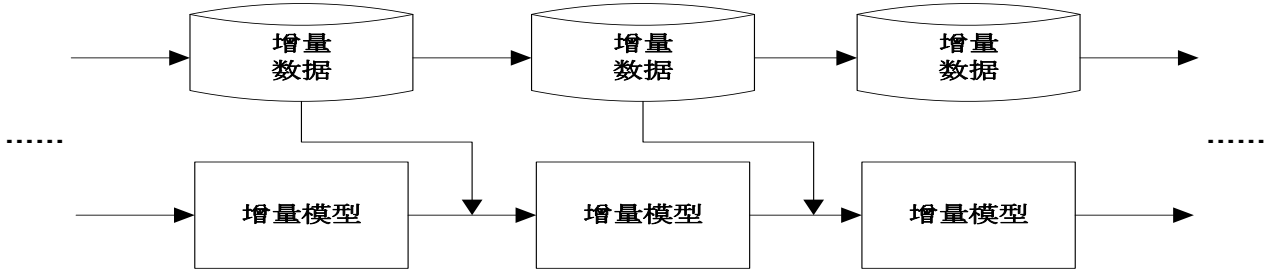
单特征：Site, Ad, Hour, Cookie, ...

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

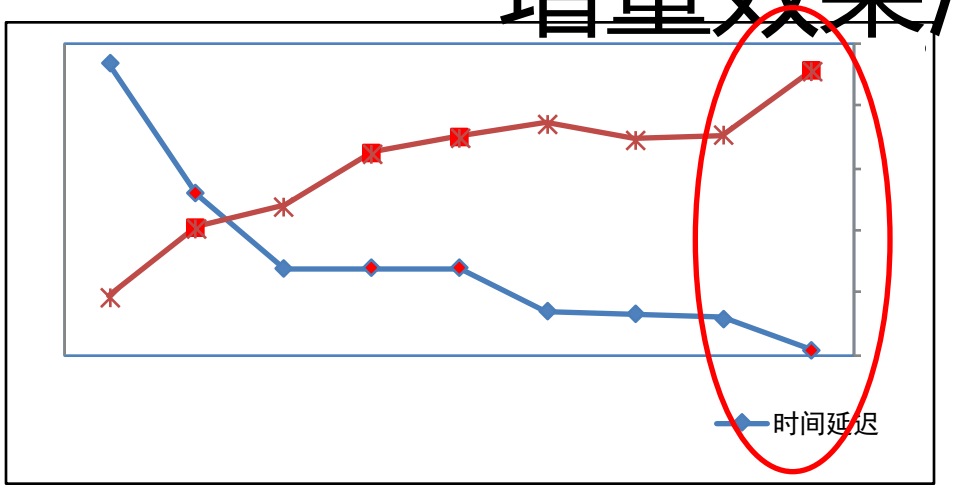
模型时效性

- 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%

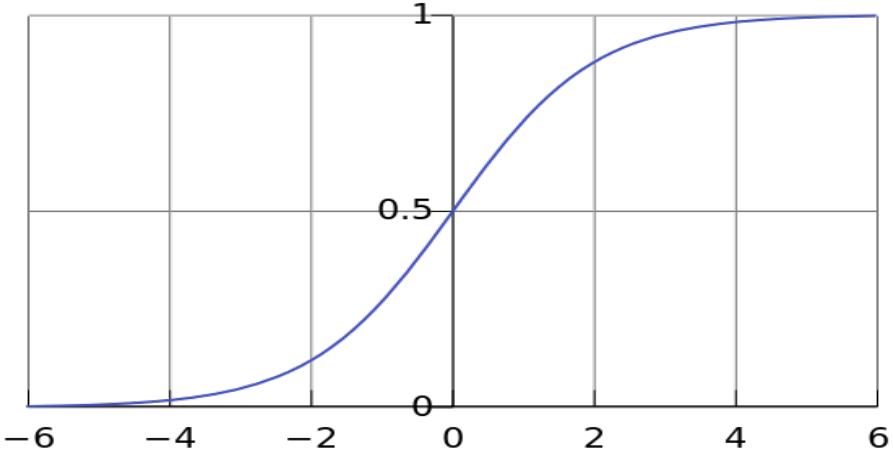
- 技术创新点：
  - ✓ 训练算法：首创SOA算法，使模型稳定性更好
  - ✓ 训练架构：批处理改为在线，节省资源80%以上
  - ✓ 在线学习平台：公司首个在大数据上实现分钟级别的在线学习

模型训练

- 线性逻辑回归模型

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

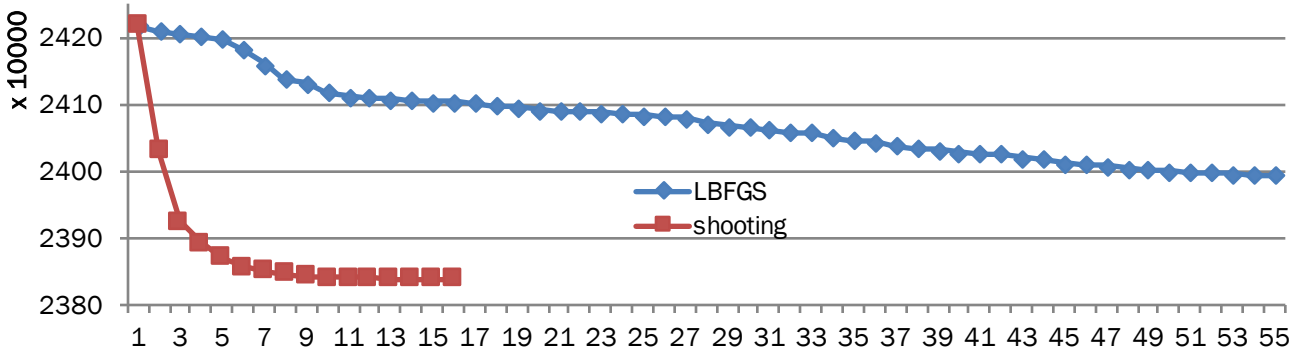
$$\ln(\frac{p_1}{p_0})=\beta_0+\sum\beta_i*x_i$$



- 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



从浅层到深度学习

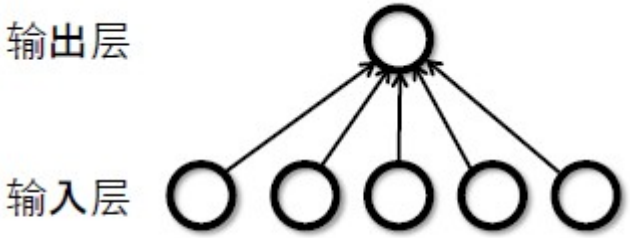
浅层模型:

千亿亿级特征

千亿级样本

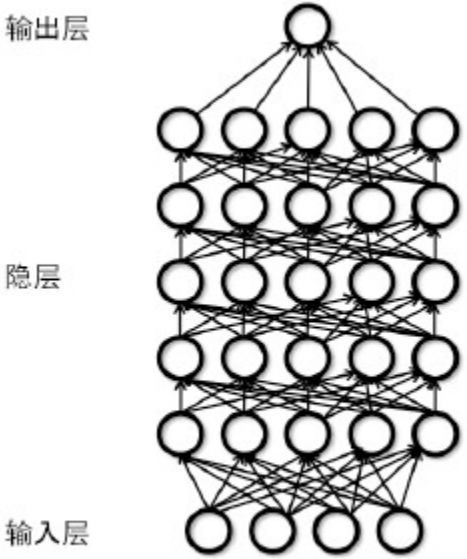
特征组合自动学习

模型在线更新



不含隐层的浅层学习模型

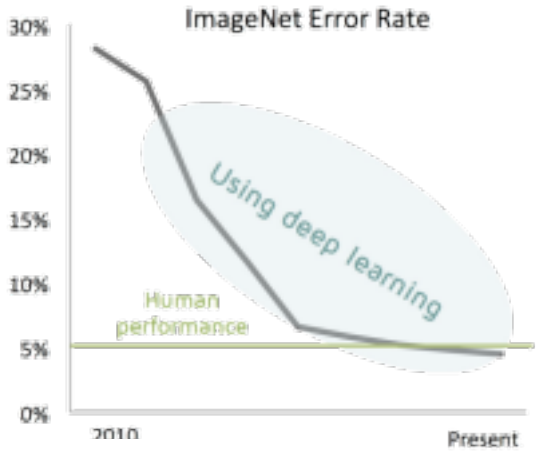
深层网络:

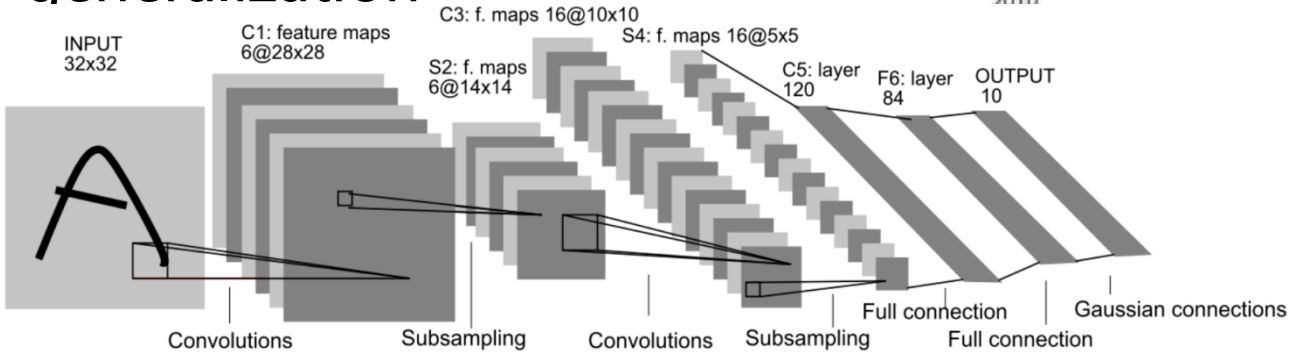


含多个隐层的深度学习模型

Deep Artificial Neural Network

- Why Go Deeper ?
  - Features: essential for successful ML
  - End-to-end hierarchical feature learning
  - Fewer parameters for better generalization
  -



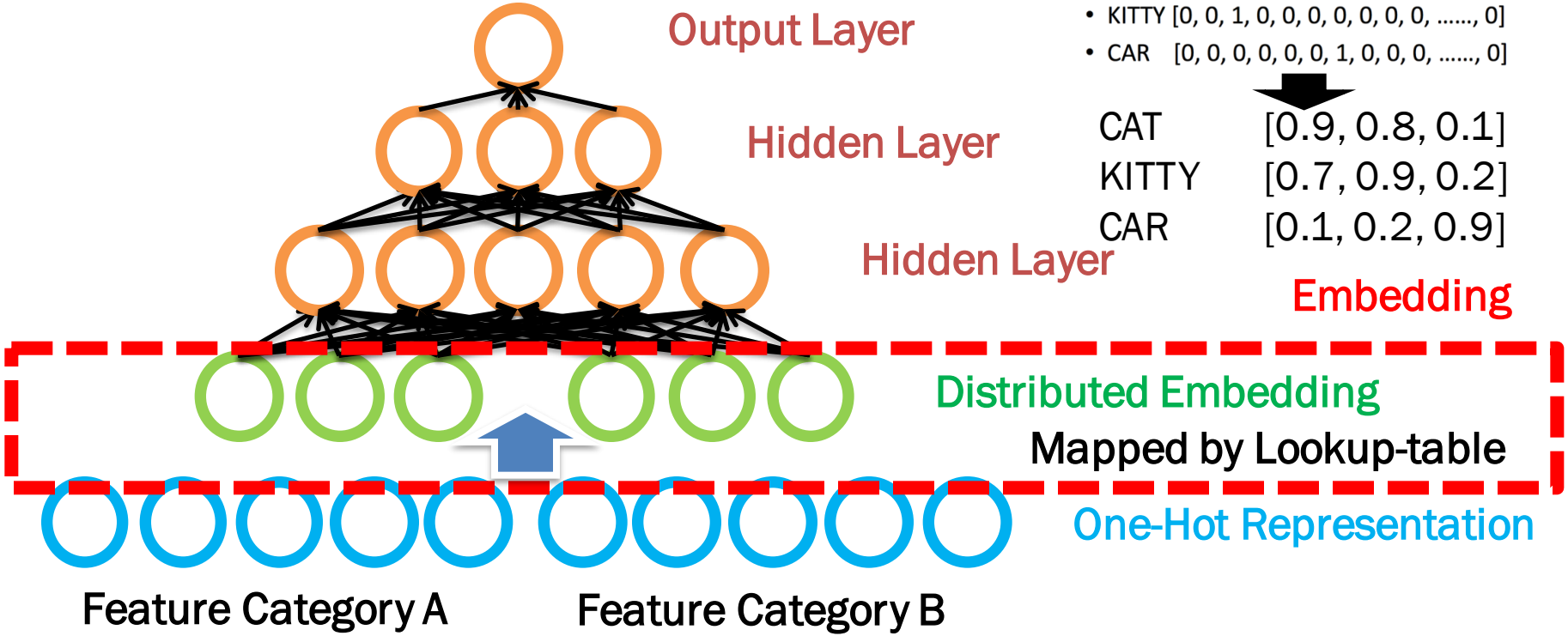


Deep Artificial Neural Network for CTR Prediction

- CAT [0, 0, 0, 0, 1, 0, 0, 0, 0, 0, ..., 0]
- KITTY [0, 0, 1, 0, 0, 0, 0, 0, 0, 0, ..., 0]
- CAR [0, 0, 0, 0, 0, 0, 1, 0, 0, 0, ..., 0]

- Difficulties of DNN applying in CTR prediction
  - **One-hot representation vs. Continuous** (Image / speech)
    - No inherited similarity
  - **Non-sequence vs. Sequence** (1D: text, speech; 2D: images; 3D: video)
    - Weight sharing is difficult
  - **Billions vs. Thousands** (text, images) of Input Dimensions
    - Huge bottom size
    - Difficult overfitting control.
  - **Sparse (Long-tailed) Features**
    - Poor generalization

Deep Artificial Neural Network for CTR Prediction

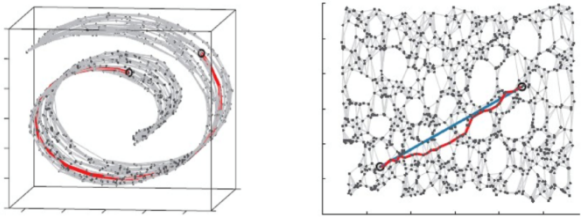


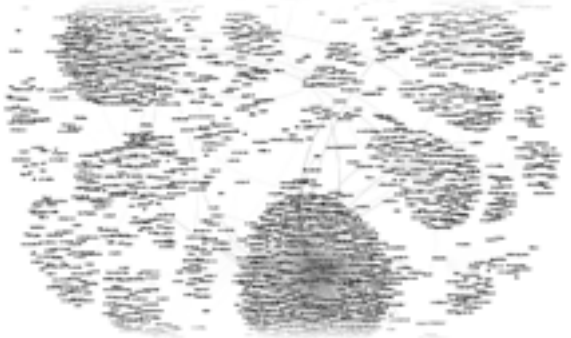
- CAT [0, 0, 0, 0, 1, 0, 0, 0, 0, 0, ..., 0]
- KITTY [0, 0, 1, 0, 0, 0, 0, 0, 0, 0, ..., 0]
- CAR [0, 0, 0, 0, 0, 0, 1, 0, 0, 0, ..., 0]

CAT	[0.9, 0.8, 0.1]
KITTY	[0.7, 0.9, 0.2]
CAR	[0.1, 0.2, 0.9]

Feature embedding for CTR Prediction

- What is Embedding?
  - Structure-preserving Mapping
  - E.g. Manifold embedding --- preserve geodesic distance
  - E.g. Graph embedding --- preserve connectivity
  - E.g. Word embedding – preserve semantic distance
  - E.g. Feature embedding – preserve feature similarity





- CAT [0, 0, 0, 0, 1, 0, 0, 0, 0, 0, ..., 0]
- KITTY [0, 0, 1, 0, 0, 0, 0, 0, 0, 0, ..., 0]
- CAR [0, 0, 0, 0, 0, 0, 1, 0, 0, 0, ..., 0]

CAT	[0.9, 0.8, 0.1]
KITTY	[0.7, 0.9, 0.2]
CAR	[0.1, 0.2, 0.9]

Feature embedding for CTR Prediction

- CAT [0, 0, 0, 0, 1, 0, 0, 0, 0, 0, ……, 0]
- KITTY [0, 0, 1, 0, 0, 0, 0, 0, 0, 0, ……, 0]
- CAR [0, 0, 0, 0, 0, 0, 1, 0, 0, 0, ……, 0]

CAT	[0.9, 0.8, 0.1]
KITTY	[0.7, 0.9, 0.2]
CAR	[0.1, 0.2, 0.9]

- Why feature embedding matters ?
  - One-hot representation:
    - Curse of dimensionality
    - Similarity / Semantic Gap
    - Discrete space: against the local smoothness properties
  - Distributed embedding:
    - Attribute sharing (e.g. attr_has_cur for cat & dog) → Similarity space
    - Continuous space → local smoothness can be expected
    - Statistically Efficient Representation → Low-dimensional
    - All above → better/easier generalization

Deep Neural Network for CTR Prediction

- How to get distributed embedding?
  - Feature Statistics
  - Unsupervised factor mining (e.g. word2vec)
  - Direct DNN training
- Difficulties:
  - Embedding Generalization
  - Scalability: **billions** of users / ads / context

Outline

- 广告背景
- 大规模机器学习
- 实战技术
- 总结展望

大数据点击率预测技术发展

一代：人工规则

二代: 40%+ 点击率提升

简单特征
小规模非线性模型

Yahoo, Facebook, Microsoft, etc.

三代: 10%+ 点击率提升

高维特征
大规模线性模型
模型实时更新

Google, Baidu, etc.

流式计算
模型在线更新

Google, Baidu, etc.

四代：百度Pulsar

20% + 点击率提升

大规模

复杂模型

实时更新

技术全球领先

百度Pulsar平台

- 数据量:**千亿样本、千亿特征**
- 模型:从浅层到深层灵活支持
- DNN:**万亿链接神经网络结构**
- 调研:自动化特征学习&参数调节
- 时效性:模型在线更新
- 评估:可视化模型&数据分析

处理上千亿样本、特征数据

自适应样本采集

自适应特征学习

自适应参数调整

自适应结构学习

Pulsar 智能化算法平台

上万台机器

Thanks !

Welcome to Join BDL!

xiafen@baidu.com