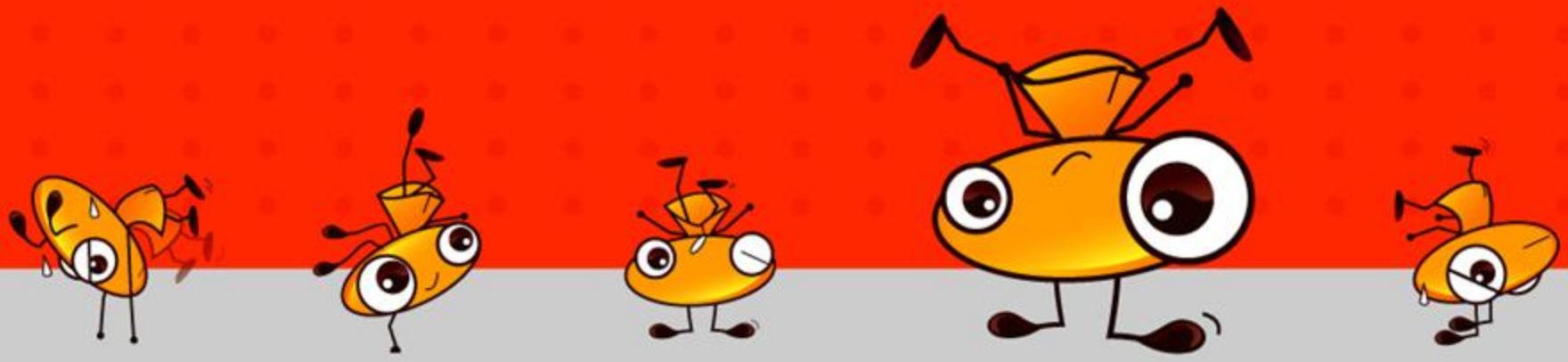
淘宝网
Taobao.com

海量数据下的非线性模型探索

盖坤
阿里巴巴



个人简介

- 花名：靖世　　姓名：盖坤
- 学校：清华大学 博士
- 研究方向: 机器学习 论文情况: NIPS,TPAMI ,AAAI,CVPR等
- 所属部门：阿里妈妈事业部-算法-基础研究
- 主要研究方向：
  - 机器学习算法
  - Rank model (CTR预估/CVR预估)

- 提纲
  - 1. 线性模型的限制
  - 2. 分片线性模型 MLR
  - 3. MLR for click model:偏移变量分解
  - 4. 迁移不同场景数据：Transfer MLR

线性模型

- CTR预估的典型做法：
  - 大规模特征+（广义）线性模型LR $p(y=1|x)=\frac{1}{1+\exp(-wx)}$
- 线性模型足够么？
  - 例子：一个曾使用的预测模式：

ad_ctr,cate_ctr,ad_pv为特征，
预测下一天ad的ctr：

$$y=\begin{cases} ad\_ctr, & if\ ad\_pv > K; \\ cate\_ctr, & if\ ad\_pv \le K. \end{cases}$$

  - 线性模型无法很好拟合。
- 当维度足够高时，线性模型是否已经够了？
  - 实际情况并不是这样。

高维非线性

- 例1：图像识别：100W像素
  - 直接线性模型？ No!
  - 几乎必须利用非线性处理：SIFT,HOG,多层NN等。
- 例2：
  - 用户维度CTR预估，特征user id，item id，标签：是否点击
  - 上亿维度的稀疏二值特征
  - 广义线性模型：$\sum_u w_u I(u = userid) + \sum_i w_i I(i = itemid) = w_{userid} + w_{itemid}$
    - 不同user 下，item上ctr的大小顺序一致　　没有个性化！
  - 特征处理：user id和item id做笛卡尔积得到交叉特征放入线性模型
    - 过拟合：用户点过的item 预估ctr高，看了没点的ctr低。
    - 单纯记忆历史行为，缺乏投放用户没点过但可能感兴趣的item的能力。

高维非线性

- CTR预估输入：高维id特征
  - 线性模型不够。人工对特征进行笛卡尔积也不一定好。
- 能否自动学到推广性能好的非线性关系？
  - 一些已有的非线性算法：
    - Tree based方法（例GBDT）
      - 每个树叶为if(user id==useri && item id == itemj)的条件判断
      - 仍然是对历史行为的记忆 缺乏推广性
    - 矩阵分解
      - 适用于两种id的情况：多种id输入？
    - Factorization machines: $f(x) = \sum_{i,j} < v_i, v_j > x_i x_j$
      - 只拟合线性关系和二次关系
      - 无法拟合其它非线性关系：例如三种特征的交叉，值的高阶变换等。

- 提纲
  - 1. 线性模型的限制
  - 2. 分片线性模型 MLR
  - 3. MLR for click model:偏移变量分解
  - 4. 迁移不同场景数据：Transfer MLR

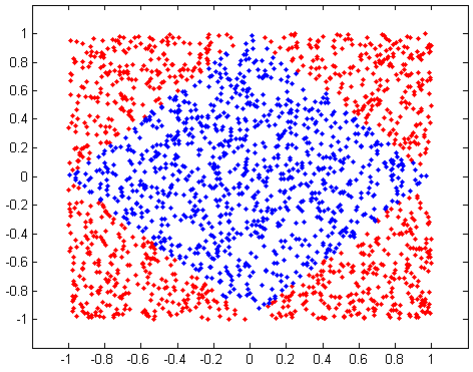
分片线性模型

- 挑战：如何从大规模数据中挖掘出推广性好的非线性模式？
- 我们的工作：提出分片线性学习算法
  - 名称:Mixture of lr (mlr)
  - 有任意强的非线性拟合能力
  - 模型复杂度可控（分片数）
    - 平衡欠拟合和过拟合
    - 保证分片内平均样本数，并用线性规律拟合，来得到好的推广性

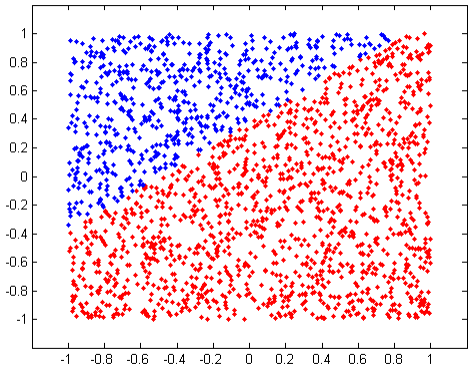
ad_ctr,cate_ctr,ad_pv为特征，
预测下一天的ctr：

$$y = \begin{cases} ad\_ctr, & if\ ad\_pv > K; \\ cate\_ctr, & if\ ad\_pv \le K. \end{cases}$$

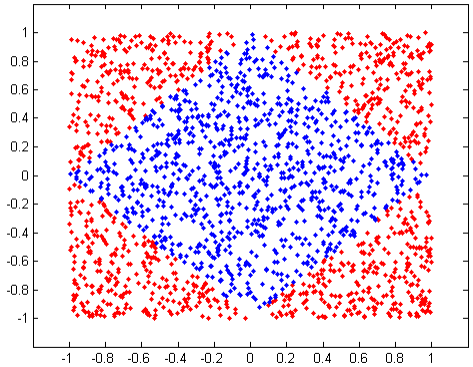
例子：分类问题



训练数据



线性模型(LR)



分片线性模型

分片线性模型形式

作用范围　　基预测器

- 分而治之　$f(x) = g\left(\sum_i \pi_i(x,\mu)\eta_i(x,w)\right)$
  - 空间划分为多个局部区域，每个区域内一个线性预测模型

逻辑回归: $p(y=1|x) = \frac{1}{1+\exp(-wx)}$

  - 分类问题：
    - 分片线性分类: $p(y=1|x) = \frac{1}{1+\exp\left(-\sum_i \pi_i(x,\mu)(w_i x)\right)}$
    - 混合逻辑回归: $p(y=1|x) = \sum_{i=1}^{m} \pi_i(x,\mu) \cdot \frac{1}{1+e^{-w_i \cdot x}}$
  - 回归问题：　$y(x) = \sum_{i=1}^{m} \pi_i(x,\mu) \cdot (w_i x)$
  - 分片（聚类）函数:中心聚类，softmax等，例:

$\pi_i(x,\mu) \propto \exp(-(x-c_i)^T diag(\mu_i)(x-c_i))$　　$\pi_i(x,\mu) \propto \exp(-(x-c_i)^2)$

$\pi_i(x,\mu) \propto \exp(\mu_i x)$

  - 目标函数：似然、误差平方和、Bregman散度等
  - 模型不限定为隐变量概率模型，算法引入分组稀疏，适用于大规模高维度数据

正则化和目标函数

- 参数矩阵:

$$\theta = [w_1, ..., w_m, \mu_1, ..., \mu_m]$$

- 特征选择：同一维度对应多个权重 — 分组稀疏正则

$$\|\theta\|_{2,1} = \sum_i \sqrt{\sum_k \theta_{ik}^2}$$

- 目标函数:

$$\min_{\theta} \quad F(\theta) = \sum_i l(f(x_i;\theta), y_i) + \lambda\|\theta\|_{2,1} + \beta\|\theta\|_1$$

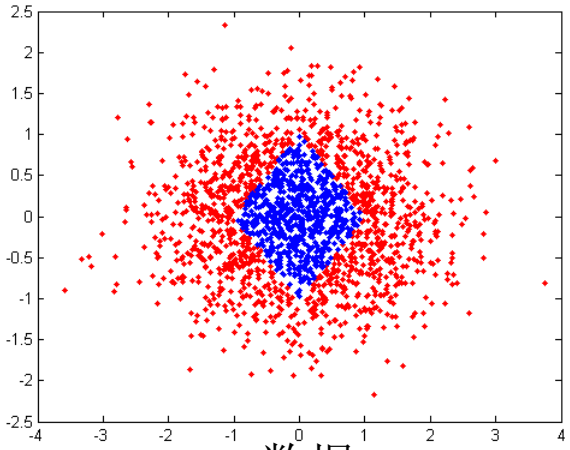
算法设计

- 目标函数：$\min_{\theta} \quad f(\theta) = \sum_i l(f(x_i;\theta), y_i) + \lambda\|\theta\|_{2,1} + \beta\|\theta\|_1$
- 难度和挑战：
  - 目标函数非凸，非光滑（不可导，不存在次梯度）
  - 超大规模数据，高维度
  - 提出针对非凸非光滑目标的快速优化方法
  - 算法实现：
    - 部署于500台MPI计算节点，目前支持约50T训练数据，千亿级别样本的训练(受限于集群总内存，支持规模随机器数增大)

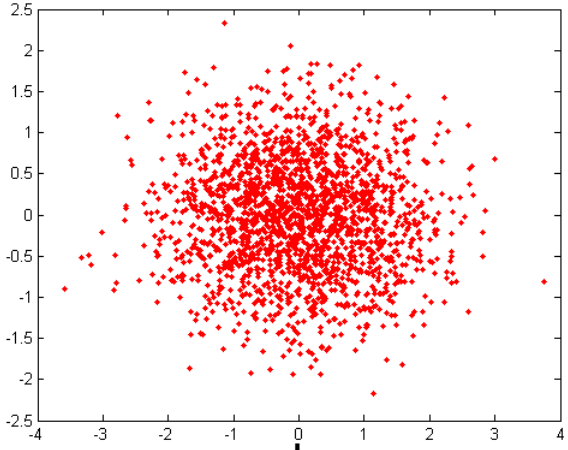
特性

- 特点
  - 分而治之
  - 可以适用于大规模高维度数据
  - 模型复杂度可控：可以线性，也可以逼近任意复杂非线性函数
  - 具有自动特征选择作用
  - 模型结构符合广告数据规律

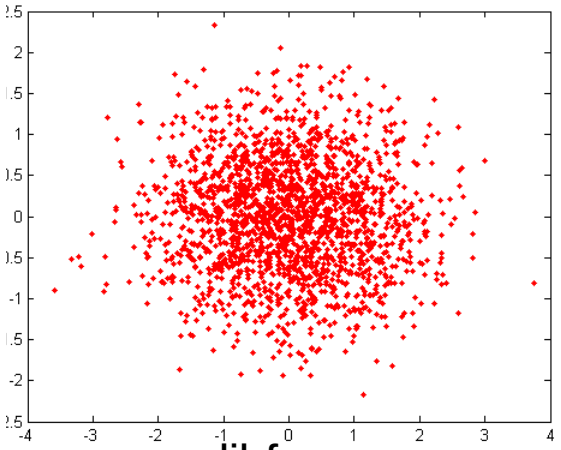
实验1



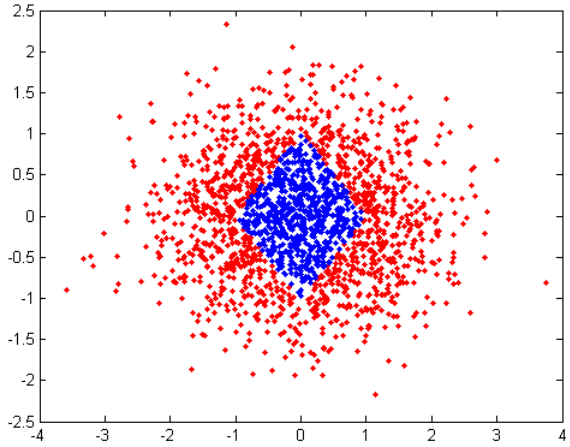
数据



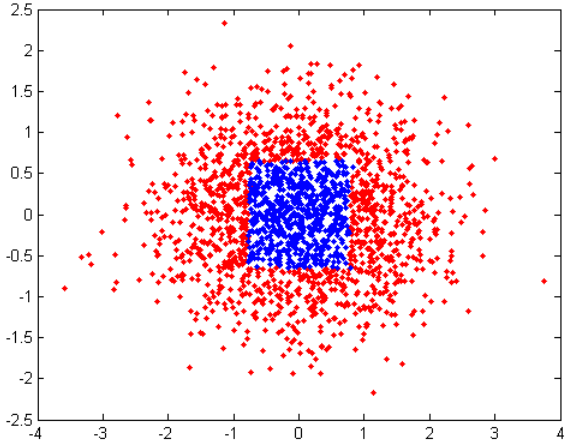
lr



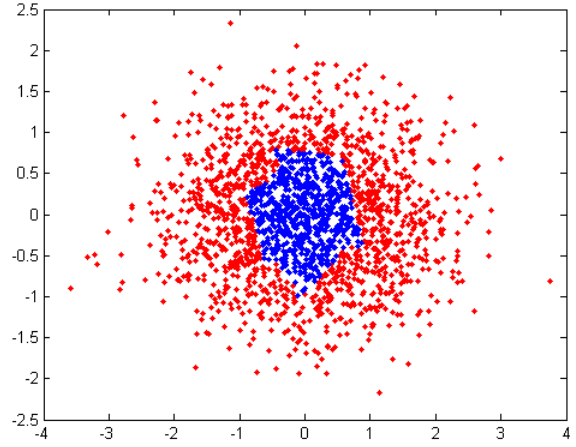
libfm



mlr(4分片)



Kmeans+lr(4 clusters)



Kmeans+lr(10 clusters)

实验2

- 三种id特征交叉实验：

特征1	特征2	特征3	类别	mlr预测概率(2分片)	Libfm预测概率（topic数20）
0	0	0	1	0.999954	0.507501
1	0	0	0	0.000050	0.494845
0	1	0	0	0.000058	0.496506
0	0	1	0	0.000045	0.494050
0	1	1	1	0.999965	0.501566
1	0	1	1	0.999969	0.491462(分错)
1	1	0	1	0.999927	0.502185
1	1	1	0	0.000038	0.520136(分错)

- mlr具有更强的非线性拟合能力，不局限于二次函数
  - 可以在更少的参数下拟合更高阶非线性信息(例如多类id交叉)

应用

- 淘宝定向营销rank
  - 2012年3月起，mlr模型开始承担主要流量
  - 试验田CTR/RPM↑30+%
- 搜索营销转化率预估（按成交收费）
  - CTR↑35%,PCVR↑30%,RPM↑25%
- 推荐融合排序
  - CTR↑8%,PCVR↑11%
- 展示广告排序
  - CTR↑30+%

- 提纲
  - 1. 线性模型的限制
  - 2. 分片线性模型 MLR
  - 3. MLR for click model:偏移变量分解
  - 4. 迁移不同场景数据：Transfer MLR

带偏移变量的MLR

- 问题：宝贝展示的页面、位置影响点击概率
- 宝贝特征x，偏移向量y(场景、页数、位置等)：
  - 学习联合概率p(x,y) — 需要x,y的大部分组合
  - 采样问题：并不是x,y的所有组合能采到样本
- 提出带偏移MLR算法： $p(x,y|\theta,w) = p_{mlr}(x|\theta)p_{lr}(y|w)$
  - y:偏移向量，包括场景、页数、位置等信息
  - 只需很少一些x,y组合
  - AUC指标↑2-8个百分点
- 应用：精品库场景CTR ↑30+%

大规模非线性ctr/cvr预估和偏移变量的分解一起优化

- 提纲
  - 1. 线性模型的限制
  - 2. 分片线性模型 MLR
  - 3. MLR for click model:偏移变量分解
  - 4. 迁移不同场景数据：Transfer MLR

Transfer MLR

- 问题：淘客搜索场景cvr预估中购买样本过少
- 思路：借鉴主搜购买数据做样本
  - 难点：样本有偏：主搜场景购买率明显高于淘客搜索
- Transfer MLR：
  - 去除不同规律，借鉴相同规律 传递宝贝的吸引力:mlr参数相近
  - 设宝贝特征x，淘客偏移向量y，主搜偏移向量z (y,z不同维度)
    - 淘客搜索：$\mathrm{pcvr}_t(x,y) = p_{mlr}(x;\theta_t)p_{lr}(y;w_t)$
    - 主搜: $\mathrm{pcvr}_s(x,z) = p_{mlr}(x;\theta_s)p_{lr}(z;w_s)$ 吸收不同场景的不同
    - 损失：$L_t(\mathrm{pcvr}_t(x,y|\theta_t,w_t)) + \lambda L_s(\mathrm{pcvr}_s(x,z|\theta_s,w_s)) + \gamma\|[\theta_s,\theta_t]\|_{2,1} \quad \text{s.t. } \|\theta_s-\theta_t\| \le \beta$
    - β→0时 $L_t(\mathrm{pcvr}_t(x,y|\theta,w_t)) + L_s(\mathrm{pcvr}_s(x,z|\theta,w_s)) + r(\theta,w_s,w_t)$
- 应用效果：淘客搜索宝贝排序pcvr ↑ 30+%

淘宝网
Taobao.com

THANK YOU