MLR: 海量数据下的非线性模型探索

靖世(盖坤)

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-广告算法-基础研究
- 主要研究方向：
  - 机器学习算法创新
  - 广告算法应用研究(Rank Model)

- 提纲
  - 1. 线性的限制和分片线性模型MLR
  - 2. 模型对比和应用效果
  - 3. MLR和topic model的关系
  - 4. MLR for click model:偏移变量分解
  - 5. 迁移不同场景数据：Transfer MLR

分片线性模型

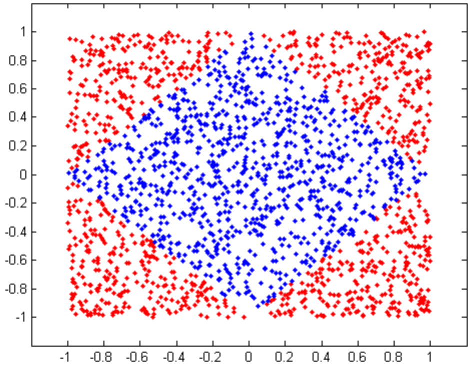
- 互联网行业机器学习算法的问题：
  - 大规模、高维度数据
  - 已有非线性学习算法复杂度高
  - 广泛使用线性模型，无法利用数据中非线性规律
    - 人工预处理，难度大，难以最优

ad_ctr,cate_ctr,ad_pv为特征，
线性模型无法拟合如下规则：

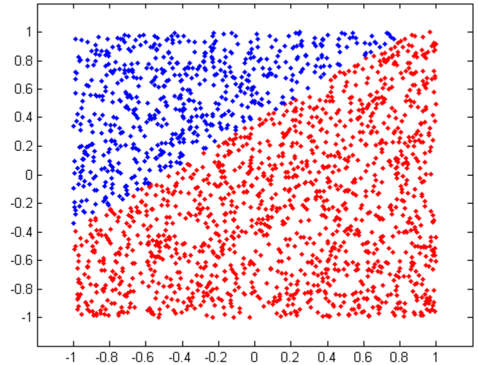
$$y = \begin{cases} ad\_ctr, & if\ ad\_pv > K; \\ cate\_ctr, & if\ ad\_pv \leq K. \end{cases}$$

- 挑战：如何从大规模高维度数据中学习**非线性**模型？
- 我们的工作：**发明新的分片线性学习算法—Mixture of Logistic Regression**

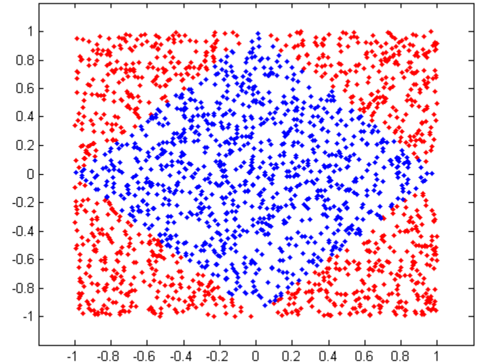
例子：分类问题



训练数据



线性模型(LR)



分片线性模型

MLR：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: $p(y=1|x)=\frac{1}{1+\exp(-\boxed{wx})}$

- 分类问题：
  - 分片线性分类: $p(y=1|x)=\frac{1}{1+\exp\left(-\sum_i \pi_i(x)(w_i x)\right)}$
  - 混合逻辑回归:

以此为例

$$p(y=1|x)=\sum_{j=1}^{m}\frac{e^{\mu_j\cdot x}}{\sum_{v=1}^{m}e^{\mu_v\cdot x}}\cdot\frac{1}{1+e^{-w_j\cdot x}}$$

- 回归问题： $y(x)=\sum_{j=1}^{m}\pi_i(x)\cdot(w_i x)$
- 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

MLR:正则化和目标函数

- MLR:

$$f(x;\theta)=\sum_{j=1}^{m}\frac{e^{\mu_j\cdot x}}{\sum_{v=1}^{m}e^{\mu_v\cdot x}}\cdot\frac{1}{1+e^{-w_j\cdot x}}$$

- 参数矩阵:

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

- 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:

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

MLR: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）
  - 提出针对**非凸非光滑目标的快速优化方法**
    - 思路: 证明存在方向导数，最速下降方向唯一，找出最速下降方向，加速为superlinear下降。
  - 算法实现：
    - 部署于200台MPI计算节点，目前支持约20T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)

MLR：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

MLR:模型对比

- MLR vs. LR
  - 例子：推荐宝贝ranking数据

测试AUC	预估ctr	预估pcvr
LR	0.700112	0.748859
MLR(m=51)	**0.713173**	**0.775776**
提升值(百分点绝对值)	**+1.3061**	**+2.6917**

- MLR vs. GBDT(boosting)
  - 例子：2012年上半年站内定向广告ranking数据

AUC	**GBDT (tree:800,depth:6)**	**MLR（m=50）**	**MLR（m=150,未收敛模型）**	**MLR(150，收敛模型)**
训练集（前一天）	0.664416	0.660369	0.664559	0.666423
测试集（后一天）	0.661497	0.665067	0.665884	0.667163

  - 此外,GBDT不适合超高维度数据

MLR：应用效果

- 定向广告质量分6期、7期
  - 试验田CTR/RPM↑30+%
  - 生产 CTR↑20+%, RPM↑10+%
- 转化率预估
  - 6-9月:淘客宝贝搜索ctr↑5+%, pcvr↑3-7%
  - 10-11月:覆盖流量上CTR↑35%,PCVR↑30%,RPM↑25%
  - 正在上线:实验流量上PCVR ↑20+%
- 个性化推荐融合排序
  - CTR↑8%,PCVR↑11%
- 其它：
  - 搜索广告top query上初步实验CTR↑2.7%, 下侧热卖宝贝CTR↑4.6%(仍需优化以得到最佳效果。对比线上ctr ↑10%)

MLR for topic model

- PLSA:

$$y(u,i) = \sum_{k=1}^{K} \Phi_{u,k}\Psi_{i,k}$$

- MLR模型

$$f(x) = \sum_{k=1}^{K} \pi_k(x)\eta_k(x)$$

  - 如果以user id 和item id作为特征：

$$y(u,i) = \sum_{k=1}^{K} \pi_k(u,i)\eta_k(u,i)$$

  - 限制π中i权重为0， η中u权重为0，则完全等价于PLSA，参数一一对应
  - MLR更灵活：参数更多
    - PLSA是二次模型。
    - MLR用分片线性思想，有拟合任意阶次函数能力
  - MLR很容易引入其它特征，提高精度，解决冷启动问题

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%
- 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

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
  - 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：$\mathrm{pcvr}_t(x,y) = \boxed{p_{mlr}(x;\theta_t)}\boxed{p_{lr}(y;w_t)}$
    - 主搜: $\mathrm{pcvr}_s(x,z) = \boxed{p_{mlr}(x;\theta_s)}\boxed{p_{lr}(z;w_s)}$
    - 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 $L_t(\mathrm{pcvr}_t(x,y|\theta,w_t)) + L_s(\mathrm{pcvr}_s(x,z|\theta,w_s)) + r(\theta,w_s,w_t)$

传递宝贝的吸引力:mlr参数相近

吸收不同场景的不同

- 应用效果：淘客宝贝排序pcvr ↑30+%

淘宝网
Taobao.com

THANK YOU