Personal Recommendation Algorithm

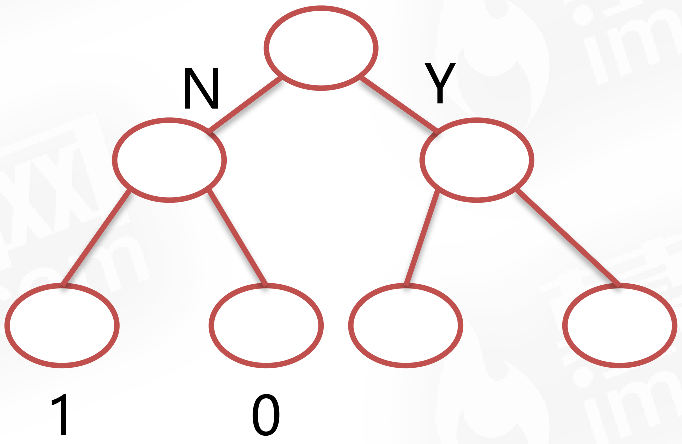
Main Flow

- GBDT(Gradient Boosting Tree)背景知识介绍
- GBDT数学原理与构建方法
- XGBoost 数学原理与构建方法

慕课网 imooc.com

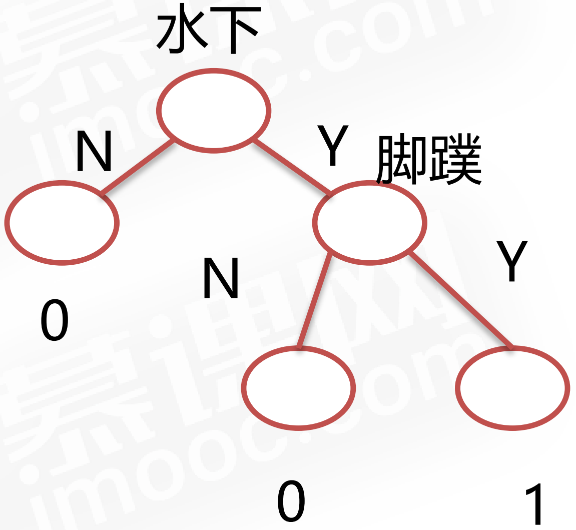
GDBT背景知识介绍

- 什么是决策树



GDBT背景知识介绍

水下生活	有脚蹼	是鱼类
1	1	1
1	1	1
1	0	0
0	1	0
0	1	0



决策树构造原理

- CART生成
- 回归树:平方误差最小化原则
- 分类树:基尼指数

回归树

- 回归树的函数表示

$$f(x)=\sum_{m=1}^{M}c_m I(x\in R_m)$$

$$\sum_{x_i\in R_m}(y_i-f(x_i))^2$$

$$c_m=ave(y_i|x_i\in R_m)$$

最优特征选取

$$\min_{j,s}\left[\min_{c1}\sum_{x_1\in R_1}(y_i-c_1)^2+\min_{c2}\sum_{x_1\in R_2}(y_i-c_2)^2\right]$$

$$R_1=\left\{x\middle|x^j\le s\right\},R_2=\left\{x\middle|x^j>s\right\}$$

$$c_1=ave\left(y_i\middle|x_i\in R_1\right),c_2=ave\left(y_i\middle|x_i\in R_2\right)$$

构建树的流程

- 遍历所有特征，特征的最佳划分对应的得分，选取最小得分的特征
- 将数据依据此选取的特征划分分成两部分
- 继续在左右两部分遍历变量找到划分特征直到满足停止条件

分类树

- 基尼指数

$$Gini(D) = 1 - \sum_{k=1}^{K} \left(\frac{|C_k|}{|D|} \right)^2$$

$$D_1 = \left\{ (x, y) \in D \middle| A(x) \geq a \right\}, D_2 = D - D_1$$

$$Gini(D, A) = \frac{|D_1|}{|D|} Gini(D_1) + \frac{|D_2|}{|D|} Gini(D_2)$$

慕课网 imooc.com

基尼指数求解

水下生活	有脚蹼	是鱼类
1	1	1
1	1	1
1	0	0
0	1	0
0	1	0

G(D,水下)=3/5*4/9+2/5*0=12/45

G(D,脚蹼)=4/5*1/2+1/5*0=4/10

Class Two

慕课网 imooc.com

Personal Recommendation Algorithm

慕课网 imooc.com

boosting

- 什么是boosting
- 如何改变训练数据的权重
- 如何组合多个基础model

Boosting Tree

- 提升树模型函数

$$f_M(x)=\sum_{m=1}^{M}T(x;\theta_m)$$

$$f_m(x)=f_{m-1}(x)+T(x;\theta_m)$$

$$\theta_m=\arg\min_{\theta_m}\sum_{i=1}^{N}L(y_i,f_{m-1}(x_i)+T(x_i;\theta_m))$$

慕课网 imooc.com

迭代损失函数

$$L(y, f(x)) = (y - f(x))^2$$

$$L(y, f_m(x)) = \left[y - f_{m-1}(x) - T(x; \theta_m)\right]^2$$

慕课网 imooc.com

提升树的算法流程

- 初始化$f_0(x)=0$
- 对m=1,2..M 计算残差r_m，拟合r_m，得到T_m
- 更新$f_m=f_{m-1}+T_m$

慕课网 imooc.com

Example

x	1	2	3	4	5	6	7	8	9	10
y	5.56	5.70	5.91	6.40	6.80	7.05	8.90	8.70	9.00	9.05

$$S=1, R_1=\{1\}, R_2=\{2,3\ldots10\}, c_1=5.56, c_2=7.50, m(s)=0+15.72=15.72$$

s	1	2	3	4	5	6	7	8	9
m(s)	15.72	12.07	8.36	5.78	3.91	1.93	8.01	11.73	15.74

Example

$T_1(x)=6.24\ x<=6; T_1(x)=8.91\ x>6 \quad f_1(x)=T_1(x)$

x	1	2	3	4	5	6	7	8	9	10
y	-0.68	-0.54	-0.33	0.16	0.56	0.81	-0.01	-0.21	0.09	0.14

$T_2(x)=-0.52\ x<=3; T_2(x)=0.22\ x>3 \quad f_2(x)=f_1(x)+T_2(x)$

梯度提升树

- 残差的数值改变

$$r_m = -\left[\frac{\partial L(y, f(x_i))}{\partial f(x_i)}\right]_{f(x)=f_{m-1}(x)}$$

慕课网 imooc.com

Class Three

慕课网 imooc.com

Personal Recommendation Algorithm

慕课网 imooc.com

XGBoost

- XGBoost模型函数

$$f_M(x) = \sum_{m=1}^{M} T(x;\theta_m)$$

$$f_m(x) = f_{m-1}(x) + T(x;\theta_m)$$

$$\arg\min_{\theta_m} \sum_{i=1}^{N} L\left(y_i, f_{m-1}(x_i) + T(x_i;\theta_m)\right) + \Omega(T_m)$$

XGBoost

- 优化目标的泰勒展开

$$f(x+\Delta x) \approx f(x) + f'(x)\Delta x + 1/2 f''(x)\Delta x^2$$

$$\min_{\theta_m} \sum_{i=1}^{N} \left[g_i T_m + 0.5 * h_i T_m^2 \right] + \Omega(T_m)$$

$$g_i = \frac{\partial L(y_i, f_{m-1})}{\partial f_{m-1}}, h_i = \frac{\partial^2 L(y_i, f_{m-1})}{\partial f_{m-1}}$$

XGBoost

- 定义模型复杂度

$$f(x)=\sum_{j=1}^{Q} c_j I\left(x \in R_j\right)$$

$$\Omega\left(T_m\right)=\partial Q+0.5\beta\sum_{j=1}^{Q} c_j^2$$

慕课网 imooc.com

XGBoost

- 目标转化

$$\min_{\theta_m} \sum_{i=1}^{N} \left[g_i T_m + 0.5 * h_i T_m^2 \right] + \Omega(T_m)$$

$$\min_{\theta_m} \sum_{i=1}^{N} \left[g_i T_m + 0.5 * h_i T_m^2 \right] + \partial Q + 0.5 \beta \sum_{j=1}^{Q} c_j^2$$

$$\min_{\theta_m} \sum_{j=1}^{Q} \left[\left(\sum_{i \in R_j} g_i \right) c_j + 0.5 \left(\sum_{i \in R_j} h_i + \beta \right) c_j^2 \right] + \partial Q$$

XGBoost

- 目标函数最优解

$$G_j = \sum_{i \in R_j} g_i, H_j = \sum_{i \in R_j} h_i$$

$$\min_{\theta_m} \sum_{j=1}^{Q} \left[G_j c_j + 0.5 \left(H_j + \beta \right) c_j^2 \right] + \partial Q$$

$$c_j = -\frac{G_j}{H_j + \beta}, obj = -0.5 \sum_{i=1}^{Q} \frac{G_j^2}{H_j + \beta} + \partial Q$$

XGBoost

- 最佳划分特征选取

$$c_j = -\frac{G_j}{H_j + \beta}, obj = -0.5\sum_{i=1}^{Q}\frac{G_j^2}{H_j + \beta} + \partial Q$$

$$Gain = \left(\frac{G_L^2}{H_L + \beta} + \frac{G_R^2}{H_R + \beta} - \frac{(G_R + G_L)^2}{H_R + H_L + \beta}\right) - \partial$$

慕课网 imooc.com

XGBoost总流程

- 初始化$f_0(x)=0$
- 对m=1,2..M 应用选择最优划分特征的方法构造树
- 更新$f_m=f_{m-1}+learning_rate*T_m$

慕课网 imooc.com

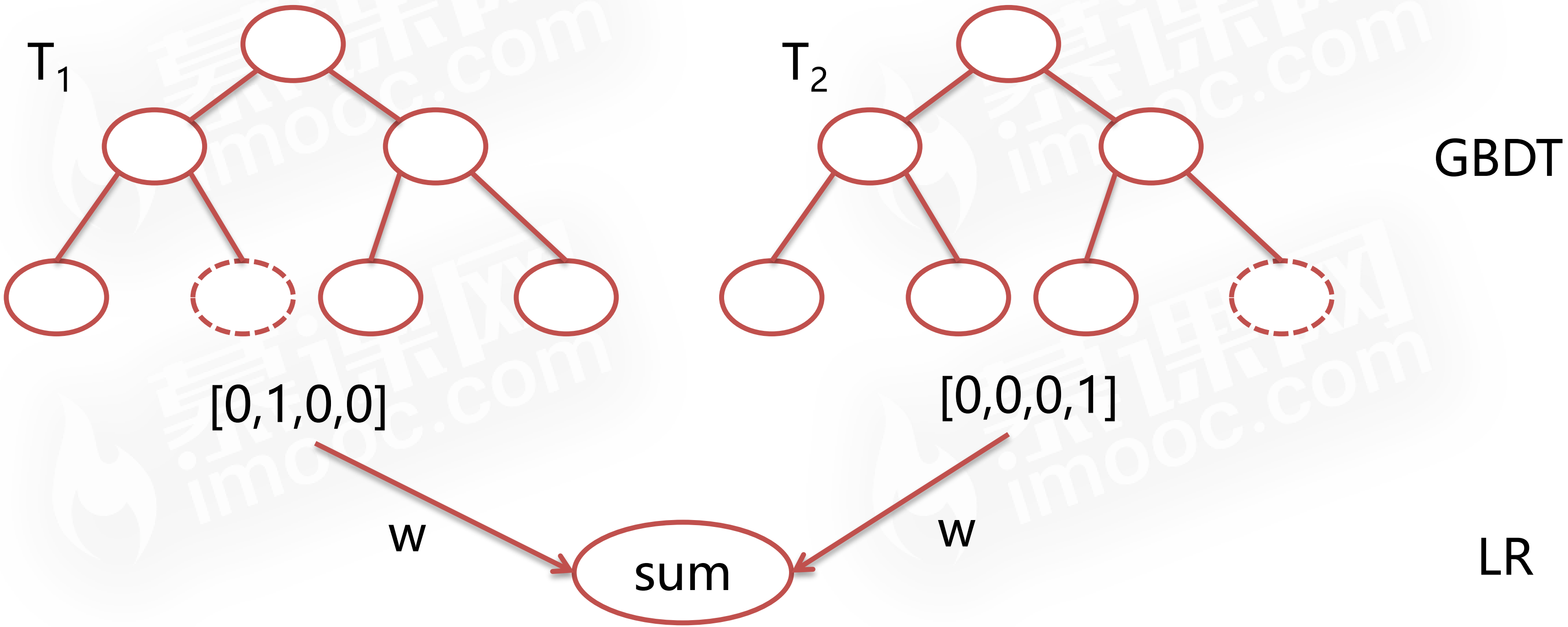
Personal Recommendation Algorithm

慕课网 imooc.com

背景知识

- Practical Lessons from Predicting Clicks on Ads at Facebook
- 逻辑回归需要繁琐的特征处理
- 树模型的feature transform能力

模型网络



优缺点总结

- 利用树模型做特征转化
- 两个模型单独训练不是联合训练