GBDT算法原理推导

一个提升树模型的数学表达为：

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

其中$T(x;\Theta_m)$为决策树表示的基模型，Θ_m为决策树参数，M为树棵数。

当确定初始提升树模型为$f_0(x)=0$，第m模型为：

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

$f_{m-1}(x)$为当前迭代模型，根据前向分步算法，可以使用经验风险最小化来确定下一棵决策树的参数Θ_m：

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

以梯度提升回归树为例，一棵回归树可以表示为：

$$T(x;\Theta)=\sum_{k=1}^{K}c_k I(x\in R_j)$$

最终模型为 $f_M(x)=\sum_{i=1}^{M}T(x;\Theta_m)$

假设回归树使用平方损失：$L(y,f(x))=(y-f(x))^2$

对应到GBRT中，损失可推导为：$L(y,f_{m-1}(x)+T(x;\Theta_m))=[y-f_{m-1}(x)-T(x;\Theta_m)]^2$

令$r=y-f_{m-1}(x)$，上式为：$L(y,f_{m-1}(x)+T(x;\Theta_m))=[r-T(x;\Theta_m)]^2$

☆使用损失函数的负梯度在当前模型的值作为回归提升树中残差近似值

$$r_{mi}=-\left[\frac{\partial L(y_i,f(x_i))}{\partial f(x_i)}\right]_{f(x)=f_{m-1}(x)} \quad \cdots\cdots 式(1)$$

给定训练集$D=\{(x_1,y_1),(x_2,y_2)\cdots(x_N,y_N)\}$，GBDT算法步骤如下：

(1)、初始化提升树模型：

$$f_0(x)=\arg\min_c\sum_{i=1}^{N}L(y_i,c)$$

(2)、对基分类器$m=1,2,\cdots M$，有：

(a)、对每个样本$i=1,2,\cdots N$，计算梯度拟合的残差。如式(1)所示

(b)、将上一步得到的残差作为样本新的真实值，并将数据$(x_i,r_{mi}),i=1,2,\cdots N$

作为下一棵树的训练数据，得到一棵新的回归树 $f_m(x)$，其对应的叶子区域为 R_{mj}，$j=1,2,\cdots,J$。其中 J 为回归树 T 的叶子结点的个数。

(c)、对叶子区域 $j=1,2,\cdots,J$ 计算最优拟合值：

$$c_{mj}=\arg\min_{c}\sum_{x_i\in R_{mj}}L(y_i,f_{m-1}(x_i)+c)$$

(d)、更新提升树模型：

$$f_m(x)=f_{m-1}(x)+\sum_{j=1}^{J}c_{mj}I(x\in R_{mj})$$

(3)、得到最终的梯度提升树

$$f(x)=f_M(x)=\sum_{m=1}^{M}\sum_{j=1}^{J}c_{mj}I(x\in R_{mj})$$