XGBoost

概述

XGBoost是一个Boosting 加法模型, 在生成树的过程中会迭代生成 m 棵树,共同组成一个模型

公式推导

建模

$$F_m(x_i) = F_{m-1}(x_i) + f_m(x_i)$$

其中

$F_m(x_i)$ 表示迭代m次生成的包含m棵树的总模型

$F_{m-1}(x_i)$ 表示迭代m-1次的总模型;这里注意相对$F_m(x_i)$来说是固定的,**可以认为是常数**.

$f_m(x_i)$ 表示第m棵树;是需要求解的变量.

i 表示特征维度中的某一维

m 表示树的数量或者迭代次数

目标函数

$$
\begin{aligned}
Obj &= \sum_{i=1}^{N} L[F_m(x_i), y_i] + \sum_{j=1}^{m} \Omega(f_j) \\
&= \sum_{i=1}^{N} L[F_{m-1}(x_i) + f_m(x_i), y_i] + \sum_{j=1}^{m} \Omega(f_j) \\
&\approx \sum_{i=1}^{N} \{L[F_{m-1}(x_i), y_i] + \frac{\partial L}{\partial F_{m-1}(x_i)} * f_m(x_i) + \frac{1}{2} * \frac{\partial^2 L}{\partial^2 F_{m-1}(x_i)} * f_m(x_i)^2\} + \sum_{j=1}^{m} \Omega(f_j) \\
&\approx \sum_{i=1}^{N} \{\frac{\partial L}{\partial F_{m-1}(x_i)} * f_m(x_i) + \frac{1}{2} * \frac{\partial^2 L}{\partial^2 F_{m-1}(x_i)} * f_m(x_i)^2\} + \Omega(f_m) \\
&= \sum_{i=1}^{N} \{g_i * f_m(x_i) + \frac{1}{2} * h_i * f_m(x_i)^2\} + \Omega(f_m) \\
&= \sum_{i=1}^{N} \{g_i * f_m(x_i) + \frac{1}{2} * h_i * f_m(x_i)^2\} + \gamma * T_m + \frac{1}{2} * \lambda * \|\vec{\omega_{\mathbf{m}}}\|_2 \\
&= \sum_{i=1}^{N} \{g_i * f_m(x_i) + \frac{1}{2} * h_i * f_m(x_i)^2\} + \gamma * T_m + \frac{1}{2} * \lambda * \sum_{j=1}^{T_m} (\omega_j^m)^2 \\
&= \sum_{i=1}^{T_m} \{(\sum_{i \in I(j)} g_i) * \omega_j^m + \frac{1}{2} * (\sum_{i \in I(j)} h_i) * (\omega_j^m)^2\} + \gamma * T_m + \frac{1}{2} * \lambda * \sum_{j=1}^{T_m} (\omega_j^m)^2 \\
&= \sum_{i=1}^{T_m} \{G_j * \omega_j^m + \frac{1}{2} * H_j * (\omega_j^m)^2\} + \gamma * T_m + \frac{1}{2} * \lambda * \sum_{j=1}^{T_m} (\omega_j^m)^2 \\
&= \sum_{i=1}^{T_m} \{G_j * \omega_j^m + \frac{1}{2} * (H_j + \lambda) * (\omega_j^m)^2\} + \gamma * T_m
\end{aligned}
$$

其中

N 表示样本的总数量

$I(j)$表示样本中落在f_m树上的第j个叶节点上的样本索引集合

$$g_i = \frac{\partial L}{\partial F_{m-1}(x_i)}$$

$$h_i = \frac{\partial^2 L}{\partial^2 F_{m-1}(x_i)}$$

$$\Omega(f) = \gamma * T + \frac{1}{2} * \lambda * \|\vec{\omega}\|_2$$

$$= \gamma * T + \frac{1}{2} * \lambda * \sum_{j=1}^{T} \omega_j^2$$

$$G_j = \sum_{i \in I(j)} g_i$$

$$H_j = \sum_{i \in I(j)} h_i$$

其中

T 表示树f的叶节点的个数

$\vec{\omega}$ 表示树f的所有叶节点输出的回归之组成的向量

γ, λ 表示超参数

极值点

$$\omega_j^* = -\frac{G_j}{H_j + \lambda}$$

树评分

树评分也就是目标函数的输出结果,结果越小,代表树的结构越好

$$Obj^* = \sum_{j=1}^{T_m}(-\frac{1}{2} * \frac{G_j^2}{H_j + \lambda} + \gamma)$$

分裂收益

$$Gain = \frac{1}{2} * [\frac{G_L^2}{H_L + \lambda} + \frac{G_R^2}{H_R + \lambda} - \frac{(G_L + G_R)^2}{H_L + H_R + \lambda}] - \gamma$$

加权分位数

$$Obj = \sum_{i}^{N}[g_i * f_m(x_i) + \frac{1}{2} * h_i * f_m^2(x_i)] + \Omega(f_m)$$

$$= \sum_{i}^{N} \frac{1}{2} * h_i * [f_m(x_i) - (-\frac{g_i}{h_i})]^2 + \Omega(f_m)$$