XGBoost 实验报告

PB20010429 侯相龙

2022 年 11 月 20 日

1 实验内容

实现 XGBoost 回归

2 实验原理

2.1 XGBoost

XGBoost 使用集成学习的原理，由多个基模型（在这里我们使用的是回归树）组成的一个加法模型。 假设第k 个基本模型是$f_k(x)$ ，那么前t 个模型组成的模型的输出为

$$\hat{y}_i^{(t)} = \sum_{k=1}^{t} f_k\left(x_i\right) = \hat{y}_i^{(t-1)} + f_t\left(x_i\right)$$

目标优化函数为：

$$Obj^{(t)} = \sum_{i=1}^{n} \operatorname{loss}\left(y_i, \hat{y}_i^{(t)}\right) + \sum_{k=1}^{t} \operatorname{penalty}\left(f_k\right)$$

其中损失函数 loss 选择平方损失函数 $\operatorname{loss}\left(y_i, \hat{y}_i^{(t)}\right) = \frac{1}{2}\left(y_i - \hat{y}_i^{(t)}\right)^2$ 利用目标函数的 Taylor 展开进行估计，并舍去常数项，我们的优化目标为：

$$\operatorname{Obj}^{(t)} = \sum_{i=1}^{n} \left[g_i f_t\left(x_i\right) + \frac{1}{2} h_i f_t^2\left(x_i\right)\right] + \operatorname{penalty}\left(f_t\right)$$

其中，g_i 为 $\operatorname{loss}\left(y_i, \hat{y}_i^{(t-1)}\right)$ 的一阶导数,h_i 为 $\operatorname{loss}\left(y_i, \hat{y}_i^{(t-1)}\right)$ 的二阶导数

2.2 回归树

对于回归树，我们令罚项为：

$$\operatorname{penalty}(f) = \gamma \cdot T + \frac{1}{2}\lambda \cdot ||w||^2$$

其中γ, λ 为我们可调整的超参数，T 为叶子数 依叶节点改变求和次序，有：

$$Obj^{(t)} = \sum_{j=1}^{T} \left[G_j w_j + \frac{1}{2}\left(H_j + \lambda\right) w_j^2\right] + \gamma T \tag{1}$$

其中，$I_{j\,j=1}^{T}$ 为叶子节点集合,$G_j = \sum_{i\in I_j} g_i, H_j = \sum_{i\in I_j} h_i$.(1) 式是关于 w 的二次函数进行优化，当 $\omega_j^* = -\frac{G_j}{H_j+\lambda}$时，有最优解：

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

3 实验步骤

3.1 数据处理

读取数据，依 training_rate=0.7 划分训练集和测试集

3.2 回归树

这里详细介绍一下**递归构建树**的算法，其余从略。

Step 1： 判断是否停止递归建树（**分配到的样本数目小于某个阈值**停止分化或**划分后树的深度大于某个阈值**），若是则退出。

Step 2： 选择最大增益的最佳的特征以及取值。具体来说，遍历所有属性及该分配到该节点所有数据相应属性的值，假设在此点划分，调用 score 函数计算增益 Gain $= Obj_P - Obj_L - Obj_R$, 根据上面推导，Obj 为：$-\frac{1}{2}\frac{G^2}{H+\lambda} + \gamma$。选择增益最大者，记录属性和值

Step 3： 判断**最大增益是否大于 0**，若大于 0 进行 Step 4，否则退出

Step 4： 将划分后样本的 x、g、h 信息分别赋给左右孩子，对于样本量大于某个阈值的子节点递归建树，并令深度 +1，转到 Step 1.

除此之外，还有寻找节点函数（二分检索），并基于此的给定 x 预测函数，此处不再赘述。

3.3 XGBoost

基于上面构造的回归树类，用 XGBoost 加法模型回归算法

Step 1: 指定**基学习器数量 M** 以及相应的参数 γ, λ 值。

Step 2: 初始化预测值 $\hat{y}$（全 0），以及基于此计算出的 loss 一阶导数 g：$y - \hat{y}$, 二阶导数 h:1 （选取的是平方损失）

Step 3: 若未达到退出条件（迭代次数：M），利用 g, h 的信息和训练数据构造基学习器——回归树；否则退出。

Step 4: 利用新学习的基学习器预测函数和上一步的预测值更新预测值 $\hat{y}$。并利用更新得到的预测值更新 g, h(方法同 Step2)

3.4 参数调试及比较

选取不同参数:基学习器数量 M、最大深度、超参数 γ, λ, 并用 $RMSE = \sqrt{\frac{1}{m}\sum_{i=1}^{m}\left(y_{\text{test}}^{(i)} - \hat{y}_{\text{test}}^{(i)}\right)^2}$ 作为指标衡量。具体见**实验结果**部分。

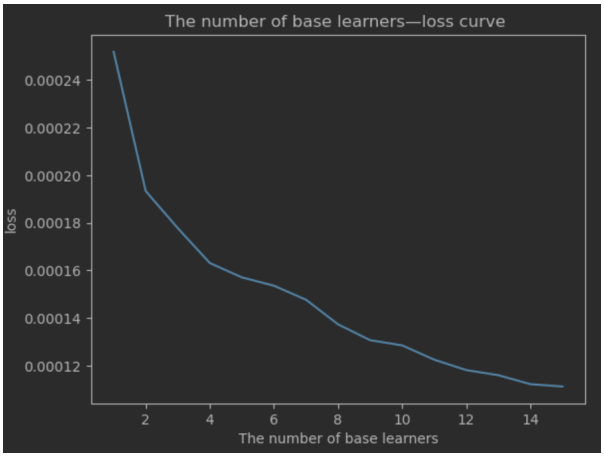
4 实验结果

4.1 最佳模型实验结果:

$\lambda = 1e-7, \gamma = 1e-9, \text{max\_depth} = 4, M = 15.$

测试结果:RMSE=1.88E-04

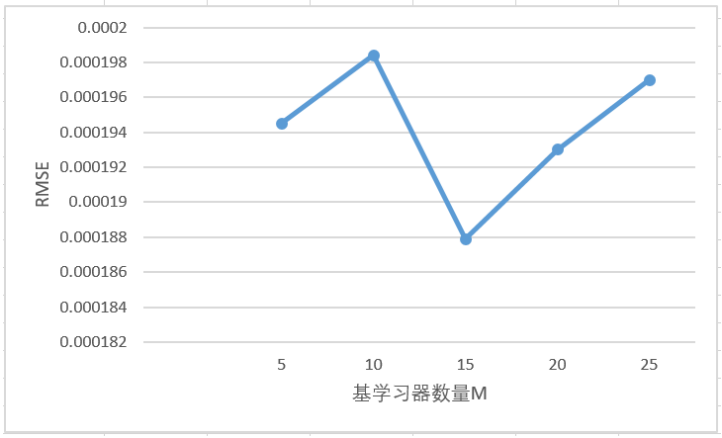
训练过程中的 loss function:



4.2 基学习器数量 M-RMSE:

$\lambda = 1e-7, \gamma = 1e-9, max\_depth = 4$

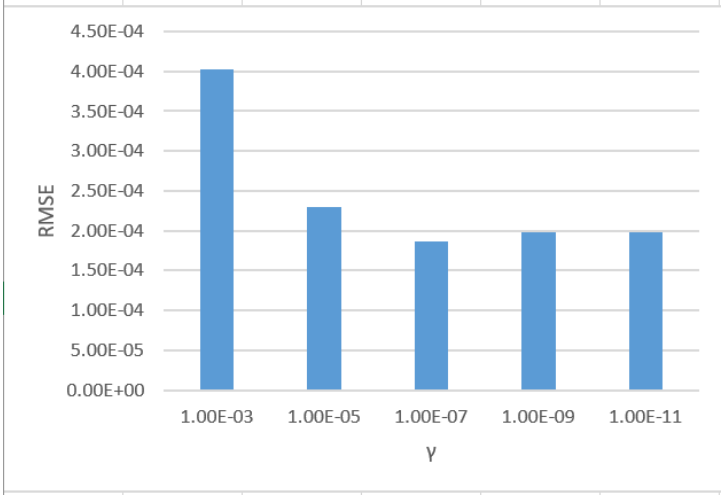
基学习器数量M	5	10	15	20	25
RMSE	1.95E-04	1.98E-04	1.88E-04	1.93E-04	1.97E-04



4.3 γ-RMSE:

$\lambda = 1e-7, M = 10, max\_depth = 4$

γ	1.00E-03	1.00E-05	1.00E-07	1.00E-09	1.00E-11
RMSE	4.03E-04	2.30E-04	1.87E-04	1.98E-04	1.98E-04



4.4 λ-RMSE:

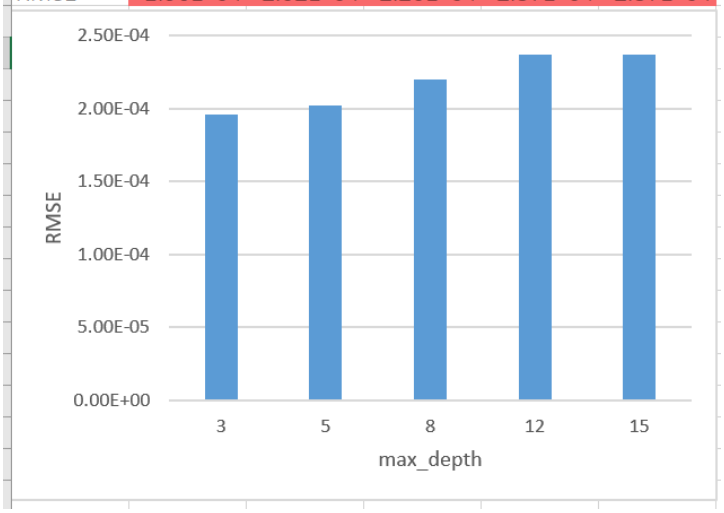
$\gamma = 1e-9, M = 4, max\_depth = 4$

λ	1.00E-03	1.00E-05	1.00E-07	1.00E-09	1.00E-11
RMSE	1.98E-04	1.98E-04	1.98E-04	1.98E-04	1.99E-04

4.5 max_depth-RMSE:

$\gamma = 1e-9, \lambda = 1e-7, M = 8$

max_depth	3	5	8	12	15
RMSE	1.96E-04	2.02E-04	2.20E-04	2.37E-04	2.37E-04



5 实验分析

i) 出现拐点的图像曲线可能是由于：随着值的变化从欠拟合变为过拟合

ii) 注意到样本标记值较小，故我们罚项中的 γ 也应该较小，否则就会" 早停"，导致欠拟合。