水滴石穿

探索，保持渴望，无所畏惧

首页 归档 关于 订阅

XGboost: A Scalable Tree Boosting System论文及源码导读

Oct 5, 2016 | 机器学习 | 3990 Hits

这篇论文一作为陈天齐，XGBoost是从竞赛pk中脱颖而出的算法，目前开源在github，和传统gbdt方式不同，XGBoost对***loss function***进行了二阶的泰勒展开，并增加了正则项，用于权衡目标函数的下降和模型的复杂度[12]。罗列下优势：

1. 可扩展性强
2. 为稀疏数据设计的决策树训练方法
3. 理论上得到验证的加权分位数略图法
4. 并行和分布式计算
5. 设计高效核外计算，进行cache-aware数据块处理

分布式训练树模型boosting方法已有[1,2,3]。

整体目标

$$L(\phi) = \sum_i l(y_i, \hat{y}_i) + \sum_k \Omega(f_k)$$

其中$L(\cdot)$为目标函数，$l(\cdot)$是损失函数，通常是凸函数，用于 刻画预测值$\hat{y}_i$和真实值y_i的差异，第二项$\Omega(\cdot)$为模型的正则化项， 用于降低模型的复杂度，减轻过拟合问题，类似的正则化方法可以在引文[4]看到。模型目标是最小化目标函数。
$L(\cdot)$为函数空间上的表达，我们可以将其转换为下面这张***gradient boosting***的方式，记$\hat{y}_i^{(t)}$为第i个样本第t轮迭代：

$$L^{(t)} = \sum_{i=1}^{n} l\left(y_i, \hat{y}_i{}^{(t-1)} + f_t(\mathbf{x}_i)\right) + \Omega(f_t)$$

对该函数在$\hat{y}_i^{(t)}$位置进行二阶泰勒展开，可以加速优化过程，我们得到目标函数的近似：

$$L^{(t)} \simeq \sum_{i=1}^{n} \left[l\left(y_i, \hat{y}^{(t-1)}\right) + g_i f_t(\mathbf{x}_i) + \frac{1}{2} h_i f_t^2(\mathbf{x}_i) \right] + \Omega(f_t)$$