# 模型背景

XGBoost，全称Extreme Gradient Boosting，即极限梯度提升。它是一种基于GBDT的算法，或者说工程实现，但它对一般的GBDT算法做了一定的优化，让其能够更快，更有效率的解决问题。该模型在陈天奇博士的论文《XGBoost：A Scalable Tree Boosting System》中正式提出。

# 模型原理

## 基本思想

在XGBoost的基本思想与GBDT相同，但是XGBoost进行了许多优化，比如：

- 利用二阶泰勒公式展开，以优化损失函数，提高计算的精确度。

- 利用正则项，简化模型，避免过拟合情况的出现。

- 采用了Blocks存储结构，以进行并行运算等。

## 目标函数推导

已知训练数据集，损失函数,正则化项,整体函数可以记为：

其中：

- 是线性空间上的表达。

- 是第个样本，是第课树。

- 是第个样本的预测值。

由于

则转换成如下形式：

接下来，三个步骤来优化XGBoost目标函数。

第一步：二阶泰勒展开，去除常数项，优化损失函数项；

第二步：正则化项展开，去除常数项，优化正则化项；

第三步：合并一次项系数，二次项系数，得到最终目标函数。

第一步：

我们了解泰勒展开如下：

我们仅取其二阶泰勒展开，并将其对使用得到：

类似的，我们取得到：

令，且记一阶导数为，二阶导数为。

得到的二阶泰勒展开：

带入目标函数可得到：

第二步：

关键在于前向计算，t-1棵树的结构已经确定。

由于是一个常数项，所以移除，目标函数可以简化为：

将正则项拆分，得到：

C为一个常数，我们将其移除，并进一步简化为：

第三步：

首先，我们重新定义一棵树，包括两部分：

- 叶子结点的权重向量为；

- 叶子结点的映射关系。

则一棵树的表达形式为：

其中，为树叶子的权重，为树的结构。

接下来定义一棵树的复杂度，它由两部分组成：

- 叶子结点的数量；

- 叶子结点的权重向量的范数。

数学语言描述如下：

其中，为叶子的总数，为叶子结点的权重向量的范数。

接着将以上内容带入目标函数中，对结点进行分组。

将属于第个叶子结点的所有样本，划入到一个叶子结点的集合中，数学描述如下：

带入目标函数得到：

将所有训练样本，按叶子结点进行分组得到：

最后合并一次项系数，二次项系数。

定义如下内容：

其中，

为叶子节点所包含样本的一阶偏导数累加之和，是一个常数。

为叶子节点所包含样本的一阶偏导数累加之和，是一个常数。

将以上内容带入目标函数，化简得到：

上式即XGBoost的最终目标函数。

## 目标函数解

由以上推导我们得到了XGBoost的目标函数，我们设每个叶子结点的目标函数为：

可以发现这是一个关于的一元二次函数。可以求其最小值。

当时候，则在处取得最小值，最小值为。

目标值最小，则树的结构最好，此时即目标函数的最优解。

同时XGBoost目标函数的的各个叶子结点的目标式子是相互独立的。即每个叶子结点的式子都达到最值点，整个目标函数也达到最值点。

则每个叶子的结点的权重都达到的时候

同时能达到最优的目标值

## 树的训练

在实际训练中，最佳分裂点是一个关键问题，我们通常会采取如下分裂方式并则取最优：

- 贪心算法

- 近似算法

- 加权分位数草图法

- 稀疏感知法