

# An Introduction to XGBoost

王社英

April 17, 2020

# Outline

XGBoost Introduction

Decision Tree Algorithm

Gradient Boosting

Tree Boosting

# Outline

## XGBoost Introduction

### Decision Tree Algorithm

ID3

C4.5

CART

### Gradient Boosting

Gradient Boosting

Gradient Boosting Boosting

### Tree Boosting

Additive Training

Model Complexity

The Structure Score

Learn the tree structure

# XGBoost Introduction

XGBoost[CG16] is a scalable end to end tree boosting system.

- a novel sparsity aware algorithm for sparse data
- weighted quantile sketch for approximate tree learning
- insights on cache access patterns, data compression and sharding

By combining these insights, XGBoost scales beyond billions of examples using far fewer resources than existing systems.

# Outline

## XGBoost Introduction

## Decision Tree Algorithm

ID3

C4.5

CART

## Gradient Boosting

Gradient Boosting

Gradient Boosting Boosting

## Tree Boosting

Additive Training

Model Complexity

The Structure Score

Learn the tree structure

# Decision Tree Algorithm

## ID3

Apache Spark is a unified analytics engine for large-scale data processing.

# Decision Tree Algorithm

## C4.5

Apache Spark is a unified analytics engine for large-scale data processing.

# Decision Tree Algorithm

## CART

Apache Spark is a unified analytics engine for large-scale data processing.



# Outline

XGBoost Introduction

Decision Tree Algorithm

ID3

C4.5

CART

Gradient Boosting

Gradient Boosting

Gradient Boosting Boosting

Tree Boosting

Additive Training

Model Complexity

The Structure Score

Learn the tree structure

# Gradient Boosting

## Gradient Boosting

### 算法5-1 Gradient Boosting算法流程

输入：训练集，损失函数  $L(y, F(x))$ , 训练轮数  $M$ 。

输出：最终模型  $F_M(x)$ 。

算法：

1) 通过常数初始化模型。

$$F_0(x) = \arg \min_{\gamma} \left( \sum_{n=1}^N L(y_i, \gamma) \right)$$

2) 对  $m = 1, 2, \dots, M$ , 执行以下步骤。

① 计算负梯度：

$$r_{im} = - \left[ \frac{\partial L(y_i, F(x_i))}{\partial F(x_i)} \right]_{F(x)=F_{m-1}(x)}, i = 1, 2, \dots, n$$

② 训练一个子模型  $h_m$ , 用来拟合  $r_{im}$ ,

③ 计算步长  $\gamma_m$ :

$$\gamma_m = \arg \min_{\gamma} \left( \sum_{n=1}^N L(y_i, F_{m-1}(x_i) + \gamma h_m(x_i)) \right)$$

④ 更新模型：

$$F_m(x) = F_{m-1}(x) + \gamma_m h_m(x)$$

3) 输出  $F_M(x)$ 。

# Gradient Tree Boosting

## Gradient Tree Boosting[xgb01]

### 算法5-2 Gradient Tree Boosting算法的执行过程

输入：训练集，损失函数  $L(y, F(x))$ ，训练轮数  $M$ 。

输出：最终模型  $F_M(x)$ 。

算法：

1) 通过损失函数最小化初始化模型：

$$F_0(x) = \arg \min_{\gamma} \left( \sum_{n=1}^N L(y_n, \gamma) \right)$$

2) 对  $m=1, 2, \dots, M$ ，执行以下步骤。

① 计算负梯度：

$$r_{im} = - \left[ \frac{\partial L(y_i, F(x_i))}{\partial F(x_i)} \right]_{F(x)=F_{m-1}(x)}, i = 1, 2, \dots, n$$

② 训练一个回归树去拟合目标值  $r_{im}$ ，树的终端区域为  $R_{jm}$  ( $j = 1, 2, \dots, J_m$ )。

③ 对  $j = 1, 2, \dots, J_m$ ，计算步长  $\gamma_{jm}$ 。

$$\gamma_{jm} = \arg \min_{\gamma} \left( \sum_{x_i \in R_{jm}} L(y_i, F_{m-1}(x_i) + \gamma) \right)$$

④ 更新模型：

$$F_m(x) = F_{m-1}(x) + \sum_{j=1}^{J_m} \gamma_{jm} I(x \in R_{jm})$$

3) 输出  $F_M(x)$ 。

# Outline

XGBoost Introduction

Decision Tree Algorithm

ID3

C4.5

CART

Gradient Boosting

Gradient Boosting

Gradient Boosting Boosting

Tree Boosting

Additive Training

Model Complexity

The Structure Score

Learn the tree structure

# Tree Boosting

## Additive Training

Apache Spark is a unified analytics engine for large-scale data processing.

# Tree Boosting

## Model Complexity

Apache Spark is a unified analytics engine for large-scale data processing.

# Tree Boosting

## The Structure Score

Apache Spark is a unified analytics engine for large-scale data processing.

# Tree Boosting

Learn the tree structure

Apache Spark is a unified analytics engine for large-scale data processing.



# Questions and Answers?

Questions and Answers?

Thank You!

# References I



Tianqi Chen and Carlos Guestrin.

Xgboost: A scalable tree boosting system.

*In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2016.*



深入理解 *XGBoost*: 高效机器学习算法与进阶.

何龙, 2020-01-01.