Mathematical Background XGBoost

by Nina Immenroth

2025-12-23

Abstract

This is a short section about the mathematical background of XGBoost. To be added as a subsection of the report on Coffee Quality Prediction for the course ML2.

Contents

1 XGBoost

XGBoost is an algorithm based on gradient boosted trees and second-order approximation (Chen 2016). It is widely used in machine learning due to its flexibility and strong empirical performance. These properties arise from a combination of algorithmic and systems-level optimizations, including efficient handling of sparse input data, approximate tree construction via weighted quantile sketches, and parallelized tree learning. In addition, careful optimization of memory access, data compression, and out-of-core computation enables the method to scale to very large datasets while maintaining computational efficiency.

This section describes the mathematical foundations of the algorithm, namely ensemble methods, regularization, and gradient boosting machines combined with second order approximation.

1.1 Tree Ensembles and Regularization

XGBoost models the prediction as an additive ensemble of K functions:

$$\hat{y}_i = \sum_{k=1}^{K} f_k(x_i), \quad f_k \in \mathcal{F},$$

where each f_k is a regression tree belonging to the function space $\mathcal{F}$ (Chen 2016; James et al. 2021). Each tree acts as a weak learner that partitions the feature space into disjoint regions and assigns a constant prediction to each leaf (James et al. 2021).

The model is trained by minimizing a regularized objective function of the form

$$\mathcal{L} = \sum_i \ell(y_i, \hat{y}_i) + \sum_{k=1}^{K} \Omega(f_k),$$

where $\ell(\cdot)$ is a differentiable loss function and $\Omega(f)$ is a regularization term that penalizes model complexity (Chen 2016). This term is defined as

$$\Omega(f) = \gamma T + \frac{1}{2}\lambda \sum_{j=1}^{T} w_j^2,$$

with T denoting the number of leaves in the tree and w_j the weight of leaf j. The regularization term encourages simpler trees and smooth leaf weights, helping to prevent overfitting.

The ensemble structure allows the model to reduce bias by combining multiple weak learners, while regularization controls variance by limiting tree complexity and the magnitude of leaf weights (James et al. 2021). This balance between bias and variance enables XGBoost to achieve strong predictive performance while maintaining good generalization.

1.2 Gradient Boosting Machines

The tree ensemble is constructed in a stage-wise manner, where weak learners are added iteratively to improve the model (Friedman 2001). At iteration t, the prediction is given by

$$\hat{y}_i^{(t)} = \hat{y}_i^{(t-1)} + f_t(x_i),$$

with $f_t \in \mathcal{F}$ denoting the newly added regression tree.

Gradient tree boosting can be interpreted as gradient descent in function space, where each new learner is chosen to minimize the loss function in the direction of the negative gradient (Friedman 2001). Specifically, the new tree at iteration t is trained to fit the negative first order gradient

$$-g_i^{(t)} = -\left.\frac{\partial \ell(y_i, \hat{y})}{\partial \hat{y}}\right|_{\hat{y}=\hat{y}_i^{(t-1)}}.$$

XGBoost extends this framework by explicitly approximating the objective function using a second-order Taylor expansion of the loss around the current predictions:

$$\ell(y_i, \hat{y}_i^{(t-1)} + f_t(x_i)) \approx \ell(y_i, \hat{y}_i^{(t-1)}) + g_i f_t(x_i) + \frac{1}{2} h_i f_t(x_i)^2,$$

where

$$h_i = \left. \frac{\partial^2 \ell(y_i, \hat{y})}{\partial \hat{y}^2} \right|_{\hat{y}=\hat{y}_i^{(t-1)}}$$

is the second order derivative of the loss (Chen 2016). By incorporating both first- and second-order information, XGBoost enables efficient optimization of tree structures and leaf weights within each boosting iteration.

2 Literature

Chen, Tianqi. 2016. "XGBoost: A Scalable Tree Boosting System." *Cornell University*.

Friedman, Jerome H. 2001. "Greedy Function Approximation: A Gradient Boosting Machine." *Annals of Statistics*, 1189–1232.

James, Gareth, Daniela Witten, Trevor Hastie, and Robert Tibshirani. 2021. *An Introduction to Statistical Learning: With Applications in r*. 2nd ed. Springer. https://doi.org/10.1007/978-1-0716-1418-1.