gbdt

Xi Chen

xchen595@163.com

Taylor formulation

- definition: A formula makes uses of infomration from a function at a point to describe its value nearby.

$$f(x) = \sum_{n=0}^{\infty} \frac{f^n(x_0)}{n!}(x - x_0)^n$$

- Iteration form:

assume $x^t = x^{t-1} + \Delta t$, then $f(x^t) = f(x^{t-1} + \Delta t) \approx f(x^{t-1}) + f^{'}(x^{t-1})\Delta t + f^{''}(x^{t-1})\frac{\Delta x^2}{2}$

Gradient Descent Method

- Iteration formula: $\theta^t = \theta^{t-1} + \Delta\theta$
- loss function taylor form at θ^{t-1}

$$L(\theta^t) = L(\theta^{t-1} + \Delta\theta) \approx L(\theta^{t-1}) + L'(\theta^{t-1})\Delta\theta$$

- We want $L(\theta^t) < L(\theta^{t-1})$, we can have $\Delta\theta = -\alpha L'(\theta^{t-1})$, i.e. $\theta^t = \theta^{t-1} - \alpha L'(\theta^{t-1})$

Newton's Method

- Second order Taylor expansion:

$$L(\theta^t) \approx L(\theta^{t-1}) + L'(\theta^{t-1})\Delta\theta + L''(\theta^{t-1})\frac{\Delta\theta^2}{2}$$

For simplicity, if θ is a scalar, then $L(\theta^t) \approx L(\theta^{t-1}) + g\Delta\theta + h\frac{\Delta\theta^2}{2}$

- In order to make $L(\theta^t)$ minimal, we let $g\Delta\theta + h\frac{\Delta\theta^2}{2}$ minimal, i.e. let $\frac{\partial\{g\Delta\theta + h\frac{\Delta\theta^2}{2}\}}{\partial\Delta\theta} = 0$

we have $\Delta\theta = -\frac{g}{h}$, so $\theta^t = \theta^{t-1} + \Delta\theta = \theta^{t-1} - \frac{g}{h}$
and matrix form: $\theta^t = \theta t - 1 - H^{-1}g$

From Parameter Space to Functional Space

- GBDT optimizes on functional space by gradient descent method
- XGBOOST optimizes on functional space by Newton's method

From Gradient Descend to Gradient Boosting

$$f^t(x) = f^{t-1}(x) + f_t(x)$$
$$f_t(x) = -\alpha_t g_t(x)$$
$$F(x) = \sum_{t=0}^{T} f_t(x)$$

Here, f_0 is a constant

From Newton's Method to Newton Boosting

$$f^t(x) = f^{t-1}(x) + f_t(x)$$

$$f_t(x) = -\frac{g_t(x)}{h_t(x)}$$

$$F(x) = \sum_{t=0}^{T} f_t(x)$$

Here, f_0 is a constant

Summary

- Boosting is an additive training
- Base classifier, f usually uses regression tree and logistic regression
- advantages: interpretability, mixture features, no-normalization, feature combinations, handling missing values, robust on outliers, feature selection, easy parallel
- disadvantages: lack of smoothness, not fit for high-dimensional sparse data