Week 7: Gradient Boosting for Regression

Kevin Lee

Department of Statistics
Western Michigan University

What is Gradient Boosting?

Gradient Boosting = Gradient Descent + Boosting

- Fit an additive model (ensemble) in a forward stage-wise manner.
- In each stage, introduce a weak leaner to compensate the shortcomings of existing weak learners.
- In Gradient Boosting, shortcomings are identified by gradients.
- Gradients tell us how to improve our model.

Gradient Boosting for Regression

Let's play a game!

- You are given $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$, and the task to fit a model $F(x)$ to minimize square loss.
- Suppose your friend wants to help you and give you a model F.
- You check his model and find the model is good but not perfect.
- There are some mistakes: $F(x_1) = 0.8$ while $y_1 = 0.9$ and $F(x_2) = 1.4$ while $y_2 = 1.3$ ...
- How can you improve this model?

Gradient Boosting for Regression

Rule of the game:

- You are not allowed to remove anything from F or change any parameter in F.
- You can add an additional model h (regression tree) to F, so the new prediction will be $F(x) + h(x)$.

Gradient Boosting for Regression

Simple solution:

- You wish to improve the model such that

$$F(x_1) + h(x_1) = y_1$$
$$F(x_2) + h(x_2) = y_2$$
$$\vdots$$
$$F(x_n) + h(x_n) = y_n$$

Gradient Boosting for Regression

Simple solution:

- Or, equivalently, you wish

$$h(x_1) = y_1 - F(x_1)$$
$$h(x_2) = y_2 - F(x_2)$$
$$\vdots$$
$$h(x_n) = y_n - F(x_n)$$

- Fit a regression tree h to data
$(x_1, y_1 - F(x_1)), (x_2, y_2 - F(x_2)), \ldots, (x_n, y_n - F(x_n))$

  Congratulations, You get a better model!

Gradient Boosting for Regression

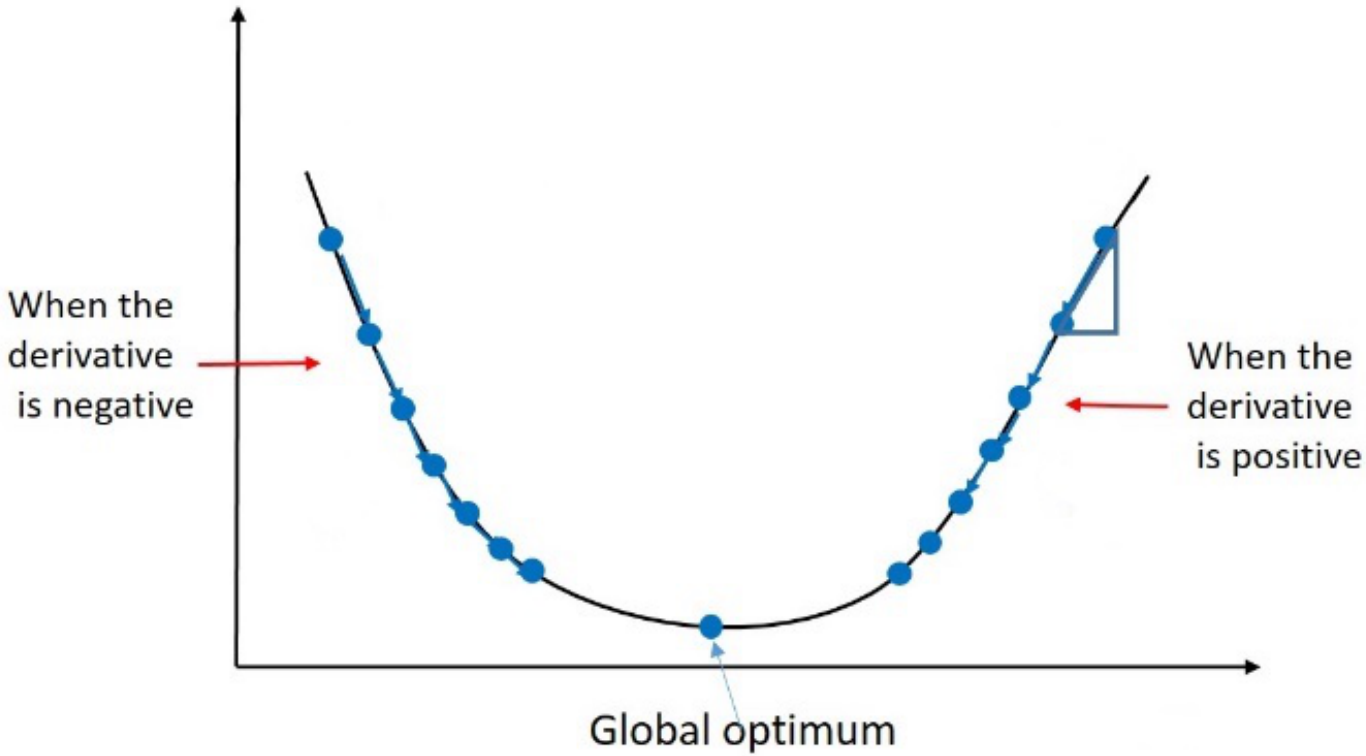
- $y_i - F(x_i)$ are called residuals. These are the parts that existing model F cannot do well.
- The role of h is to compensate the shortcoming of existing model F.
- If the new model $F + h$ is still not satisfactory, we can add another regression tree!
- How is this related to gradient descent?

Gradient Descent

- Gradient descent is a first-order iterative optimization algorithm for finding the minimum of a function.
- To find a local minimum of a function using gradient descent, one takes steps proportional to the negative of the gradient of the function at the current point.
- If, instead, one takes steps proportional to the positive of the gradient, one approaches a local maximum of that function; the procedure is then known as gradient ascent.

$$a_{n+1} = a_n - \gamma \nabla F(a_n)$$

Gradient Descent



Gradient Boosting for Regression

How is this related to gradient descent?

- Let's assume we have loss function, $L(y, F(x)) = (y - F(x))^2/2$
- We want to minimize $J = \sum_{i=1}^{n} L(y_i, F(x_i))$ by adjusting $F(x_1), F(x_2), \ldots, F(x_n)$.
- We can treat $F(x_i)$ as parameters and take derivatives.

$$\frac{\partial J}{\partial F(x_i)} = \frac{\partial \sum_{i=1}^{n} L(y_i, F(x_i))}{\partial F(x_i)} = \frac{\partial L(y_i, F(x_i))}{\partial F(x_i)} = F(x_i) - y_i$$

- So we can interpret residuals as negative gradients!

$$y_i - F(x_i) = -\frac{\partial J}{\partial F(x_i)}$$

Gradient Boosting for Regression

How is this related to gradient descent?

- For regression with square loss,

  residual $<=>$ negative gradient

  fit h to residual $<=>$ fit h to negative gradient

  update F based on residual $<=>$ update F based on negative gradient

- So we actually updating our model using gradient descent!
- The concept of gradients is more general and useful than the concept of residuals.

Gradient Boosting for Regression

Loss Functions for Regression Problem

- Why do we need to consider other loss functions? Isn't square loss good enough?
- Square loss is:
  - Easy to deal with mathematically
  - Not robust to outliers. Pay too much attentions to outliers. Try hard to incorporate outliers into the model and degrade the overall performance.

Gradient Boosting for Regression

Loss Functions for Regression Problem

- Absolute loss (more robust to outliers)

$$L(y, F(x) = |y - F(x)|$$

- Huber loss (more robust to outliers)

$$L(y, F(x) = \begin{cases} \frac{1}{2}(y - F(x))^2 & |y - F(x)| \leq \delta \\ \delta(|y - F(x)| - \frac{1}{2}\delta) & |y - F(x)| > \delta \end{cases}$$

Gradient Boosting for Regression

Loss Functions for Regression Problem

y_i	0.5	1.2	2	5*
$F(x_i)$	0.6	1.4	1.5	1.7
Square loss	0.005	0.02	0.125	5.445
Absolute loss	0.1	0.2	0.5	3.3
Huber loss($\delta = 0.5$)	0.005	0.02	0.125	1.525

Gradient Boosting for Regression

Summary

- Fit an additive model (ensemble) in a forward stage-wise manner.
- In each stage, introduce a new regression tree h to compensate the shortcomings of existing model.
- The shortcomings are identified by negative gradients.
- For any loss function, we can derive a gradient boosting algorithm.
- Absolute loss and Huber loss are more robust to outliers than square loss.

Reference

- "A Gentle Introduction to Gradient Boosting" by Cheng Li, Northeastern University