Introduction to Gradient Boosting

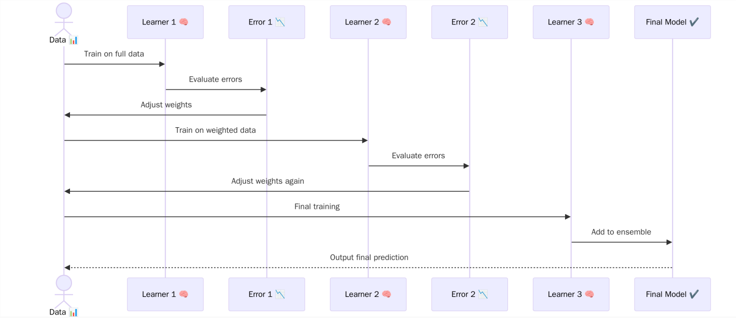
Building Strong Models from Weak Learners

May 29, 2025

What is Gradient Boosting?

- Gradient Boosting is an ensemble machine learning technique that builds a strong model from an ensemble of weaker models.
- The method works by sequentially adding models that correct the residual errors of prior models.
- Instead of fitting a single model to the data, it builds the model stage-by-stage, focusing on the data points where earlier models performed poorly.
- Each new model is trained to predict the residuals (errors) of the previous combined model.

Gradient Boosting Visualization



Block diagram representing the step-by-step refinement in Gradient Boosting.

Block Diagram of Gradient Boosting

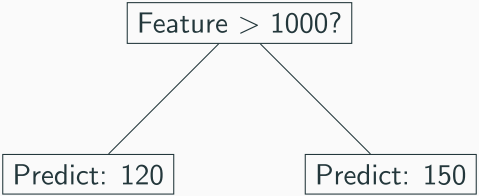
Model 1 → Residuals → Model 2 → Improved Prediction

- **Model 1:** A weak learner (e.g., shallow decision tree) makes an initial prediction.
- **Residuals:** Errors between predictions and actual targets are computed.
- **Model 2:** A new model is trained to predict these residuals.
- **Improved Prediction:** The new model's output is added to the original to refine the prediction.
- This process continues iteratively, reducing error step-by-step.

What is a Weak Learner?

- A model that performs slightly better than random guessing.
- Example: a decision stump (a tree with one split).
- Individually weak, but powerful when combined.

Example of a Weak Learner (Decision Tree)



- Weak learners are shallow trees or simple models.
- They are combined to form a powerful ensemble.


Model 1
Residuals
Model 2
Improved Prediction

Recap: Linear Regression

- Predicts values using:

$$\hat{y} = a + b \cdot x = 75 + 0.05 \cdot x$$

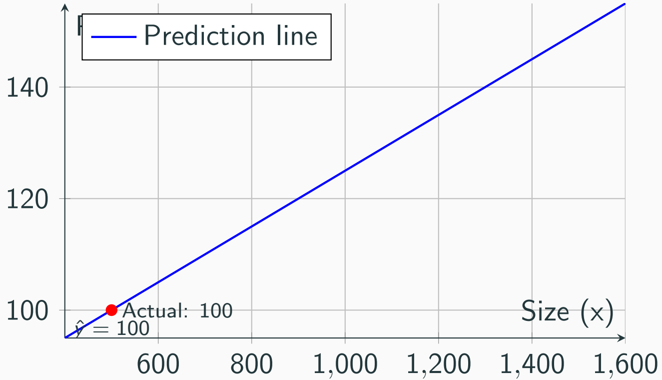
- Example: for a house size of $x = 1000$,

$$\hat{y} = 75 + 0.05 \cdot 1000 = 125$$

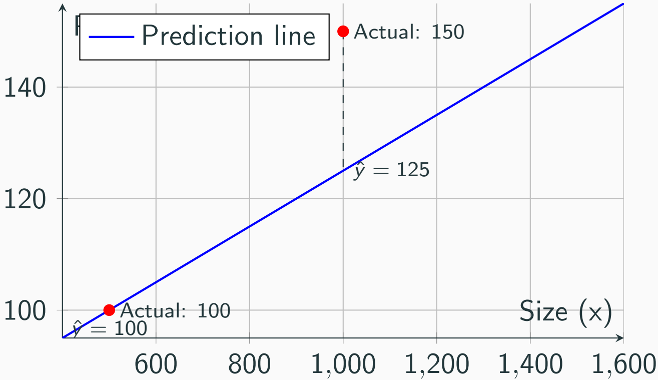
- Residual: difference between actual and predicted value:

$$r_i = y_i - \hat{y}_i = 150 - 125 = 25$$

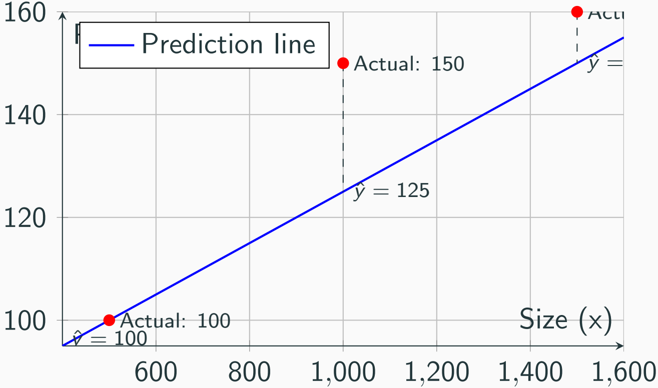
Plot: Price vs Size



Plot: Price vs Size



Plot: Price vs Size



Learning Iteratively

$$\hat{y}^{(t)} = \hat{y}^{(t-1)} + \eta \cdot h_t(x)$$

- η: learning rate
- $h_t(x)$: weak learner at step t

Toy Example - House Prices

Size (sq ft)	Price (1000s)
500	100
1000	150

Model 1 predicts: 120 for both

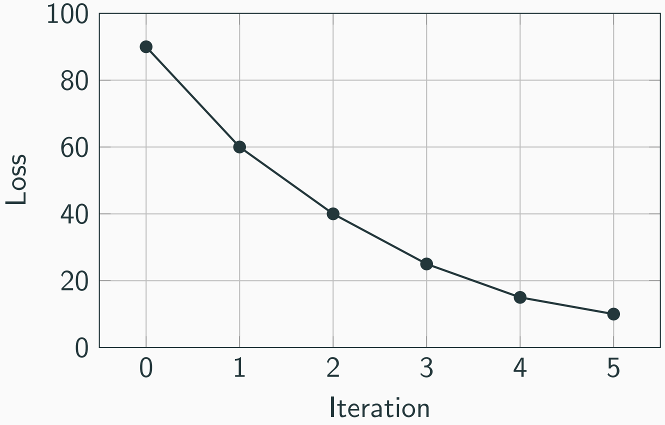
Residuals: -20 and +30

Loss Function

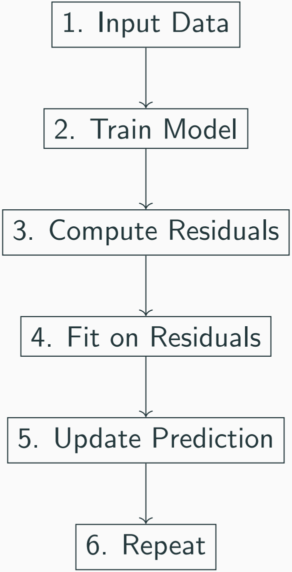
$$L = \frac{1}{n} \sum (y_i - \hat{y}_i)^2$$

- Measures how well predictions match targets
- Gradient Boosting minimizes this iteratively

Residual Reduction Over Iterations



Gradient Boosting Process Flow



Conclusion

- Builds strong models from weak learners
- Effective for structured data
- Widely used in practice: XGBoost, LightGBM, CatBoost

Q&A

Questions?