Ensembles: Boosting Models

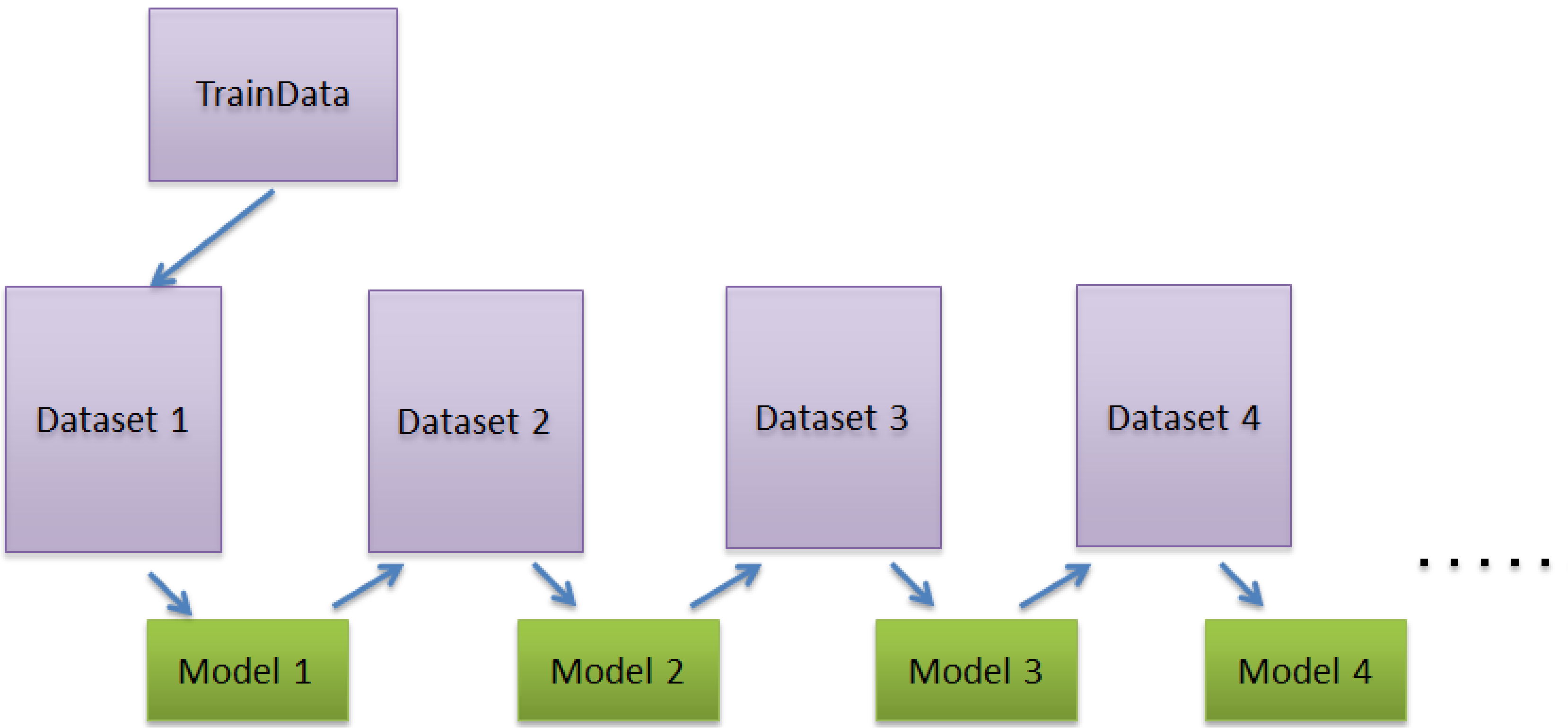
Boosting: Mixture of Experts

Similar to bagging, but uses a more sophisticated method for constructing its diverse training sets.

Main ideas:

- Train the next classifier on examples that previous classifiers made errors on.
- Assign each classifier a confidence value that depends on its accuracy.

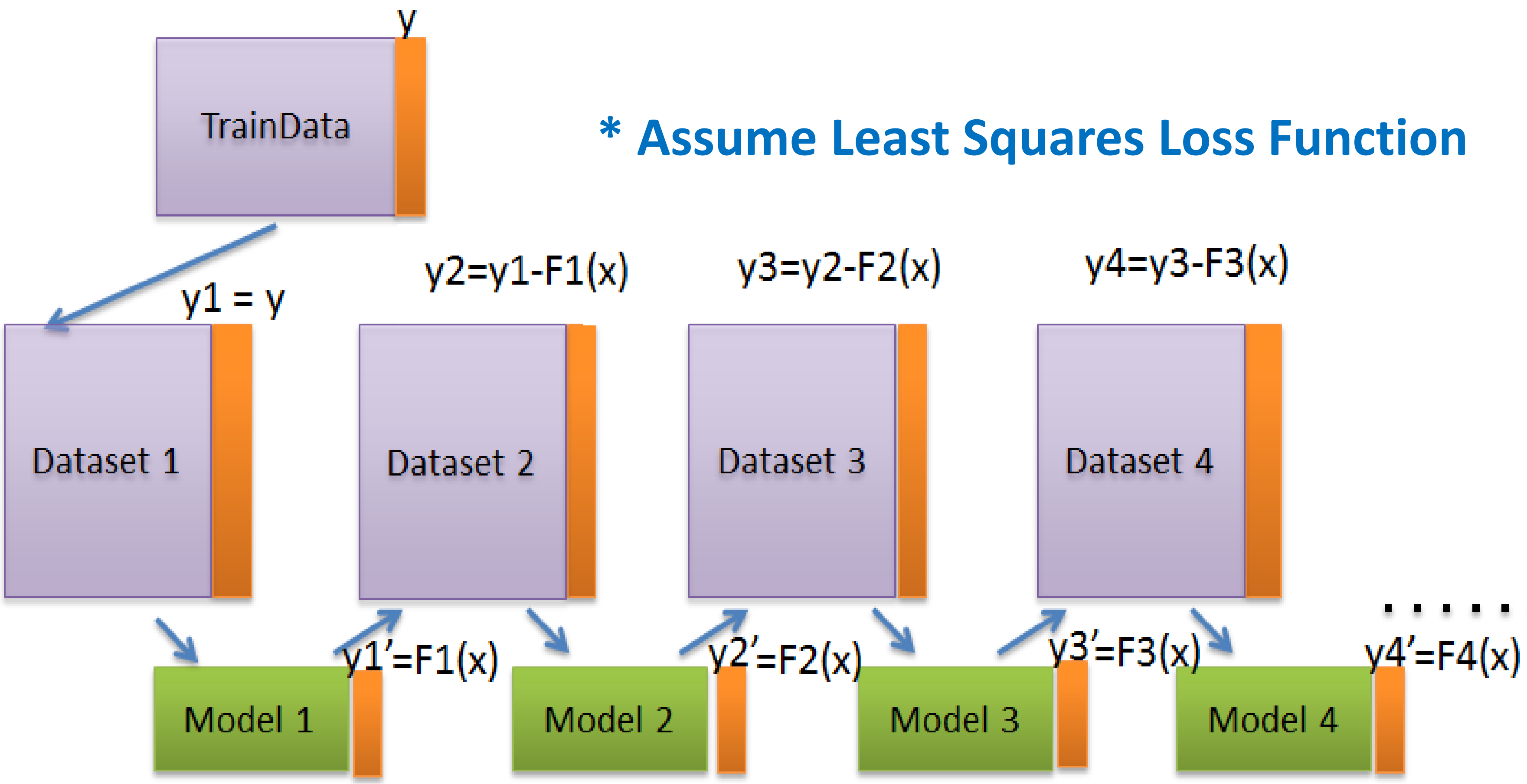
Boosted Models



*Each model corrects the mistakes or shortcomings of its predecessor.

Gradient Boosting

Gradient Boosting Idea



Same dataset is used for each model but each dataset is associated with different values for target variable.

Gradient Boosting: Informal Description

- $y_i - F1(x_i)$ [$F1(x_i)$ in short y_i'] are called residuals of model F1. These are the parts that existing model F1 cannot do well.
- The role of F2 is to compensate the shortcomings of existing model F1.
- If the new model F1+ F2 is still not satisfactory, we can add another regression model F3, etc.,

Boosting Idea vs Gradients

For any model/stage F,

- Minimize J by adjusting $F(x_1), F(x_2), \ldots, F(x_n)$

$$J = \sum_i \left(y_i - F(x_i)\right)^2$$

- We can treat $F(x_i)$ as parameters and take derivatives: $$\frac{\partial J}{\partial F(x_i)} = F(x_i) - y_i$$

- We can interpret residuals as negative gradients: $$y_i - F(x_i) = -\frac{\partial J}{\partial F(x_i)}$$

Gradient Boosting

- The benefit of formulating GB algorithm using gradients is that it allows us to consider other loss functions and derive the corresponding algorithms in the same way.

- Why do we need to consider other loss functions for regression?

Generalizing Gradient Boosting

Loss Functions for Regression: Squared Loss

✓ Easy to deal with mathematically

× Not robust to outliers

Outliers are heavily punished because the error is squared.

Example:

y_i	0.5	1.2	2	5*
$F(x_i)$	0.6	1.4	1.5	1.7
$L = (y - F)^2/2$	0.005	0.02	0.125	5.445

Consequence?

Pay too much attention to outliers. Try hard to incorporate outliers into the model. Degrade the overall performance.

Loss Functions for Regression: Absolute & Huber Losses

- Absolute loss (more robust to outliers)

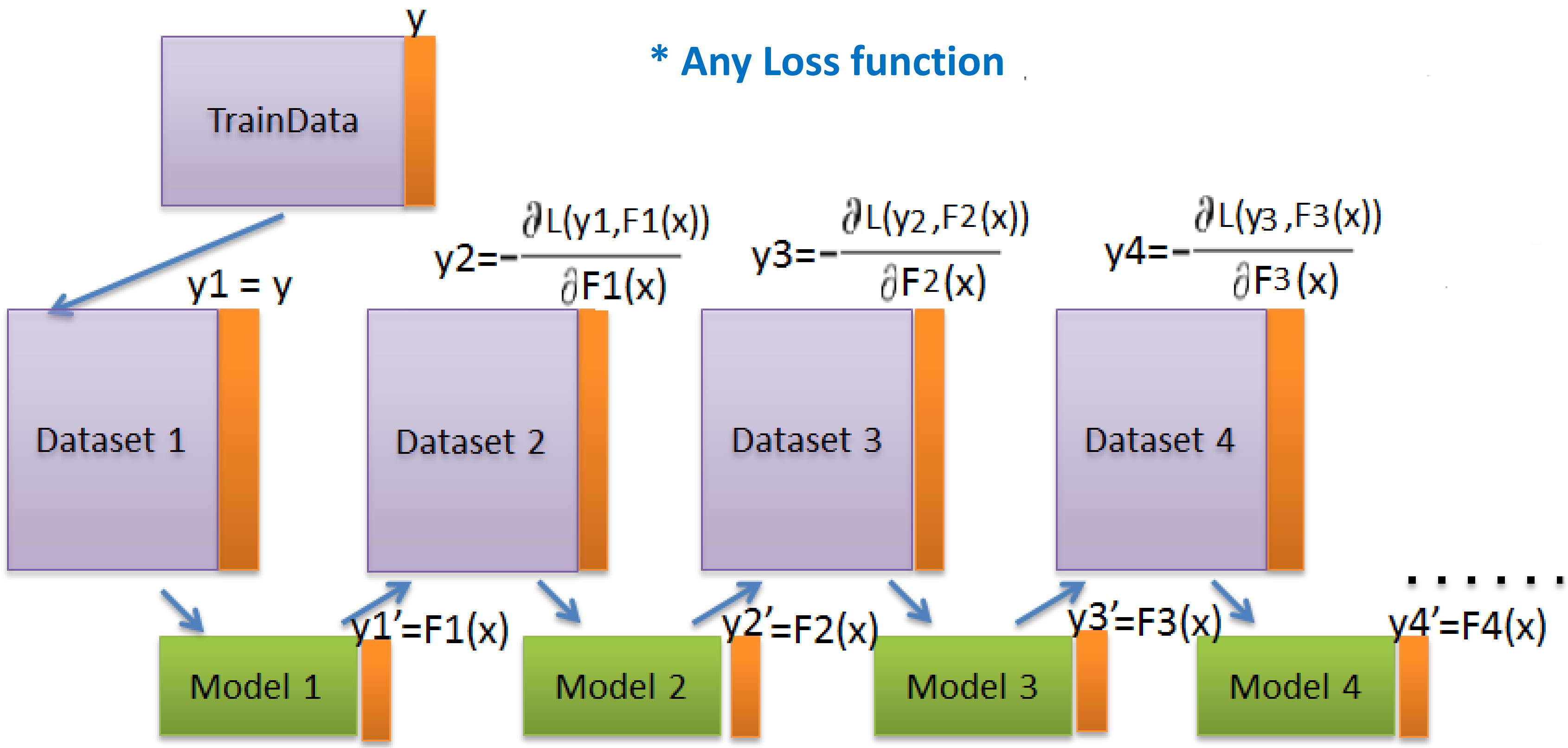
$$L(y, F) = |y - F|$$

- Huber loss (more robust to outliers)

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

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

Generalized Gradient Boosting Idea



In general, Gradients are used to provide the modified values for target variable.

Gradient Boosting vs Ada Boosting

- Fit an additive model (ensemble) $\sum_t \rho_t h_t(x)$ in a forward stage-wise manner.
- In each stage, introduce a weak learner to compensate the shortcomings of existing weak learners.
- In Gradient Boosting, "shortcomings" are identified by gradients.
- Recall that, in Adaboost, "shortcomings" are identified by high-weight data points.
- Both high-weight data points and gradients tell us how to improve our model.