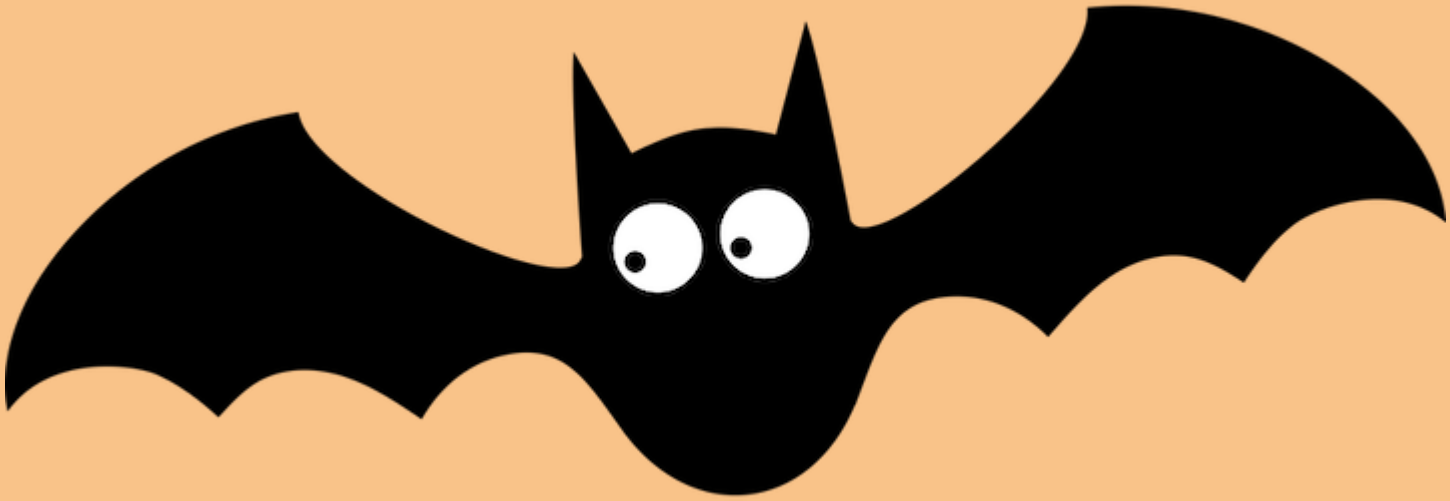


Gradient Boosting

Son Nguyen

Gradient Boosting

① Combine of multiple "weak" models (usually simple trees)

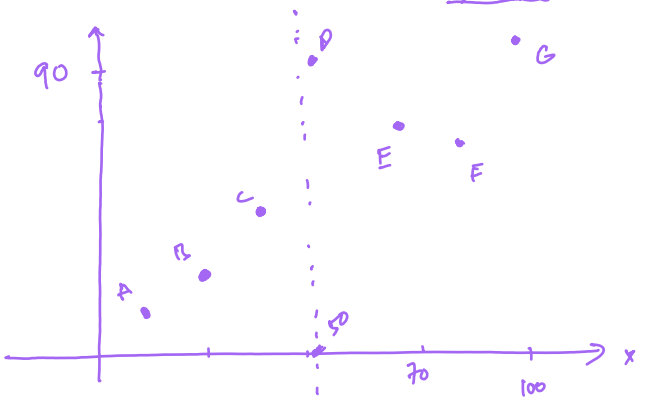
Example :

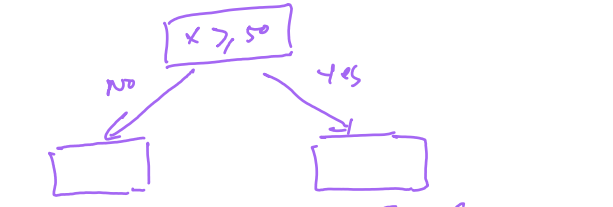
tip today → kept tomorrow

x	y
70	80
60	90 ←
30	20
40	40
80	75
90	100
10	15

This is a regression (not classification) b/c y is continuous / numeric

② How decision trees work for regression





A, B, C
predict by
the average
y values of
A B C

D, E, F, G
predict by
the average
values of D E F G

Model 1 is train on the original data

x	y	y_1	e_1
70	80	100	-20
60	90	20	70
30	20	30	-10
40	40	70	-30
80	75	100	-25
90	100	90	10
10	15	25	-10

y_1 : prediction of Model 1 (M1)

e_1 : errors of M1 : $y - y_1$

target of model 2

$e_1 - \hat{e}_1 = e_2$

x	e_1	$\hat{e}_1$	e_2
70	-20	-15	-5
60	70	80	-10
30	-10	-5	-5
40	-30	-30	0
80	-25	-20	-5
90	10	15	-5
10	-10	-10	0

Data of Model 2

Model 2 is designed to predict the error of M1

$\hat{e}_1$: prediction of M2

target of M3

x	e_2	$\hat{e}_2$	e_3
70	-5	-7	2
60	-10	-5	-5
30	-5	-6	1
40	0	2	-2
80	-5	-8	3
90	-5	-2	-3
10	0	-1	1

→ .... → ....

Let say we stop at M3, the final prediction of the committee of M1, M2, M3 is

Final prediction = $y_1 + \hat{e}_1 + \hat{e}_2$

x	y	y_1	e_1
70	80	100	-20
60	90	20	70
30	20	30	-10
40	40	70	-30
80	75	100	-25
90	100	90	10
10	15	25	-10

→

learning rate ↓ ($L = 70\%$)

x	e_1	$e_1 \cdot L$
70	-20	$-20 \cdot L = -14$
60	70	$70 \cdot L = 49$
30	-10	$-10 \cdot L = -7$
40	-30	$-30 \cdot L = -21$
80	-25	$-25 \cdot L =$
90	10	$10 \cdot L = 7$
10	-10	$-10 \cdot L = -7$

target for M2

x	e_2	e_2	e_3
70	-5	-7	2
60	-10	-5	-5
30	-5	-6	1
40	0	2	-2
80	-5	-3	3
90	-5	-2	-3
10	0	-1	1

notice:

① L is usually from .01 to .1

② Larger L tends to overfit the data

③ Smaller L requires greater numbers of "weak" models in the gradient boosters.

⑧ Evaluating Regression Models

True	Prediction	Error
y_1	$\hat{y}_1$	$y_1 - \hat{y}_1$
y_2	$\hat{y}_2$	$y_2 - \hat{y}_2$
y_3	$\hat{y}_3$	$y_3 - \hat{y}_3$
y_4	$\hat{y}_4$	$y_4 - \hat{y}_4$
y_5	$\hat{y}_5$	$y_5 - \hat{y}_5$

True	Baseline model
y_1	$\bar{y}$
y_2	$\bar{y}$
y_3	$\bar{y}$
y_4	$\bar{y}$
y_5	$\bar{y}$

$$\bar{y} = \frac{y_1 + y_2 + y_3 + y_4 + y_5}{5}$$

① MAE : Mean Absolute Error

$$MAE = \left(|y_1 - \hat{y}_1| + \cdots + |y_5 - \hat{y}_5| \right) / 5 = \Sigma |y - \hat{y}| \, / 5$$

② MSE : Mean square Error

$$= \left((y_1 - \hat{y}_1)^2 + \cdots + (y_5 - \hat{y}_5)^2 \right) / 5 = \frac{\Sigma (y - \hat{y})^2}{5}$$

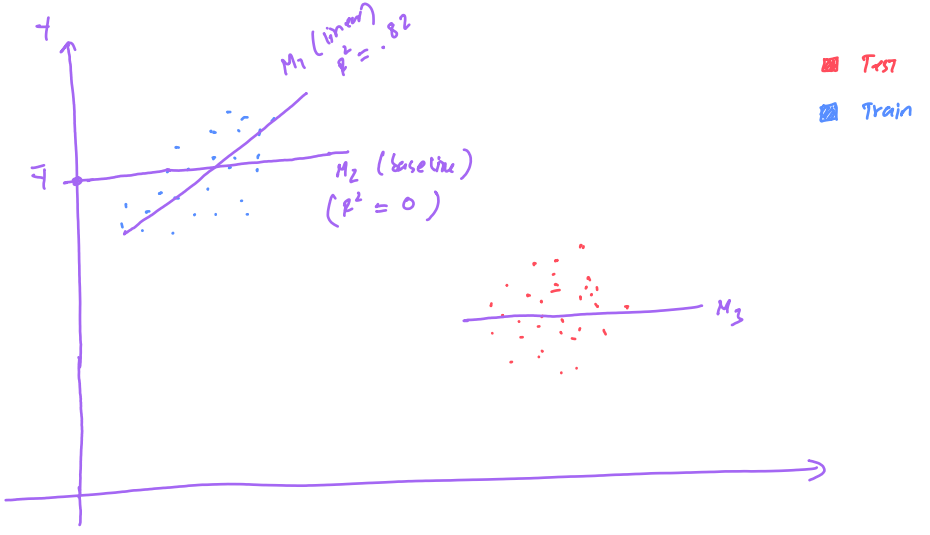
③ R-squared

$$R^2 = 1 - \frac{MSE \text{ (of model M)}}{MSE \text{ (of baseline model)}}$$

⑧ $R^2 = 1$: we have a "perfect" model.

⑧ $R^2 = 0$: M is just as good as baseline model.

⑧ $R^2 < 0$: M is not as good as the baseline model.

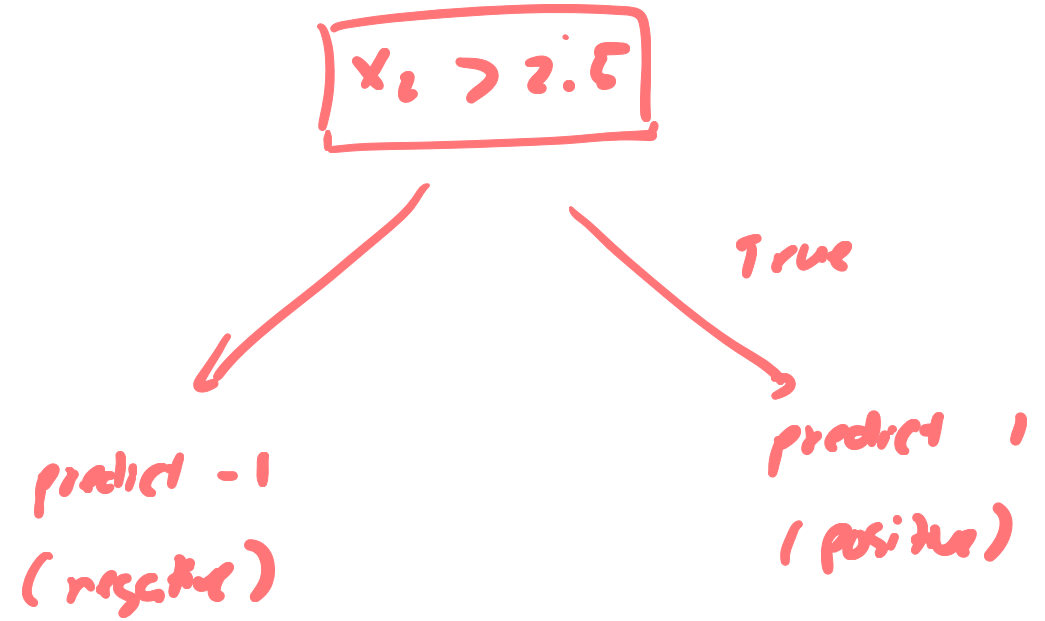


On Test data: R-Square of M1 is negative

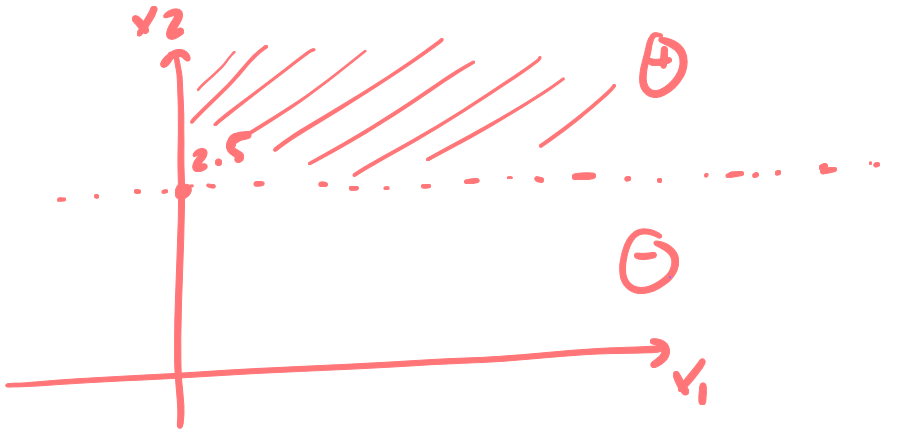
R-quare of M3 is 0

(*) $I(x_2 > 2.5) = \begin{cases} 1 & \text{if } x_2 > 2.5 \\ -1 & \text{if } x_2 < 2.5 \end{cases}$

The same as



The same as



Drawing Decision Boundary of the AdaBoost consisting of the 3 following Stumps.

Stump 1 : $I(y_2 > 2.5)$ Error $\varepsilon_1 = .2$

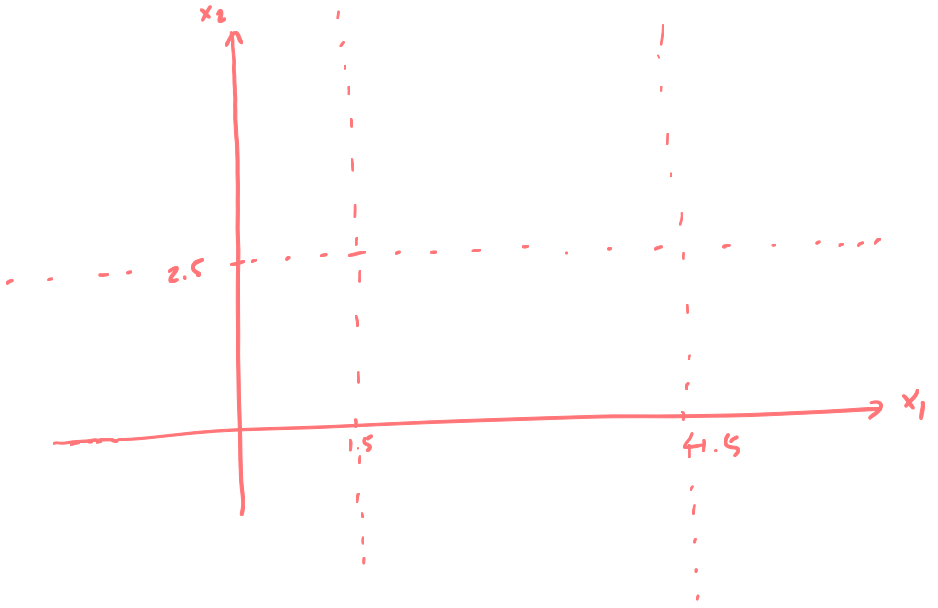
voting power $\alpha_1 = .693$

Stump 2 : $I(y_1 < 1.5)$ Error $\varepsilon_2 = .1875$

$\alpha_2 = .733$

Stump 3 : $I(y_1 < 4.5)$ Error $\varepsilon_3 = .115$

$\alpha_3 = 1.018$



(x) sign function

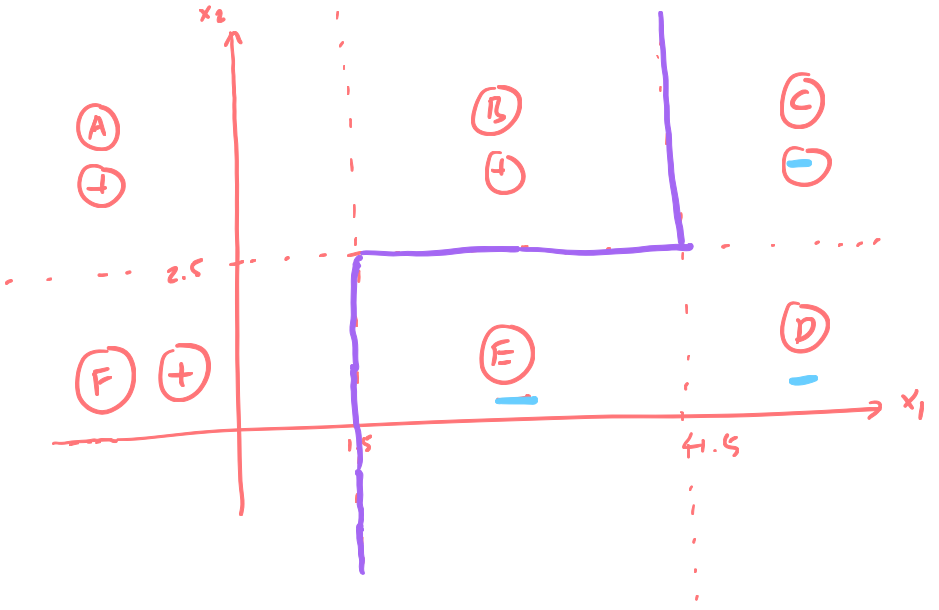
$sign(x)$ → 1 if $x \geq 0$

-1 if $x < 0$

Example : $sign(-6) = -1$; $sign(.67) = 1$

$sign(2022) = 1$ ....

$sign(x) = I(x \geq 0)$



For Region A :

$$\text{sign}\left(\alpha_1 \cdot \text{Stump 1} + \alpha_2 \cdot \text{Stump 2} + \alpha_3 \cdot \text{Stump 3}\right)$$

$$= \text{sign}\left[\alpha_1 \underbrace{I(x_2 \geq 2.5)}_{=1} + \alpha_2 \cdot \underbrace{I(x_1 < 1.5)}_{} + \alpha_3 (x_1 < 4.5)\right]$$

$$= \text{sign}\left(.693 * 1 + .733 * 1 + 1.018 * 1\right)$$

$$= \text{sign}\left(2.441\right) = 1$$

For Region B :

$$= \text{sign}\left[\alpha_1 \underbrace{I(x_2 \geq 2.5)}_{1} + \alpha_2 \cdot \underbrace{I(x_1 < 1.5)}_{-1} + \alpha_3 \underbrace{(x_1 < 4.5)}_{1}\right]$$

$$= \text{sign}(.693 - .733 + 1.018) = \text{sign}(.978) = 1$$

For C :

$$\text{sign}\left[\alpha_1 \underbrace{I(x_2 \geq 2.5)}_{1} + \alpha_2 \cdot \underbrace{I(x_1 < 1.5)}_{-1} + \alpha_3 \underbrace{(x_1 < 4.5)}_{-1}\right]$$

$$= \text{sign}(.693 - .733 - 1.016) = \text{sign}(-1.056) = -1$$