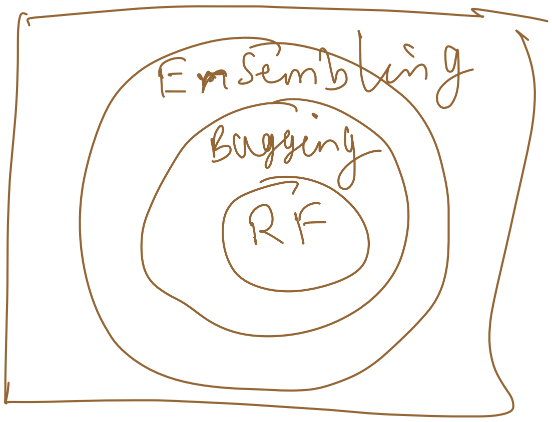
Previous Lecture (Sept 13)

1) Quizzes
2) Pruning & recap of best depth selection
3) Ensembles & bagging
4) Random Forest & Combining decision trees.
5) Randomness in model
6) Validating RF (Random Forest)
7) Overall performance
8) OOB Score (Out-of-Bag)
9) Bias Variance Trade off
10) Reducing variance
11) Code
12) Optimizing RF
13) Hyper-parameter tuning
14) Computing Feature Importance

Today's class

1) Quizzes + Recap
2) Intro to Boosting
3) Boosting Intuition – How to combine Base Learners?
4) What happens at train & test time
5) GBDT Intuition
6) Sklearn implementation



Error

Loss Function

Regression

(Residual) Error = $y^{(i)} - \hat{y}^{(i)}$

truth

pred

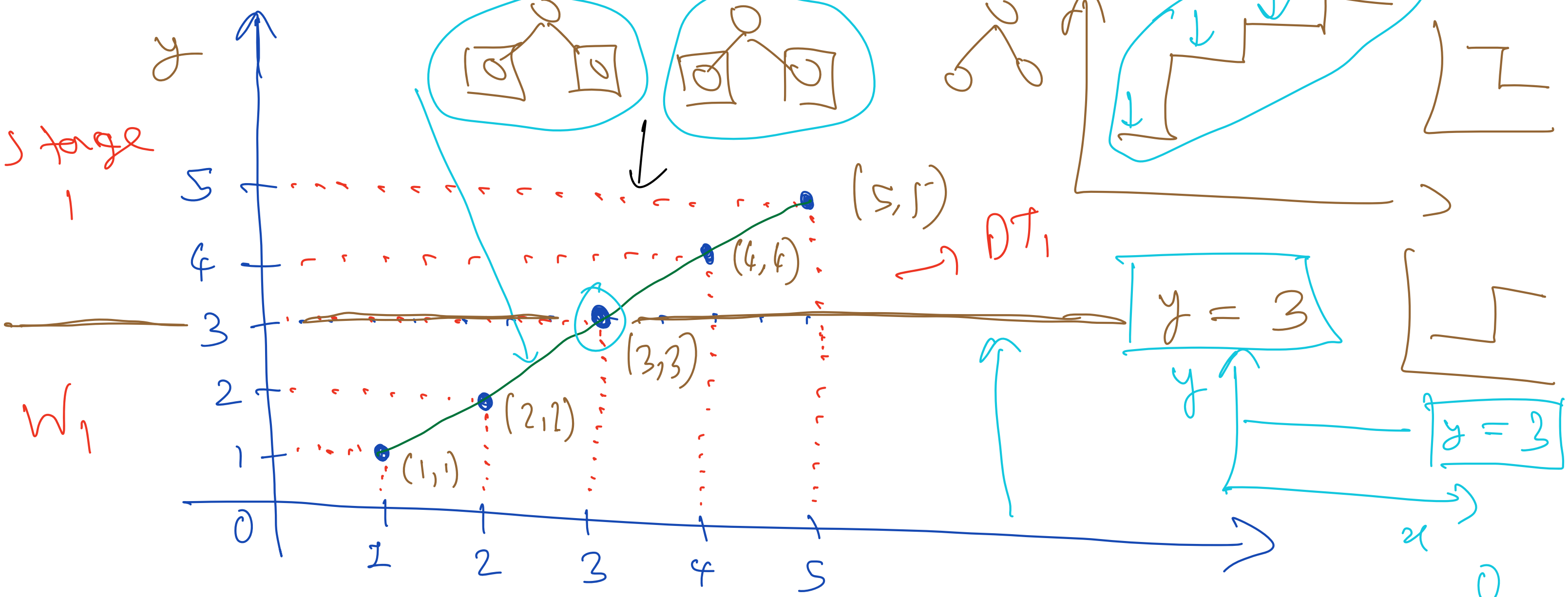
all samples

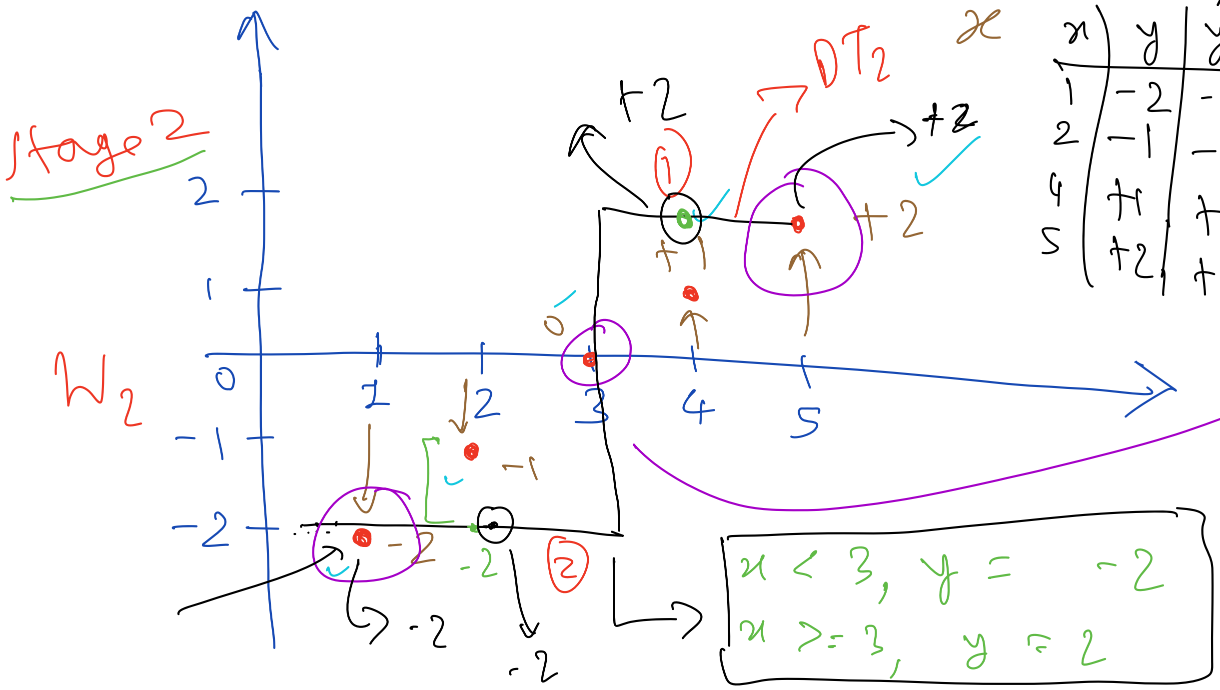
$$\text{Loss Fn (MSE)} = \sum_{i=1}^{m} \left(y^{(i)} - \hat{y}^{(i)} \right)^2$$

1 sample

$$\text{Loss Fn} \left(y^{(i)}, \hat{y}^{(i)} \right) = \left(y^{(i)} - \hat{y}^{(i)} \right)^2$$

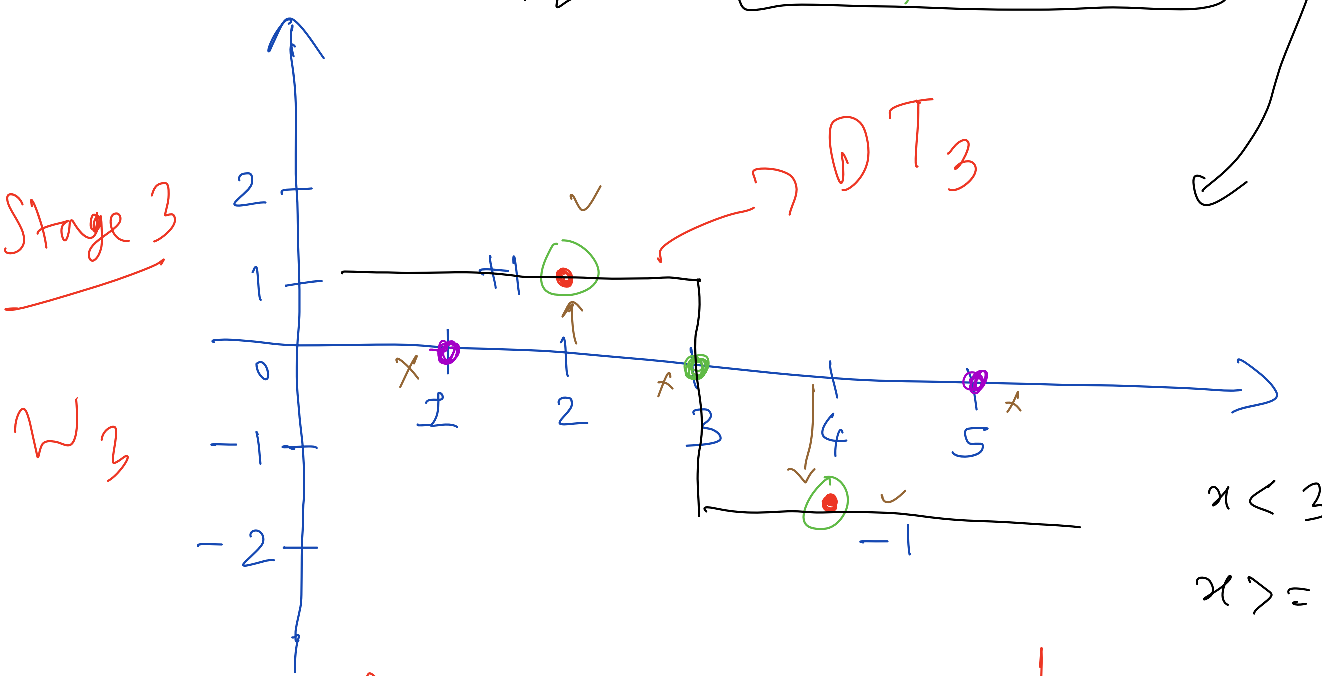
Tree 2 Tree 1 Tree M





x	y	$\hat{y}$	$e = y - \hat{y}$
1	-2	-2	0
2	-1	-2	+1
4	+1	+2	-1
5	+2	+2	0

$x = 3$



$x < 3, \; y = 1$

$x >= 3, \; y = -1$

$$\hat{y}_q = \boxed{W_1} \cdot DT_1 + W_2 \cdot DT_2 + W_3 \cdot DT_3$$

$2 \quad \gamma \cdot DT_1 + \gamma \cdot DT_2 + \gamma \cdot DT_3$ X

Bagging : Base Learners → Low Bias, High Var

Boosting : Base Learners → High Bias, Low Var

x_q Stage 0, Stage 1

$$f_1(x_q) = h_0(x_q) + \gamma_1 h_1(x_q)$$

↓

pred for x_q

$$f_0(x_q) = h_0(x_q)$$

$$err_{1,x_q} = y_q - \boxed{f_0(x_q)}$$ → after stage 0 o/p

$$err_{2,x_q} = y_q - f_1(x_q)$$

O/P after stage 1

$$err_{3,x_q} = y_q - \boxed{f_2(x_q)}$$

$$\boxed{f_1(x_q) = h_0(x_q) + \gamma_1 h_1(x_q)}$$

$$f_2(x_q) = h_0(x_q) + \gamma_1 h_1(x_q) + \gamma_2 h_2(x_q)$$

Stage 0 : $\{x^{(i)}, y\} \rightarrow h_0(x^{(i)})$

Stage 1 : $\{x^{(i)}, err_1^{(i)}\} \rightarrow h_1(x^{(i)})$

Stage 2 : $\{x^{(i)}, err_2^{(i)}\} \rightarrow h_2(x^{(i)})$

Stage 3 : $\{x^{(i)}, err_3^{(i)}\} \rightarrow h_3(x^{(i)})$

$h_0 \quad , \quad h_1 \quad , \quad h_2 \quad , \quad \gamma_3$

DT_1

$$\boxed{f_1(x_i) = h_0(x_i) + \gamma_1 h_1(x_i)}$$

$y_i \quad h_0(x_i)$

$$\boxed{\gamma_1} h_1(x_i) = y_i - h_0(x_i)$$

$$err_1(x_i) = y_i - f_1(x_i)$$

$$= y_i - [h_0(x_i) + \gamma_1 h_1(x_i)]$$

min

$h_0(x_i), \quad h_1(x_i), \quad \gamma_1$

$\boxed{\gamma_2}\, h_2(x_i)$

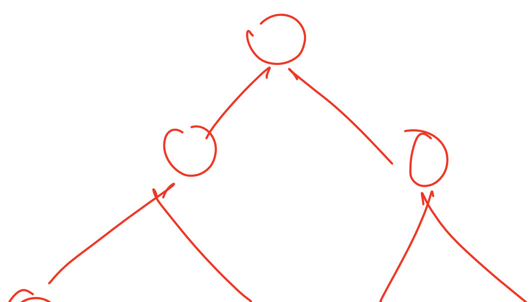
Stage 2

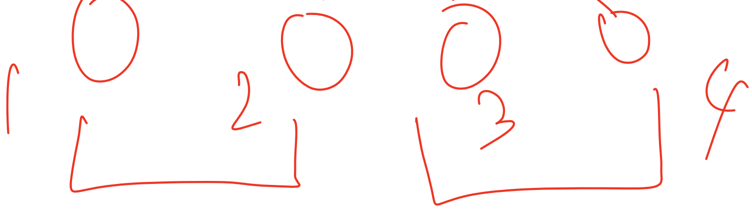
$$err_3(x_i) = y - \left[h_0(x_i) + \gamma_1 h_1(x_i) + \gamma_2 h_2(x_i) \right]$$

min

learning

DT $\rightarrow$





1000 condition f_1 to

τ_1

τ_2

τ_3

⋮

τ_{1000}

τ_1 to τ_{250} → Core 1

τ_{251} to τ_{500} → Core 2

τ_{501} to τ_{750} → Core 3

τ_{751} to τ_{1000} → Core 4

$\boxed{\text{Resume} \rightarrow 10:40}$

$$err^{(i)} = \underbrace{\boxed{y^{(i)} - \hat{y}^{(i)}}}$$

truth pred

Regression MSE $(x^{(i)})$

$$L_{(x^{(i)})} = \left(y^{(i)} - \hat{y}^{(i)}\right)^2$$

$\hookrightarrow (-1) \times \hat{y}^{(i)}$

$$\frac{\partial L_{(x^{(i)})}}{\partial \hat{y}^{(i)}} = 2\left(y^{(i)} - \hat{y}^{(i)}\right)(-1)$$

$$= -2 \times \underset{\text{residual}^{(i)}}{err^{(i)}}$$

$$err^{(i)} = -\frac{1}{2}\frac{\partial L(x^{(i)})}{\partial \hat{y}^{(i)}}$$

residual

$$-\frac{1}{2}\frac{\partial L(x^{(i)})}{\partial \hat{y}^{(i)}} = \text{pseudo-err}$$

or

pseudo-residual

for classification, L is log-loss

$$x_q = h_0(x_q) + 0.5\, h_1(x_q)$$

$\boxed{0,1}$ 0.5 −0.2

$0.5 + 0.5(-0.2)$

$$= 0.5\,(1 - 0.2)$$

$$= 0.5 \times 0.8 = \boxed{0.4}$$

$\boxed{10.0}$

class 0

sigmoid

class 0 or 1

Bagging

NN

Log Reg

Naive Bayes Classifier

KNN classifier

Boosting

Stage 0	Stage 1	Stage 2
NN	log Reg	Naive Bayes

all of them should be slightly underfitting models

Stage 3

KNN classifier