Recap

Ensembles ✓

1. Bagging
2. Boosting → (Random forest) ✓ → Base Learners + Aggregate
3. Stacking
4. Cascading

Row + Col Sampling

→ Bagging (Random forest) -

→ Hyperparameters ✓

→ Hyperparameter tuning ✓

Agenda

Boosting

KNN

?

Random forest :-

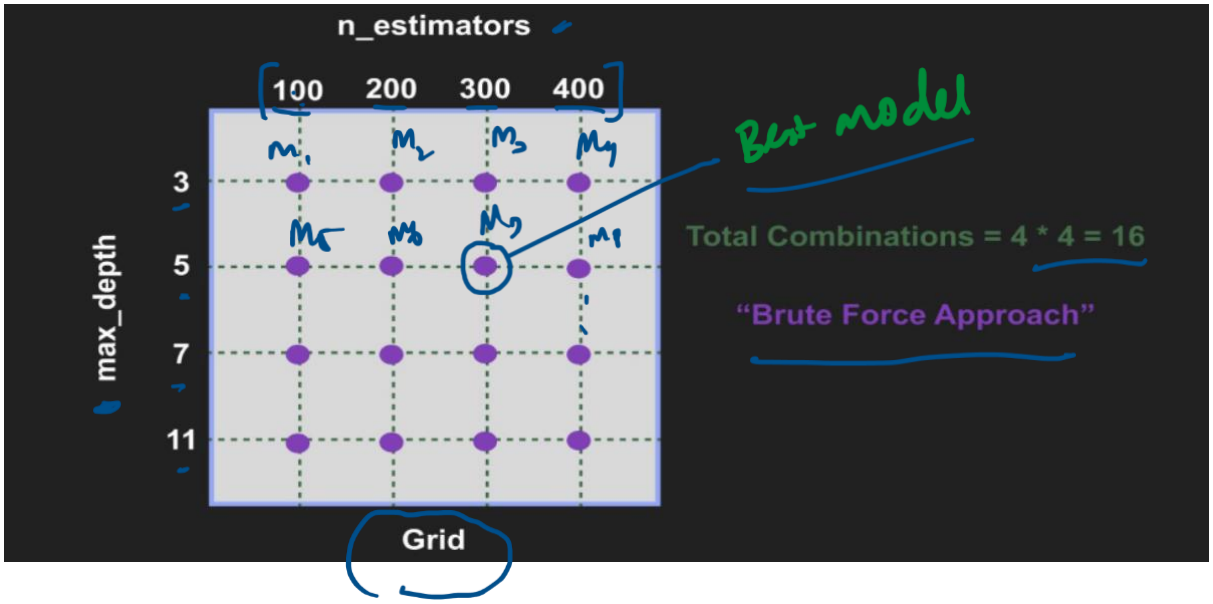
Tree Parameters

1. max no. of leafnodes
2. class wt - - -

- max_depth ✓
- n_estimators ✓
- max_features ✓

+ tree parameters)

Hyper Parameter



n_estimators - [100 200 300 400] (4)

max_depth - [3 5 7 9] (4)

(4 x 16)

	max depth	max dep	max dep
n estimators	100, 3	100, 5	100, 7 ...

n_estimators - [100 200 300 400 500] 5

max_depth - [5 6 7] 3

Max_feature - [4 5 6 7] x4

63

60 different models

	n_est	max_depth	max_feat	Accuracy
m_1	100	4	5	0.81
			6	0.82
m_2	100	4		
⋮				

3 hyperparam: a x b x c — Grid.

Grid Search + CV

GreadSearch CV — hyperparameter tuning + Cross validation

→ Randomised Search

RandomSearchCV

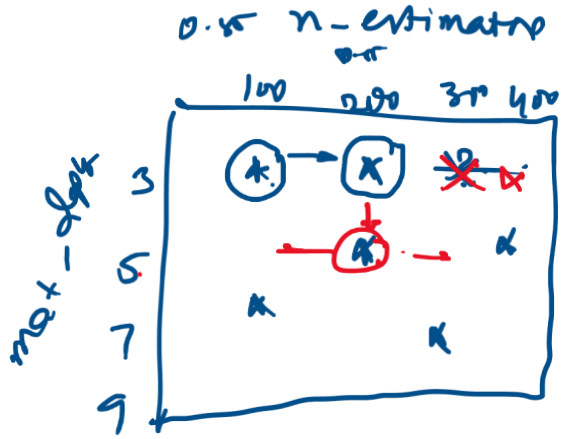
n – (5) count
max – (5)
max_f (5)

cv – 10

1250 x

n_estimators 100 200 300 400

ccp_alpha - 0.1 — 0.5 increment by 0.05



Grid.

When Grid Search Space is Very Large.

→ Reduces Speed ✓

↳ faster computation

150

n_estimators [100 200 300]

(100 125 150 175 200 225)

Boosting

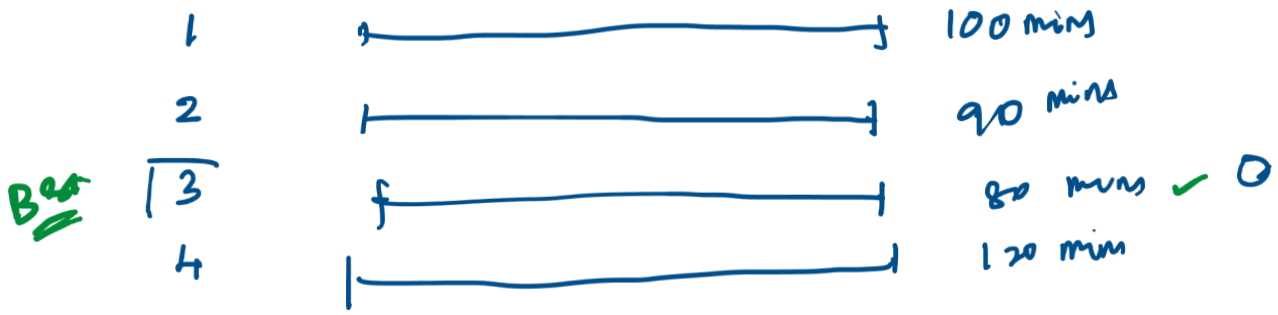
Bagging — Base learner + [Aggregation]

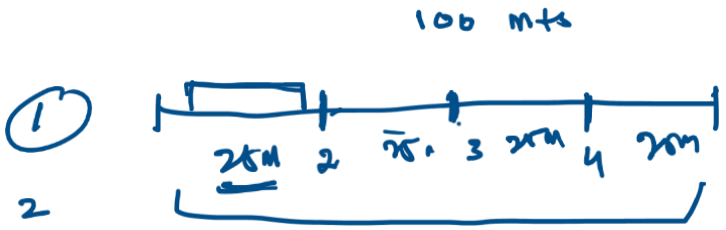
diversity

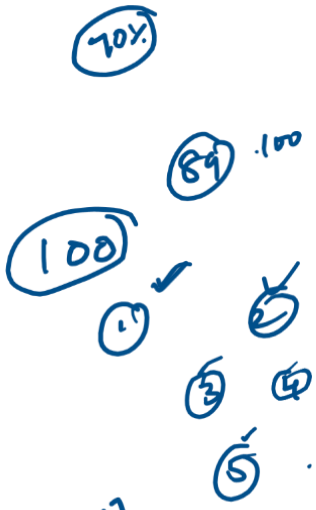
Base learner: low ✓ | high ✓ variance

Aggregation: low Variance

→ (Row sampling + col sampling)







Boosting — Base learners (weak learners)

High bias, low variance.

Boosting — Base learners (weak learners) + additive combination

High Bias

x_1 Height	x_2 Gender	y Weight
1.6	M	52
1.5	F	55
⋮	⋮	⋮

$\bar{y}$ 60

$[x^{(i)}, y^{(i)}]$

$[n \times d]$

Regression $[x^{(i)}]$ $x \in R^d$ $[y^{(i)}]$ $y \in R^1$

$n \times d$.

$y^{(i)} \rightarrow \bar{y}$

Base model $\hat{y} = \bar{y}$ (Bar model)

Mean model Huge Bias.

$M^0 - \hat{y} = mean(\bar{y})$

$M \rightarrow f(x)$ $M^0 - mean(y)$ ✓

$\hat{y} = \bar{y}$ (mean model)

$M^1 \rightarrow err^0 = y^{(i)} - \bar{y}$

$y^{(i)} = M^0(\bar{y}) + err^0$

$82 = 60 + 22$ (?) ↓

hypothesis (x)

$M^0 — h_0(x) - \bar{y}$

m' $\{ x^{(i)}, err^0 \}$

weight avg

82 = 60 + 22 ?

m' - $h_1(x)$ -

additive combining.

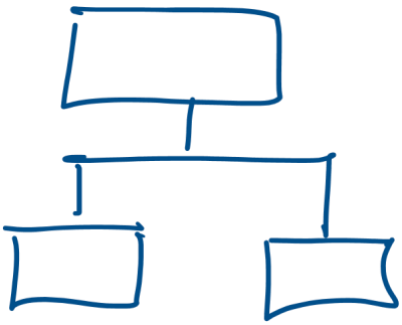
Mean model M_1 ,

$M = h_0(x) + h_1(x^{(i)}) + err$

weak model

Tree Shallow depth - max-depth = 2

max_depth = 1 ✓



One Split tree

Decision stumps

M

high - far from reality ↓ Accuracy ↑ error

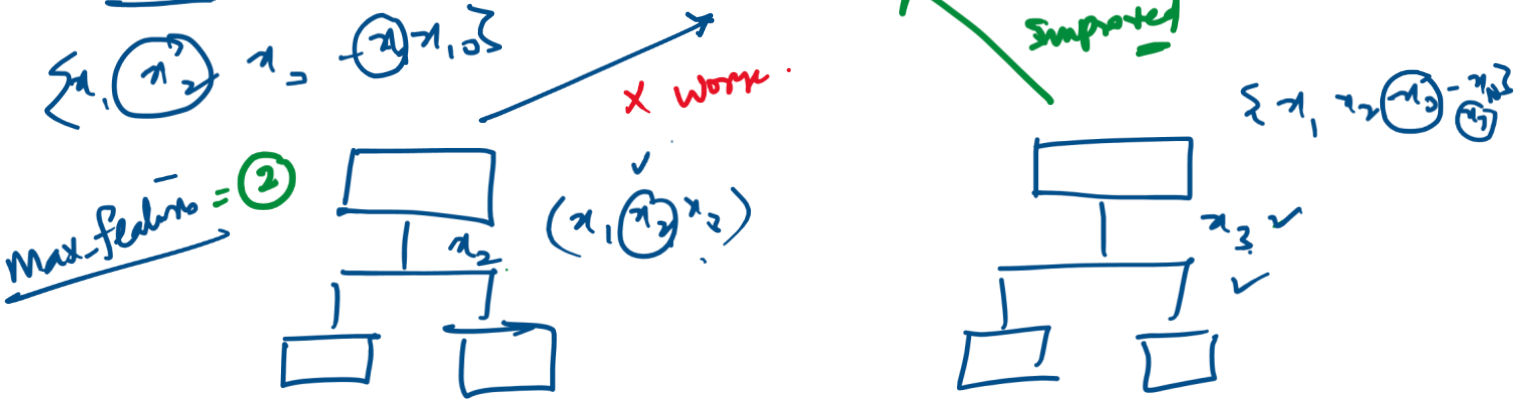
low - close to reality ↑ Accuracy ↓ error

M' $h_1(x)$ $\{x^{(i)}, err^{(i)}\}$

$h_1(x)$ high bias — Decision Stumps
Shallow trees

Baseline

M
$F(x) = h_0(x) + (h_1(x))$ ↑ ✓ $h_1(x)$?



Regularization

$h_0(x) + \gamma_1 h_1(x)$

γ_1 – more
γ_1 – less

γ_1 - hyper Parameter — learn parameter

$M\ f(x) = h_0(x) + \gamma_1 h_1(x) + \gamma_2 h_2(x)$
$+ \gamma_3 h_{13}(x) + \cdots \gamma_m h_m(x)$

learn combine $\gamma_1 \quad \gamma_2 \quad \gamma_3 \cdots \gamma_m$

$$err^h = F_{h-1}(x) - err^{h-1}$$

$h-1$ - models. $F_{h-1} = h_0(x) + \gamma_1 h_1(x) + \gamma_2 h_2(x) + \cdots \gamma_{h-1} h_{h-1}(x)$

$$\underbrace{F_k(x)}_{k+1} = h_0(x) + \underline{\gamma_1} h_1(x) + \underline{\gamma_2} h_2(x) + \cdots \underline{\gamma_k} h_k(x)$$

$h_{k+1}(x) \rightarrow \{ x^{(i)}, err^k \}$

$$err^k = y^{(i)} - F_k(x)$$ ✓ $F_k(x)$

$\gamma_1 \quad \gamma_2$

Gradient descent?

$$F_2(x) = h_0(x) + \gamma_1 h_1(x) + \gamma_2 h_2(x)$$

$err = y^{(i)} - F_2(x)$ Samall 1

Regression loss fn — "MSE"

$$F_m(x) = h_0(x) + \sum_{i=1}^{K} \gamma_i h_i(x)$$

K - hyperparameters γ_i - Parameter

→ $h_2(x)$ → without $h_1(x)$

$h_2(x)$ $\{x^{(i)}, err^{(i)}\}$

→ Boosting additive, Sequential learning

Combining, learn weights that Optimizes the error.

$$F_k(x) = h_0(x) + \gamma_1 h_1(x) + \gamma_2 h_2(x) \\ - - + \gamma_k h_k(x)$$

$$err^K = (y^{(i)} - F_K(x)) \downarrow$$

GBDT [Gradient Boosting DeCision Trees]

$$F_K(x) = h_0(x) + \gamma_1 h_1(x) + \cdots \gamma_K h_K(x)$$

$\gamma_1 - \cdots \gamma_K$ ✓

Loss function MSE

$$L(y^{(i)}, \hat{y}^i) = \sum_{i=1}^{n} (y^{(i)} - \hat{y}^{(i)})^2$$

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

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

error
Residual

-ve gradient ↓ ~ Residual ↓

$$\gamma_1^1 = \gamma_0 - \eta \cdot \frac{\partial L}{\partial \gamma}$$

-ve gradient ↓ ~ error ↓

gradient → 0 ✓ gradients

pseudo residual → -ve gradient

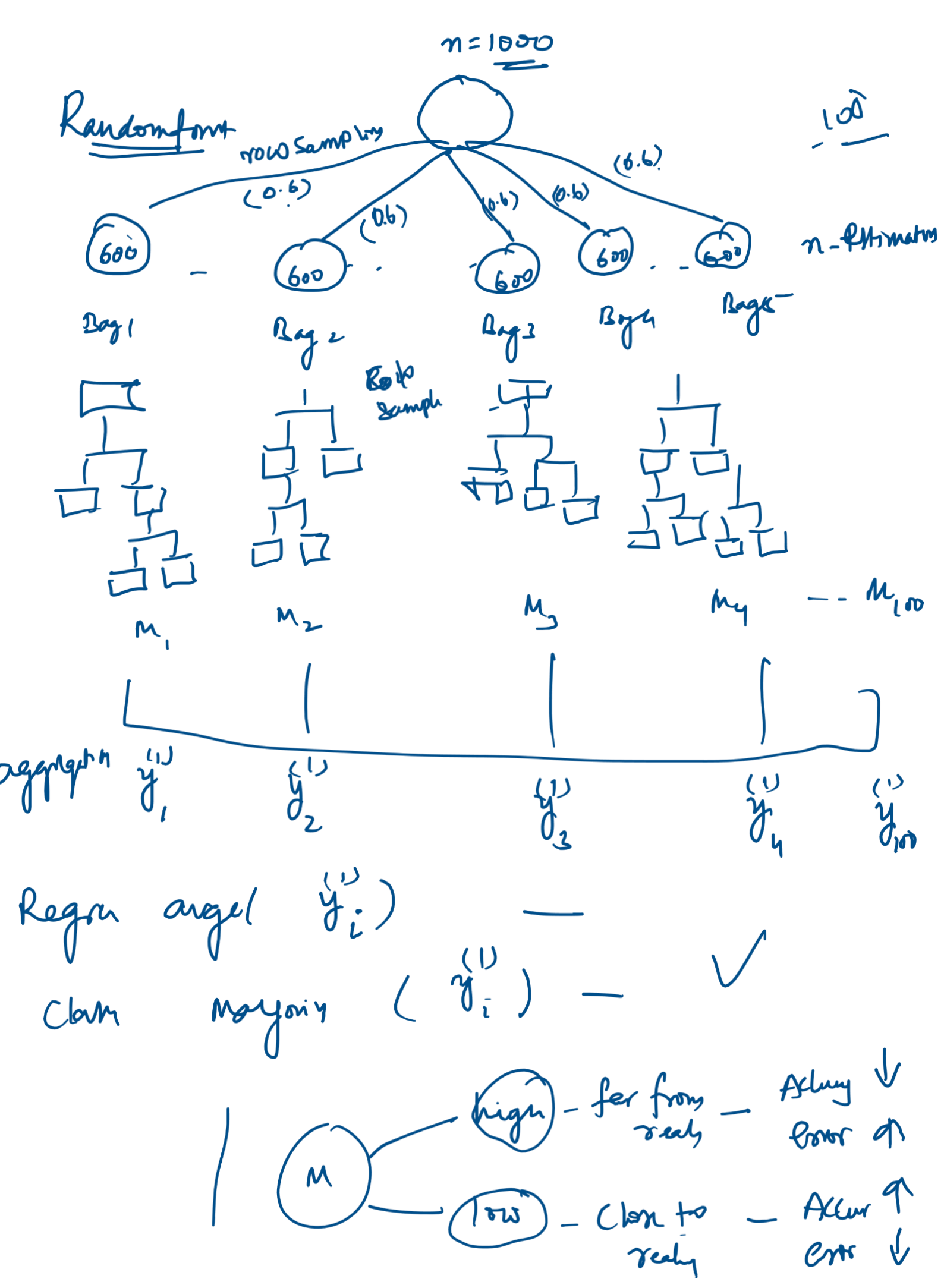
loss function — differentiable. -ve gradient ↓

Custom loss function - differentiable ✓

minimize — pseudo-error, -ve gradients

Regression — MSE —

Classification — log loss -


n = 1000
Random forest
row sampling
(0.6)
(0.6)
(0.6)
(0.6)
(0.6)
600
600
600
600
600
100
n - estimators
Bag 1
Bag 2
Bag 3
Bag 4
Bag 100
Row sample
M1
M2
M3
M4
M100
aggregation
y(1)1
y(1)2
y(1)3
y(1)4
y(1)100
Regression avg(y(1)i)
Class Majority (y(1)i)
high
far from real
Accuracy ↓
Error ↑
M
low
close to real
Accuracy ↑
Error ↓

n_jobs = -1 ✓

CPU 4 core, 8 threads

n_jobs = 4