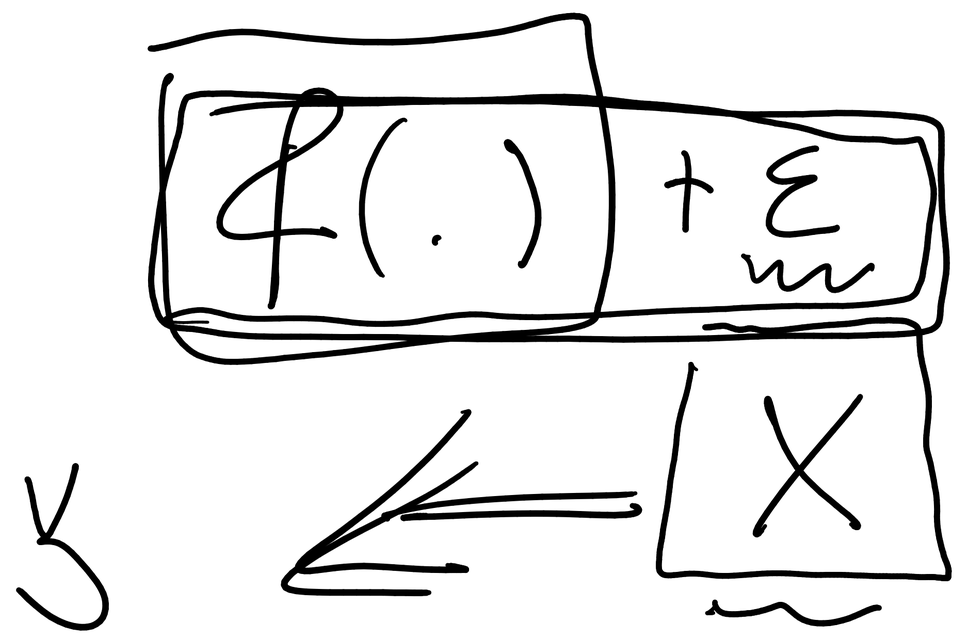
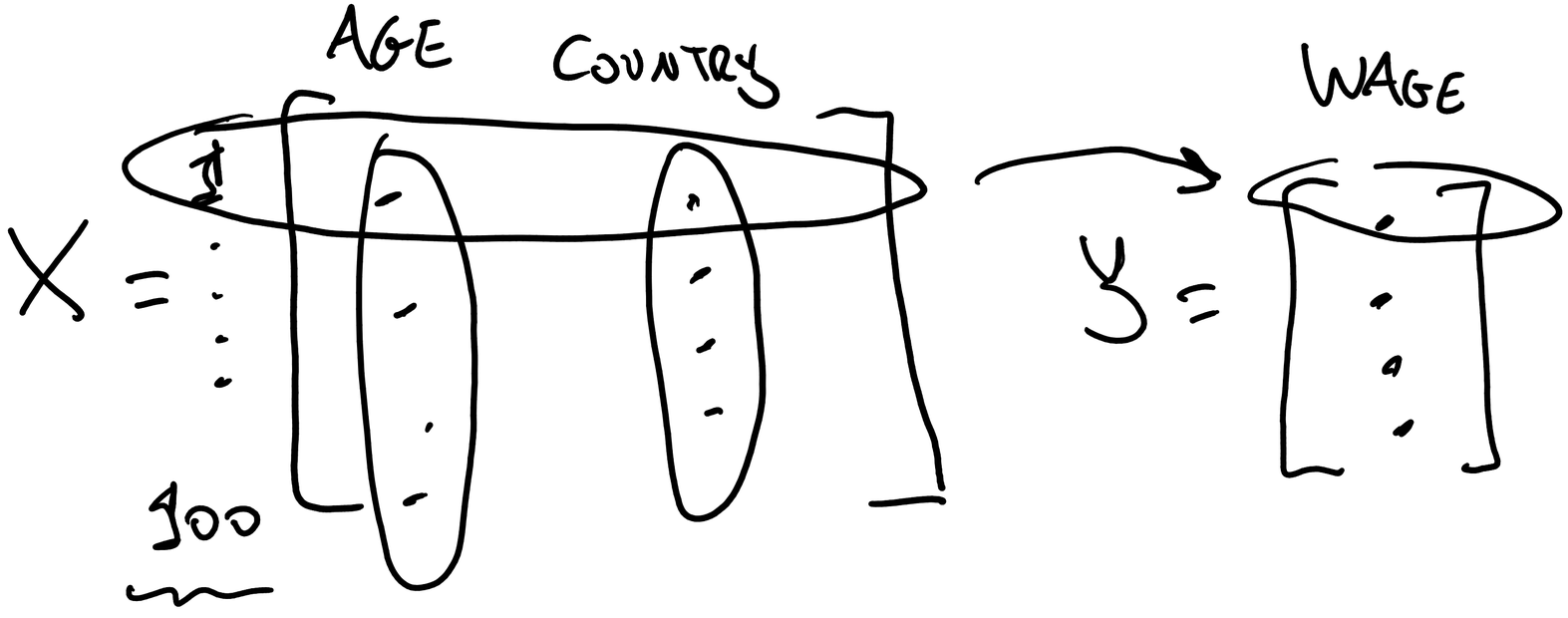
ML FUNDAMENTAL

MODEL →



TARGET

FEATURES



LINEAR REGR.

$$f(x) = b + w_1 x_1 + \dots + w_p x_p$$

$(w_1, \dots, w_p)$

↳ TRAINING

- LOSS FUNCTION

- GRADIENT DESCENT $(2, 1.5, \dots)$

- OVERFITTING

- REGULAR.

REGRESSION

$y \in \mathbb{R}$

$y \in \mathbb{N}$

$y \begin{cases} A \\ B \end{cases}$

$y \in \{0, 1\}$

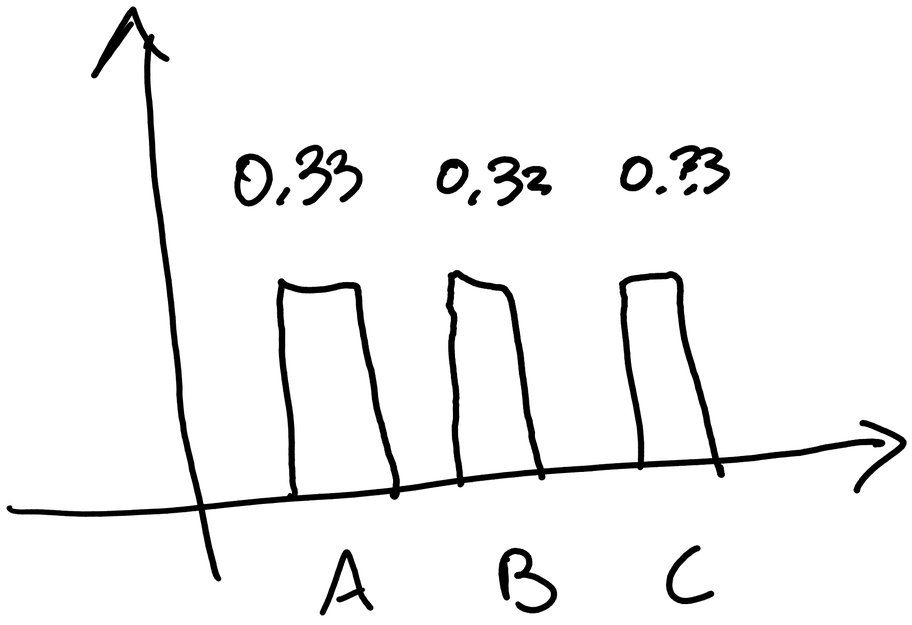
$(y_i - \hat{y}_i)^2$

CLASSIFICATION

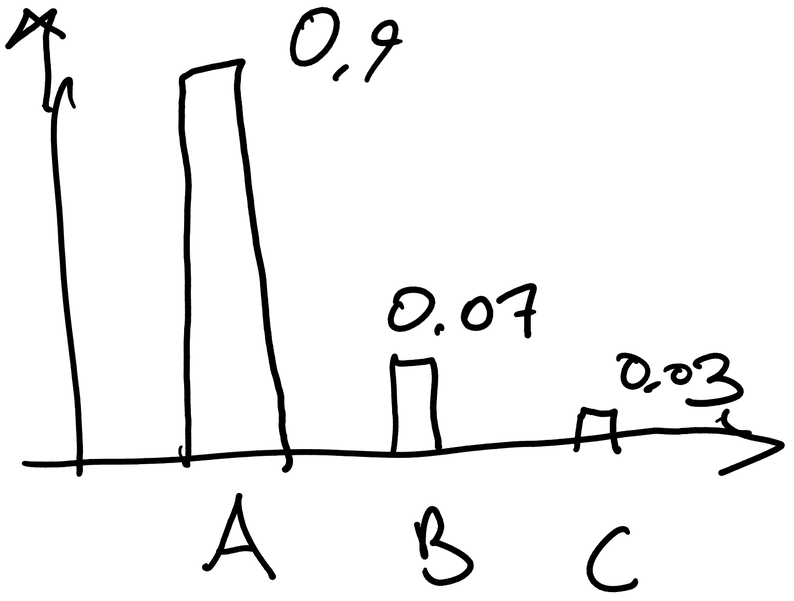
$y_i = 0$ $\quad$ $\hat{y}_i = 1$

$(0 - 1)^2$

1



2

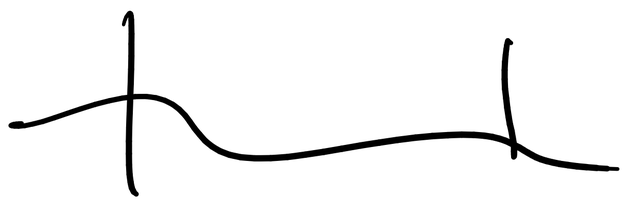


1 $y = \begin{cases} A \\ B \\ C \end{cases}$

$$\hat{f}(x) = 0.93$$

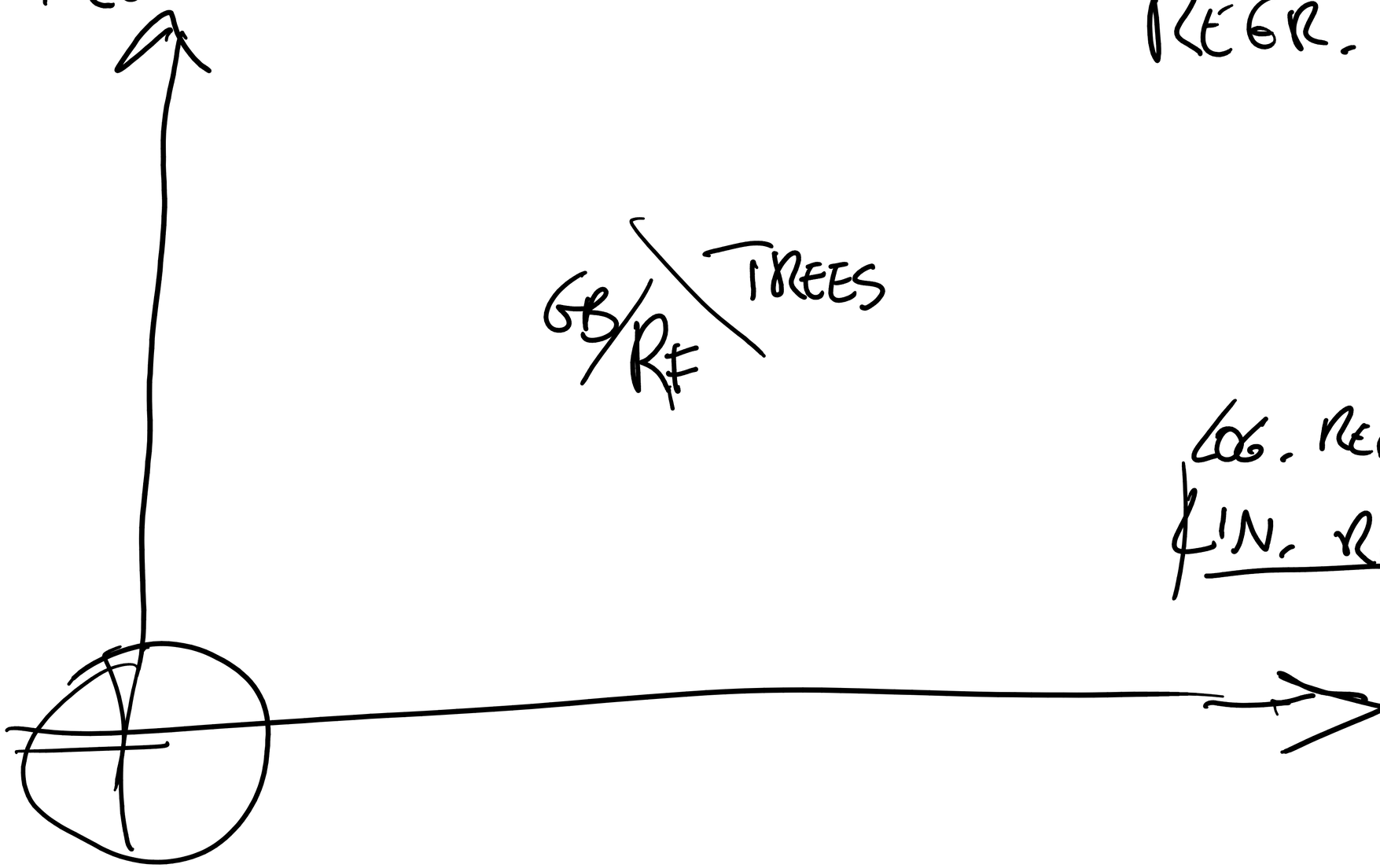
$$y = \begin{cases} A \rightarrow 0 \\ B \rightarrow 1 \end{cases}$$

$$(y_i - \hat{y}_i)^2 = 1 - 0 = ?$$



$$\hat{y}_i = \begin{cases} 1 \\ 0 \end{cases} \rightarrow p \in [0, 1]$$

VARIANCE



BIAS

REGR. / CLASS

GB / RF / TREES

LOG. REGR.
LIN. REGRES

TREES

$X = [\ ()\ ()\ ()\ ]$

X_1

$X = \begin{cases} 1 \\ \vdots \\ \vdots \\ \vdots \\ 100 \end{cases} [\ \ ]$

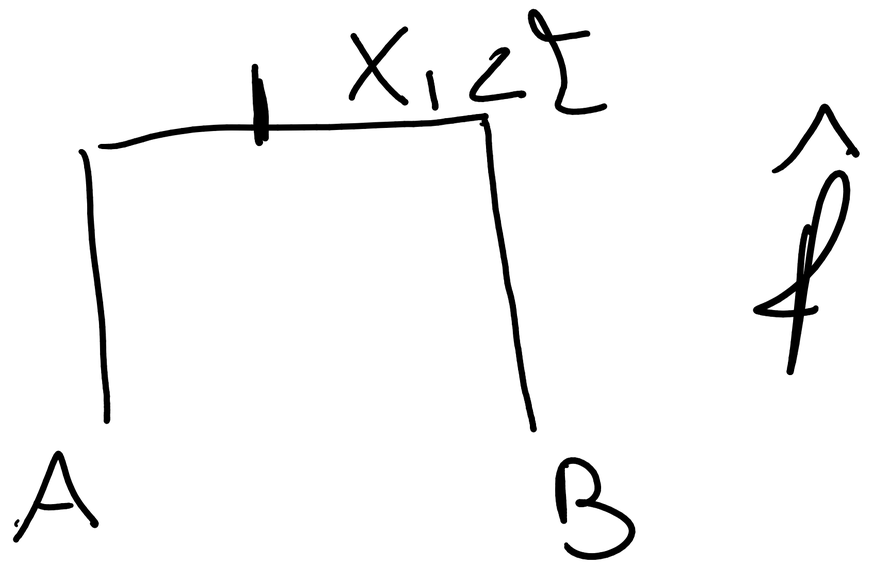
$y = [\ \ ]$

IF $X_1 < t$:

THEN $\hat{y} = A$

ELSE

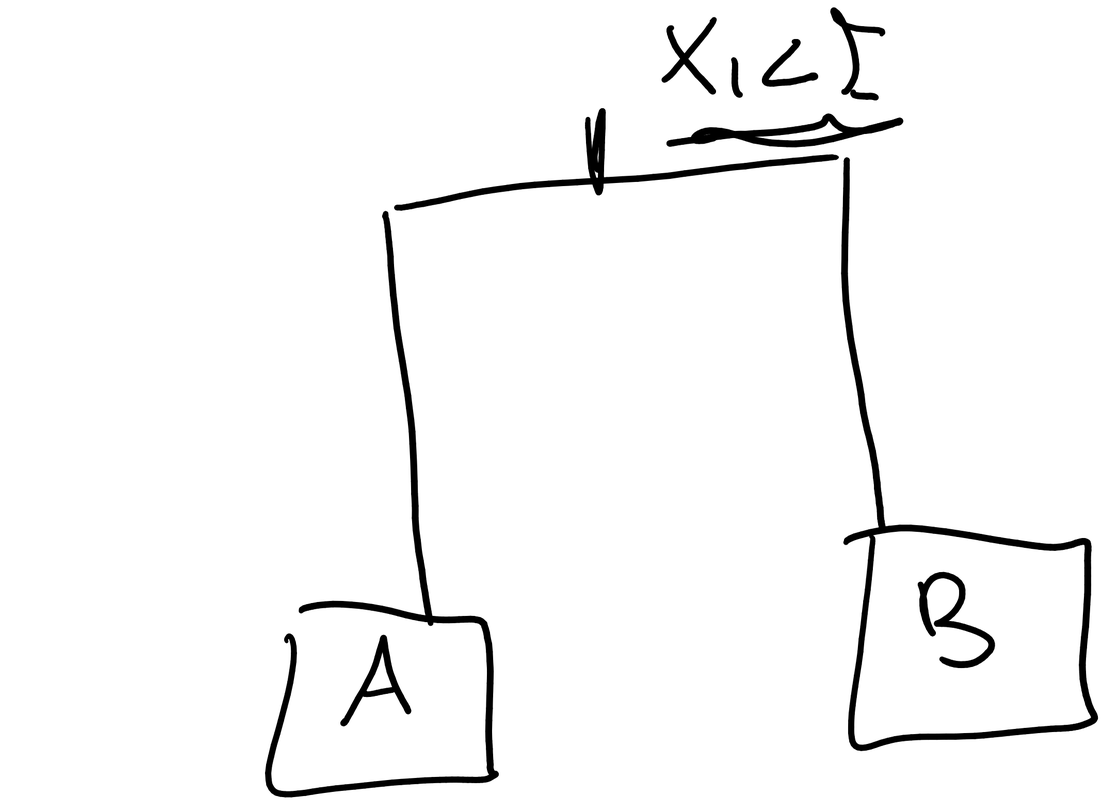
$\hat{y} = B$



$\hat{f}$

$t = 20$

$$X = \begin{matrix} 1 \\ \vdots \\ 100 \end{matrix} \begin{bmatrix} X_1 \\ 30 \\ 21 \\ \vdots \\ 1 \end{bmatrix} \rightarrow \hat{f} \Rightarrow \hat{Y} = \begin{bmatrix} B \\ B \\ \\ A \end{bmatrix} \quad Y = \begin{bmatrix} \\ \\ \\ \end{bmatrix}$$



$\begin{pmatrix} A & A & A & B \\ & & & \\ A & A & A & \end{pmatrix}$

REAL VALUES

$\begin{pmatrix} B & B & B & A \\ & & & \\ A & A & B & \end{pmatrix}$

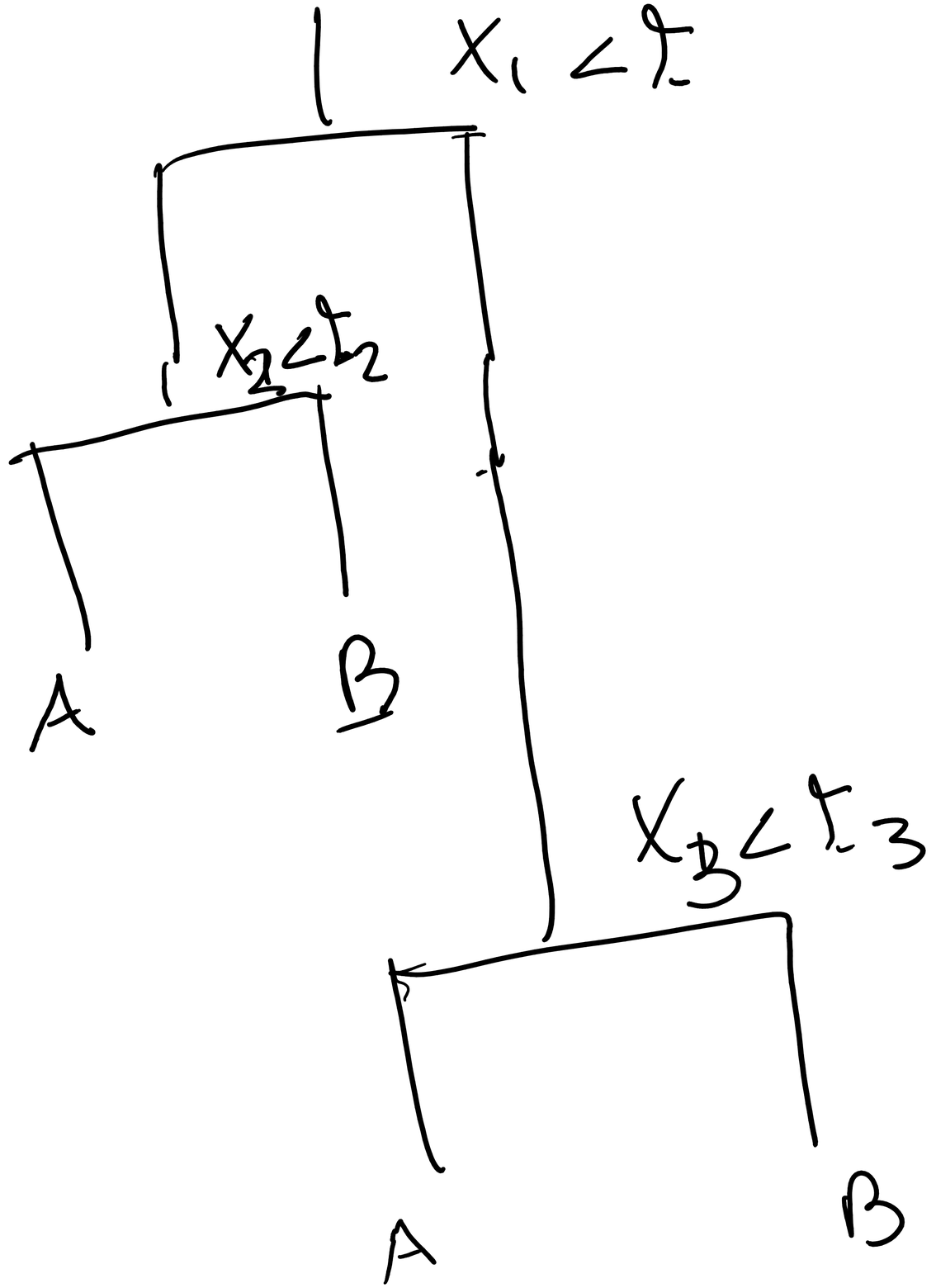
REAL VALUES

ENTROPY_left

ENTROPY_Right

$\frac{m!}{m \cdot n}$ (entropy_left + entropy_right)

GREEDY ALG.


$X_1 < t_1$
$X_2 < t_2$
A
B
$X_3 < t_3$
A
B

1
SPECIFICATION

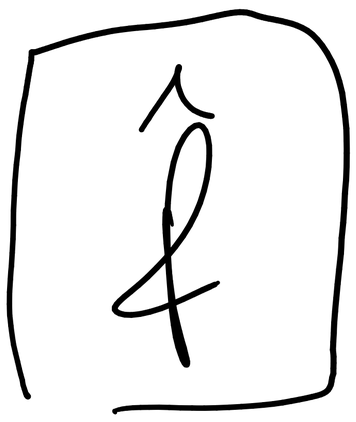
2
TRAINING

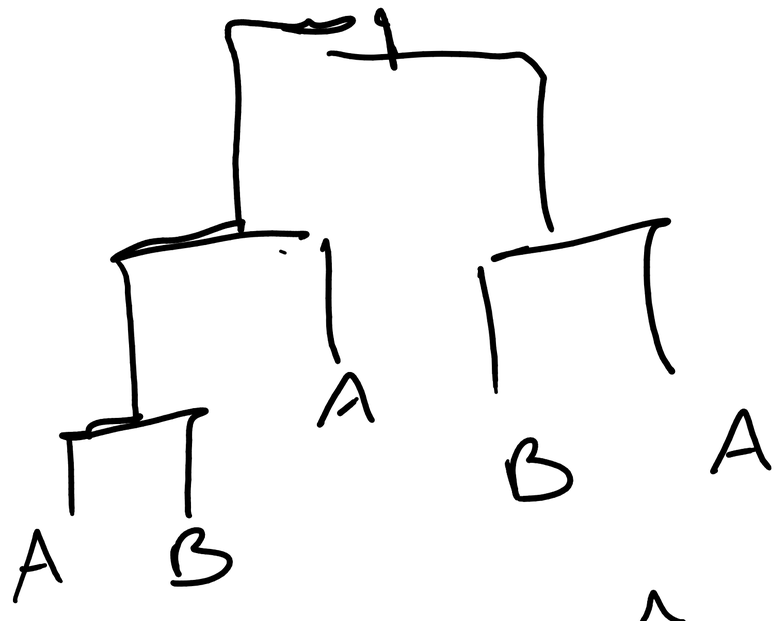
3
EVALUA.
(PRED.)

TREE

ENTR.

ACCURACY





ACCURACY $\hat{y}, y$

$\hat{f}(X_{TEST})$

$\hat{y} = \begin{pmatrix} A \\ \vdots \\ B \end{pmatrix}$

ENSEMBLING OF MODELS

$\hat{f}_1, \ldots, \hat{f}_B$

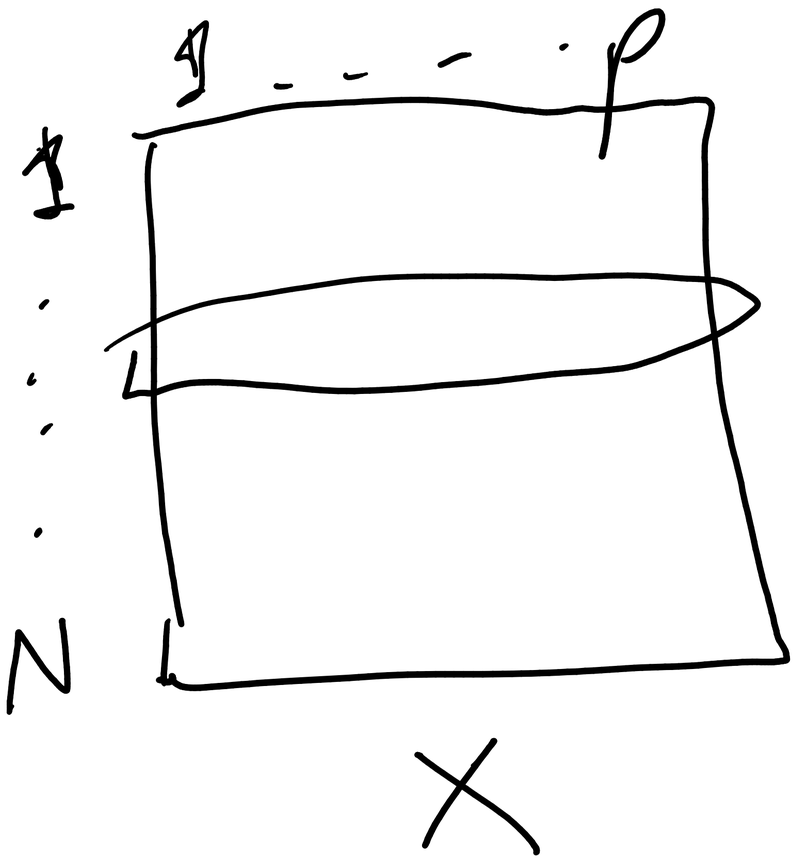
PARALLEL

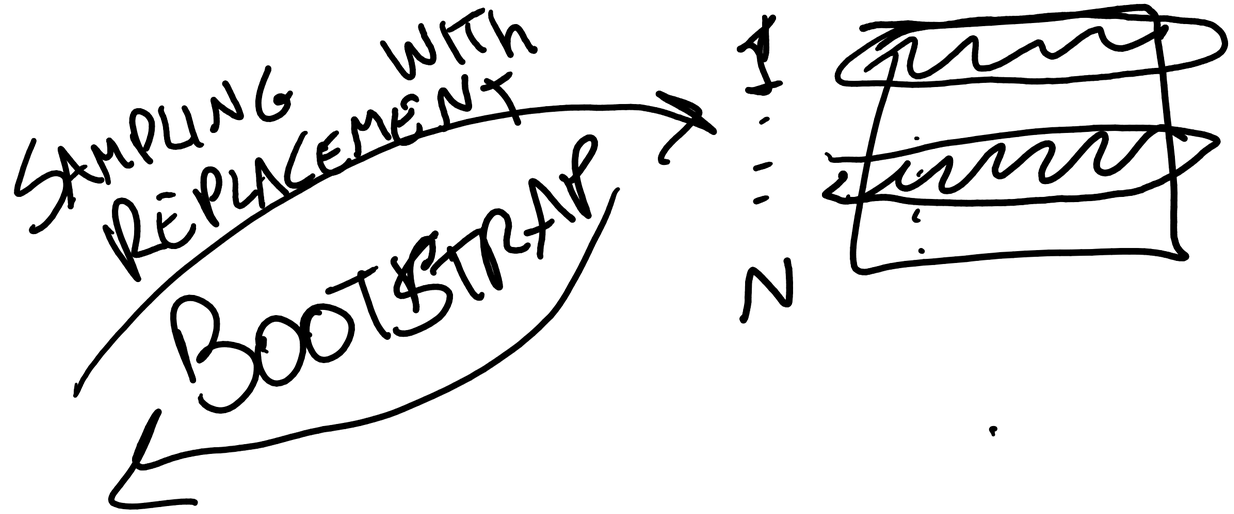
$\hat{f}_1, \ldots, \hat{f}_B$

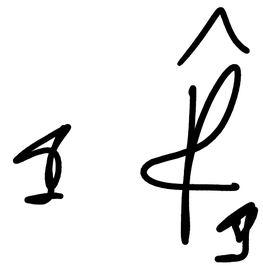
SEQUENTIAL

$\hat{f}_B(\hat{f}_{B-1}(\ldots(\hat{f}_1)\ldots))$

$\hat{f}_1 \rightarrow \hat{f}_2 \rightarrow \ldots$

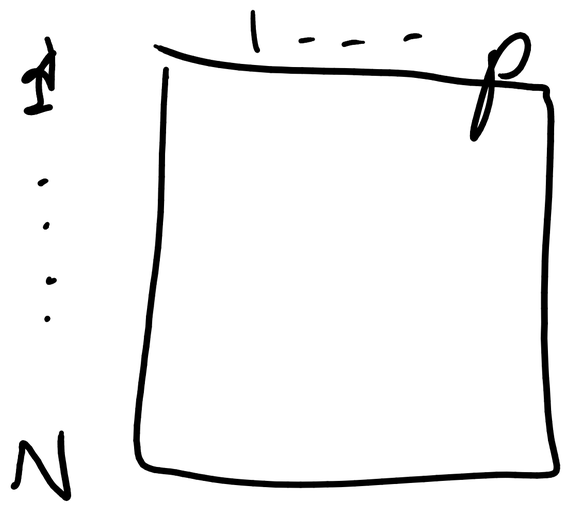






BAGGING

(BOOTSTRAPPING AND AGGREGATING)



$$(X_1, \dots, X_n)$$

$$X_i \sim N(\mu, \sigma^2)$$

$$\bar{X} \sim N\left(\mu, \frac{\sigma^2}{n}\right)$$

RANDOM FOREST

- BAGGING (1, ..., B datasets)
- FOR EACH DATASET:
  - → FIT A TREE
    - FOR EACH SPLIT, USE SUBSET OF FEATURES
- AGGREGATE PRED.

PREDICTIONS WITH ENSEMBLING (CLASSIFICATION)

$X^* = (x^*_1, \ldots, x^*_p)$

$\hat{f}_1(x^*) = 0.7 \Rightarrow B$

$\vdots \quad \vdots \quad \vdots$

$= 0.3 \Rightarrow A$

$\hat{f}_B(x^*) = 0.9 \Rightarrow B$

A

B

FINAL PREDICTION

AVERAGE IN CASE OF REGRESSION

DEPTH	# OF TREES
↑ => BIAS ↓	↑ => VARIANCE ↓

BOOSTING

X_{TRAIN} y_{TRAIN}

$\hat{f}_1(X_{TRAIN}) = \hat{y}_1$, $\underline{e_1} = |\hat{y}_1 - y_{TRAIN}|$

$\hat{f}_2(X_{TRAIN}) = \hat{e}_1$ $e_2 = |\hat{e}_1 - e_1|$

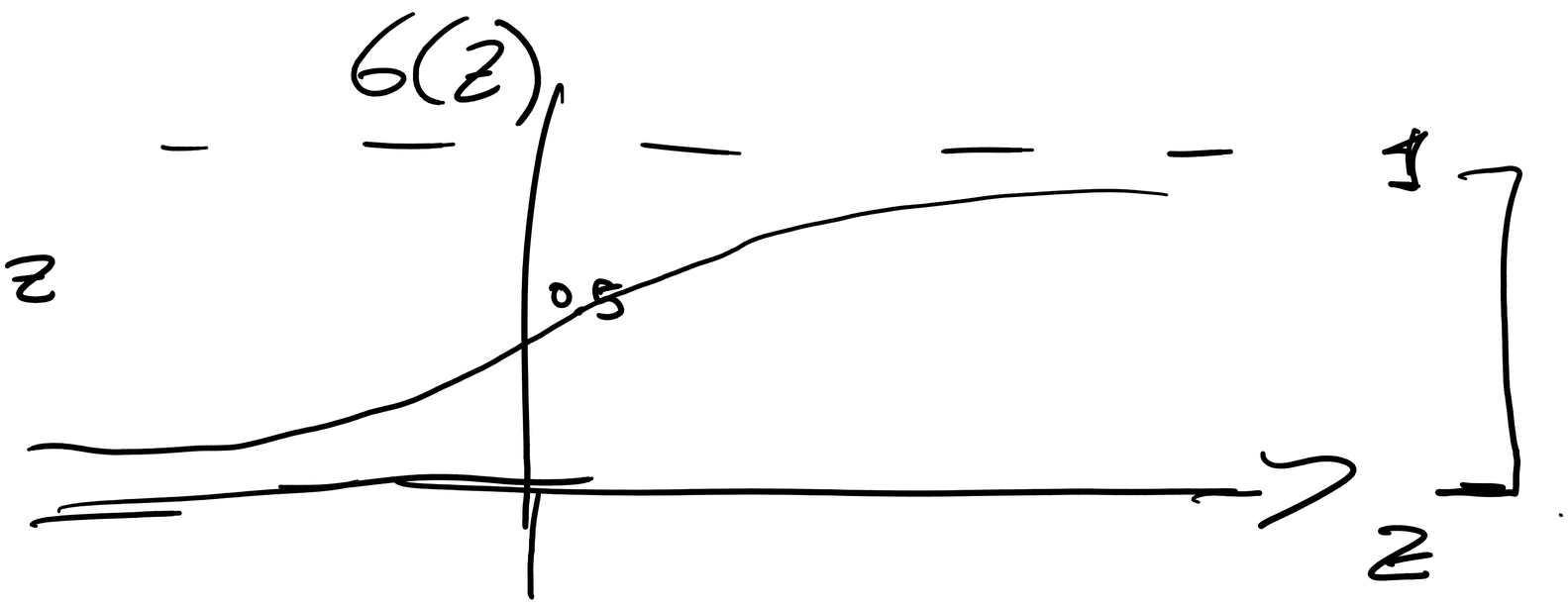
$\hat{f}_B(X_{TRAIN}) = \hat{e}_{B-1}$

LOGISTIC REGRESSION

$$\underset{\sim}{y} = b + w_1 x_1 + \dots + w_p x_p$$

$$y \in \mathbb{R} \Rightarrow y \in [0,1]$$

$$\sigma(z) = \frac{1}{1 + e^{-z}}$$



LOGISTIC REGR.

$$p = \frac{1}{1 + e^{-(b + w_1 x_1 + \dots + w_p x_p)}}$$

LOSS FUNCTION : BINARY CROSS ENTROPY

$\hat{y}$ \ y	0	1
0	$-\log(1-0) = 0$	$-\log(1-1) = +\infty$
1		$-\log(1) = 0$

$\hat{p} = 1$

$\hat{p} = 0$

EVALUATION METRICS (CLASS.)

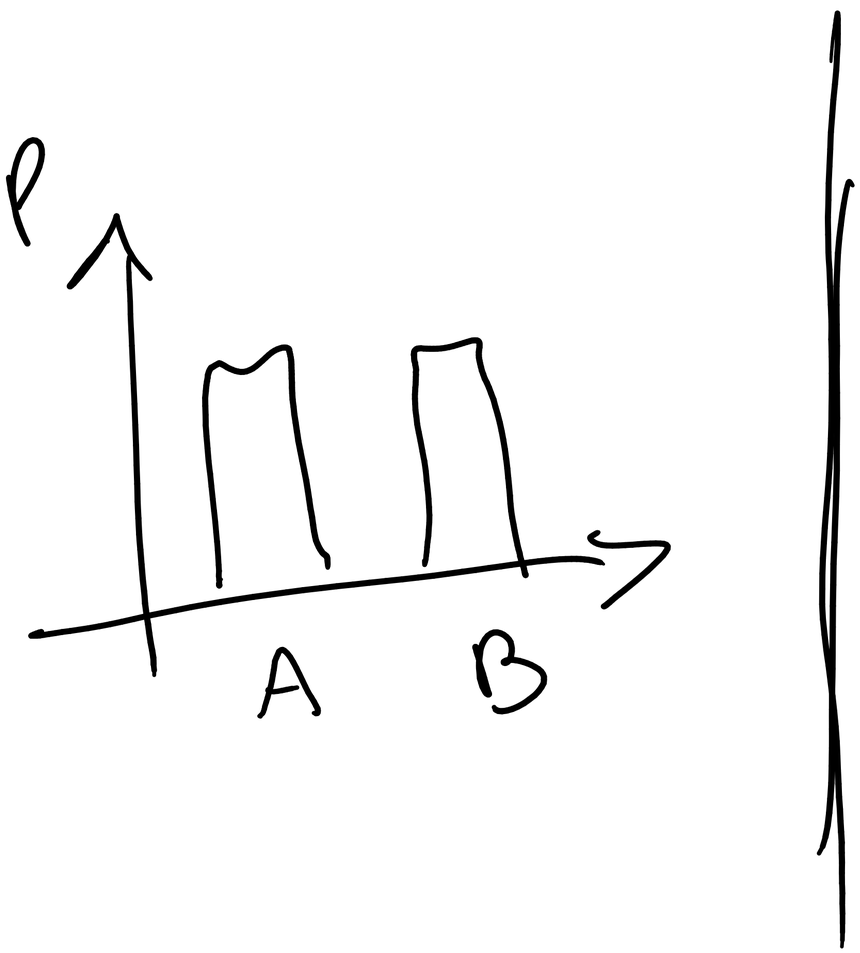
$\hat{f}_{TREE}$

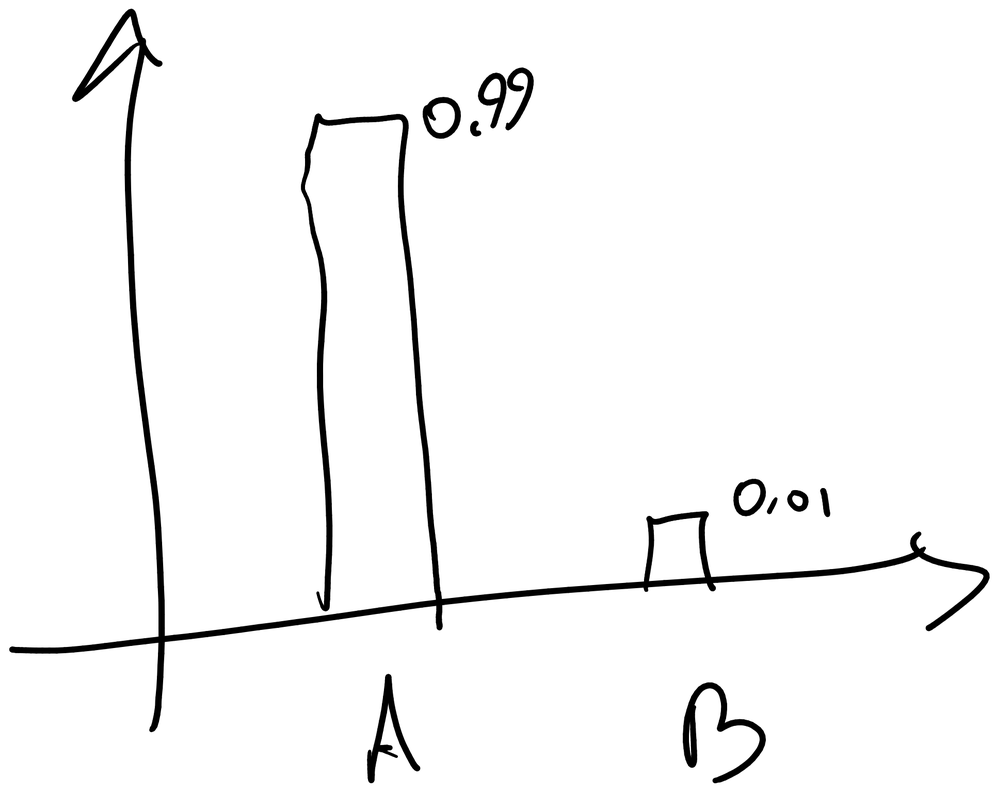
$\hat{f}_{RF}$ — X_{TEST} → ?

$\hat{f}_{GB}$ / (y_{TEST})

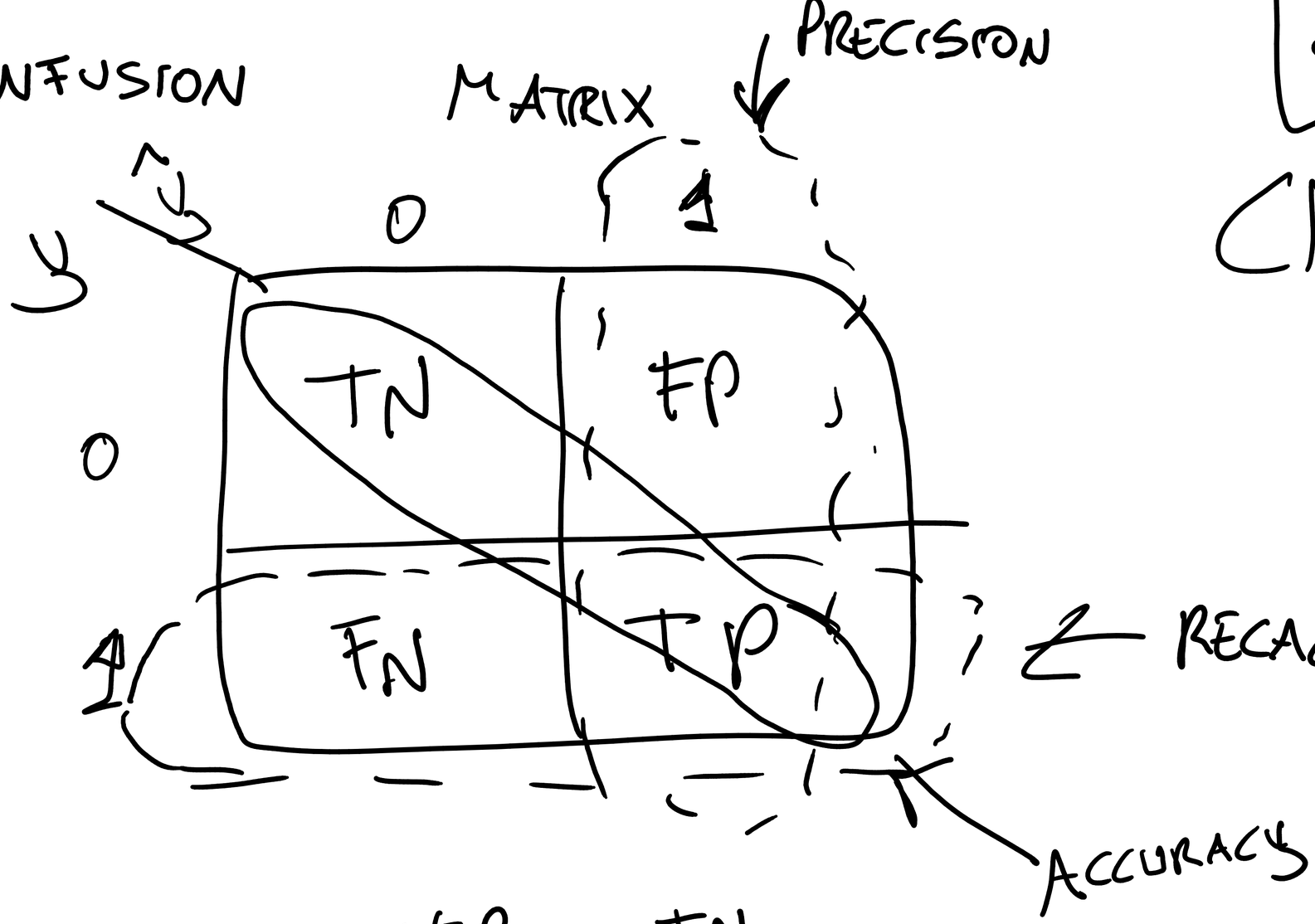
$y = \begin{cases} A \\ B \end{cases}$ BINARY CLASS.

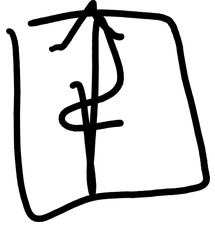
- UNBACANCED SAMPLE





— CONFUSION MATRIX





$CM(\Sigma)$

$$\hat{y} = \begin{cases} 0 & \hat{p} < \Sigma \\ 1 & \hat{p} \geq \Sigma \end{cases}$$

- ACCURACY : $\dfrac{TP + TN}{N}$
- PRECISION : $\dfrac{TP}{TP + FP}$
- RECALL : $\dfrac{TP}{FN + TP}$

$$F_1\text{-SCORE} = \frac{2}{\frac{1}{PREC} + \frac{1}{REC}}$$