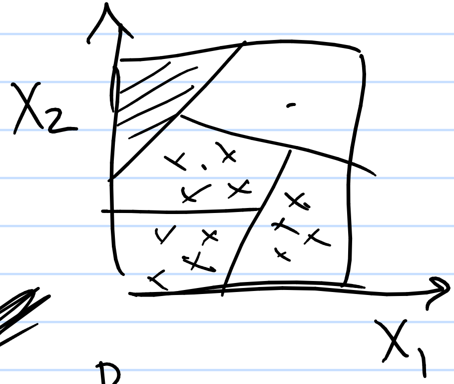
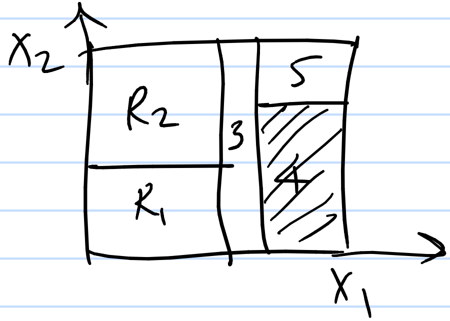
Tree model: partitioning the input space.

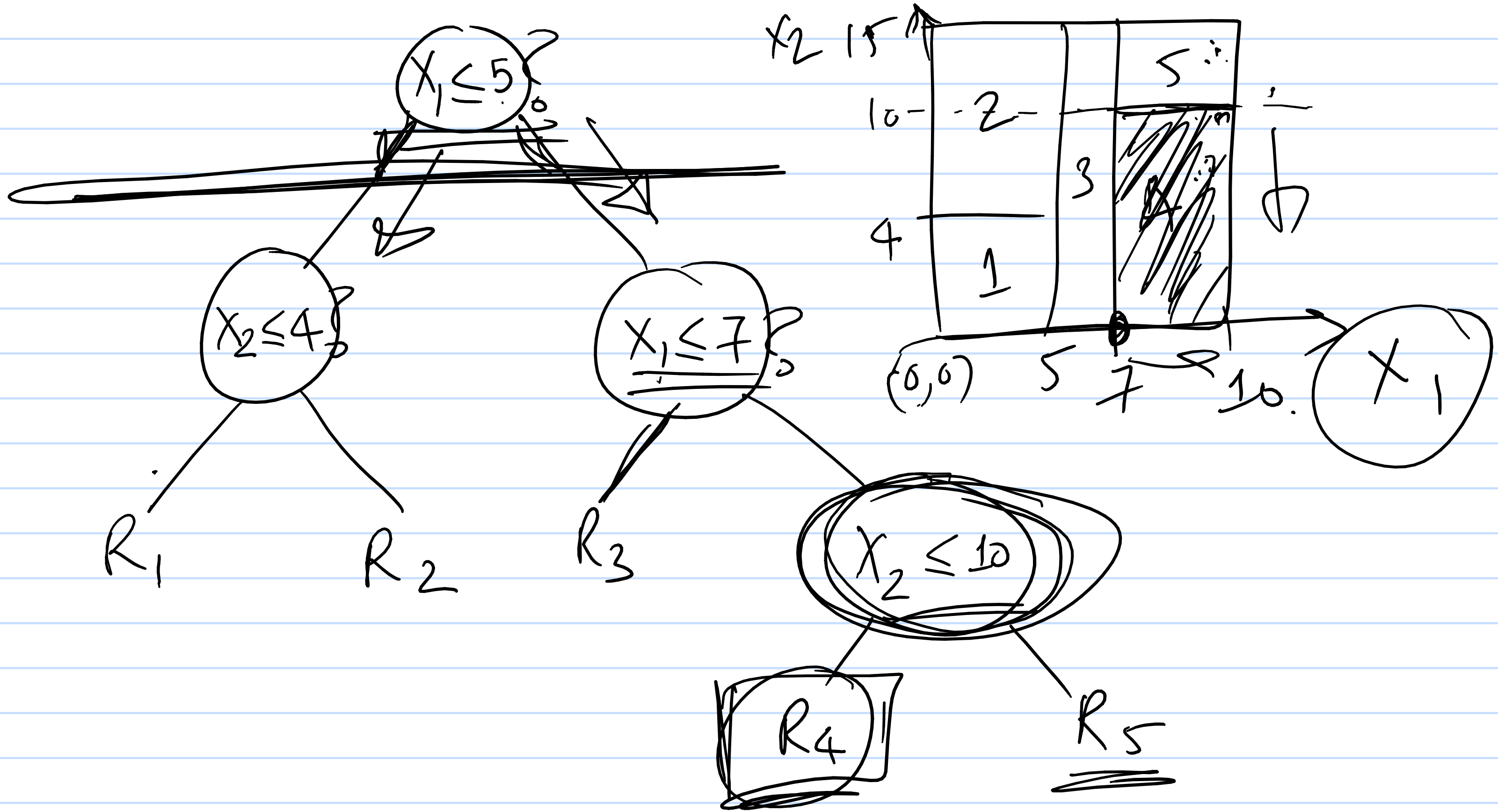
$$f(x) = \sum_{j=1}^{J} c_j \, \mathbb{1}[x \in R_j]$$

$x \in \mathbb{R}^p$



Regular.





$X_1 \le 5$?
$X_2 \le 4$?
$X_1 \le 7$?
$X_2 \le 10$
R_1
R_2
R_3
R_4
R_5
X_2
15
10
4
1
2
3
4
5
(0,0)
5
7
10
X_1

$$\{R_j, c_j\}_{j=1}^{J} = \theta$$

CART

(1) Regression tree

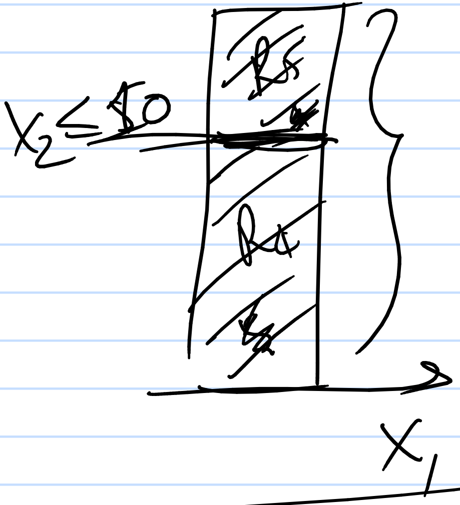
X_k, S

$$X_k \leq S$$

$$\boxed{\min_{\theta} \frac{1}{N} \sum_{i=1}^{N} \Big(y_i - \underbrace{f(x_i)}_{\sum_{j=1}^{J} c_j \mathbb{1}[x_i \in R_j]}\Big)^2}$$

$X_1 \dots X_p$

X_k



$1, s = 10$

$$= \min_{c_4} \sum_{x_i \in R_4(s)} (y_i - c_4)^2 + \min_{c_5} \sum_{x_i \in R_5(s)} (y_i - c_5)^2 =$$

$$\frac{d}{db}(a-b)^2 = -2(a-b)$$

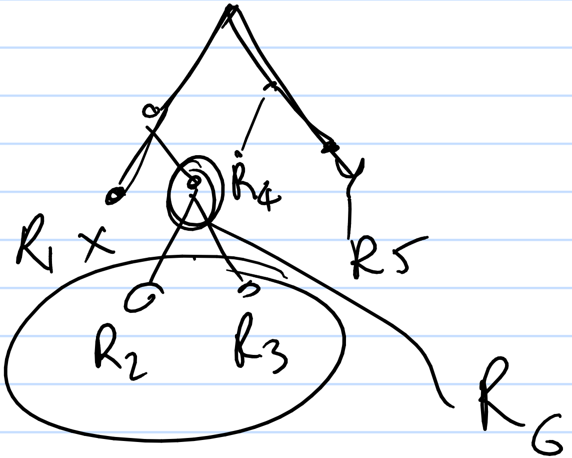
$$-\sum_{x_i \in R_4} \cancel{2}(y_i - c_4) = 0$$

$$\boxed{\frac{1}{\#}\sum y_i = c_4}$$

Post processing of a tree : pruning

$$\underbrace{\sum_{i:\ x_i \in R_j} \left(y_i - c_j \right)^2}$$

Node impurity function



R_j: $\hat{p}_1 = \dfrac{\#\{y_i = 1\}}{\# R_j}$

$\vdots$

$\hat{p}_3$

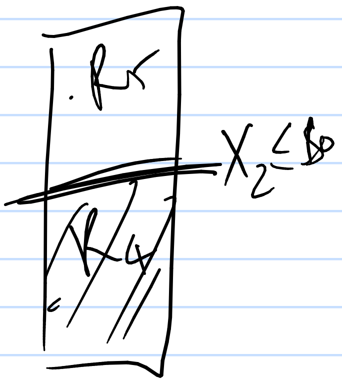
$K(j) = \arg\max(\hat{p}_1, \hat{p}_2, \hat{p}_3)$

Score

R_{j_1}: $1 - \hat{p}_{j_1 k(j_1)}$

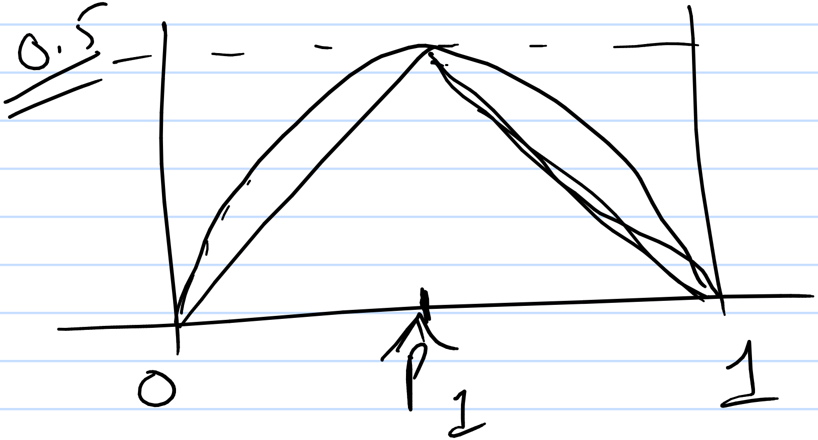
R_{j_2}: $1 - \hat{p}_{j_2 k(j_2)}$

1 S



Gini Index

Cross entropy



$$\boxed{0 \le \hat{p}_1 \le 1} \qquad \hat{p}_1 + \hat{p}_2 = 1$$

$$0 \le \hat{p}_2 \le 1$$

Issues

(1) Categorical features

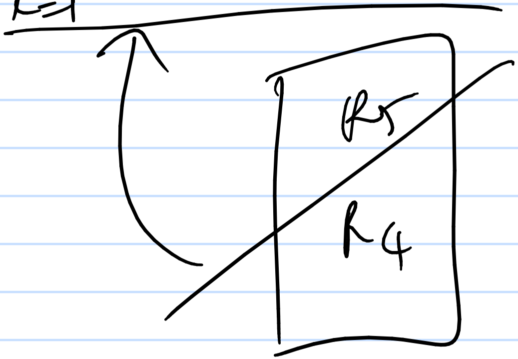
d f c

d f c

d f c

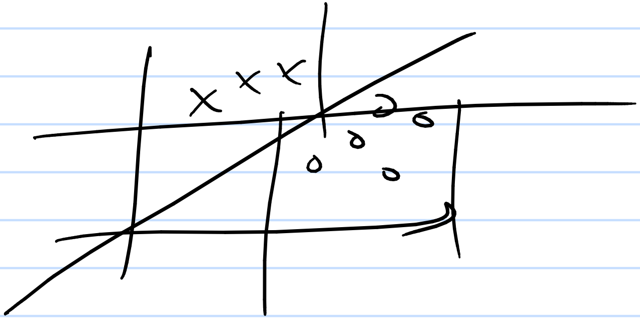
(2) $X_1 \leq S$

$$\sum_{k=1}^{p} \alpha_k X_k \leq S$$



(3) Stability / Robustness

(4) Smoothness



Missing data. // Corruptions / noise

$$X = \begin{bmatrix} & & na \\ na & & \\ & na & \end{bmatrix}_{N \times p} \quad Y = \begin{bmatrix} \\ na \\ \\ \end{bmatrix}_{N \times 1}$$

Selection bias

① Missing at random

$$P(R \mid X, Y) = P(R \mid X^{\text{observed}}, Y^{\text{observed}}) \quad R = \begin{bmatrix} 0 & 0 & 1 & 0 \\ 1 & 0 & & 0 \\ & & & 0 \\ & 0 & 1 & 0 \end{bmatrix}$$

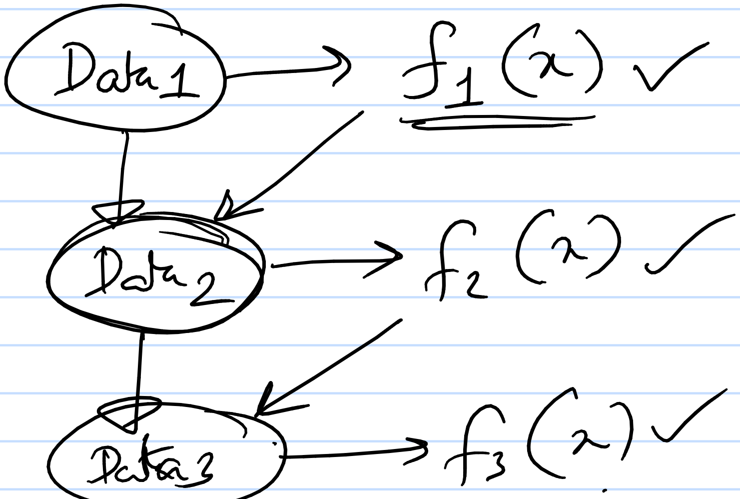
② Missing completely at random.

$$P(R \mid X, Y) = P(R)$$

Forward Stagewise Additive Models.

(1) Adaboost.M1 $\longrightarrow$ $f(x) = \text{sign}\left(\sum_{m=1}^{M} \alpha_m \underbrace{f_m(x)}_{\text{weak classifiers } \theta_m}\right)$

Binary classification +1, −1



weights.

$p = 3$

$$\vec{X} = \begin{bmatrix} 1 & 2 & .5 \\ -1 & .3 & 2 \\ 0.6 & 1.3 & 5 \\ .4 & 1 & 2 \end{bmatrix} \quad \begin{matrix} w_1 \\ w_2 \\ w_3 \\ w_4 \end{matrix} \quad \vec{y} = \begin{bmatrix} 10 \\ 5 \\ 4 \\ 1 \end{bmatrix}$$

$$\boxed{\sum_{i=1}^{N} \left(f_2(x_i) - y_i\right)^2 w_i}$$

Adaboost

$w_i \leftarrow w_i \exp(\alpha_1)$ if i was misclassified

α_1 = depends on overall misclassification error

1. FSAM

2. Adaboost $\in$ FSAM

   Why does it work?

3. GBDT $\in$ FSAM

FSAM.

1. $f_1(x) = 0$

for $m = 1, \ldots M$:

(a) $\theta_m = \underset{\theta}{\operatorname{argmin}} \sum_{i=1}^{N} L\left(y_i, f_{m-1}(x_i) + b(x_i; \theta)\right)$

(b) $f_m(x) = f_{m-1}(x) + b(x; \theta_m)$

Adaboost $\in$ FSAM

approximate.

Binary $\{-1,1\}$ Loss function $L(y, f(x)) = e^{-y\,f(x)}$ in FSAM

$$y_i,\quad f_{m-1}(x_i),\quad b(x_i;\theta)$$

$$\exp(-y_i(f_{m-1}(x_i) + b(x_i;\theta)))$$

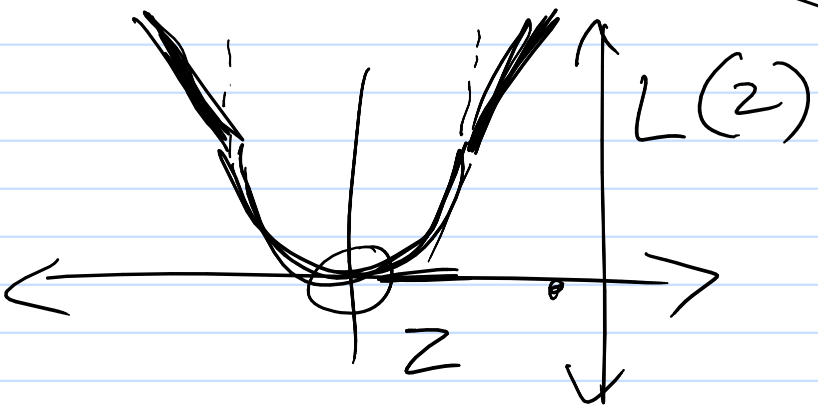
$$= \boxed{\exp(-y_i f_{m-1}(x_i))} \cdot \exp(-y_i\, b(x_i;\theta))$$

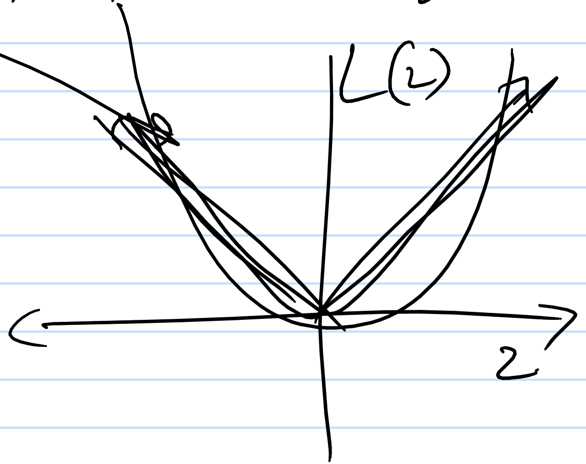
$\approx w_i$

Reg.

$$L(y, f(x)) = \begin{cases} (\underbrace{y - f(x)}_{z})^2 & \text{if } f(x) \text{ is small} \\ |y - f(x)| & \text{if } f(x) \text{ is large.} \end{cases}$$

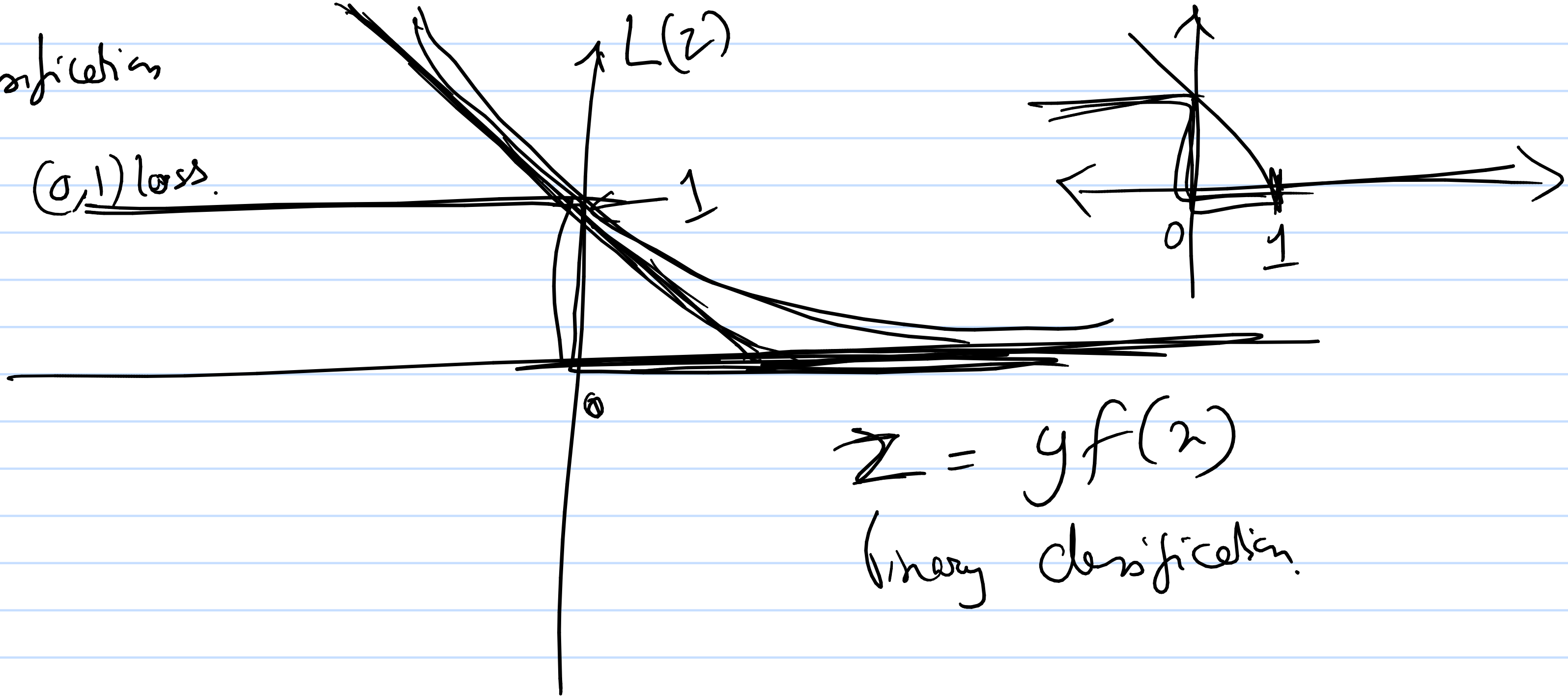
Huber loss function





Classification

(0,1) loss.



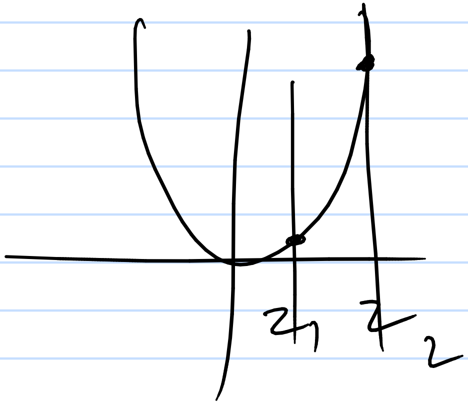
$$z = y f(x)$$

binary classification.

GBDT :

$\{x_i, y_i\}_N$

$f_{m-1}(x)$



$$g_{im} = \left.\frac{\partial L(y_i, z)}{\partial z}\right|_{z = f_{m-1}(x_i)}$$

$$f_m(x) = f_{m-1}(x) + T(x; \theta_m)$$

$$X \qquad \cancel{Y_{N\times 1}} \qquad - \begin{bmatrix} g_{1m} \\ \vdots \\ g_{Nm} \end{bmatrix}_{N\times 1}$$

$$\frac{\partial}{\partial z}(y_i - z)^2 = \left. \underline{\underline{-2(y_i - z)}} \right|_{z = f_{m-1}(x_i)} = g_{im}$$

$$= \underline{\underline{10}}$$

$$\min_{\theta} \sum_{i=1}^{N} L\left(y_i, f_{m-1}(x_i) + \underline{\underline{b(x_i;\theta)}}\right)$$

$$z_i$$

$$\min_{z_1, \ldots z_n} L\left(y_i, \underline{\underline{f_{m-1}(x_i)}} + z_i\right)$$

$$\boxed{-\underbrace{\frac{\partial L(y_i, z_i)}{\partial z_i}}\Bigg|_{z_i = f_{m-1}(x_i)}} = -g_{im}$$

$$f_m(x_i) = \underline{\underline{f_{m-1}(x_i)}} - \underbrace{\frac{\partial L(y_i, z_i)}{\partial z_i}}\Bigg|_{z_i = f_{m-1}(x_i)}$$