Lecture 22

Bagging example

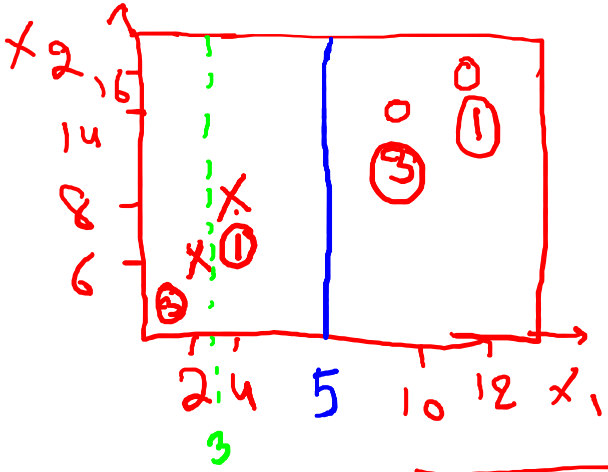
$$D = \begin{array}{cccccccc} \multicolumn{4}{c}{\overleftrightarrow{\quad 1 \quad}} & \multicolumn{4}{c}{\overleftrightarrow{\quad 0 \quad}} \\ \end{array}$$

$$D = \left\{ \begin{matrix} 1 & 2 & 3 & 4 & 10 & 11 & 12 & 13 \\ 5 & 6 & 7 & 8 & 14 & 15 & 16 & 17 \end{matrix} \right\} \text{ - Original 2D dataset}$$

$$-D_1 = \left\{ \begin{matrix} 2 & 2 & 10 & 12 & 10 & 10 & 2 & 4 \\ 6 & 6 & 14 & 16 & 14 & 14 & 6 & 8 \end{matrix} \right\}$$

$$\underbrace{\begin{array}{cccc} 1 & 1 & 0 & 0 \\ \multicolumn{4}{c}{\left\{ \begin{matrix} 1 & 3 & 11 & 13 \\ 5 & 7 & 15 & 17 \end{matrix} \right\}} \end{array}}_{\text{Val}}$$

Build a tree



$D_2, D_3 \ldots D_n$

$\boxed{G_{right}: \frac{1}{5}\cdot\frac{4}{5} + \frac{4}{5}\cdot\frac{1}{5}}$

Suppose one split is at 3 & other at 5

Gini index for '3' $\Rightarrow 2p(1-p)$

Where 'p' is prob. of one class

$G_{1/left} \rightarrow 2\cdot 1(1-1)$

$\cancel{G_{2/right} \rightarrow 2\cdot\frac{1}{3}\cdot\frac{2}{3} = \frac{4}{9}}$

Gini overall =

$\cancel{\frac{1}{4}\cdot 0 + \frac{3}{4}\cdot 4/9}$

$= \cancel{1/3}$

$G_{left/right}$ for split at '5' is ? $= 0$

Tree is obtaind with split at 5

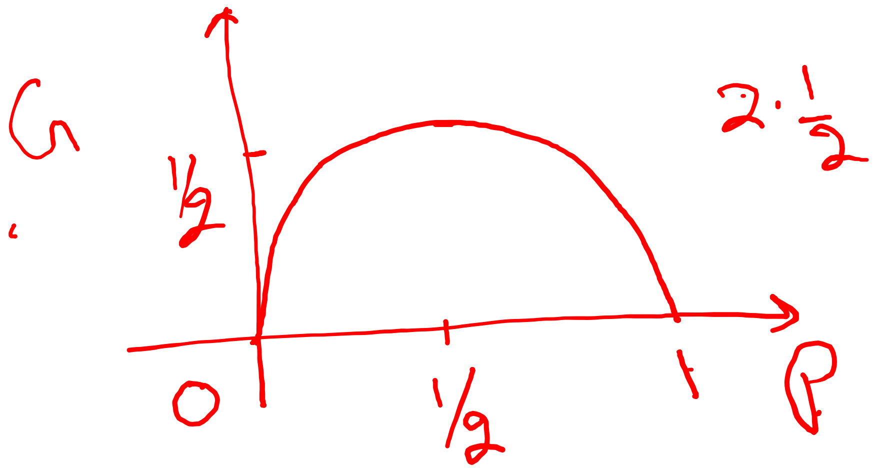
*) Compute OOB → samples $\left\{ \begin{matrix} 1 & 3 & 11 & 13 \\ 5, & 7, & 15, & 17 \end{matrix} \right\}$

Prediction for these samples are all correct. Why?

OOB error for this tree $= 0$

OOB → Out of Bag

OOB error for split at 3? $= 1/4$



$2 \cdot \frac{1}{2}\left(1 - \frac{1}{2}\right) = \frac{1}{2}$

$G = 2p(1-p)$
$= 2 \cdot 1 \cdot 0 = 0$

If a node had ~~ca~~ samples of single class only, then
Prob. for that class $\rightarrow$ 1
" " other class $\rightarrow$ 0

$\rightarrow$ overall Gini $\rightarrow$ $P_L G_{left} + P_R G_{right}$

$0 + \frac{5}{8} \cdot \frac{8}{25} = \frac{1}{5}$

For right node with split at '3':

There are 5 samples, $\left\{ \begin{matrix} 4 \\ 8 \end{matrix} \right\}$ class 1

~~and~~ remaining class → 0

Prob. of class 1: → $1/5$

$$G_{right} = 2 \cdot \frac{1}{5} \frac{4}{5} = 8/25$$

Orig. dataset $\rightarrow D$ of 'n' samples.

for $i = 1$ to B

$\rightarrow$ Bootstrap dataset D_i of 'n' sample

$\rightarrow$ Obtain Tree using $D_i \rightarrow T_i$

$\rightarrow$ Compute OOB error for samples absent in D_i but present in D

end

~~Obtain~~ For a given test sample x_*, predict

$$\frac{1}{B} \sum_{i=1}^{B} T_i(x_*) \quad // \text{regression}$$

$D_1 \rightarrow$ learn $h_1(x) : x \in \mathbb{R}^d \rightarrow \{1, 0\}$

$\rightarrow$ Compute OOB error on D_1 using h_1

$\hookrightarrow$ val of

$D_2 \rightarrow$ learn $h_2(x)$

compute OOB error on D_2 using h_2

⋮

$h_n(x)$

Given any test sample $(x^*) \rightarrow$ compute $h_1(x^*), h_2(x^*) \ldots$

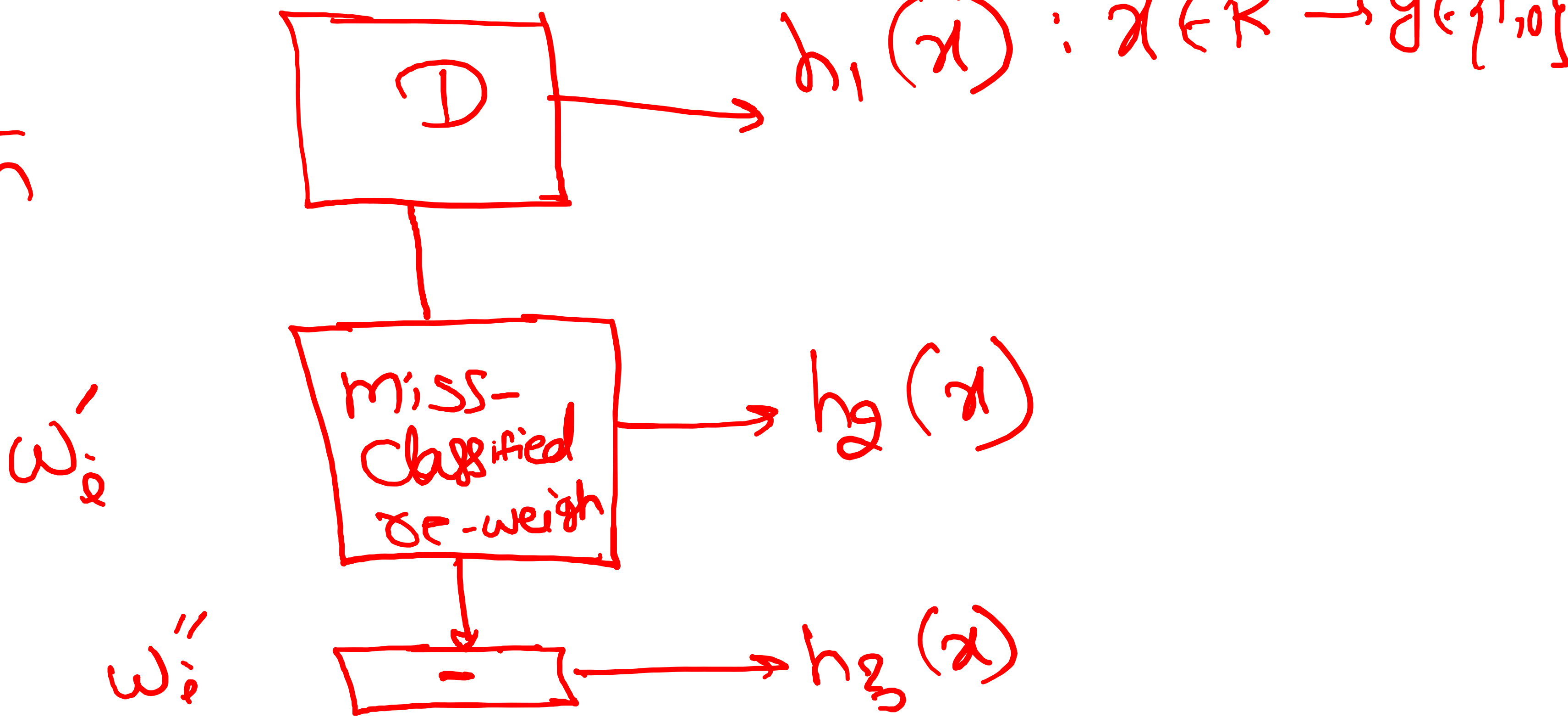
$h_n(x^*)$

* Boosting → we can use to boost the performance of weak classifiers

weak classifiers perform slightly better than random.

$D = \{x_i, y_i\}_{i=1}^{n}$ $\quad y_i \in \{1, 0\}$

weights $w_i = \frac{1}{n}$



Let's say we learn 'm' number of classifiers.

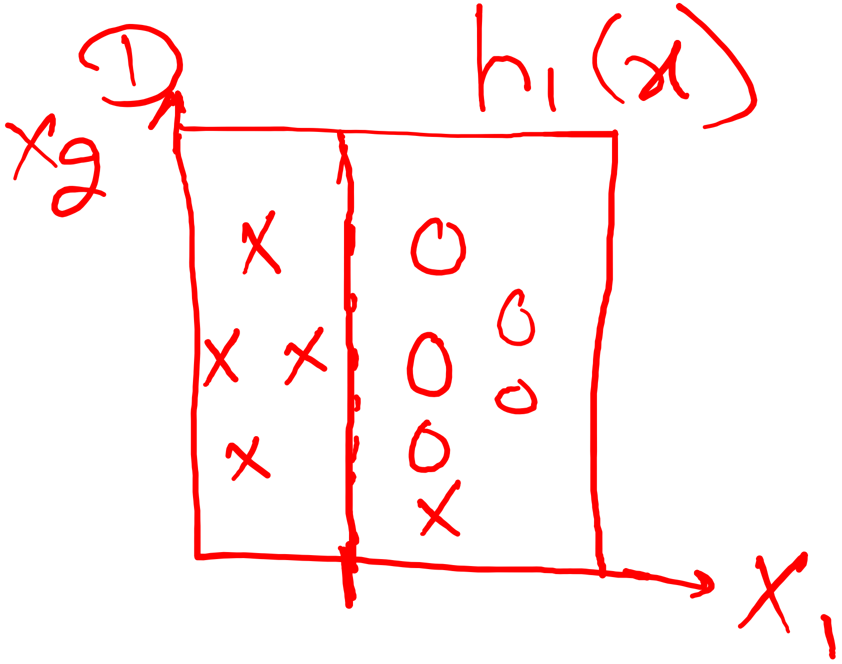
Final classifier: $f(x) = \text{sign}\left[\sum_{j=1}^{m} \alpha_j h_j(x)\right]$

(1) for $j = 1$ to m

learn $h_j \leftarrow \underset{H}{\arg\min}\ L_j$

$H \rightarrow$ is the class of classifiers

$L_j \rightarrow$ is a loss. $L_j = \dfrac{\sum_{i=1}^{n} w_i \mathbb{I}(y_i \neq h_j(x_i))}{\sum_i w_i}$



$w_i = \frac{1}{10} \quad \forall i$

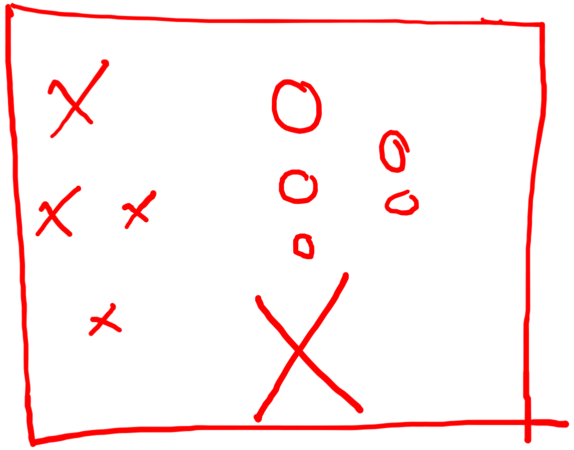
$H \rightarrow$ comprised of all tree classifiers.

$$L_1 = \frac{1/10}{1}$$

2 $\alpha_j = \frac{1}{2} \log\left\{\frac{1 - L_j}{L_j}\right\}$

3 $w_i \leftarrow w_i e^{2\alpha_j} \quad \forall$ misclassified sampled.

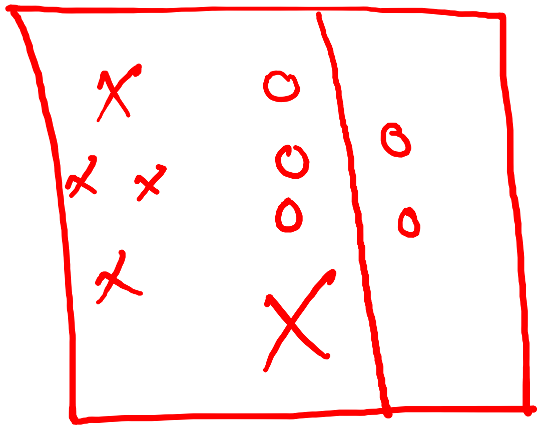
1)



Weigh for cross.

$$\frac{1}{10} e^{2 \cdot \frac{1}{2} \log 9} = \frac{9}{10}$$

$h_2(x)$



$$L_j = \frac{\frac{1}{10}+\frac{1}{10}+\frac{1}{10}}{\frac{9}{10}+\frac{9}{10}} = 3/18$$

2) $\alpha_j = \frac{1}{2} \log \frac{1-3/18}{3/18} = \frac{1}{2} \log 5$

$$w_i \leftarrow \frac{1}{10} e^{2\cdot\frac{1}{2}\log 5} = \frac{1}{2}$$

D