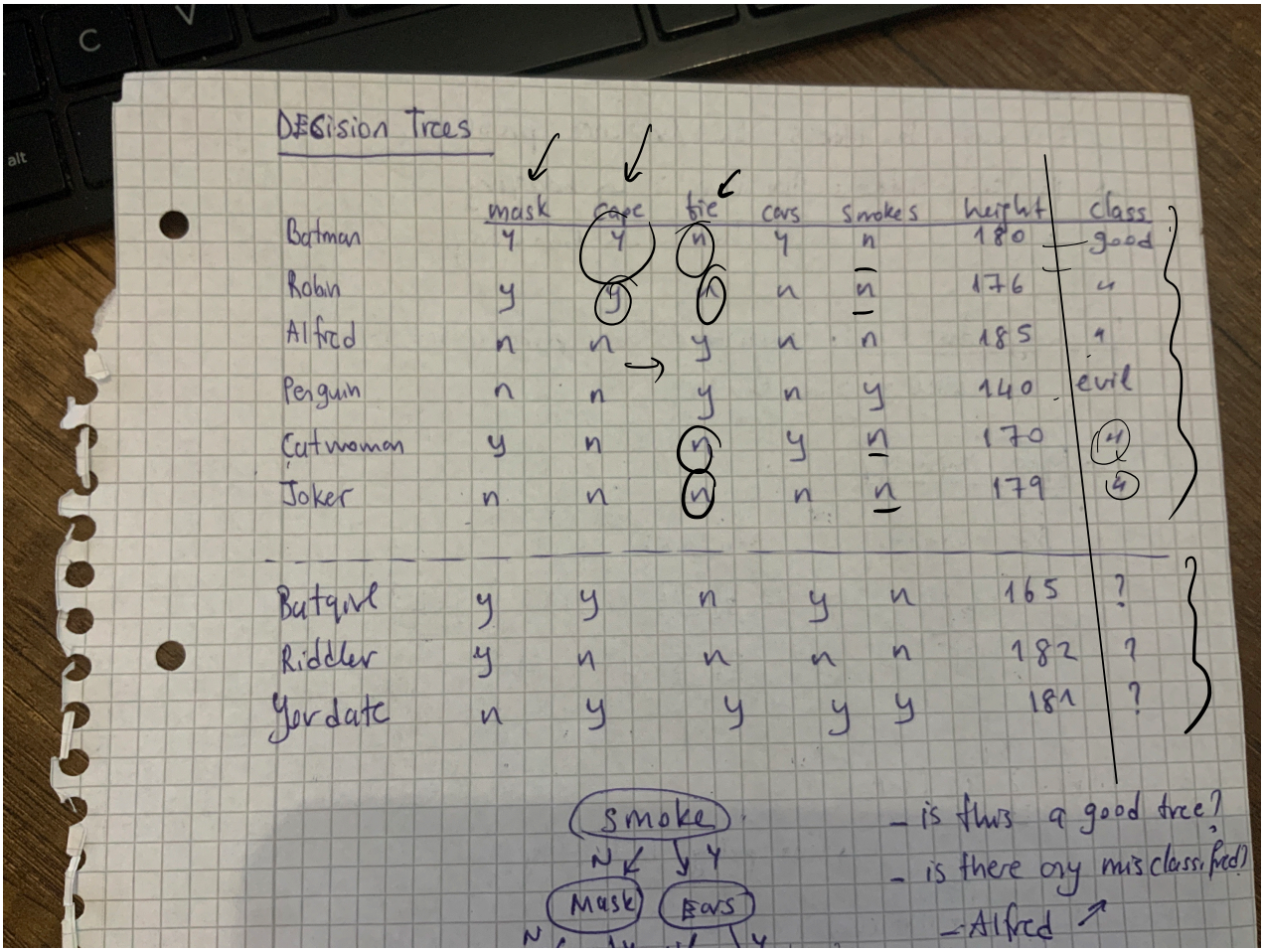
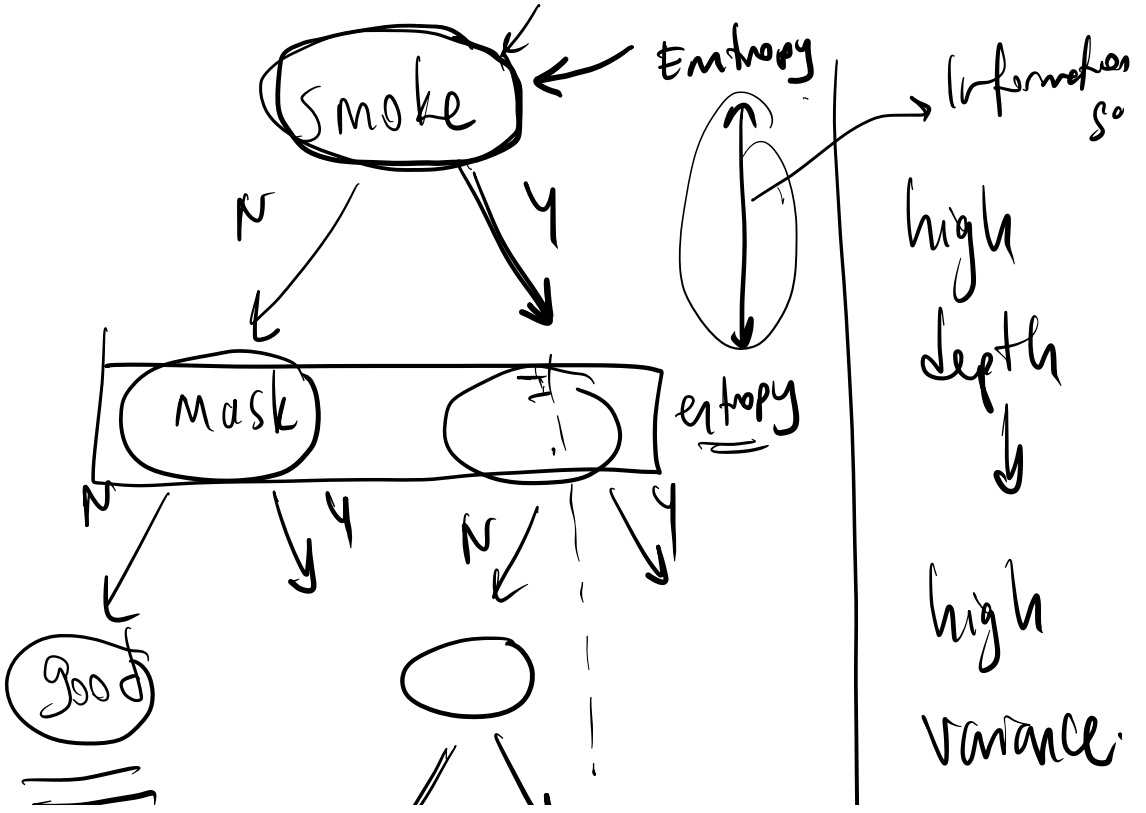
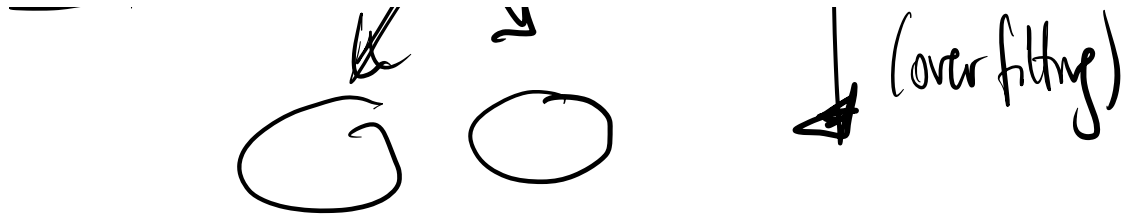
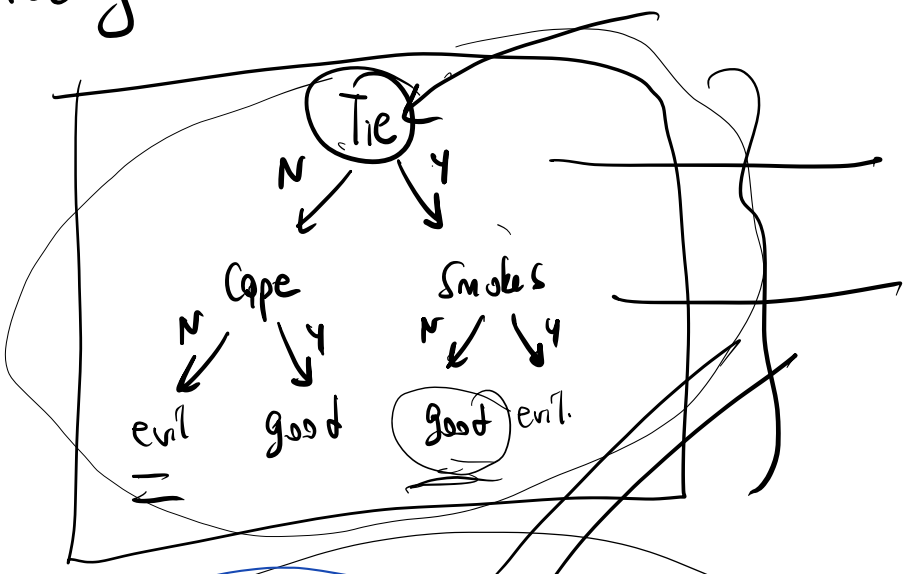

DEcision Trees
mask cape tie cars smokes height class
Batman y y n y n 180 good
Robin y y n n n 176 "
Alfred n n y n n 185 "
Penguin n n y n y 140 evil
Catwoman y n n y n 170 "
Joker n n n n n 179 "
Batgirl y y n y n 165 ?
Riddler y n n n n 182 ?
Yordale n y y y y 181 ?
smoke
N Y
Mask Cars
- is this a good tree?
- is there any misclassified?
- Alfred

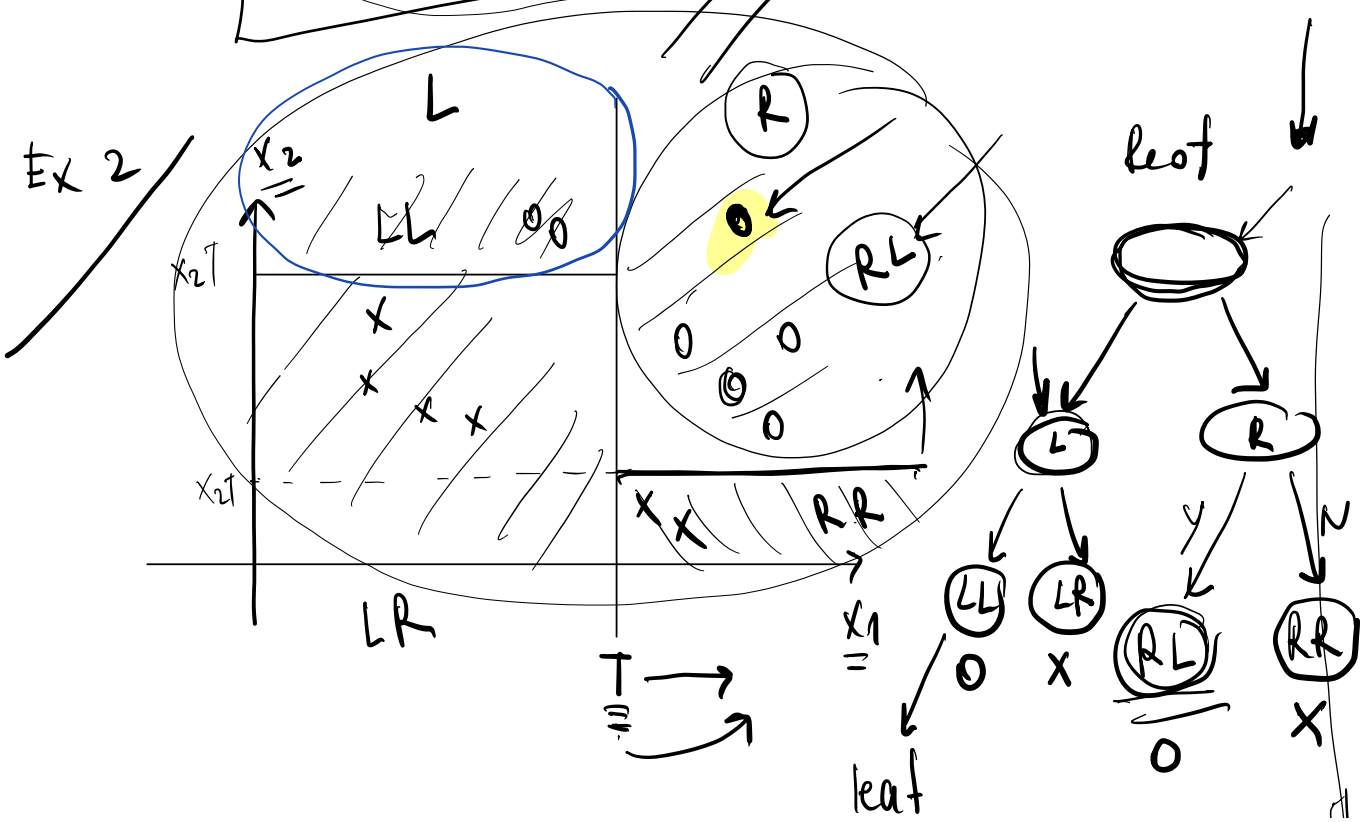

smoke
N
Y
Mask
H
Good
Entropy
entropy
Information
so
high
depth
high
variance



* find smallest tree that classify everything correctly.



Ex 2



How to split a tree node?

Data: $S = \{(x_1, y_1), \ldots (x_n, y_n)\}$, $y_i \in \{1, \ldots, c\}$

c is the # of classes.

Impurity functions: measures how pure a set is in terms of label (same label)

Gini impurity

↓

0 (perfect purity, all elements in the subset belong to the same class).

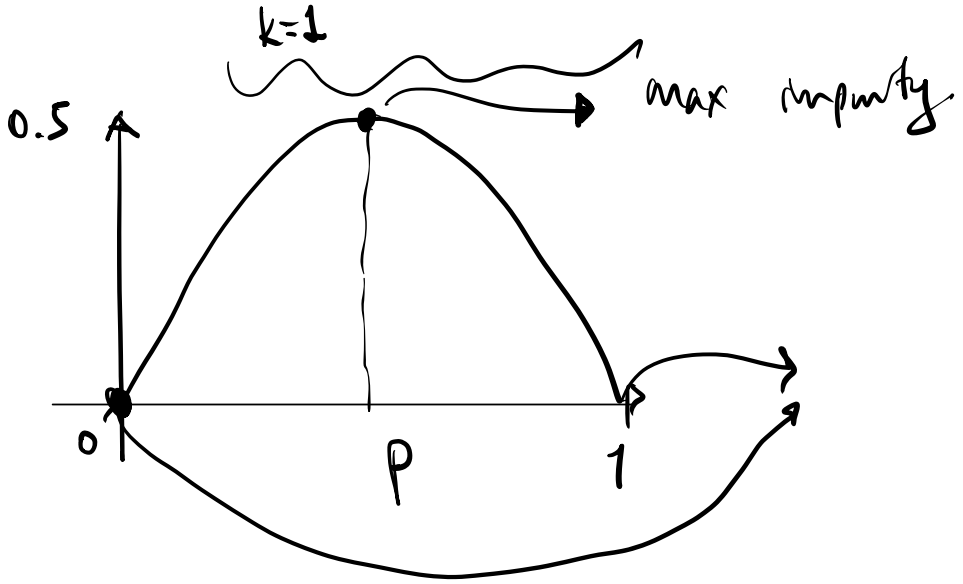
0.5 (complete impurity, where instances are evenly dist. across the classes)

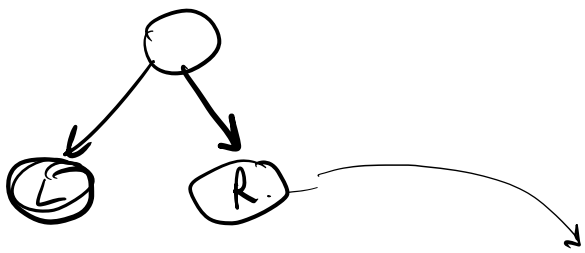
lets $S_k \subseteq S$ where $S_k = \{(x, y) \in S : y = k\}$

$S = S_1 \cup S_2 \ldots \cup S_c$

$$p_k = \frac{|S_k|}{|S|}$$ ← fraction of inputs in S with label k

$$G(S) = \sum_{k=1}^{c} p_k (1 - p_k)$$





$$G^T(S) = \frac{|S_L|}{|S|} G^T(S_L) + \frac{|S_R|}{|S|} G^T(S_R)$$

Ex/ (10) instances of class "Yes"

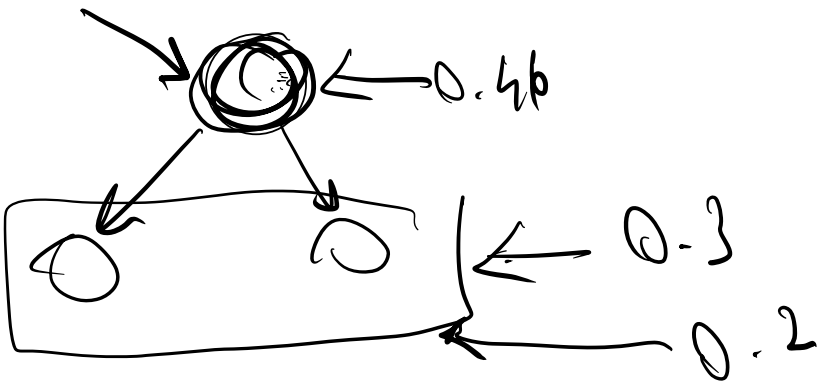
(6) " " " "No".

$$P_{yes} = \frac{10}{10+6} = \frac{10}{16} \quad , \quad P_{No} = \frac{6}{16}$$

$$Gini(S) = 1 - (p^2_{yes} + p^2_{No})$$

$$= 1 - \left(\left(\frac{10}{16}\right)^2 + \left(\frac{6}{16}\right)^2\right)$$

$$\simeq 0.46875$$



Entropy

$$\text{Entropy}(S) = -\sum_{i=1}^{c} p_i \log_2(p_i)$$

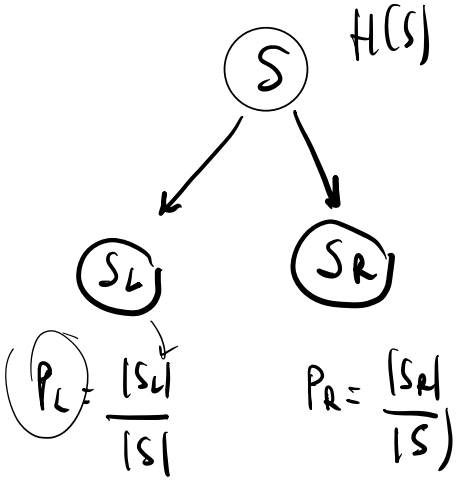
Ex/ 9 samples of class → "YES"
5 samples of " → "NO"

$p_{yes} = 9/14 \qquad p_{No} = 5/14$

$$\text{Entropy}(S) = -\left(p_{yes} \log_2(p_{yes}) + p_{No} \log_2(p_{No})\right)$$

$$= -\left(9/14 \log_2 9/14 + 5/14 \log_2 5/14\right)$$

$$\simeq 0.939$$



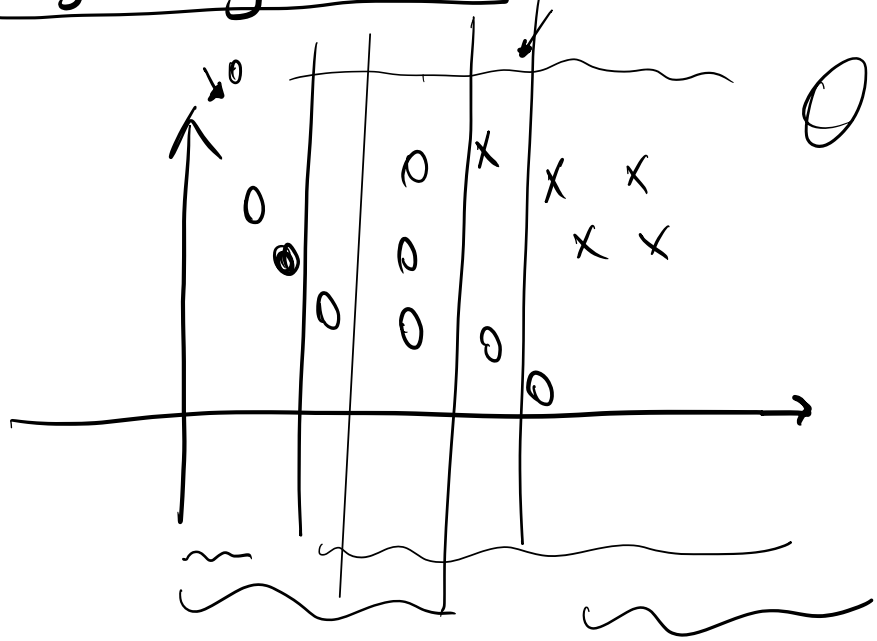
$$H(\text{Tree}) = P_L \cdot H(S_L) + P_R H(S_R)$$

- How do you find the optimal split?

↳ NP-hard

// greedy approach ←

→ try every feature.

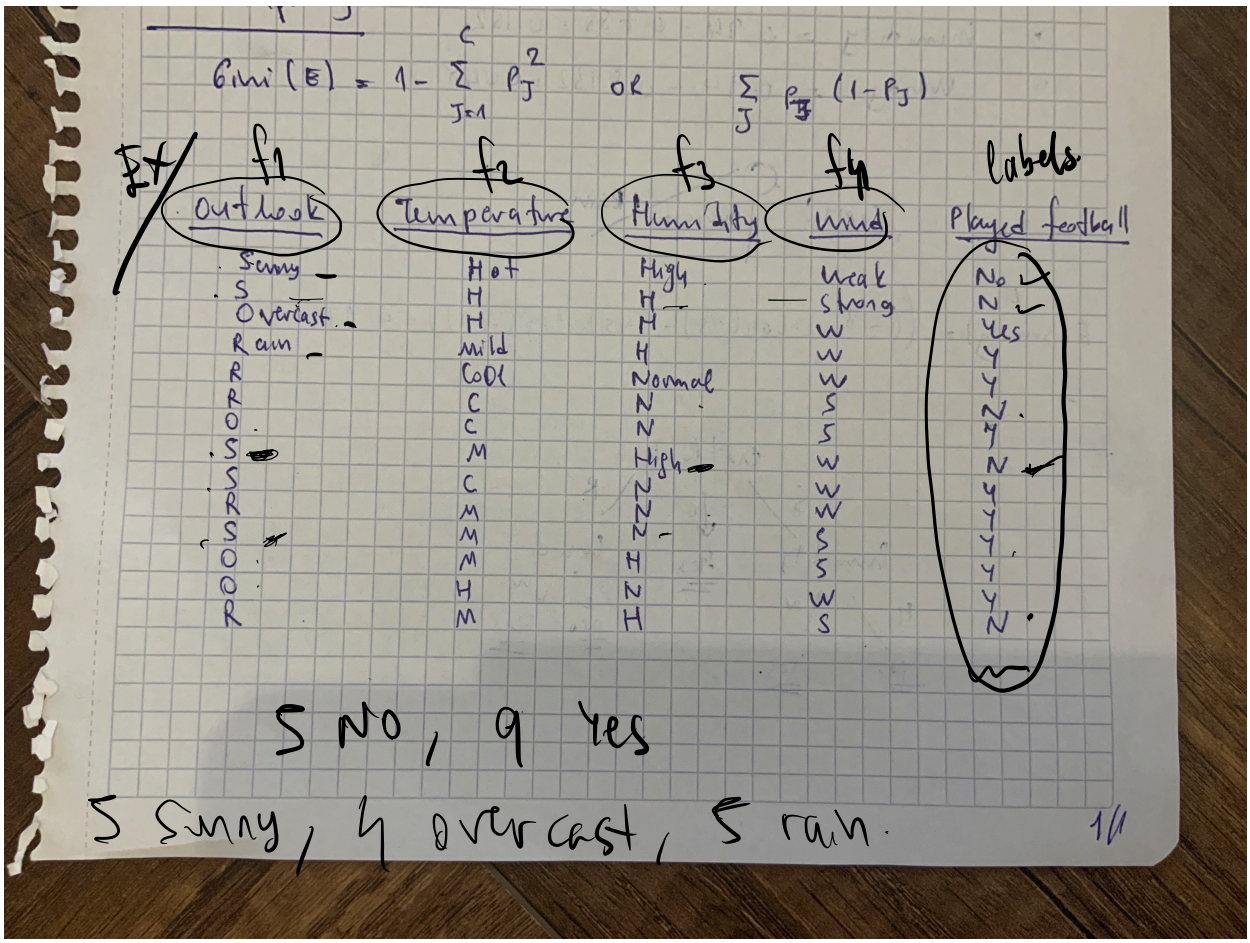


$O(n \cdot D \cdot K)$

Information Gain.

$IG(S, A) = H(S) - H(S, A)$

Adv of D.T.



$$E(S) = -\left[\frac{9}{14} \cdot \log\left(\frac{9}{14}\right) + \left(\frac{5}{14}\right) \cdot \log\left(\frac{5}{14}\right)\right]$$

$$= 0.94$$

$$E(S, \text{Outlook}) = \frac{5}{14} \cdot E(3,2) + \frac{4}{14} \cdot E(4,0)$$

Sunny

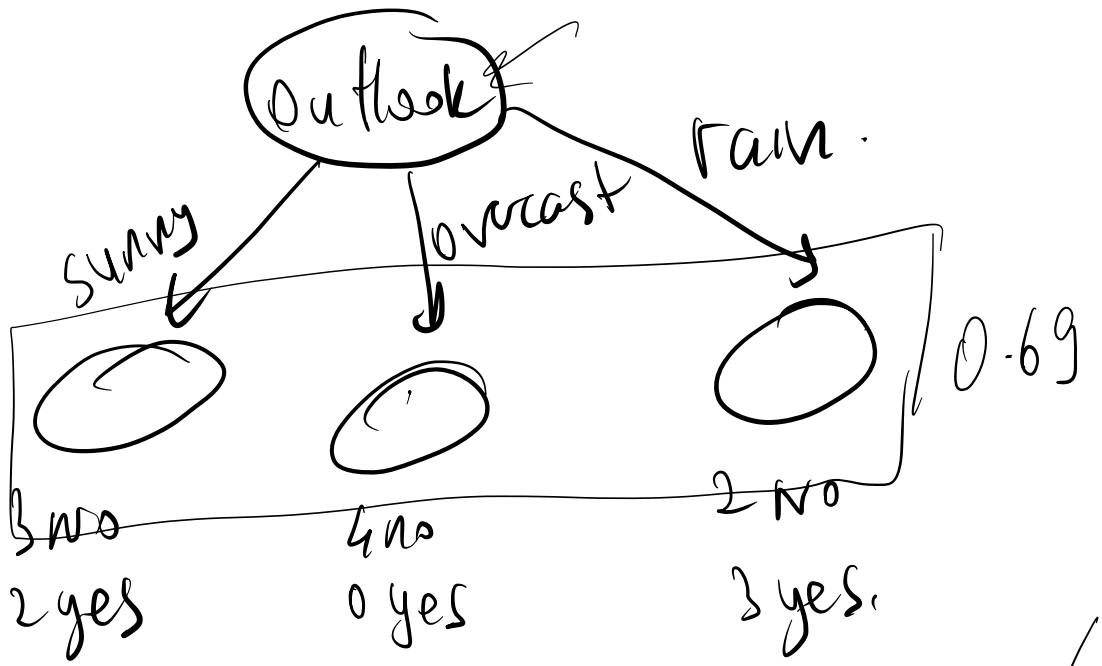
No Yes

0

$+ \quad 5/14 \; * \; E(2,3)$

$\Rightarrow 5/14\left(-(3/5)\cdot\log(3/5) - (2/5)\log(2/5)\right) + 4/14\,(0) +$

$5/14\left(-(2/5)\log(2/5) - (3/5)\log(3/5)\right) = 0.693$

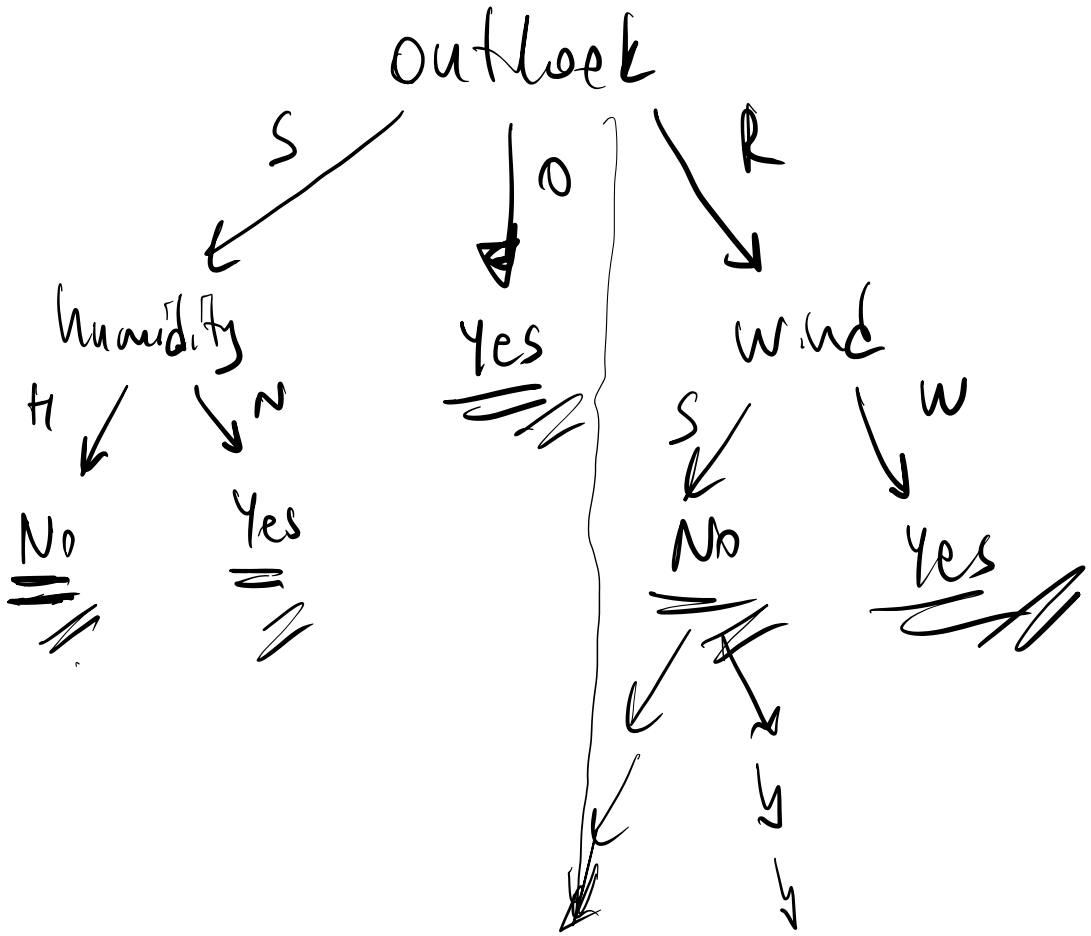


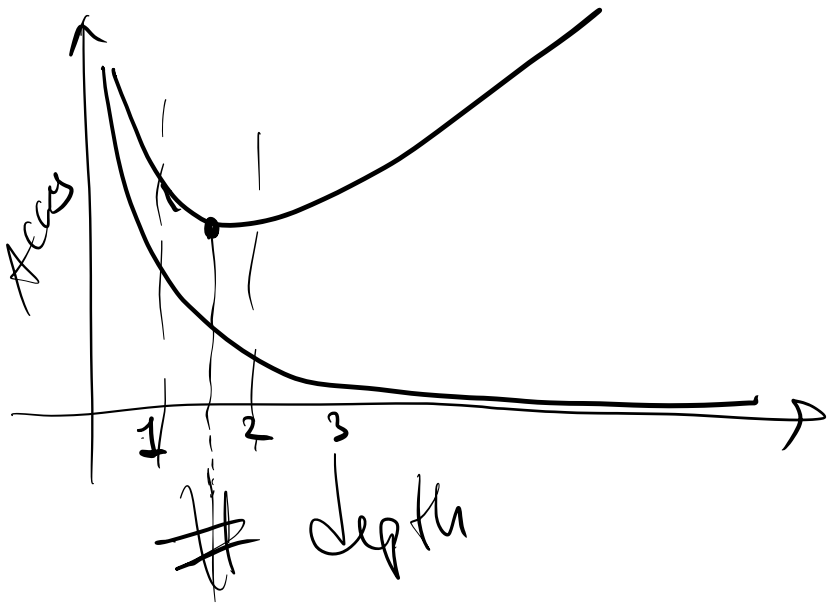
$IG(S, \text{Outlook}) = 0.94 - 0.693 = 0.247$

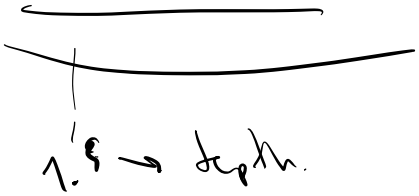
$IG(S, \text{temperature}) = 0.94 - 0.911 = 0.029$

$IG(S, \text{humidity}) = 0.94 - 0.788 = 0.152$

$IG(S, windy) = \boxed{0.94} - \boxed{0.8932} = 0.048$



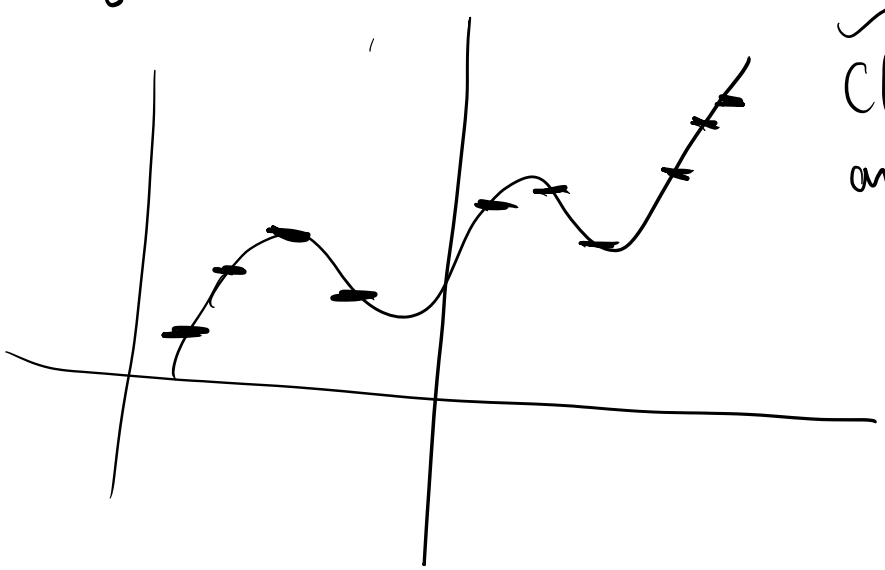




Regression trees ↙

CART

Classification and regression trees.



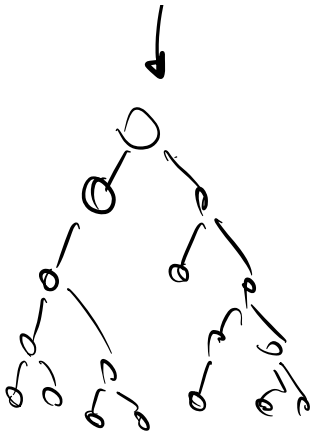
Ensemble learning

Bagging

- D.T face with variance ↑

$$E\left[\left(h_D(x) - \bar{h}(x)\right)^2\right]$$

variance → actual / expected.



we want $h_D \to \bar{h}$

weak law of large numbers

$$\frac{1}{m} \sum_{i=1}^{m} \underline{x_i} \to \bar{x} \text{ as } m \to \infty$$

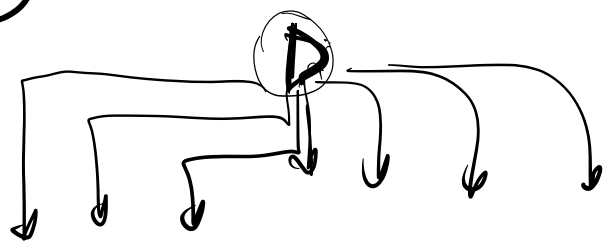
Assume we have "m" training sets $D_1, D_2 \ldots D_m$ drawn from P^m.

Initial

$D \to h_D$ $\quad \hat{h} = \frac{1}{m} \sum_{i=1}^{m} h_{D_i} \to \bar{h}$ as $m \to \infty$

D.

Bagging (Bootstrap Aggregation).



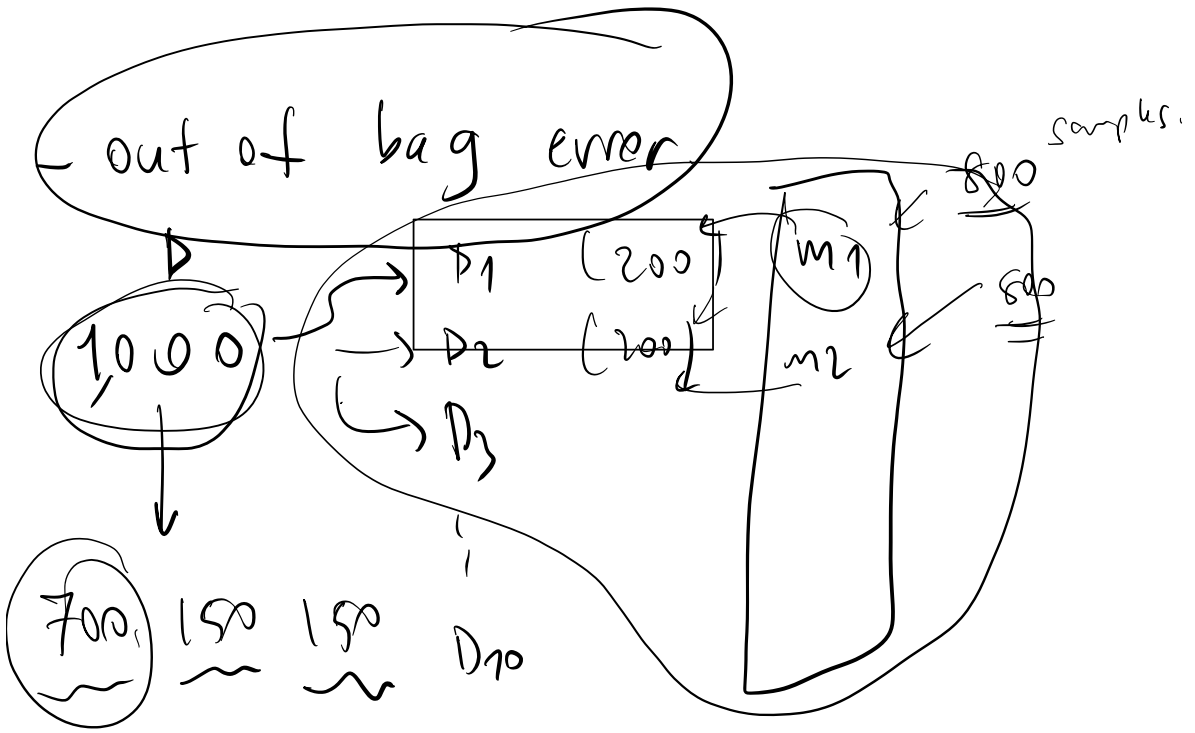
D_1 D_2 D_3 D_4 D_5 D_6 -- D_m

- drawing uniformly with replacement

let $\theta(x, y | D)$ be a prob. dist. that picks a training sample (x_i, y_i) from D uniformly at random).

1. Sample m data sets $D_1 \dots D_m$ from D with replacement.
2. for each D_j train a classifier h_j.

3. The final classifier is $h(x) = \frac{1}{m} \sum_{j=1}^{m} h_j(x)$



Random forest (Bagging)

original dataset

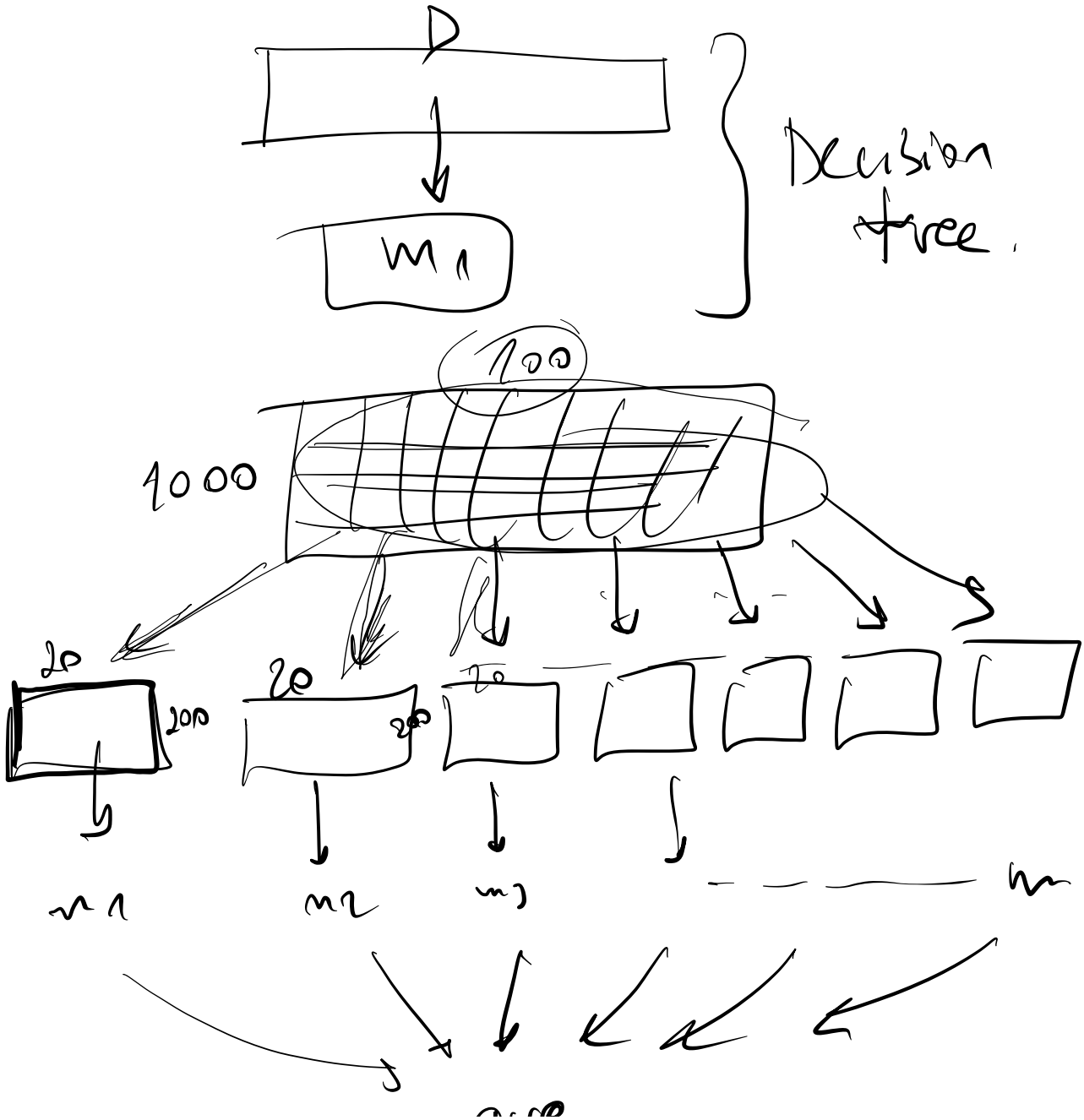
1. Sample "m" data sets $D_1 \dots D_m$ from D with replacement.

2. For each D_j train a full decision tree $h_j()$ max-depth ∞

with small modification: before each split randomly

Sub sample $k \leq d$ features (without replacement) and only consider those for your split.

3. final classifier is $h(x) = \frac{1}{m} \sum_{j=1}^{m} h_j(x)$



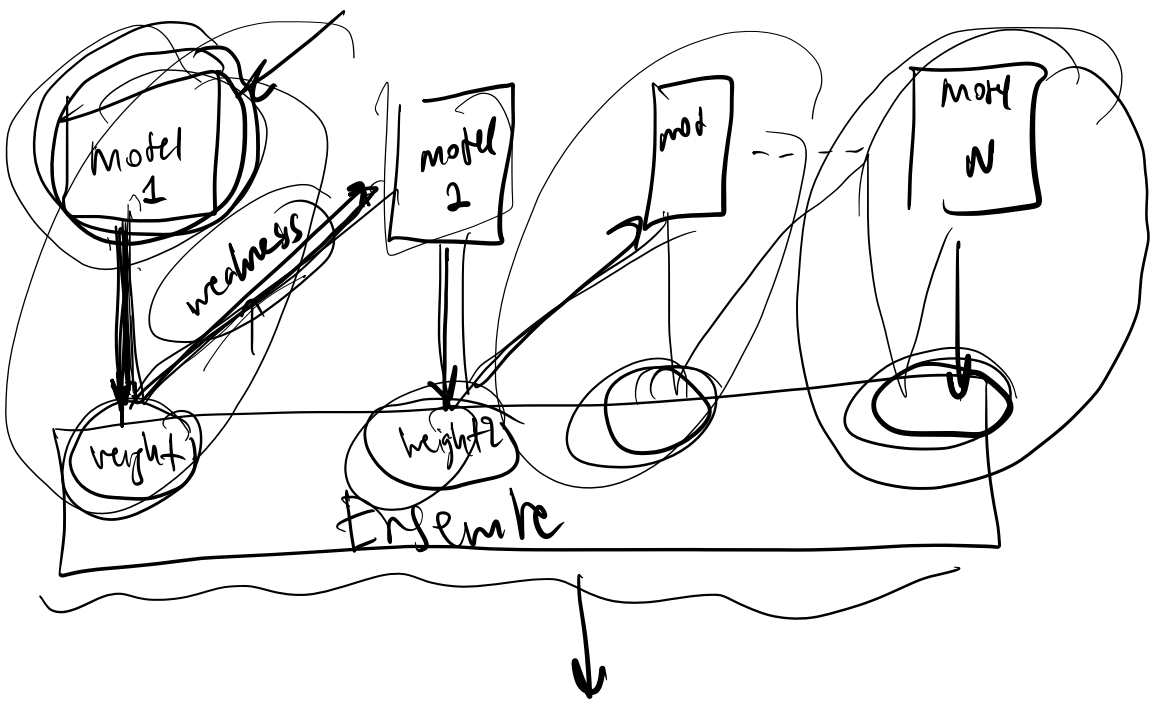
avg -

Boosting.

- high bias

$$h(x) = \sum_{J=1}^{m} \alpha_J h_J(x).$$

→ Can we combine weak learners? (Strong learner)



Adaptive boosting : AdaBoost.

Gradient boosting:

XGBoost