TREE MODELS

8/7/2023

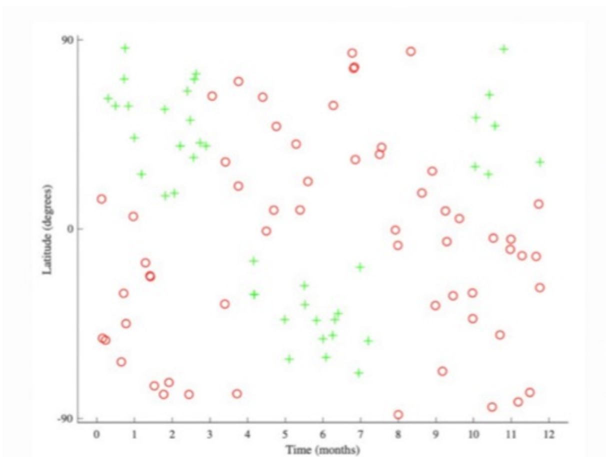
$\mathbb{1}[x_1 \geq 0]$

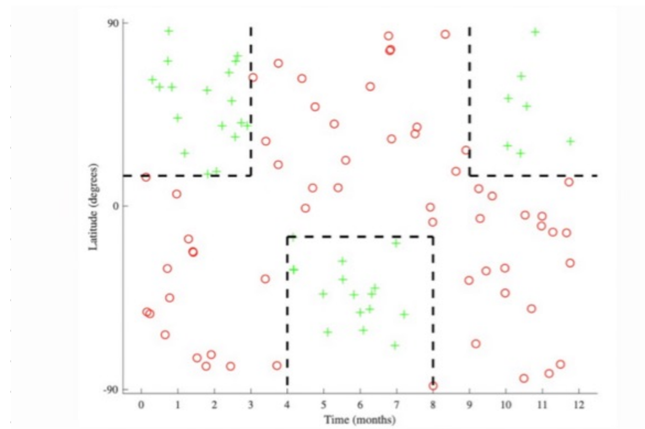
$\mathbb{1}[x_2 \leq 0]$ $\qquad$ $\mathbb{1}[x_2 \geq 0]$

$\mathbb{1}[x_1 \geq 1,1]$

$\mathbb{1}[x_2 \geq 0,5]$ $\qquad$ $\mathbb{1}[x_2 \geq 1]$

F: 1 $\quad$ T: 0 $\qquad$ F: 1 $\quad$ T: 0





d NUMBER OF FEATURES

n NUMBER OF EXAMPLES

$$\mathcal{X} = \bigcup_{i=1}^{n} R_i$$

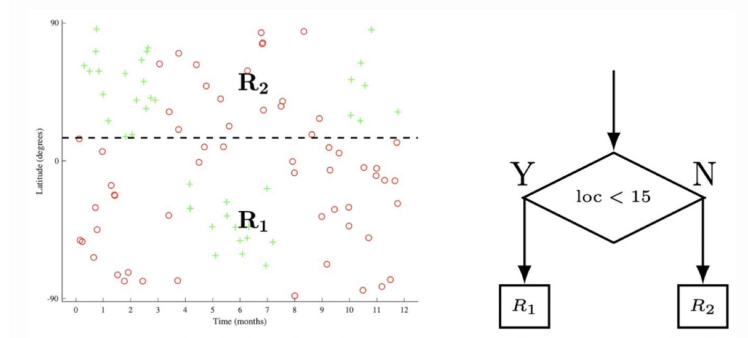
$$R_j \cap R_i = \emptyset \quad (i \neq j)$$

$R_i \cup R_j = R_p$ PARENT (Ex $\mathbb{R}^d$)

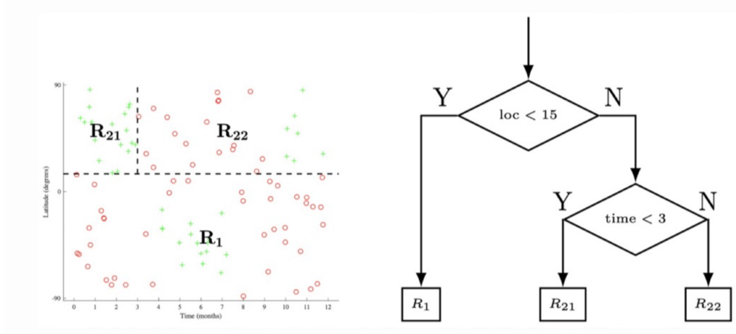
WE CAN DEFINE A REGION IN THE FOLLOWING WAY

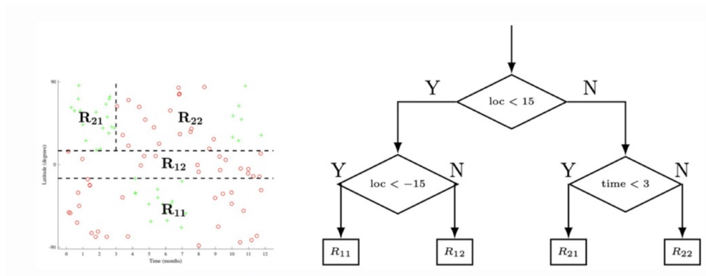
$$R_1 = \{ X \mid X_j < t,\ X \in R_p \} \longrightarrow$$

$$R_2 = \{ X \mid X_j \geq t,\ X \in R_p \}$$



AND CONTINUE TO SPLIT THE SPACES IN ORDER TO CLASSIFY, IN THIS CASE





WE CAN DEFINE A LOSS FUNCTION AS A SET FUNCTION ON A REGION $\mathbb{R}$
GIVEN A SPLIT OF A $\mathbb{R}_p$ INTO $\mathbb{R}_1$, $\mathbb{R}_2$ WE CAN COMPUTE $\mathcal{L}(\mathbb{R}_p)$
AS WELL AS THE CARDINALITY WEIGHTED-LOSS OF THE CHILDREN,

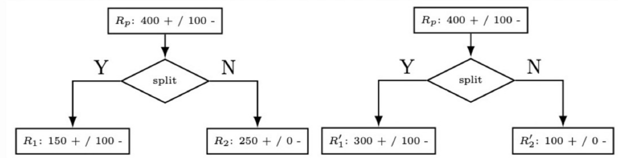
WE SELECT THE LEAF REGION, FEATURE, THRESHOLD THAT MINIMIZE THE LOSS

$$J(\mathbb{R}_p) - \frac{|\mathbb{R}_1| \, J(\mathbb{R}_1) + |\mathbb{R}_2| \, J(\mathbb{R}_2)}{\mathbb{R}_1 + \mathbb{R}_2}$$

FOR A CLASSIFICATION PROBLEM, WE FOCUS ON $\mathcal{L}_{\text{MISSCLASS}}$. FOR A REGION $\mathbb{R}$
LET $\hat{p}_c$ BE THE PROPORTION OF EXAMPLES IN $\mathbb{R}$ THAT ARE OF CLASS c.
MISSCLASSIFICATION LOSS ON $\mathbb{R}$ CAN BE EXPRESSED AS

$$J_{\text{MISSCLASS}}(\mathbb{R}) = 1 - \max_c (\hat{p}_c)$$

THE N OF EXAMPLES THAT WOULD BE MISSCLASSIFIED IF WE PREDICTED THE MAJORITY CLASS IN $\mathbb{R}$



- The first split is isolating out more of the positives, but we note that:

$$L\left(R_p\right)=\frac{\left|R_1\right| L\left(R_1\right)+\left|R_2\right| L\left(R_2\right)}{\left|R_1\right|+\left|R_2\right|}$$
$$=\frac{\left|R_1'\right| L\left(R_1'\right)+\left|R_2'\right| L\left(R_2'\right)}{\left|R_1'+\right| R_2'\mid}=100$$

- Thus, not only can we not only are the losses of the two splits identical, but neither of the splits decrease the loss over that of the parent.

WE NEED TO DEFINE A MORE SENSITIVE LOSS. LET'S FOCUS ON CROSS-ENTROPY

$$J_{CE}(R) = -\sum_c \hat{p}_c \log_2 \hat{p}_c$$

$\hat{p} \log_2 \hat{p} = 0$ IF $\hat{p} = 0$, N OF BITS NEEDED TO SPECIFY THE OUTCOME OR CLASS GIVEN A KNOWN DISTRIBUTION

$J_{\text{REDUCTION}}$ FROM $P \longrightarrow C$ = INFORMATION GAIN

ALGORITHM

TREE FITTING

$$\begin{cases} \text{FIND } j \in \{1, \dots, d\}, \theta \in \mathbb{R} \\ \underset{j,\theta}{\operatorname{argmax}}\, G(j,\theta) \quad (G: \text{"gain"}) \end{cases}$$

$S \to$ left $\mathbb{1}\{x_j^{(i)} \leq \theta\}$ righ $\mathbb{1}\{x_j^{(i)} > \theta\}$

RETURN S

PROS

- WELL-INTERPRETABLE
- ROBUST TO OUTLIERS
- HANDLES MIX OF DISCRETE AND CONTINUOUS FEATURES
- ROBUST TO MONOTONE TRANSFORMATION
- CAN FIT QUICKLY

CONS

- GENERALIZE POORLY
- HIGHLY UNSTABLE

ENSEMBLE LEARNING

$$F_1, \dots, F_m$$

$$f(x) = \sum_{i=1}^{n} \beta_i F_i(x) \qquad \beta_j \in \mathbb{R}, \quad \beta_j = \frac{1}{m}$$

THIS IS USEFUL TO RESOLVE UNSTABILITY PROBLEMS OF CERTAIN SUPERVISED ALGORITHMS

THE TRAIN IS DETERMINISTIC, IF I FEED THE SAME INPUTS AND AVG, WE GET THE SAME RESULT

BAGGING (BOOTSTRAP AGGREGATION)

$$\{S\} = n$$

$$\uparrow$$

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

EX $\{1, 3, 5, 7, 7, 8\}$

$$P(X=1) = \frac{1}{6}$$

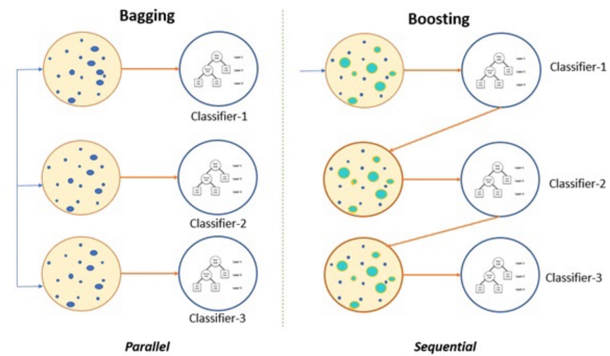
$$P(X=7) = \frac{1}{3}$$

A BOOTSTRAP IS A SUBSET S' WE TRAIN AND MAKE DIFFERENT PREDICTIONS. THEN LINEARLY COMBINE THEM

Y : # OF UNIQUE DATAPOINTS FROM S

$X_1, \dots, X_n$: WHETHER THE i-TH DATAPOINT IS IN A BOOTSTRAPPED SET

$$\mathbb{E}[Y] = \sum_{i=1}^{n} [X_i] = \sum_{i=1}^{n} P(X_i = 1)$$



$$P(X_i = 0) = \left(1 - \frac{1}{n}\right)^n$$

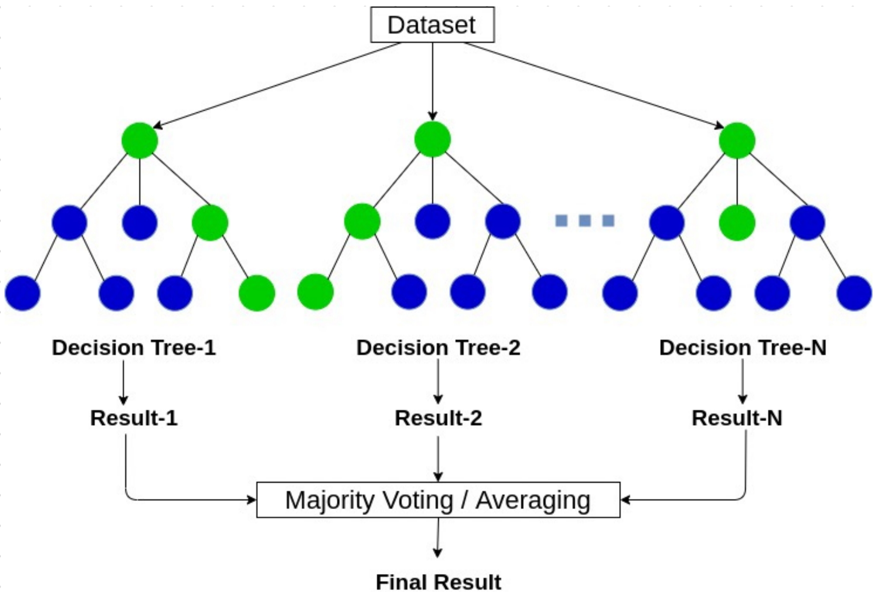
$$P(X_i = 1) = 1 - \underbrace{\left(1 - \frac{1}{n}\right)}_{\approx 0,632}^{n} = c^{-1}$$

$$\lim_{n \to +\infty} \left(1 + \frac{c}{n}\right)^n = e^c$$

RANDOM FOREST

- → BAGGING
- → SAMPLE A SUBSET OF FEATURES FOR SPLITS

WE HAVE MANY TREES



BOOSTING

$$f_{i-1} = \sum_{j=1}^{i-1} \beta_j F_j \rightarrow f_i = f_{i-1} + F_j = \sum_{j=1}^{i} \beta_j F_j$$

WE ARE BASICALLY COMBINING BAD CLASSIFIERS INTO ONE GOOD, BY AMALGAMATE THEM.

EX IN CLASSIFIERS

$$\delta, \varepsilon, \gamma \in (0,1)$$

TOL OF UNCERTAINTY — TOL OF ERROR — IMPROVEMENT IN ACCURACY

γ- WEAK LEARNERS VS STRONG LEARNERS

LOSS $P(Acc(f) > \frac{1}{2} + \gamma) \geq 1 - \delta$

$P(J(f) < \varepsilon) \geq 1 - \delta$ if $> m(\delta, \varepsilon)$

MATTER OF TIME UNTIL GETS TO PERF. CLASS.

MORE WEAK INTO STRONG

FORWARD STAGEWISE ADDITIVE MODELING (ALG)

AT ITER i, LOOK FOR F_i (PARAMETRIZED BY θ)

$$\beta_i, \theta_i = \underset{\beta,\theta}{\arg\min}\ \mathcal{L}(\underbrace{f_{i-1}}_{\text{CONSTANT}} + \beta F_i)$$

$$= \arg\min \sum_{i=1}^{n} \ell\left(y^{(i)}, f_{i-1}(x^{(i)}) + \beta F_i(x^{(i)}; \theta)\right)$$

$$F_i = F_i(\cdot\ ; \theta_i)$$

$$f_i(x) = f_{i-1}(x) + \beta_i F_i(x)$$

$$F_i = \underset{F}{\arg\min}\ \mathcal{L}(f_{i-1} + F) = \underset{F}{\arg\min} \sum_{i=1}^{n} \ell\left(y^{(i)}, f_{i-1}(x^{(i)}) + F(x^{(i)})\right)$$

WE ARE OPTIMIZING W.R.T FUNCTIONS, WE IMAGINE FUNCTIONS AS CONTINUOUS DIMENSIONAL VECTORS

LOSS FOR GB

$$\mathcal{L}(f) = \sum_{i=1}^{n} l\left(y^{(i)}, f(x^{(i)})\right) = \sum_{j=1}^{K} \sum_{x^{(i)} \in R_j} l(y^{(i)}, w_j)$$

XGB

N OF LAYERS

PREDICTION

$$\mathcal{L}(f) = \sum_{i=1}^{n} l\left(y^{(i)}, f(x^{(i)})\right) = \sum_{j=1}^{K} \sum_{x^{(i)} \in R_j} l(y^{(i)}, w_j) + \gamma J + \frac{1}{2} \lambda \sum_{j=1}^{J} w_j^2$$

WE "DISCRETIZE" AND OBTAIN

$$f = \begin{bmatrix} f_{i-1}(x^{(1)}) \\ \vdots \\ f_{i-1}(x^{(n)}) \end{bmatrix} \xrightarrow{\text{AND OBTAIN}} g = \begin{bmatrix} \dfrac{\partial \ell(y^{(1)}, f_{i-1}(x^{(1)}))}{\partial f_{i-1}(x^{(1)})} \\ \vdots \\ \dfrac{\partial \ell(y^{(n)}, f_{i-1}(x^{(n)}))}{\partial f_{i-1}(x^{(n)})} \end{bmatrix} \rightarrow$$

DERIVATIVE OF LOSS W.R.T. PREDICTION

WE ATTEMPT TO RE-CREATE GRADIENT DESCENT U.R. IN A CONTINUOUS SPACE

$$f = f - \alpha g$$

WE WANT TO FIT A MODEL FOR g

$$F_i \approx \underset{F}{\operatorname{argmin}} \sum_{j=1}^{n} \left(-g_j - F(x^{(j)})\right)^2$$

DECISION - TREES FIT THE MODEL VERY NICELY, SINCE THEY ARE THE QUINTESSENTIAL OF WEAK LEARNERS (WE NEED QUICK LEARNERS)

YOU LEARN THE GRADIENT AND ADD TREE BY TREE

VARIANTS:

→ EXTREME GRADIENT BOOST
→ NG BOOST