Binary Classification and Regression Tree

- heart attack patients
- covariates → minimum systolic blood pressure within 24 hours of hospital admission

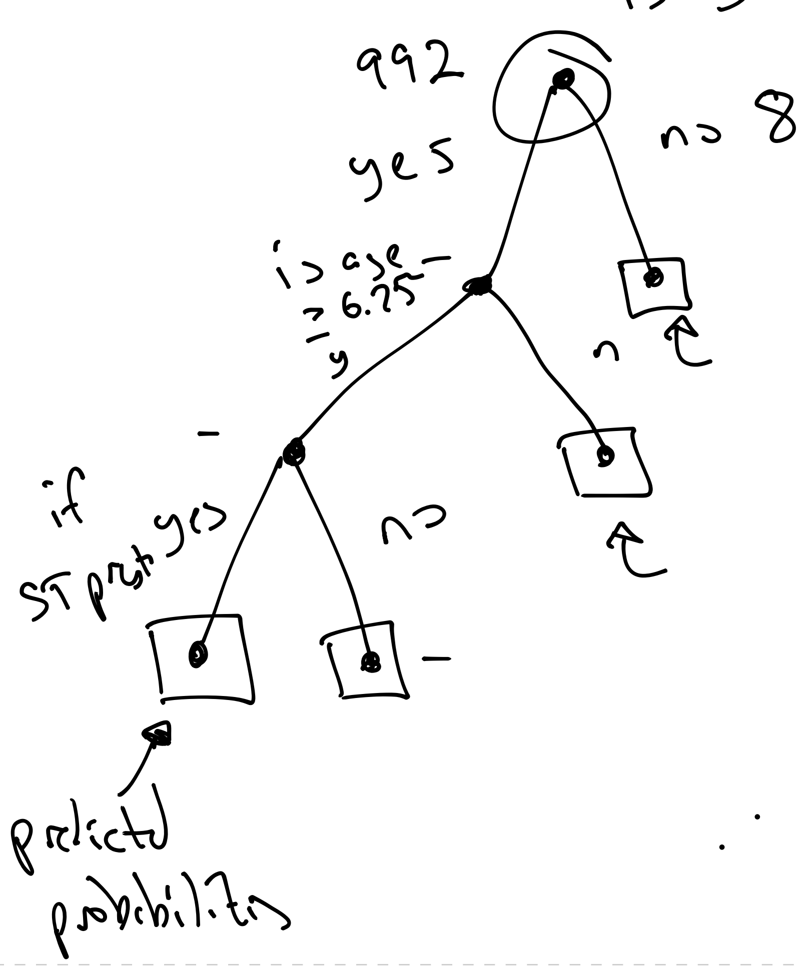
age

sinus tachychardia

↳ elevated heart rate

- whether they had death within 30 days of hospital admission

is blood pressure ≥ 91



1000

min_samples_split

10

- age (continuous variable)
- blood pressure (continuous)
- ST (binary)

...

given some loss function,
we choose by brute force,
the split in the data
that produce the minimal
loss at the next step,
and do this recursively
throughout the tree until
we hit a termination condition

max_depth
min_samples_split
min_samples_leaf

age	level of education
50	0 some HS
50	1 HS
	2 some college

3 college

S1 4 graduate degree

≤ 2

S2 ≤ 3

→ cross-validation and find optimal parameter for max_depth, min_samples_split and min_samples_leaf

→ what metric do trees optimize?

↳ regression
squared error loss

N_1 and N_2

ind in

$\bar{y}_{N_1}$ $\bar{y}_{N_2}$

$$\sum_{i \in N_1} (y_i - \bar{y}_{N_1})^2$$

$$+ \sum_{i \in N_2} (y_i - \bar{y}_{N_2})^2$$

age ≤ 60

mea age $\leq 60 = \bar{y}_{60-}$

mea age $> 60 = \bar{y}_{60+}$

$$\sum_{ppl \leq 60} (y_i - \bar{y}_{60-})^2 +$$

$$\sum_{ppl > 60} (y_i - \bar{y}_{60+})^2$$

age ≤ 70

$$\sum_{ppl \leq 70} (y_i - \bar{y}_{70-})^2$$

$$+ \sum_{ppl > 70} (y_i - \bar{y}_{70+})^2$$

Classification Tree

metric: Gini impurity

how often a randomly chosen item in a node will be incorrectly labeled if it is given a random label according to the distribution of labels in the class

two classes: n individuals m are in class 0

$n-m$ are in class 1

$p = \frac{m}{n}$ $\quad (1-p) = \frac{n-m}{n}$

$$P(\text{misclassified}) = P(\text{label class } 1 \mid \text{class } 0)\, P(\text{class } 0)$$
$$+ P(\text{label class } 0 \mid \text{class } 1)\, P(\text{class } 1)$$

$$= 2p(1-p)$$

---

① Boosting

↳ fit of weak models that learn to predict the errors in the previous term in the sequence

(2) Bagging

↳ bootstrap aggregation

↳ generate a bunch of bootstrap samples

fit a weak model on each of them

(3) Stacking

↳ train a bunch of models in parallel

↳ fit another machine learning model to their predictions

Bagging

- bagged trees

generate something like

n bootstrap samples of data set

fit decision tree on every bootstrap sample and average predictions

- feature bagging → Random Forests

generate n bootstrap samples

at every node in every tree,

only considers splits along

a random subset of
variables

n_estimators = 100
max_depth
min_samples_split
min_samples_leaf
max_features

Extremely randomized tree:

① generate BS samples
② at each node, only consider
random subset of factors

③ generate a split randomly for the
factors and only consider
those

Boosting

$F_1, \dots, F_m$

$$\underline{F_i(\vec{x})} = \underline{F_{i-1}(\vec{x})} + \underline{h_i(\vec{x})}$$

h_i has been fit
on the term

$y - F_{i-1}(x)$

$h_i(x)$ tries to
predict
(error)

residual

term of F_{i-1}

① $F_0(x) = \bar{Y}$

M is the # of estimator

② For $m = 1, \dots, M$

$$r_{im} = y_i - F_{m-1}(x_i)$$

fit an estimator $h_m : x_i \to r_{im}$

$$F_{m+1}(x) = F_0(x) + \sum_{j=1}^{n} v \cdot \underline{\underline{h_j(x)}}$$

decision tree

$v \in [0, 1]$

max.depth = 3.

choose best. learning-rate

height and eight