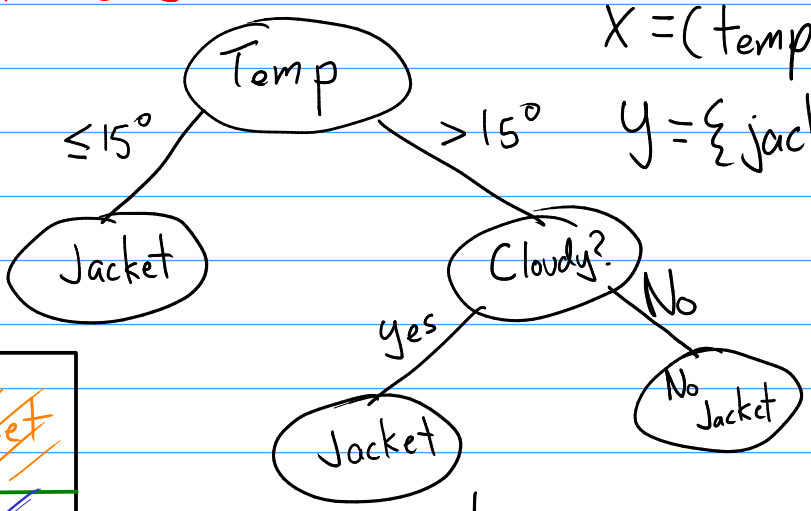
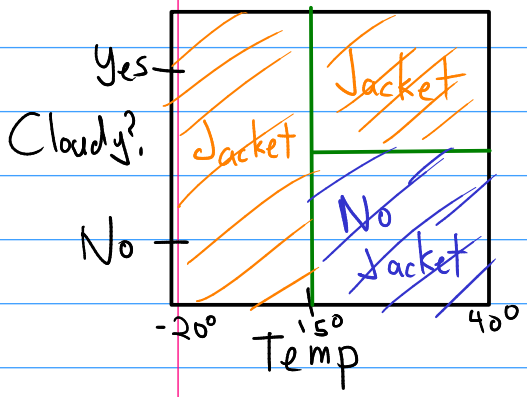
Decision Trees



$X = (\text{temp}, \text{cloudy})$

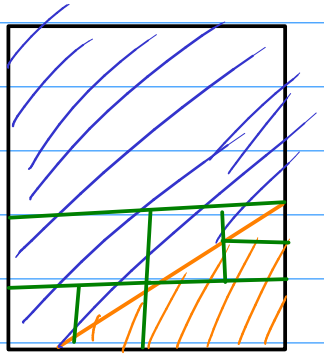
$y = \{\text{jacket}, \text{no jacket}\}$



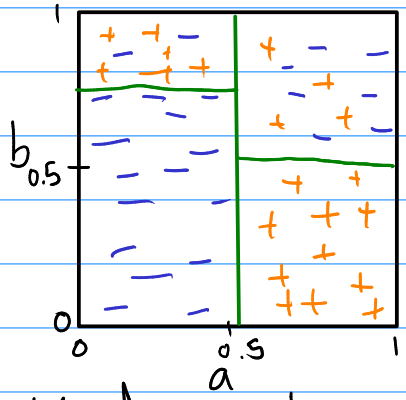
- Simple
- Interpretable

- Classification or Regression
- Tends to overfit

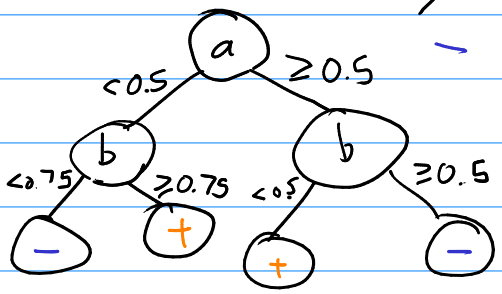
- Can handle nonlinear fns
- Maybe not linear ones



$X = (a, b)$

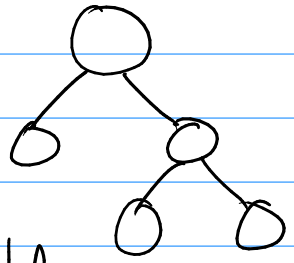


(0.3, 0.7)

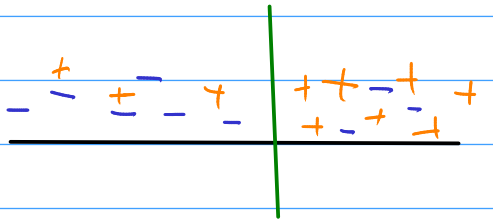


Prediction: Walk down tree, use leaf's prediction

Building a tree.
Start with one
Recursively split
Split(leaf):
Choose variable + threshold,
create two leaves w/ partition of points



Intuition: Pick split which makes leaves "pure"



Loss fn ℓ for a leaf
Small for "pure" nodes, large for mixed

$$t^* = \arg\min_t \underbrace{\ell(\{(x_i, y_i) : x_i \le t\})}_{\text{cost of left}} + \underbrace{\ell(\{(x_i, y_i) : x_i > t\})}_{\text{cost of right}}$$

$\le n$ values for t. (Fewer for categorical)

Example losses: $\ell(S)$, $S = \{(x_i, y_i)\}$

$\hat{p}_c$ = fraction of S with label $c = \frac{1}{|S|} \cdot \sum_{i \in S} \mathbb{1}\{y_i = c\}$

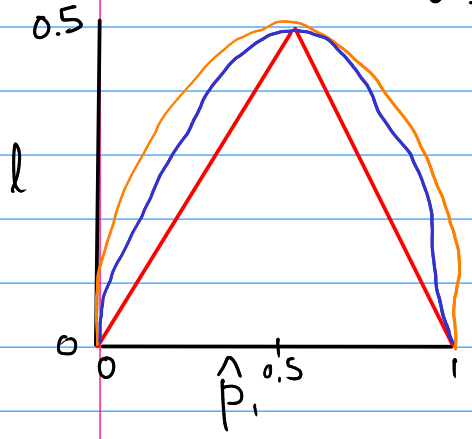
$\hat{y} = \arg\max_c \hat{p}_c$

$\hat{p}_0 = \frac{2}{3}$ 0 0 0
$\hat{p}_1 = \frac{1}{3}$ 1
$\hat{y} = 0$

- Misclassification Error: $\ell(S) = 1 - \hat{p}_{\hat{y}}$

If all 0's $\to \hat{p}_0 = 1 \to \ell(S) = 1 - 1 = 0$
50-50 $\to \hat{p}_0 = \hat{p}_1 = \frac{1}{2} \to \ell(S) = 1 - \frac{1}{2} = \frac{1}{2}$.

- Entropy: $\ell(S) = -\sum_{\text{Classes } c} \hat{p}_c \log \hat{p}_c$



- Gini index

$$\ell(S) = \sum_{\text{classes } c} \hat{p}_c (1 - \hat{p}_c)$$

(Regression: $\ell(S) = \min_p \sum_i (y_i - p)^2$

$= \sum_i (y_i - \bar{y})^2$, $\bar{y}$ = mean y_i's $\in S$)

Which variable to split on? Try all!

$$(j^*, t^*) = \underset{j,t}{\operatorname{argmin}} \; \ell(\{(x_i, y_i) : x_{ij} \le t\}) + \ell(\{(x_i, y_i) : x_{ij} > t\})$$

Pick dim + threshold to minimize loss after split.

age	Smokes	Cancer?
10	No	0
18	yes	0
25	No	0
35	yes	0
50	No	1
55	yes	1
70	yes	1
80	No	0
85	yes	1
90	yes	1

Gini index $\sum_c \hat{p}_c(1-\hat{p}_c)$

Split on smokes? $[\hat{p}_0, \hat{p}_1]$

No: $[\frac{3}{4}, \frac{1}{4}]$ $\quad \frac{3}{4}(\frac{1}{4}) + \frac{1}{4}(\frac{3}{4}) = \frac{3}{8}$

yes: $[\frac{2}{3}, \frac{1}{3}]$ $\quad \frac{2}{3}(\frac{1}{3}) + \frac{1}{3}(\frac{2}{3}) = \frac{4}{9}$

$\frac{3}{8} + \frac{4}{9} \approx 0.82$

Split on age?

≤ 35: $[1, 0]$ $\quad 0$

> 35: $[\frac{1}{6}, \frac{5}{6}]$ $\quad 2 \cdot \frac{1}{6}(\frac{5}{6}) = \frac{5}{18} < 0.33$

Stop?

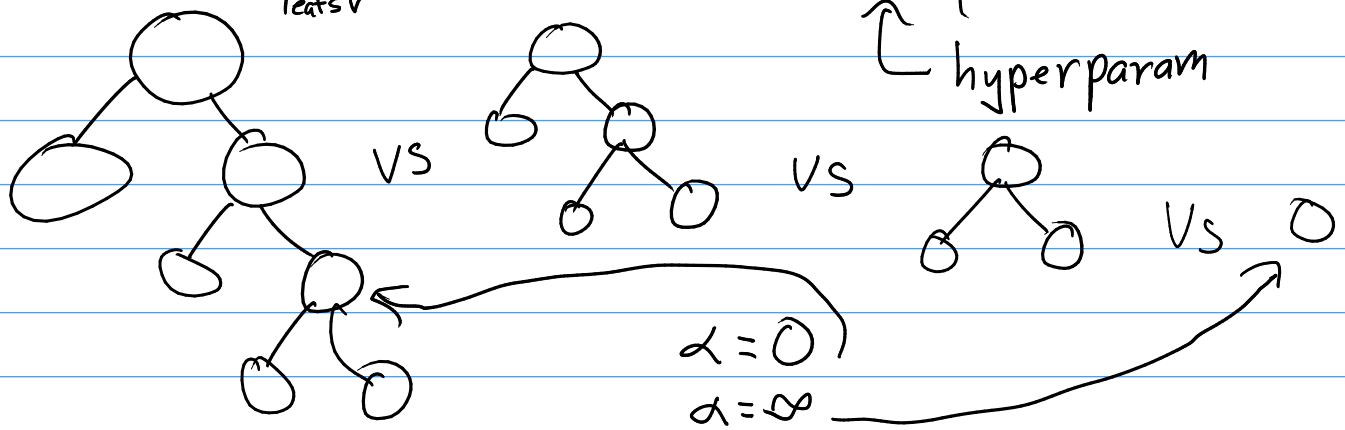
- Depth
- Running time
- Sparse leaves
- Leafs are homogeneous
- Small improvements only

$$\Delta = \ell(S) - (\ell(S_L) + \ell(S_R))$$

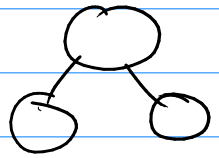
Pruning: Grow tree "fully", regularize over subtrees

$$\min \sum_{\text{leafs } v} \text{"error" in leaf } v \; + \alpha \cdot (\#\text{ of leaves})$$

↑ hyperparam



Decision stump



Fast
Not very good