DECISION TREES

INPUT : Dataset $\{ (x_1, y_1) \cdots , (x_n, y_n) \}$

OUTPUT : DECISION TREE

Decision tree

QUESTION?

Y / \ N

(1)   (0)

PREDICTION: Given X_{test}, traverse through the tree to reach a leaf node. Predict y_{test} = answer in leaf node
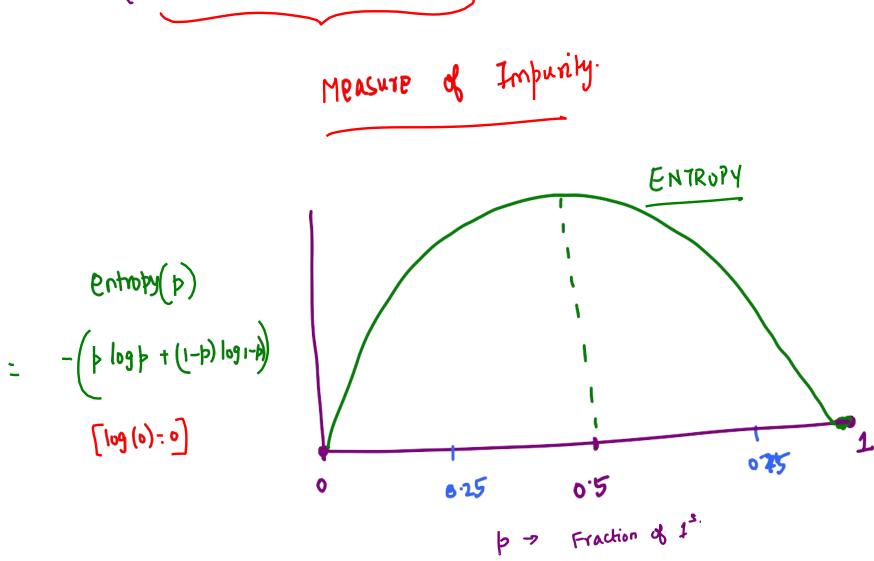
QUESTION:

A question is a (feature, value) pair.

Eg: height $\leq$ 180 cm?

• How to measure "goodness" of a Question?

$$D \qquad \{ (x_1, y_1) \quad \cdots \quad (x_n, y_n) \}$$

Y

N

D_{yes} $\{ (x_5, y_5), (x_8, y_8) \cdots \}$

$\{ (x_1, y_1), (x_2, y_2), \cdots \}$ D_{No}

$$\{ y_1, y_2, \ldots, y_n \}$$

$$y_i \in \{0, 1\}.$$

Measure of Impurity.

ENTROPY

entropy(p)

$$= -\left(p \log p + (1-p) \log 1-p \right)$$

$$[\log(0) = 0]$$

0   0.25   0.5   0.75   1

$p \to$ Fraction of $1's$.

$$\text{Information gain} \binom{\text{feature,}}{\text{value}} = \text{ENTRODY}\left(D\right) - \left(\frac{\gamma \, \text{ENTROPY}\left(D_{yes}\right) + (1-\gamma) \, \text{ENTROPY}\left(D_{no}\right)}{2}\right)$$

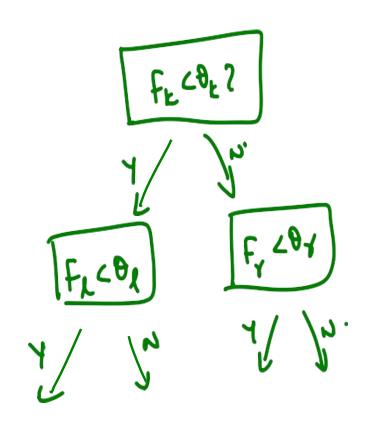$$\gamma = \frac{|D_{yes}|}{|D|} \qquad\qquad (1-\gamma) = \frac{|D_{no}|}{|D|}$$

ALGORITHM

→ Discretize each feature in $[min, max]$ range.

→ Pick Question $(f_k \leq \theta)$ that has highest information gain

→ Repeat for D_{yes} & D_{no}.

$f_k < \theta_k$?

Y   N

$F_\lambda < \theta_\lambda$ $F_Y < \theta_Y$

Y   N   Y   N

Points

- Depth is a hyper-parameter.

- Can also stop growing if node is "sufficiently" smooth.

Decision Boundary

$f_1 \leq 5$

$weight$
f_2

7

3 5 height

X

$f_1 \leq 5$

Y   N

$F_2 \leq 7$ $f_2 \leq 7$

Y   N   Y   N

$f_1 \leq 3$ ①   ⓪

Y   N

⓪   ①

f_1