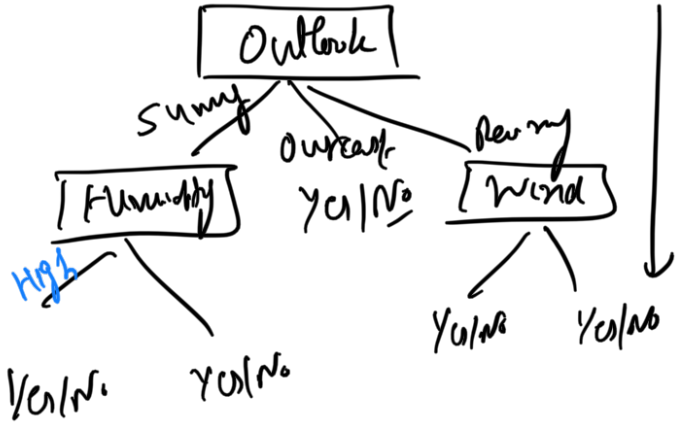
CoL 774
Machine Learning
NOV 27, 2020

Logistics
① Make extra class - Sunday Nov 29 8 am - 9 am

Last Class:- Decision Trees:-



Boolean := 2^{2^n}

↳ Axis || Rectangles

Learning Decision Trees from Data

$\{x^{(i)}, y^{(i)}\}_{i=1}^{m}$] Inductive Bias!

Grow trees which are "small" in size

Decision Tree Learning Algorithm:-

$D \leftarrow \{x^{(i)}, y^{(i)}\}_{i=1}^{m}$

GrowTree(D) {

//Leaf creation

if $\forall i, (x^{(i)}, y^{(i)}) \in D, y^{(i)} = 0$

Return createLeaf(0);

if $\forall i (x^{(i)}, y^{(i)}) \in D, y^{(i)} = 1$

Return createLeaf(1);

//split

$x_j \leftarrow$ ChooseBestAttrToSplit(D);

n_j = NewNode(x_j, NULL);

$\forall \ell \in \{1 \dots K\}$ ← # of attribute values

$x_j \in \{v_1 \dots v_K\}$

$D_\ell = \{x^{(i)}, y^{(i)}\} \in D$ s.t. $x_j^{(i)} = v_\ell$

n_j.add(GrowTree(D_ℓ), v_ℓ);

}

Binary classification problem

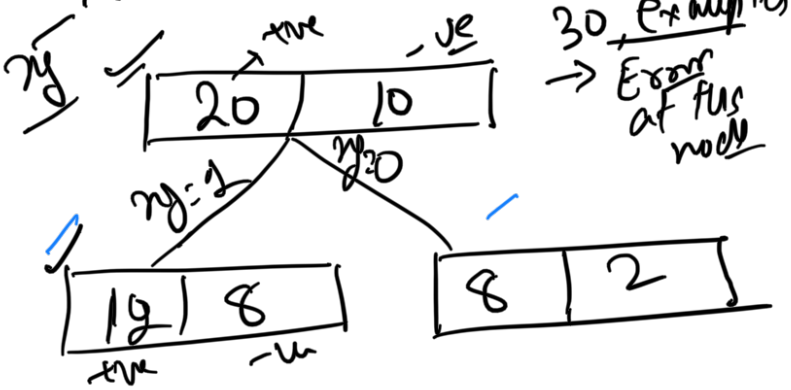
Data points such that value of x_j attribute is v_ℓ

3

return n_j;

3

How do we choose the Attribute to split on?



30 Examples → Error of this node

$i = 1 \ldots 30$

Given this kind of characteristics how good attribute x_j is to split the data?

"Metrics" → Score

"Pure"

$E_{x_j} = 10$

$E[x_j = v_1] = 8$ $\quad E[x_j = v_2] = 2$

↳ $E_{x_j\text{-split}} = 8 + 2 = 10$

No Reduction on Error "progress"

↳ Entropy :- "Measure Randomness in the System"

↳ Consider a Random Variable

$Y \in \{1 \ldots r\}$

$P(Y = k) = p_k$ $\quad \sum_{k=1}^{r} p_k = 1$

Entropy $H(Y) = -\sum_{k=1}^{r} p_k \log p_k$ $\quad -E[\log P(Y=k)]$

Assume a simple Boolean case:-

$p_0 + p_1 = 1$

p_0 p_1 $\quad Y \in \{0, 1\}$

$$H(Y) = -[p_0 \log p_0 + p_1 \log p_1]$$

$$= -[p \log p + (1-p)\log(1-p)]$$

$p \to 0 \Rightarrow -\infty$

$p \log p$:- as $p \to 0$ $\quad p \log p \to 0$

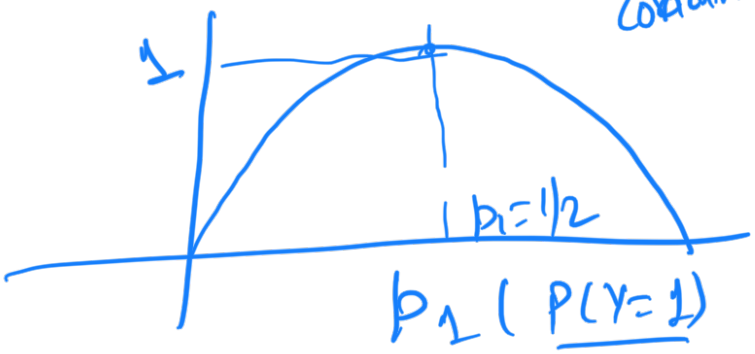
↳ Aside: $\log p = \frac{1/p}{1/p}$ $\quad \frac{d}{dp}[\log p] = \frac{1/p}{-1/p^2}$

$$= -p \to 0$$

$p \log p$:- as $p \to 1$ $\quad p \log p \to 0$

⇒ The (max) value of $-[p \log p]$ is obtained at 0.5

concave function



$H(Y)$

$Y \in \{0, 1\}$

] Deep connection with Information Theory:-

Entropy: "bits" required to transmit a message

$Y \in \{1, \ldots, K\}$

↳ $H(Y) = \sum_{k=1}^{K} -p_k \log p_k$ is maximum

when $p_k = 1/K \quad \forall k$

↳ Amount of randomness in the system

Our goal:- To reduce the entropy of class label.

$H(Y) = -\frac{2}{3}\log\frac{2}{3} - \frac{1}{3}\log\frac{1}{3}$

How do we characterize a split?

$$H(Y|X_j) = \sum_x P(X_j=x) H(Y|X_j=x)$$

well define entropy of the distribution $P(Y|X_j=x)$

Which attribute to split on?

$\forall j$: split on the attribute which minimizes $H(Y|X_j)$

$$j^* = \arg\min_j H(Y|X_j)$$

↓ Attribute to split on

$$MI(Y,X) = H(Y) - H(Y|X)$$

Symmetric quantity

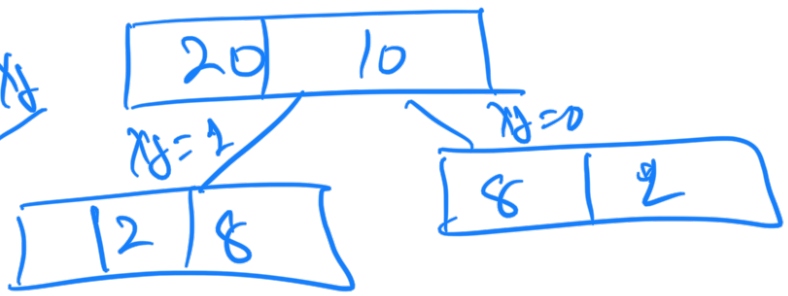
$$= MI(X,Y)$$

$$= H(X) - H(X|Y)$$

$$\hookrightarrow j^* = \arg\max_j MI(Y, X_j)$$

$$= \arg\max_j H(Y) - H(Y|X_j)$$

$\forall j$



$H(Y) = -\frac{2}{3}\log\frac{2}{3} - \frac{1}{3}\log\frac{1}{3}$

$=$ ①

$H(Y|X_3=1) = 3/5 \log 5/3 + 2/5 \log 5/2$ — ②

$H(Y|X_3=0) = 4/5 \log 5/4 + 1/5 \log 5/1$ — ③

$MI(Y, X_3) = H(Y) - H(Y|X_3)$

$= H(Y) - \left[H(Y|X_3=1)\, 2/3 + H(Y|X_3=0)\, 1/3 \right]$

compute these values (HW)

$= ??$ ≥ 0

"Progress"

⇒ A better metric than using plain error classification / Gain

Entropy - Most popular metric for computing the attribute to split on in decision trees.