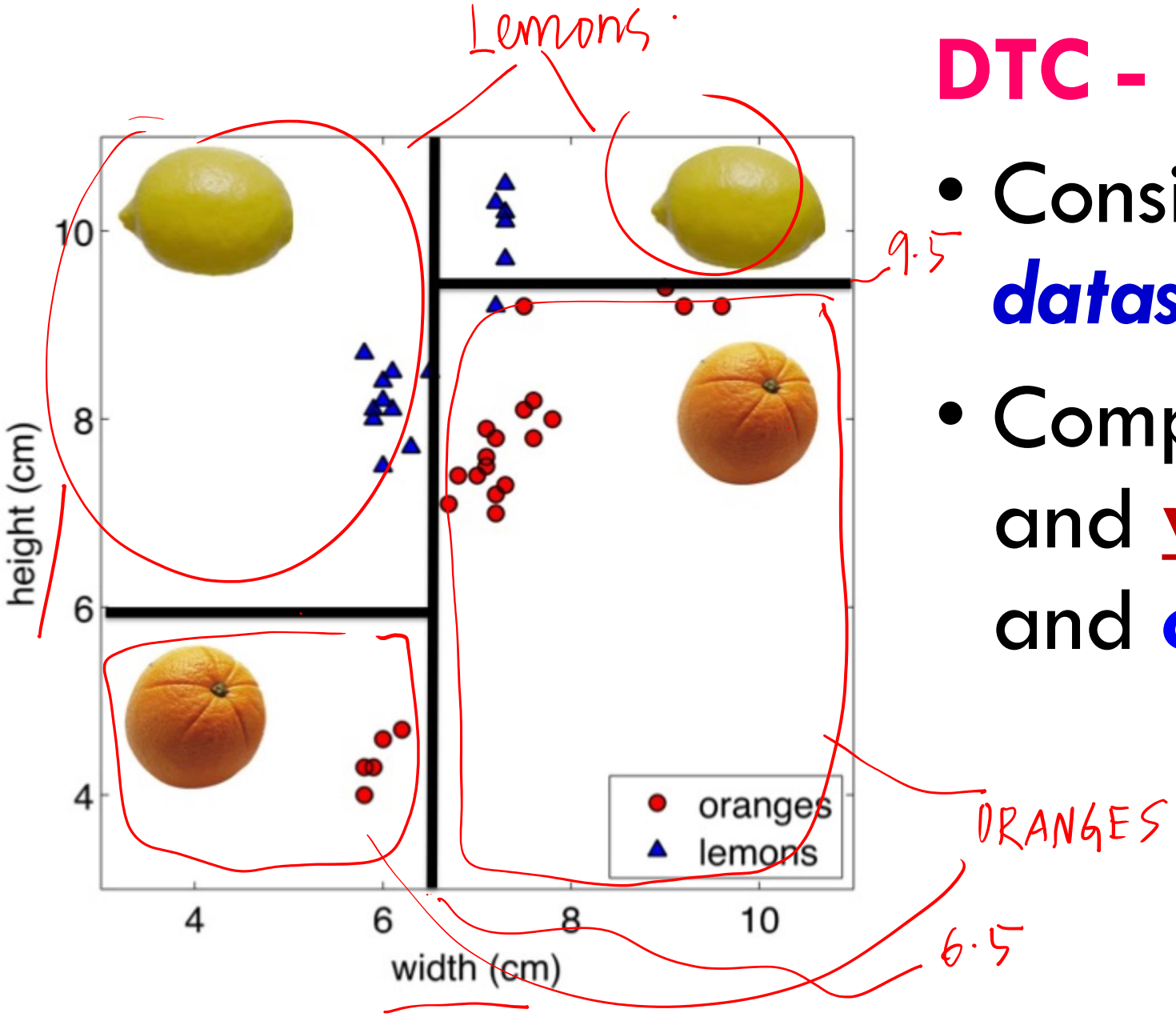
Elective Module:

Advanced ML Techniques

Chapter 10

Decision Tree Classifiers

DTC

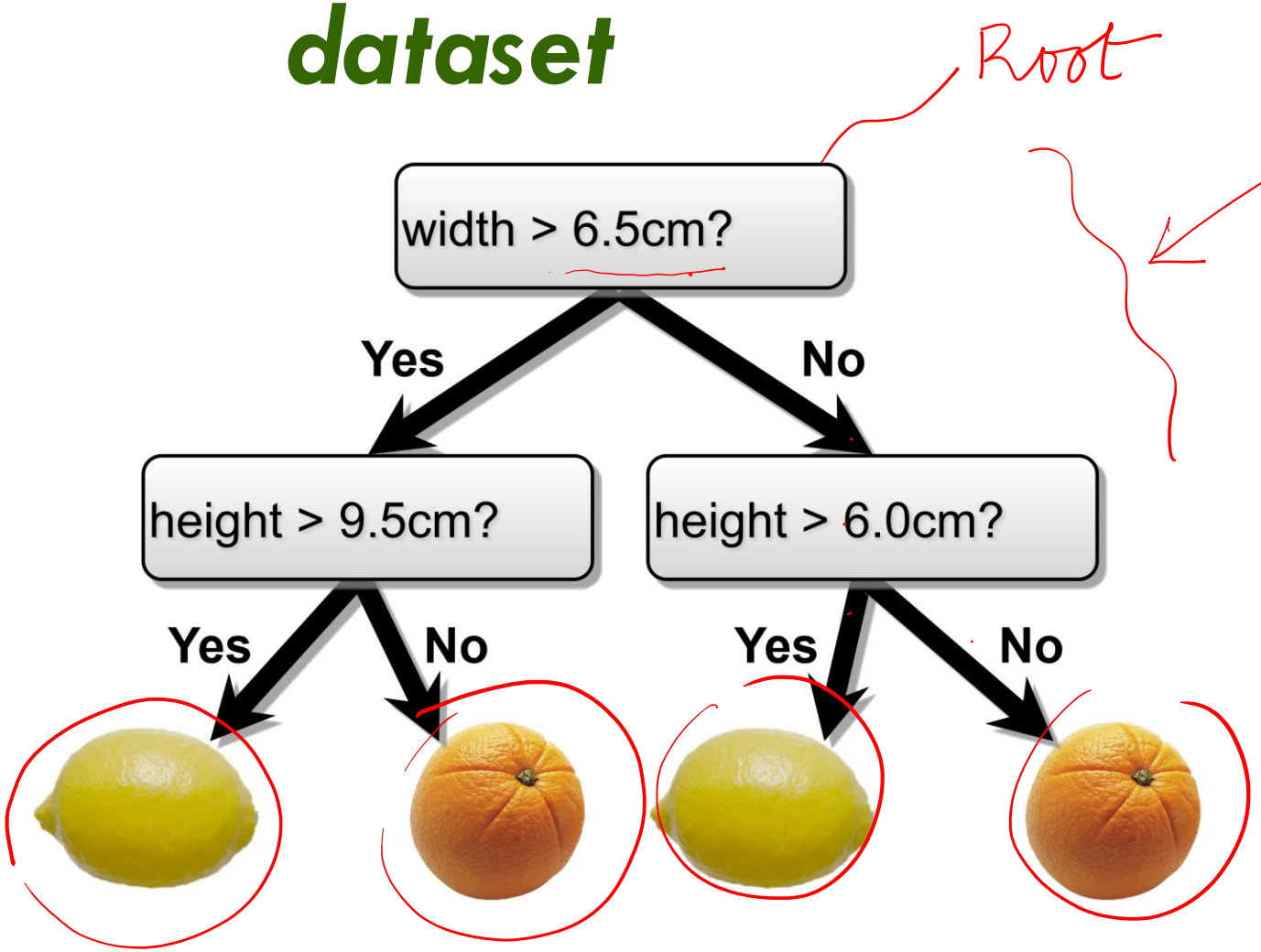


DTC - Example

- Consider the simple ***dataset*** shown
- Comprises of **heights** and **weights** of ***lemons*** and ***oranges***

DTC - Example

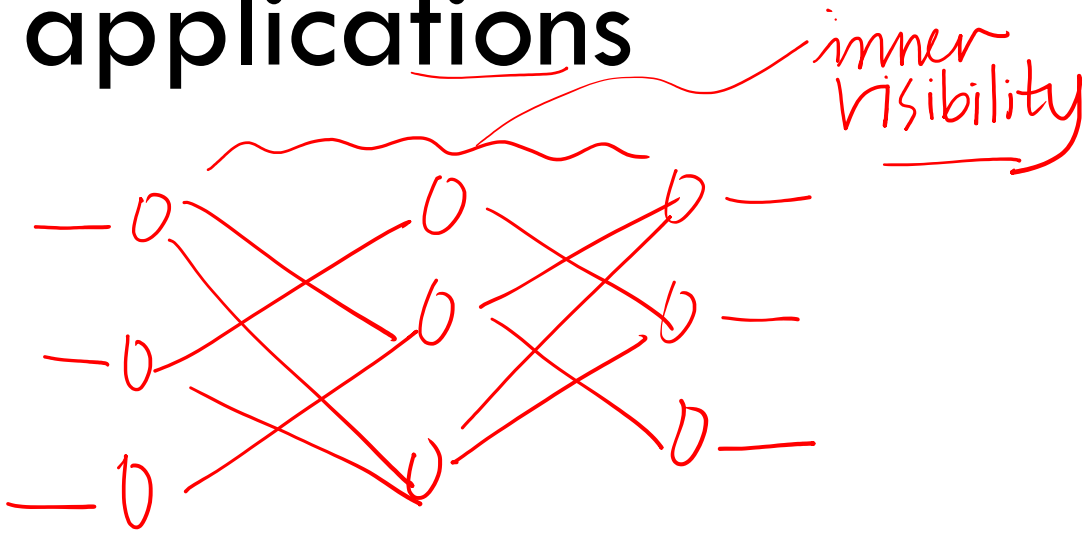
- DTC below is built using the given ***dataset***



Decision Tree Classifiers

- **Advantages**: ***interpretable***, ***intuitive*** (in contrast to Neural Nets)
- Popular in **medical diagnosis** applications

DTCs very popular in medicine

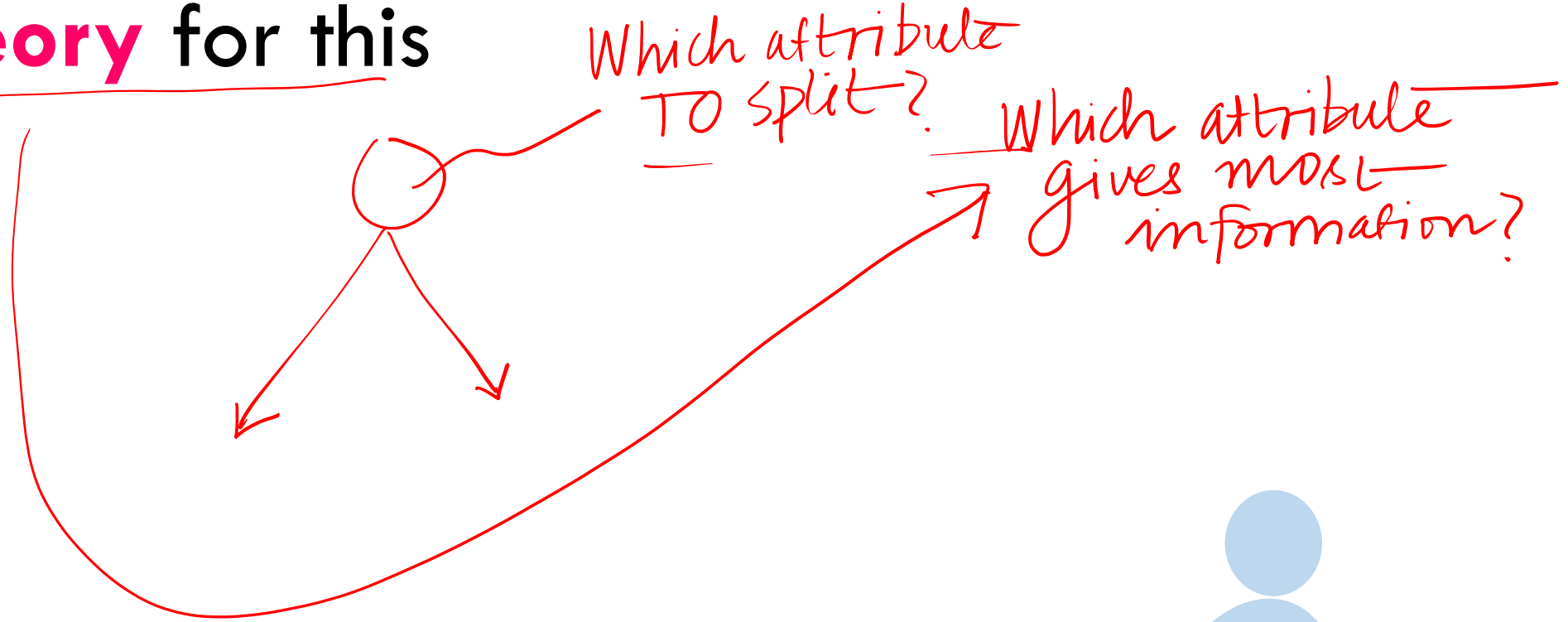


Decision Tree Classifiers

- DTCs are well-suited to model ***discrete outcomes***

Learning Decision Trees — Build DTC.

- How to choose the ***best attribute***?
- One can use principles of **information theory** for this



Entropy

intrinsic information

- Consider a **source** X with symbols x_i and probabilities $p(x_i)$

X → x_1 $P(x_1)$
x_2 $P(x_2)$
x_3 $P(x_3)$.
⋮

Entropy

$\sum_i p(x_i) = 1$

Entropy = 0 if $p(x_i) = 1$, $p(x_j) = 0, j \neq i$

- The **entropy** $H(X)$ of this source is defined as

$$H(X) = p(x_1)\log_2\left(\frac{1}{p(x_1)}\right) + p(x_2)\cdot\log_2\left(\frac{1}{p(x_2)}\right) + \cdots$$

$$= \sum_i p(x_i)\log_2 \cdot \frac{1}{p(x_i)}$$

$$= -\sum_i p(x_i)\log_2(p(x_i))$$

Entropy or information content.

Entropy ≥ 0.

Entropy maximum: $p(x_i) = \frac{1}{n}$: Equiprobable symbols.

Entropy

- The **entropy** $H(X)$ of this source is defined as

$$H(X) = \sum_{i=1}^{n} p(x_i) \log_2 \frac{1}{p(x_i)}$$

$$= -\sum_{i=1}^{n} p(x_i) \log_2 p(x_i)$$

Shannon: Father of information Theory

Entropy-Example

- Example: Consider the binary event ***like*** or ***dislike*** ice cream (IC), ($\overline{\text{IC}}$)

$$X = \{\text{IC}, \overline{\text{IC}}\}$$

$$P(\text{IC}) = \frac{3}{4}, P(\overline{\text{IC}}) = \frac{1}{4}$$

Like ice cream

DO NOT like ice cream

$1 - \frac{3}{4} = \frac{1}{4}$.

Information or entropy of Random variable.

Entropy-Example

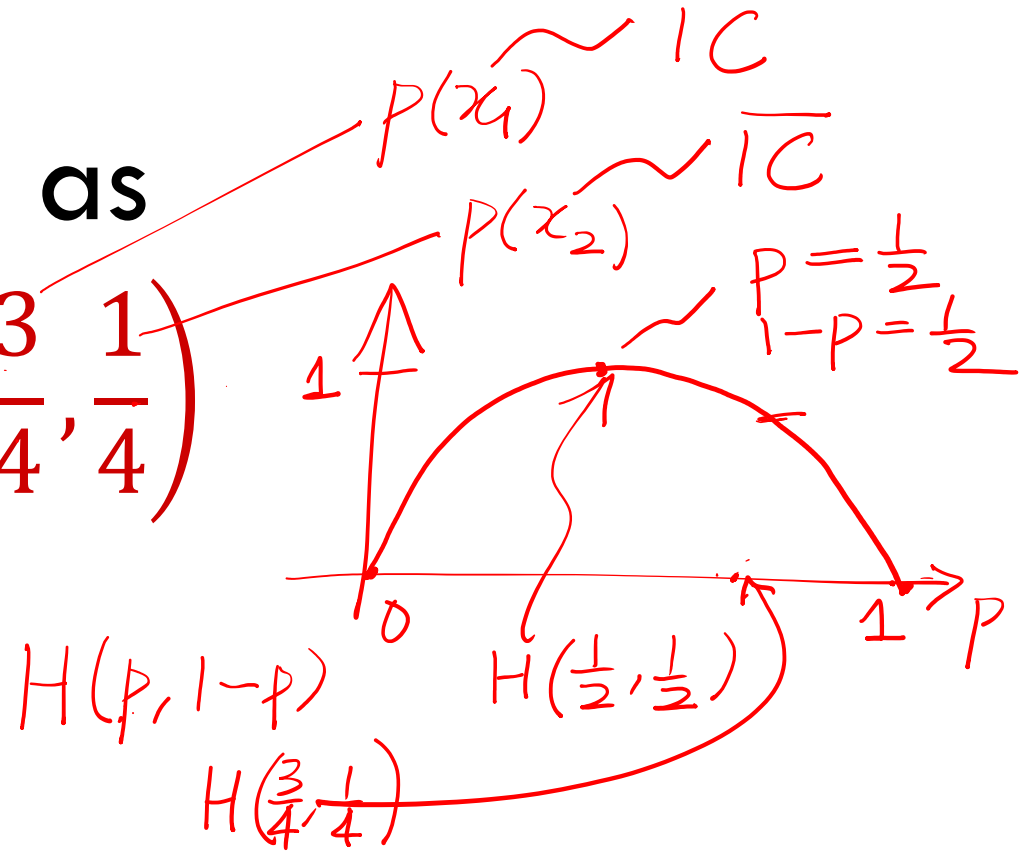
- The ***entropy*** is given as

$$H(X) = H\left(\frac{3}{4}, \frac{1}{4}\right)$$

$$= \frac{3}{4} \cdot \log_2 \frac{4}{3} + \frac{1}{4} \cdot \log_2 4$$

$$= 0.811 = H(X)$$

ice cream RV



Entropy-Example

- The ***entropy*** is given as

$$H(X) = H\left(\frac{3}{4}, \frac{1}{4}\right)$$

$$= \frac{3}{4} \times \log_2 \frac{4}{3} + \frac{1}{4} \times \log_2 4 \approx 0.811$$

Conditional entropy

- Consider **two sources**: X with symbols x_i and Y with symbols y_j

$H(X|Y)$

Conditional Entropy $X|Y$.

Conditional Entropy.

Conditional entropy

- The ***conditional entropy*** $H(X|Y)$ is defined as

$$H(X|Y) = \sum_j p(y_j) H(X|Y=y_j)$$

Weighing with probability.

Calculate entropy of X for each value of $Y = y_j$

Weighted average

conditional Entropy

Conditional entropy

- The ***conditional entropy*** $H(X|Y)$ is defined as

$Y \in \{y_1, y_2, .., y_m\}$

$$H(X|Y) = \sum_{j=1}^{m} p(y_j) H(X|Y = y_j)$$

Conditional entropy — Example.

- Consider the table below showing ***joint probabilities*** of

$$X = \{\mathrm{IC}, \overline{\mathrm{IC}}\}, Y = \{\mathrm{CHOC}, \overline{\mathrm{CHOC}}\}$$

Ice cream
Choc

like ice cream and chocolate
like chocolate but NOT ice cream

	IC	$\overline{\mathbf{IC}}$
CHOC	$\frac{1}{2}$	$\frac{1}{8}$
$\overline{\mathbf{CHOC}}$	$\frac{1}{4}$	$\frac{1}{8}$

Conditional entropy

- What is $H(X|Y)$?

$$H(X|Y) = P(CHOC) \times H(X|Y=CHOC) + P(C\overline{HO}C) \times H(X|Y=C\overline{HO}C)$$

$$P(CHOC) = \frac{1}{2} + \frac{1}{8} = \frac{5}{8}$$

$$P(C\overline{HO}C) = \frac{1}{4} + \frac{1}{8} = \frac{3}{8}$$

	IC	$\overline{\textbf{IC}}$
CHOC	$\frac{1}{2}$	$\frac{1}{8}$
$\overline{\textbf{CHOC}}$	$\frac{1}{4}$	$\frac{1}{8}$

Conditional entropy

- What is $H(X|Y)$?

$$H(X|Y) = P(\text{CHOC}) \times H(X|\text{CHOC}) + P(\overline{\text{CHOC}})\, H(X|\overline{\text{CHOC}})$$

	IC	$\overline{\textbf{IC}}$
CHOC	$\frac{1}{2}$	$\frac{1}{8}$
$\overline{\textbf{CHOC}}$	$\frac{1}{4}$	$\frac{1}{8}$

Conditional entropy

$P(A|B) = \frac{P(A \cap B)}{P(B)}$

- $H(X|Y = \text{CHOC}) = ?$

$$P(\text{IC}|\text{CHOC}) = \frac{P(\text{IC} \cap \text{CHOC})}{P(\text{CHOC})} = \frac{\frac{1}{2}}{\frac{1}{2}+\frac{1}{8}} = \frac{\frac{1}{2}}{\frac{5}{8}} = \frac{4}{5}$$

$$P(\overline{\text{IC}}|\text{CHOC}) = \frac{\frac{1}{8}}{\frac{1}{2}+\frac{1}{8}} = \frac{\frac{1}{8}}{\frac{5}{8}} = \frac{1}{5}$$

	IC	$\overline{\textbf{IC}}$
CHOC	$\frac{1}{2}$	$\frac{1}{8}$
$\overline{\textbf{CHOC}}$	$\frac{1}{4}$	$\frac{1}{8}$

Conditional entropy

- $H(X|Y = \text{CHOC}) = ?$

$$P(\text{IC}|\text{CHOC}) = \frac{4}{5}, P(\overline{\text{IC}}|\text{CHOC}) = \frac{1}{5}$$

	IC	$\overline{\textbf{IC}}$
CHOC	$\frac{1}{2}$	$\frac{1}{8}$
$\overline{\textbf{CHOC}}$	$\frac{1}{4}$	$\frac{1}{8}$

Conditional entropy

$$H(X|\text{CHOC}) = H\left(\frac{4}{5}, \frac{1}{5}\right)$$

$$= \frac{4}{5}\log_2 \frac{5}{4} + \frac{1}{5}\log_2 5 = 0.722$$

Entropy in X given person likes chocolate

	IC	$\overline{\text{IC}}$
CHOC	$\frac{1}{2}$	$\frac{1}{8}$
$\overline{\text{CHOC}}$	$\frac{1}{4}$	$\frac{1}{8}$

Conditional entropy

$$\mathrm{H}(X|\mathrm{CHOC}) = H\left(\frac{4}{5}, \frac{1}{5}\right)$$

$$= \frac{4}{5} \times \log_2\left(\frac{5}{4}\right) + \frac{1}{5} \times \log_2(5) \;=\; 0.722$$

	IC	$\overline{\textbf{IC}}$
CHOC	$\frac{1}{2}$	$\frac{1}{8}$
$\overline{\textbf{CHOC}}$	$\frac{1}{4}$	$\frac{1}{8}$

Conditional entropy

- $H(X|Y = \overline{\text{CHOC}}) = ?$

$$P(\text{IC}|\overline{\text{CHOC}}) = \frac{P(\text{IC} \cap \overline{\text{CHOC}})}{P(\overline{\text{CHOC}})} = \frac{\frac{1}{4}}{\frac{1}{4}+\frac{1}{8}} = \frac{\frac{1}{4}}{\frac{3}{8}} = \frac{2}{3}$$

$$P(\overline{\text{IC}}|\overline{\text{CHOC}}) = \frac{\frac{1}{8}}{\frac{1}{4}+\frac{1}{8}} = \frac{\frac{1}{8}}{\frac{3}{8}} = \frac{1}{3}.$$

	IC	$\overline{\textbf{IC}}$
CHOC	$\frac{1}{2}$	$\frac{1}{8}$
$\overline{\textbf{CHOC}}$	$\frac{1}{4}$	$\frac{1}{8}$

Conditional entropy

- $H(X|Y = \overline{\text{CHOC}}) = ?$

$$P(\text{IC}|\overline{\text{CHOC}}) = \frac{2}{3}, P(\overline{\text{IC}}|\overline{\text{CHOC}}) = \frac{1}{3}$$

	IC	$\overline{\textbf{IC}}$
CHOC	$\frac{1}{2}$	$\frac{1}{8}$
$\overline{\textbf{CHOC}}$	$\frac{1}{4}$	$\frac{1}{8}$

Conditional entropy

$$H(X|\overline{CHOC}) = H\left(\frac{2}{3}, \frac{1}{3}\right)$$

$$= \frac{2}{3} \cdot \log_2 \frac{3}{2} + \frac{1}{3} \cdot \log_2 3 = 0.918$$

	IC	$\overline{IC}$
CHOC	$\frac{1}{2}$	$\frac{1}{8}$
$\overline{CHOC}$	$\frac{1}{4}$	$\frac{1}{8}$

Conditional entropy

$$\mathrm{H}(X|\overline{\mathrm{CHOC}}) = H\left(\frac{2}{3}, \frac{1}{3}\right)$$

$$= \frac{2}{3} \times \log_2\left(\frac{3}{2}\right) + \frac{1}{3} \times \log_2(3)$$

$$= 0.918$$

	IC	$\overline{\text{IC}}$
CHOC	$\frac{1}{2}$	$\frac{1}{8}$
$\overline{\text{CHOC}}$	$\frac{1}{4}$	$\frac{1}{8}$

Conditional entropy

- Finally, the **conditional entropy** is given as,

$$H(X|Y)$$

$$= P(\text{CHOC}) \times H(X|\text{CHOC}) + p(\overline{\text{CHOC}}) \times H(X|\overline{\text{CHOC}})$$

$$= \frac{5}{8} \times 0.722 + \frac{3}{8} \times 0.918$$

$$= 0.7955$$

Conditional entropy

- Finally, the **conditional entropy** is given as,

$$H(X|Y)$$

$$= P(\mathrm{CHOC}) \times H(X|\mathrm{CHOC}) + p(\overline{\mathrm{CHOC}}) \times H(X|\overline{\mathrm{CHOC}})$$

$$= \frac{5}{8} \times 0.722 + \frac{3}{8} \times 0.918 = 0.7955$$

Information gain

- The **information gain (IG)** is defined as

$$\mathrm{IG}(X|Y) = H(X) - H(X|Y) \geq 0$$

$$= 0.811 - 0.7955 = 0.0155$$

information gained by observing Y..

Information gain

- The **information gain (IG)** is defined as

$$\mathrm{IG}(X|Y) = H(X) - H(X|Y)$$
$$= 0.811 - 0.7955 = 0.0155$$

Mutual information

- This is also known as the ***Mutual Information (MI)***

Mutual information

Larger mutual info ⇒ more information conveyed.

DTC Feature Selection

- Choose the ***feature***…that maximizes the **information gain**!

Conditional Entropy

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

Attribute

Choose attribute that maximizes information gain.

DTC Example

- Consider the **table** shown below
- **Customer decisions** to wait or not at restaurants

1.	Alternate: whether there is a suitable alternative restaurant nearby.
2.	Bar: whether the restaurant has a comfortable bar area to wait in.
3.	Fri/Sat: true on Fridays and Saturdays.
4.	Hungry: whether we are hungry.
5.	Patrons: how many people are in the restaurant (values are None, Some, and Full).
6.	Price: the restaurant's price range ($, $$, $$$).
7.	Raining: whether it is raining outside.
8.	Reservation: whether we made a reservation.
9.	Type: the kind of restaurant (French, Italian, Thai or Burger).
10.	WaitEstimate: the wait estimated by the host (0-10 minutes, 10-30, 30-60, >60).

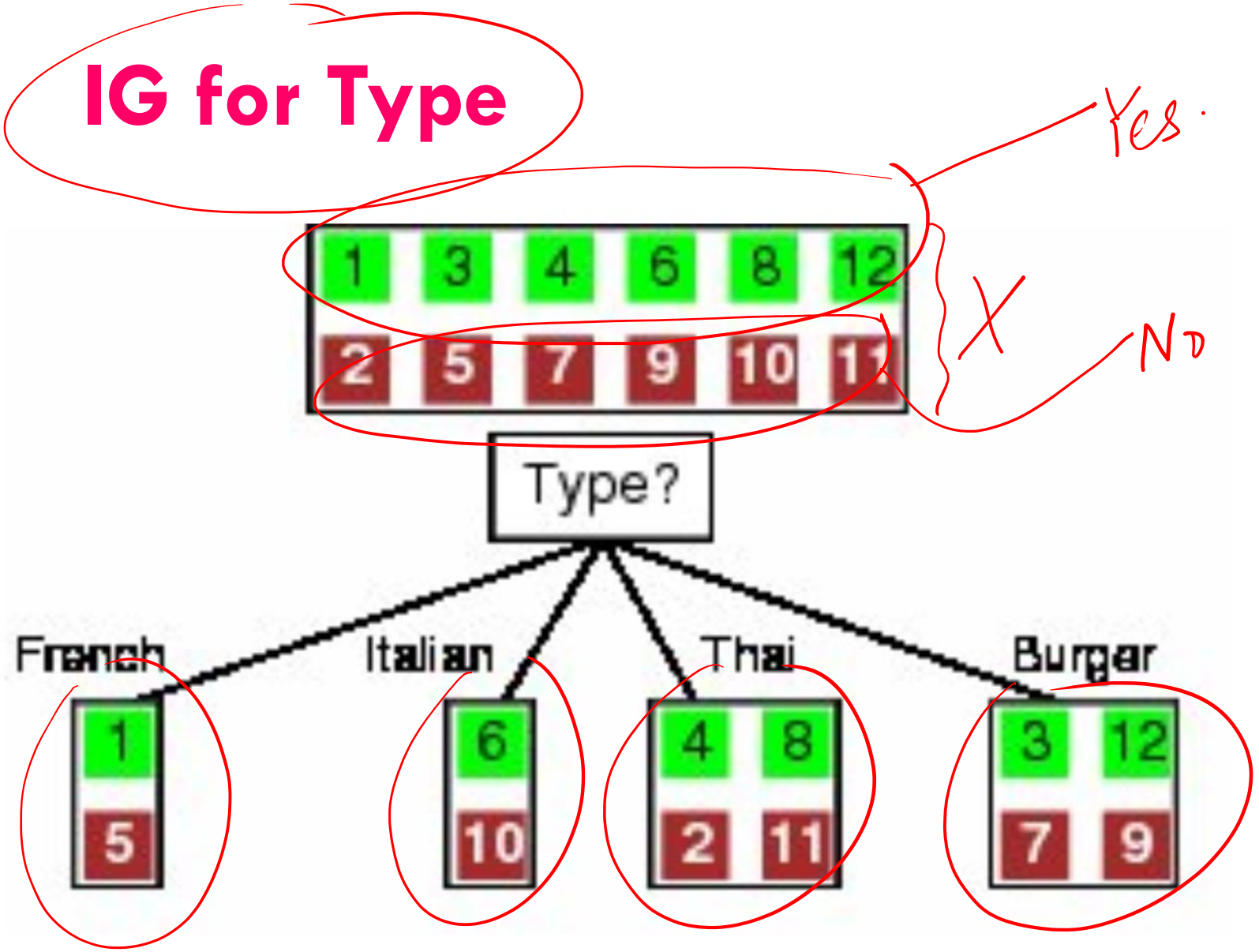
DTC Example

- **Table columns**

Example	Input Attributes										Goal
	Alt	*Bar*	*Fri*	*Hun*	*Pat*	*Price*	*Rain*	*Res*	*Type*	*Est*	*WillWait*
$\mathbf{x}_1$	*Yes*	*No*	*No*	*Yes*	*Some*	*$$$*	*No*	*Yes*	*French*	*0–10*	$y_1 =$ *Yes*
$\mathbf{x}_2$	*Yes*	*No*	*No*	*Yes*	*Full*	*$*	*No*	*No*	*Thai*	*30–60*	$y_2 =$ *No*
$\mathbf{x}_3$	*No*	*Yes*	*No*	*No*	*Some*	*$*	*No*	*No*	*Burger*	*0–10*	$y_3 =$ *Yes*
$\mathbf{x}_4$	*Yes*	*No*	*Yes*	*Yes*	*Full*	*$*	*Yes*	*No*	*Thai*	*10–30*	$y_4 =$ *Yes*
$\mathbf{x}_5$	*Yes*	*No*	*Yes*	*No*	*Full*	*$$$*	*No*	*Yes*	*French*	*>60*	$y_5 =$ *No*
$\mathbf{x}_6$	*No*	*Yes*	*No*	*Yes*	*Some*	*$$*	*Yes*	*Yes*	*Italian*	*0–10*	$y_6 =$ *Yes*
$\mathbf{x}_7$	*No*	*Yes*	*No*	*No*	*None*	*$*	*Yes*	*No*	*Burger*	*0–10*	$y_7 =$ *No*
$\mathbf{x}_8$	*No*	*No*	*No*	*Yes*	*Some*	*$$*	*Yes*	*Yes*	*Thai*	*0–10*	$y_8 =$ *Yes*
$\mathbf{x}_9$	*No*	*Yes*	*Yes*	*No*	*Full*	*$*	*Yes*	*No*	*Burger*	*>60*	$y_9 =$ *No*
$\mathbf{x}_{10}$	*Yes*	*Yes*	*Yes*	*Yes*	*Full*	*$$$*	*No*	*Yes*	*Italian*	*10–30*	$y_{10} =$ *No*
$\mathbf{x}_{11}$	*No*	*No*	*No*	*No*	*None*	*$*	*No*	*No*	*Thai*	*0–10*	$y_{11} =$ *No*
$\mathbf{x}_{12}$	*Yes*	*Yes*	*Yes*	*Yes*	*Full*	*$*	*No*	*No*	*Burger*	*30–60*	$y_{12} =$ *Yes*

Alternative
NOT Friday/Saturday
Yes hungry
No Rain
cuisine
Wait : Yes!
10 Attributes.
TYPE or PATRONS?
Which is better for DTC?

IG for Type



Yes. $Pr(Yes) = \frac{1}{2}$

$Pr(No) = \frac{1}{2}$

X

No

IG for Type

Information Gain

- IG for the **TYPE feature** is given as

$$H(X) - H(X|TYPE)$$

French

Italian

Thai

Burger

IG for Type

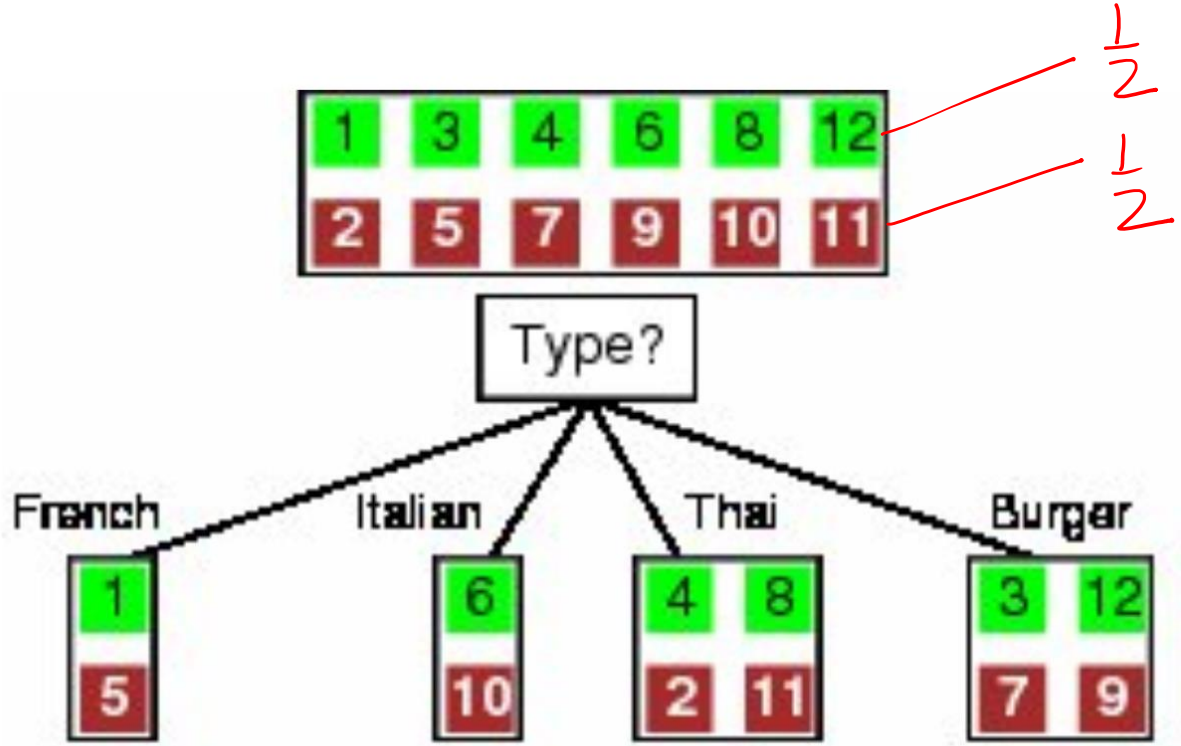
- IG for the **TYPE feature** is given as

$$\mathrm{IG}(TYPE) = H(X) - H(X|TYPE)$$

IG for Type

$$H(X) = H\left(\frac{1}{2}, \frac{1}{2}\right) = \frac{1}{2}\log_2 2 + \frac{1}{2}\log_2 2$$

$$= \frac{1}{2} \times 1 + \frac{1}{2} \times 1 = 1$$



IG for Type

$$H(X) = H\left(\frac{1}{2}, \frac{1}{2}\right) = 1$$

IG for Type

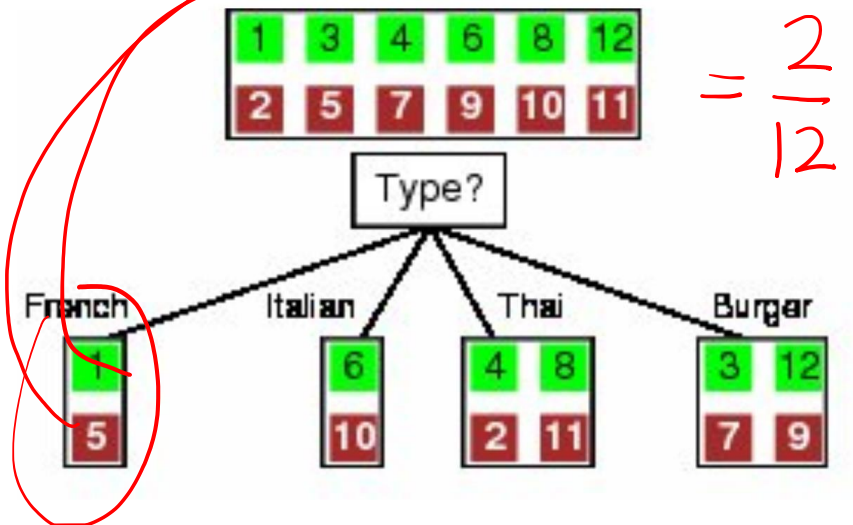
Conditional Entropy / Type

$$H(X|TYPE)$$

$$= P(\text{Fr}) \times H(X|\text{Fr}) + P(\text{It}) \times H(X|\text{It}) + P(\text{Th}) \times H(X|\text{Th}) + P(\text{Bu}) \times H(X|\text{Bu})$$

$$= \frac{2}{12} \times H\left(\frac{1}{2}, \frac{1}{2}\right) + \frac{2}{12} \times H\left(\frac{1}{2}, \frac{1}{2}\right) + \frac{4}{12} \times H\left(\frac{1}{2}, \frac{1}{2}\right) + \frac{4}{12} \times H\left(\frac{1}{2}, \frac{1}{2}\right)$$

No / Yes



$$= \frac{2}{12} + \frac{2}{12} + \frac{4}{12} + \frac{4}{12} = 1$$

$H(X|TYPE)$.

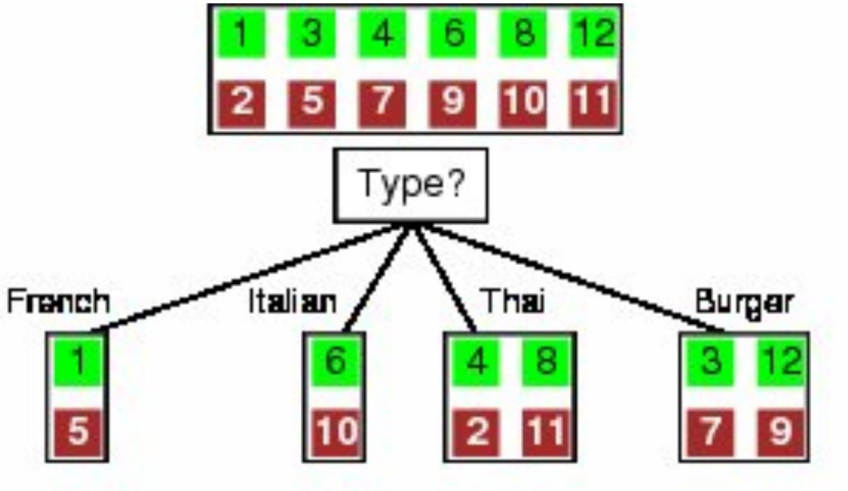
IG for Type

$$H(X|TYPE)$$

$$= P(\text{Fr}) \times H(X|\text{Fr}) + P(\text{It}) \times H(X|\text{It}) + P(\text{Th}) \times H(X|\text{Th}) + P(\text{Bu}) \times H(X|\text{Bu})$$

$$= \frac{2}{12} \times 1 + \frac{2}{12} \times 1 + \frac{4}{12} \times 1 + \frac{4}{12} \times 1$$

$$= 1$$



IG for Type

- IG for the **TYPE feature** is given as

$$\mathrm{IG}(TYPE) = H(X) - H(X|TYPE)$$

$$= 1 - 1 = 0$$

→ NOT suitable for DTC

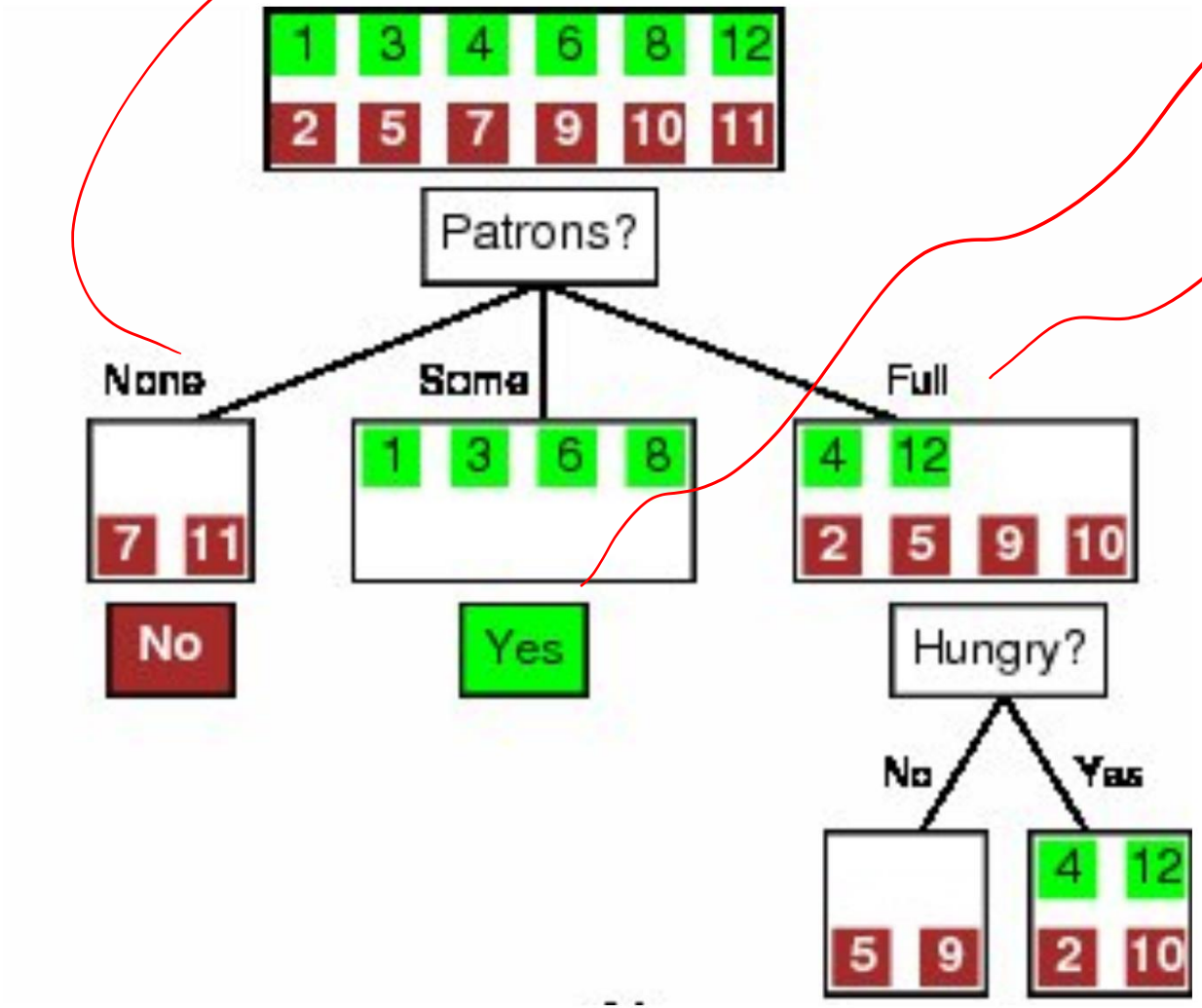
IG for Type

- IG for the **TYPE feature** is given as

$$\mathrm{IG}(TYPE) = H(X) - H(X|TYPE)$$
$$= 1 - 1 = 0$$

Information Gain = 0.

IG for Patrons



None
Some
Full..

IG for Patrons

- IG for **PATRONS** feature is given as follows

$$H(X) - H(X|\text{PATRONS})$$

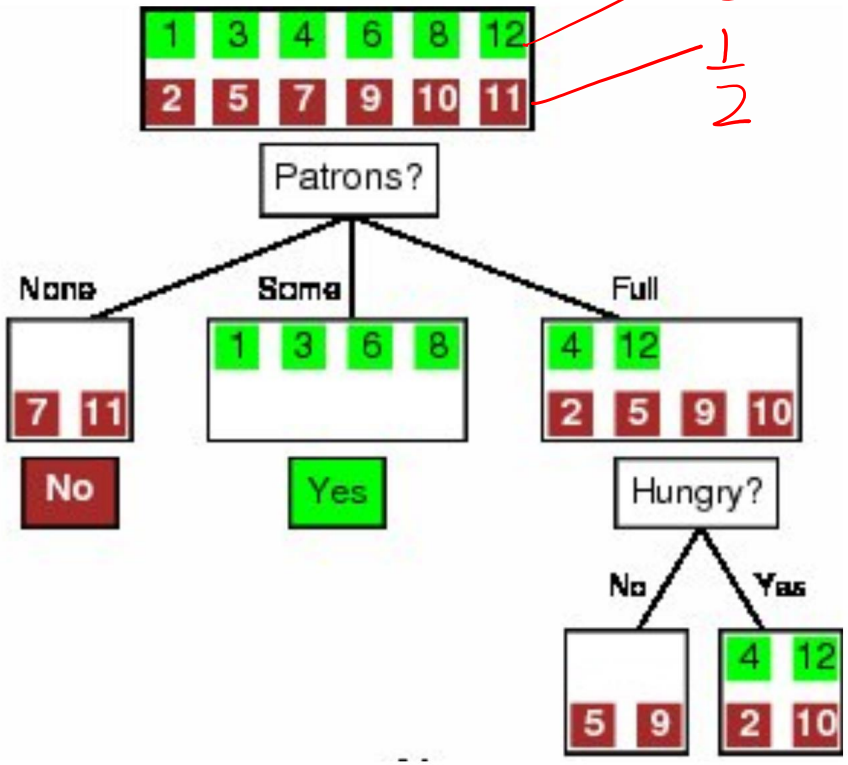
IG for Patrons

- IG for **PATRONS** feature is given as follows

$$H(X) - H(X|PATRONS)$$

IG for Patrons

$$H(X) = H\left(\frac{1}{2}, \frac{1}{2}\right) = 1$$



$\frac{1}{2}$

$\frac{1}{2}$

$= H\left(\frac{1}{2}, \frac{1}{2}\right) = 1 = H(X).$

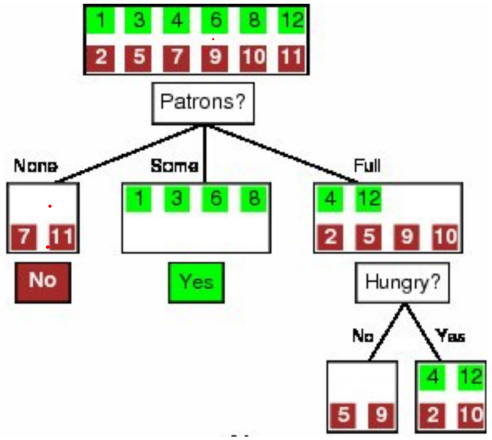
IG for PATRONS

$$H(X|PATRONS)$$
$$= P(\text{None}) \times H(X|\text{None}) + P(\text{Some}) \times H(X|\text{Some}) + P(\text{Full}) \times H(X|\text{Full})$$

$$= \frac{2}{12} \times H(0,1) + \frac{4}{12} \times H(1,0) + \frac{6}{12} \times H\left(\frac{1}{3}, \frac{2}{3}\right)$$

$$= \frac{1}{2} \cdot \left(\frac{1}{3} \cdot \log_2 3 + \frac{2}{3} \cdot \log_2 \frac{3}{2}\right)$$

$$\approx 0.46 \sim H(X|PATRONS).$$



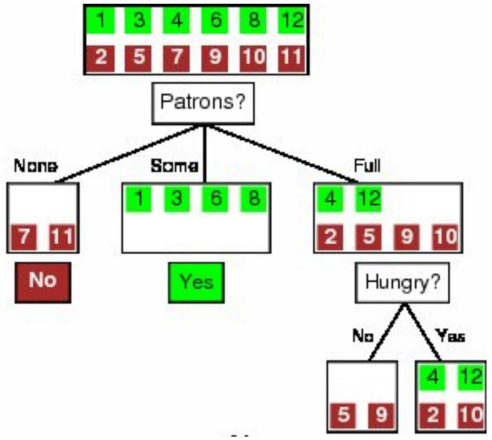
IG for PATRONS

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

$$= P(\text{None}) \times H(X|\text{None}) + P(\text{Some}) \times H(X|\text{Some}) + P(\text{Full}) \times H(X|\text{Full})$$

$$= \frac{2}{12} \times 0 + \frac{4}{12} \times 0 + \frac{1}{2} \times H\left(\frac{1}{3}, \frac{2}{3}\right)$$

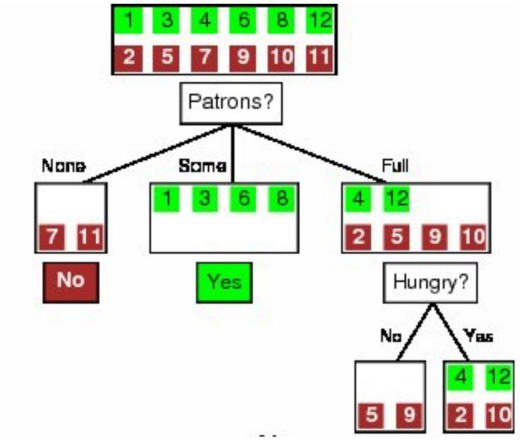
$$= 0.46$$



IG for Type

$$H(X) - H(X|PATRONS)$$

$$= 1 - 0.46 = 0.54$$

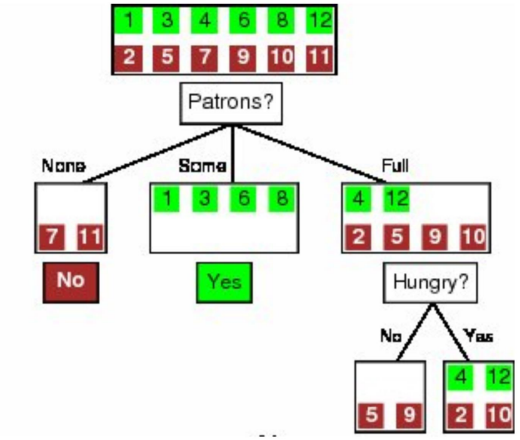


IG for Type

$$H(X) - H(X|PATRONS)$$

$$= 1 - \left(\frac{2}{12} \times 0 + \frac{4}{12} \times 0 + \frac{1}{2} \times H\left(\frac{1}{3}, \frac{2}{3} \right) \right)$$

$$= 1 - 0.46 = 0.54$$



IG for Type

- Since

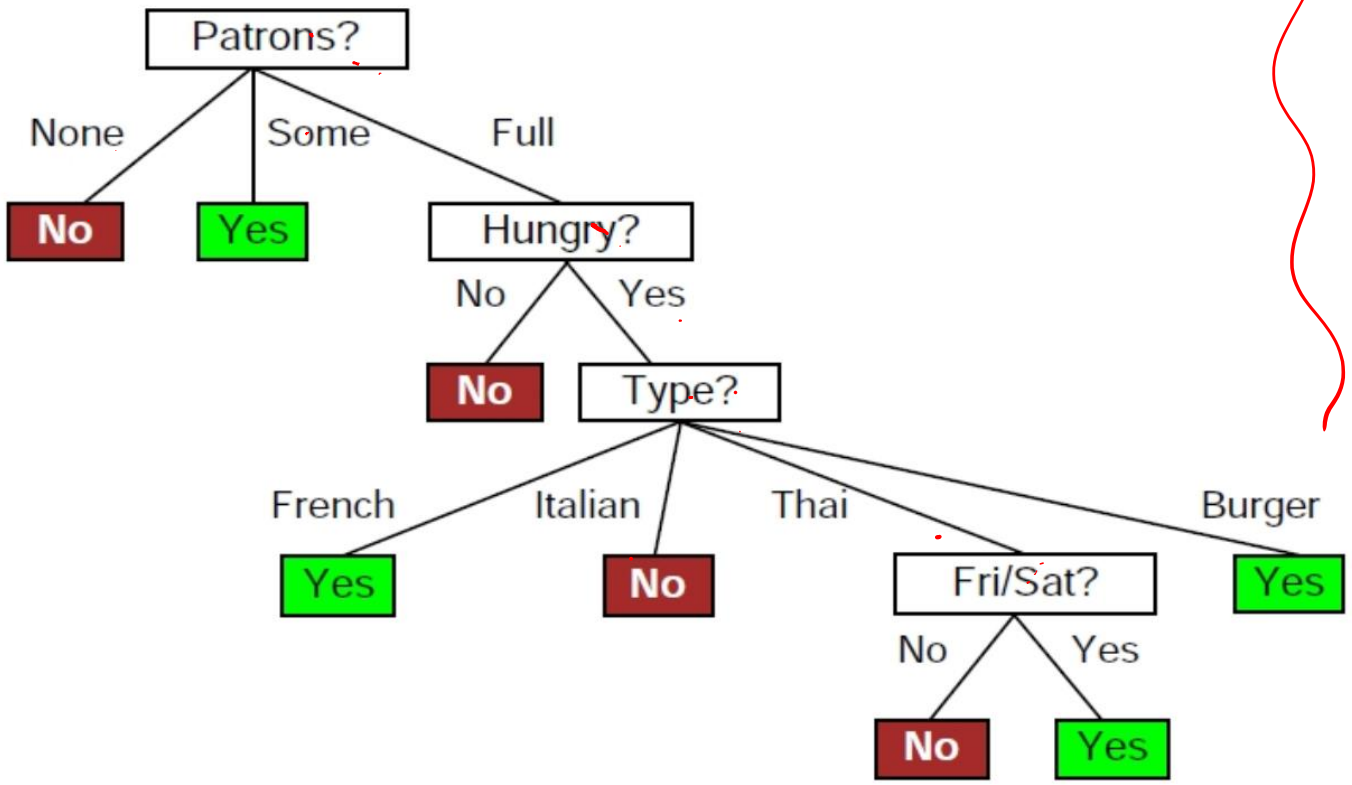
$$\mathrm{IG}(\mathrm{PATRONS}) = 0.54 > 0 = \mathrm{IG}(\mathrm{TYPE})$$

Choose PATRONS as attribute for DTC.

IG for Type

- We choose **PATRONS** as the feature to split

Final DTC



DTC.

Instructors may use this white area (14.5 cm / 25.4 cm) for the text. Three options provided below for the font size.

Font: Avenir (Book), Size: 32, Colour: Dark Grey

Font: Avenir (Book), Size: 28, Colour: Dark Grey

Font: Avenir (Book), Size: 24, Colour: Dark Grey

Do not use the space below.