Chapter 8

Decision Tree Classifiers

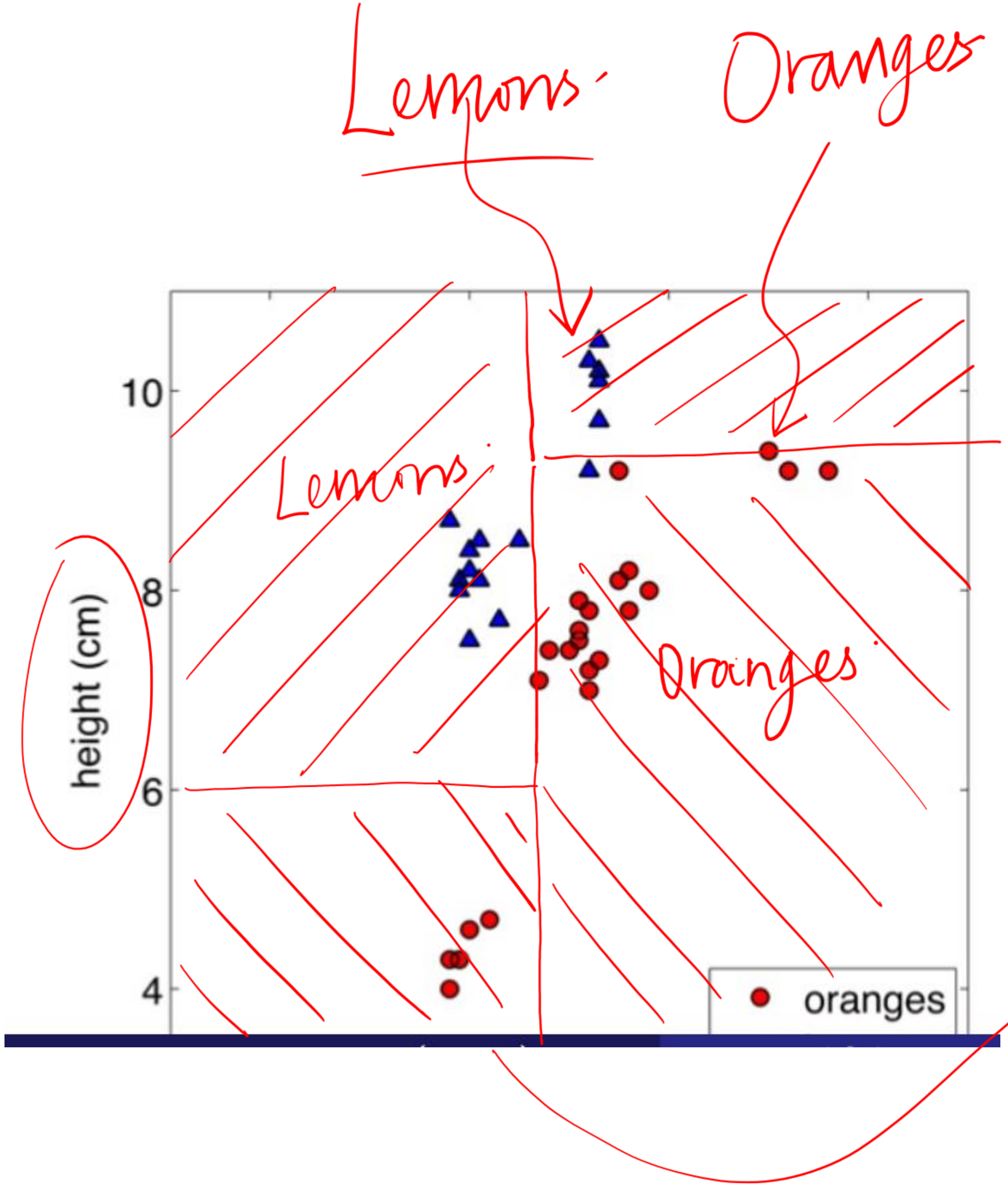
DTC -

Decision Trees -

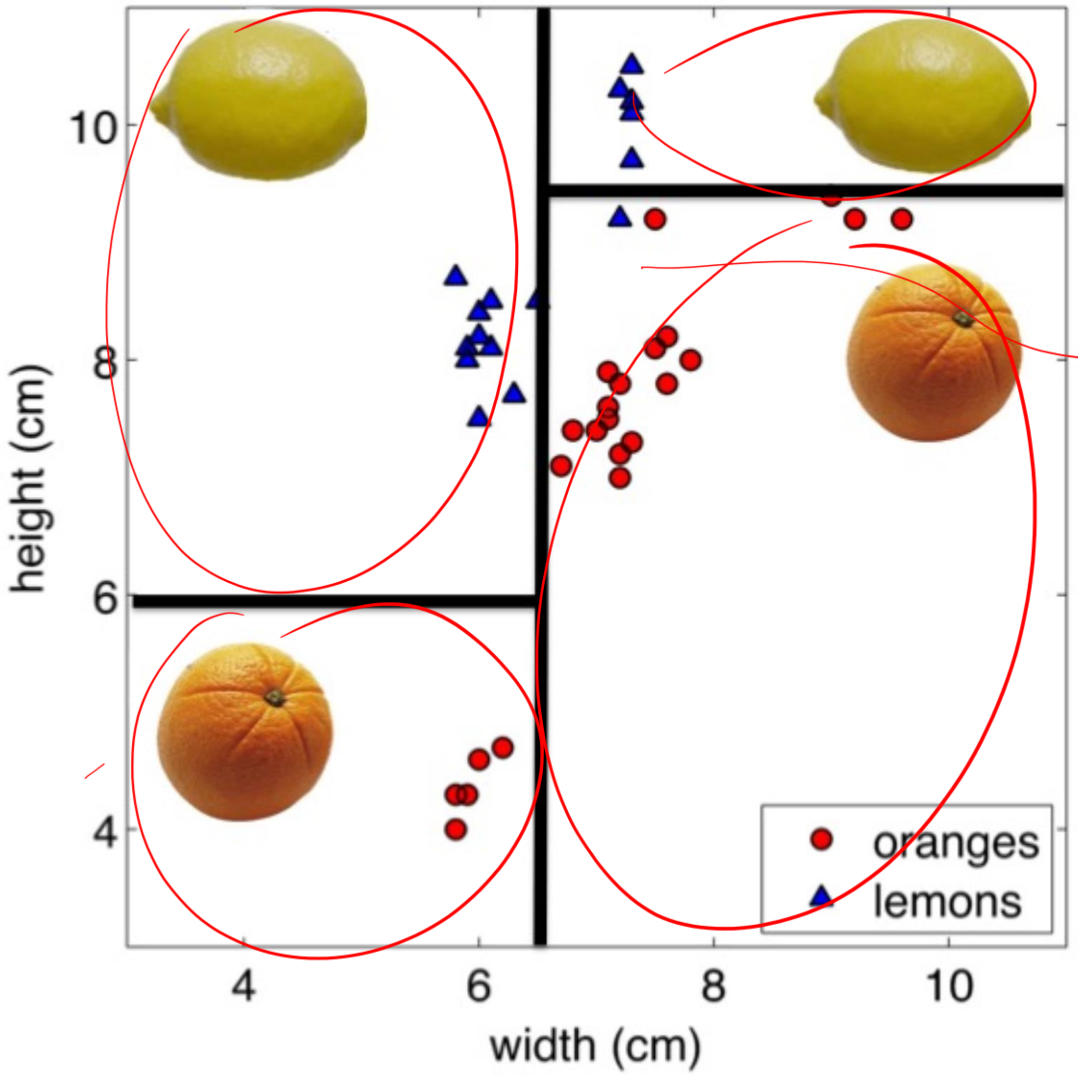
DTC - Example

2 Features

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



DTC - Example



- Dataset can be partitioned as shown

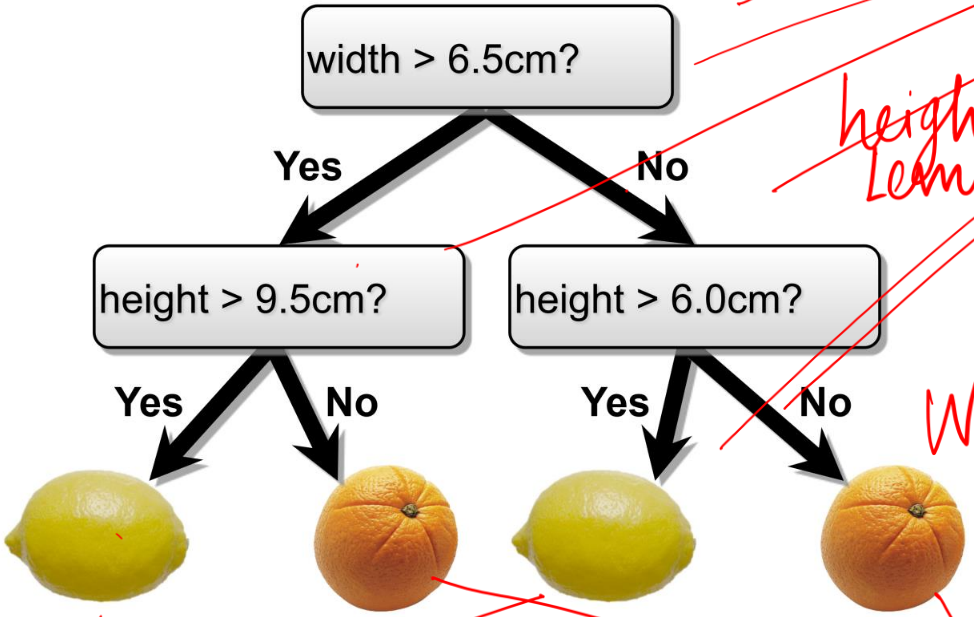
Highly non-linear partition

DTC - Example

Decision Tree Classifier

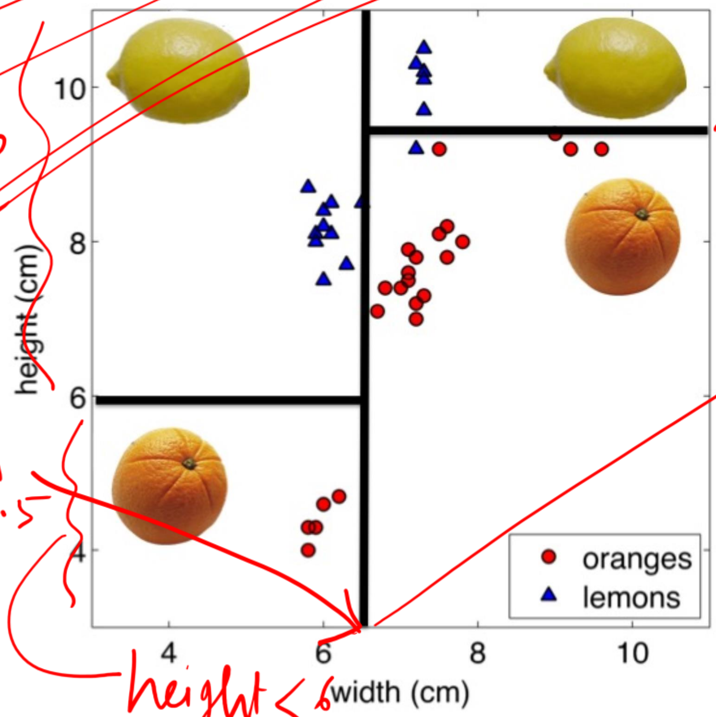
- DTC below is built using the given ***dataset***

DTC. Nodes Branches.



height>6 Lemon

Width < 6.5



height > 9.5 ⇒ Lemon

height < 9.5 orange

Width > 6.5

height < 6 Orange

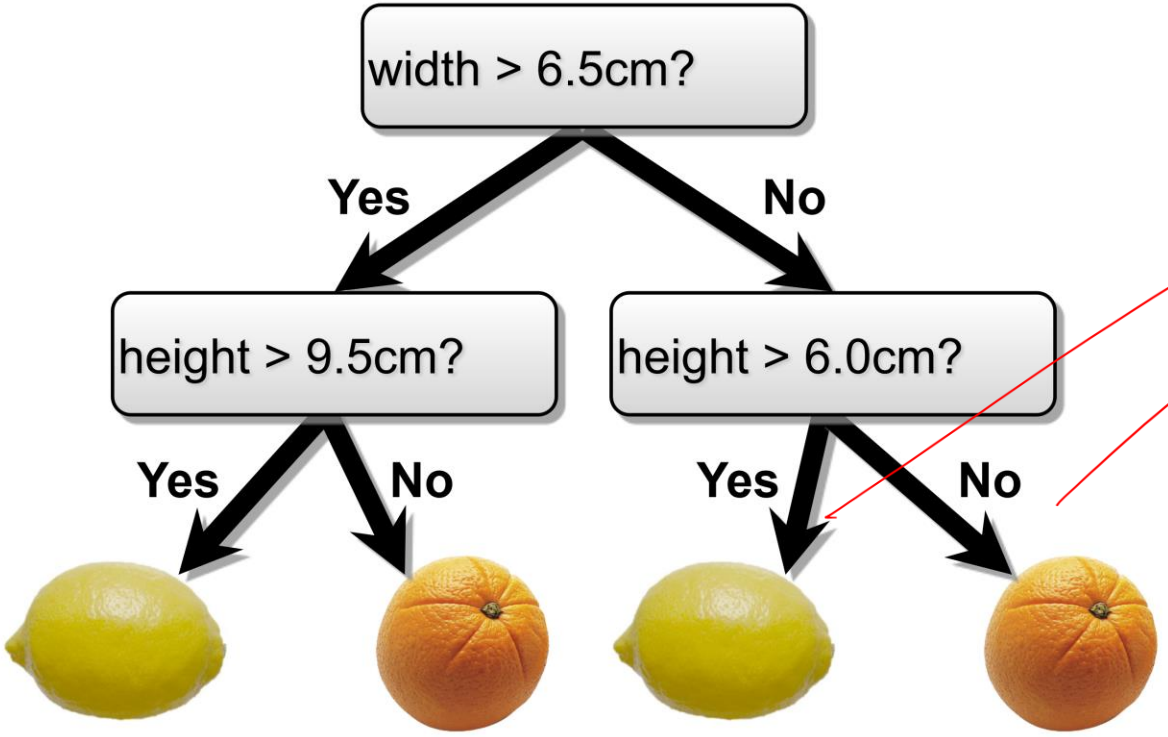
Lemon

Orange

Decision Tree Classifiers

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

Lemon
Orange
Binary

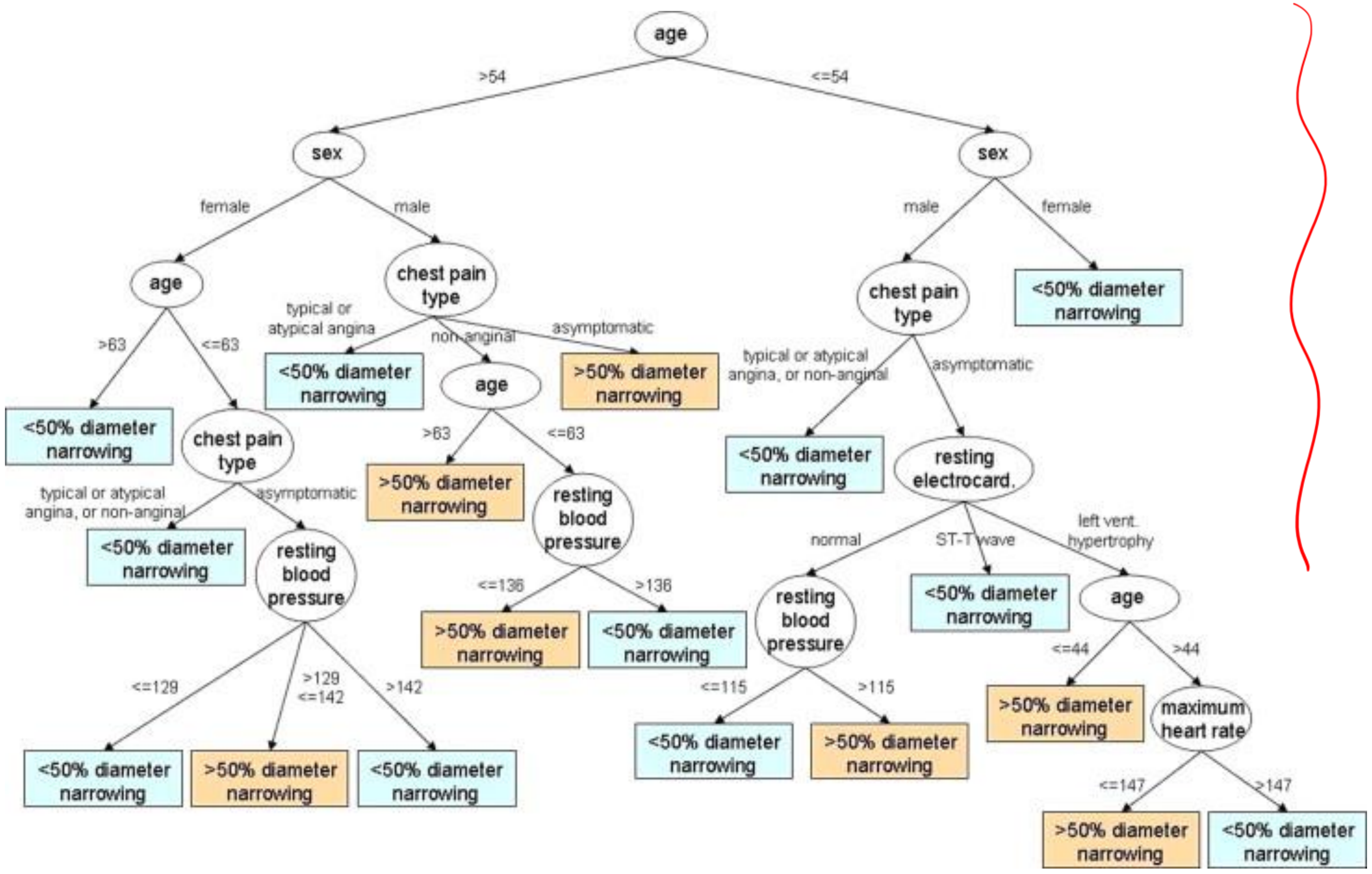


Decision Tree Classifiers

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

Decision Tree Classifiers

Decision tree to identify patients with heart disease



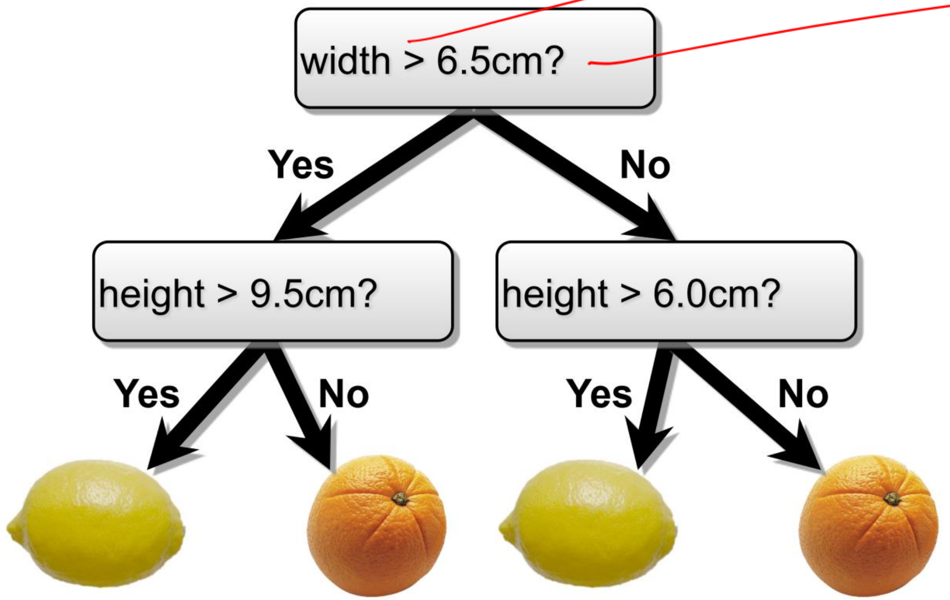
DTC
For Heart Disease

Learning Decision Trees

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

Width ?

How to choose best Feature at each stage?



Entropy

- Consider a ***event*** X with outcomes x_i and probabilities $p(x_i)$

Entropy

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

$x_1, x_2, \ldots, x_n$

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

Entropy

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

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

Entropy-Example

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

x_1

x_2

IC − likes

$\overline{\mathrm{IC}}$ − dislikes

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

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

Entropy-Example

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

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

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

$$= 0.811$$

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

$H\left(\frac{3}{4}\right)$

Entropy.

Entropy-Example

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

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

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

DTC Example

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

W

$\overline{W}$

10 Features-

Dataset

Alternative

Bar

Friday

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

DTC Example

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

Hungry or not

Are there patrons in restaurant

How expensive

Is there reservation

cuisine

Estimated wait Time

Decision
Wait
Not wait

DTC Example

- **Table columns**

Features -

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).

Entropy for Response (Will Wait)

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

$$= \frac{1}{2}\log_2(2) + \frac{1}{2}\log_2(2)$$

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

Entropy

W $\quad p(W) = \frac{6}{12} = \frac{1}{2}$

$\bar{W}$ $\quad p(\bar{W}) = \frac{6}{12} = \frac{1}{2}$

Entropy for Response (Will Wait)

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

Entropy of customer response

Entropy Note:

$$H(0,1) = H(1,0) = 0$$

Sunrise

$p(East) = 1$

$p(West) = 0.$

$H(0,1) = 0$

$0 \cdot \log_2 \frac{1}{0} + 1 \log_2 1.$

0

$\lim \to 0^-$

Entropy:

$0 \leq H(p, 1-p) \leq 1$

$\underbrace{H\left(\frac{1}{2}, \frac{1}{2}\right) = 1}_{\text{maximum}}$

Conditional entropy

- Consider **two events**: X with outcomes x_i and Y with outcomes y_j

Conditional entropy

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

$y_1, y_2, \dots, y_m$

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

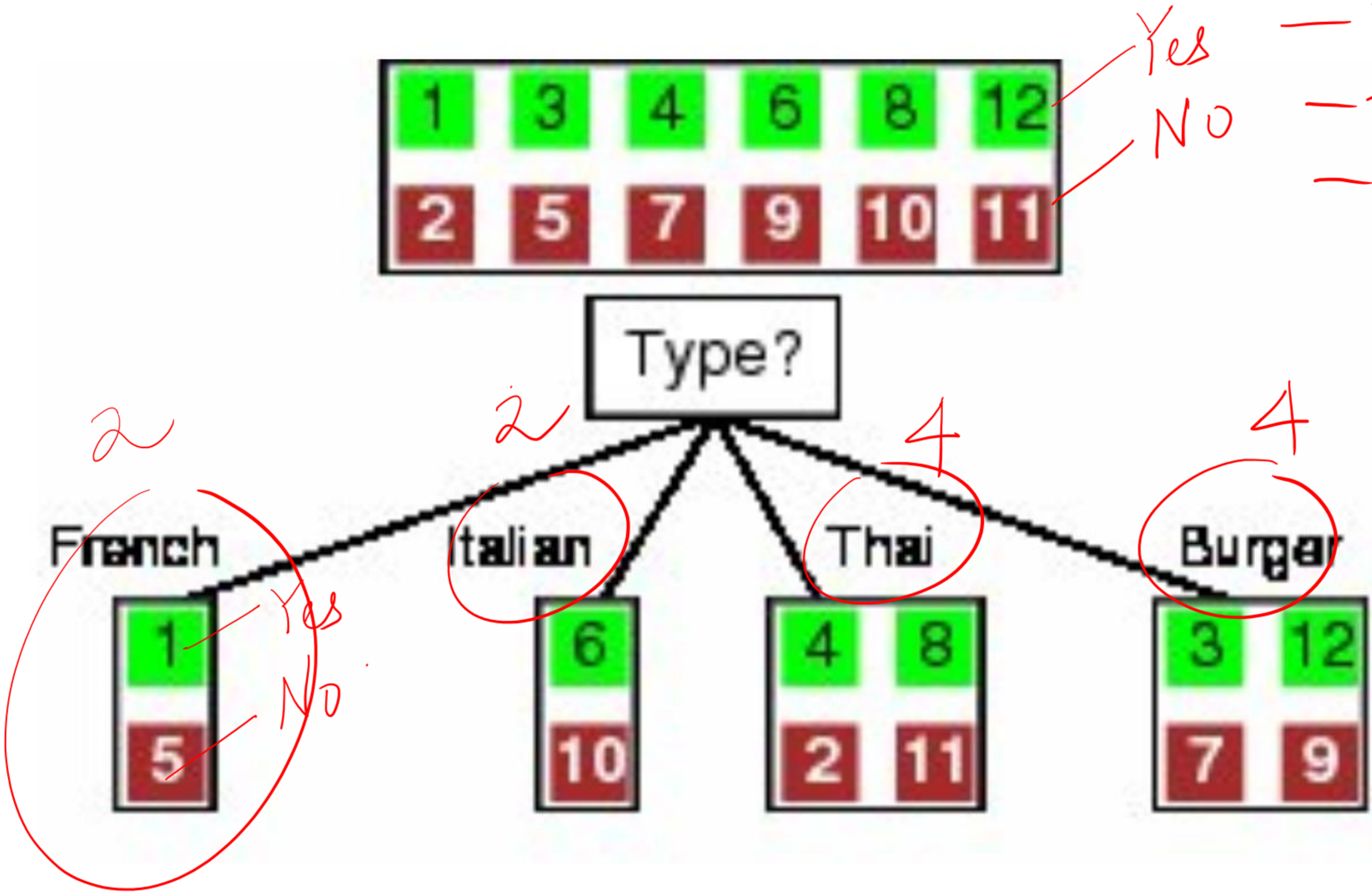
Conditional entropy (CE)

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

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

CE for Type

Cuisine
- French
- Italian
- Thai
- Burger



CE for Type

$$H(X|TYPE)$$

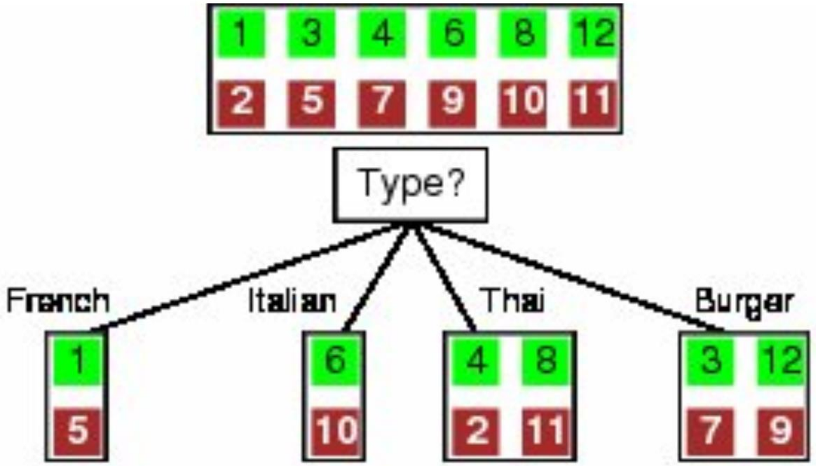
Final Decision

4 Possible values

$= H(X|Fr) \cdot Pr(Fr) + H(X|It) \cdot Pr(It)$

$+ H(X|Th) Pr(Th) + H(X|Br) \cdot Pr(Br).$

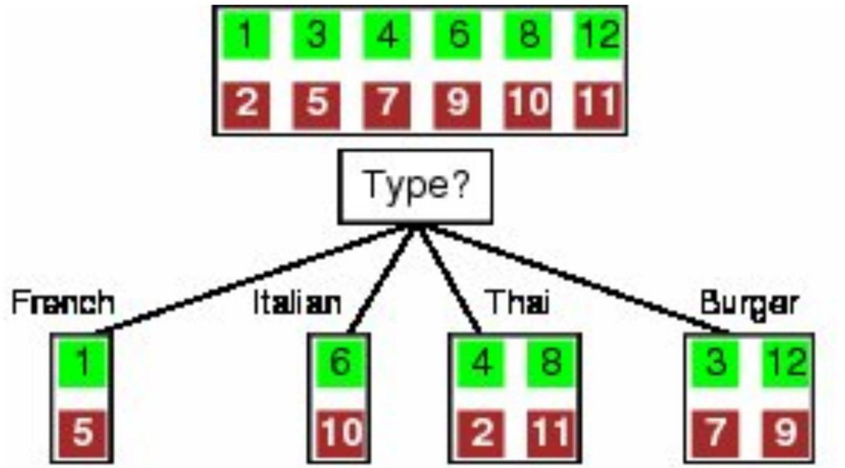
conditional entropy
of Final decision given type.



CE for Type

$$H(X|TYPE)$$

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

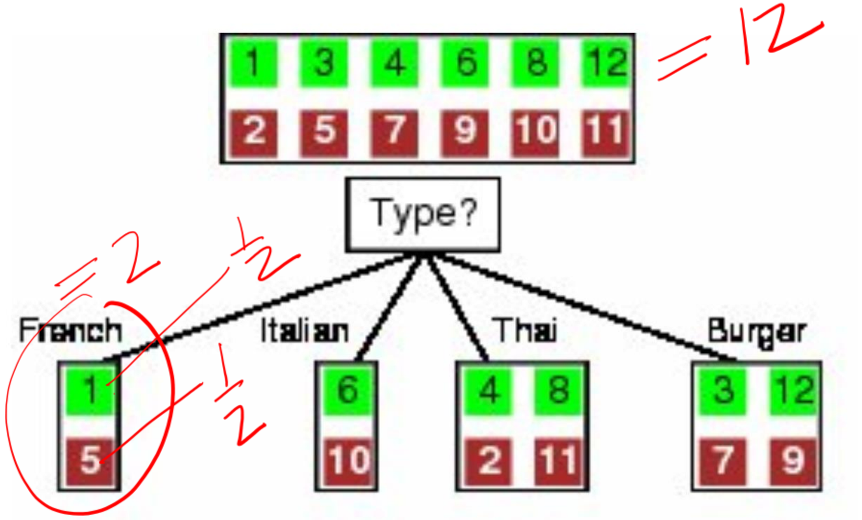


Type

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

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

$$P(\mathrm{Fr}) = \frac{2}{12} = \frac{1}{6}$$

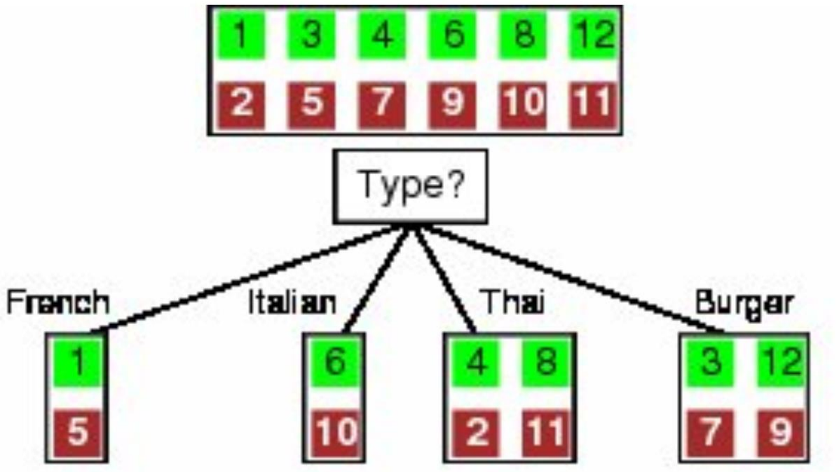


Type

Entropy of X given French

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

$$P(\mathrm{Fr}) = \frac{2}{12} = \frac{1}{6}$$



CE for Type

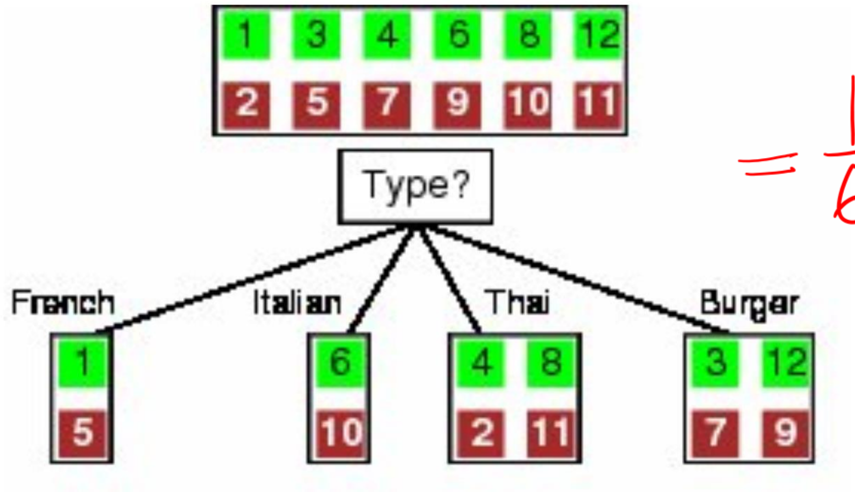
$$H(X|TYPE) = 1$$

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

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

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

$$= \frac{1}{6} + \frac{1}{6} + \frac{1}{3} + \frac{1}{3} = 1.$$



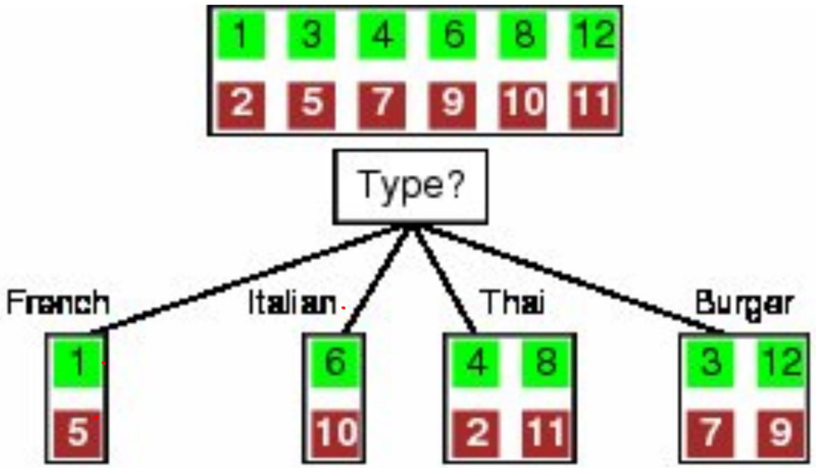
CE for Type

$$H(X|TYPE) = 1$$

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

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

$$= 1$$



Information gain

Of a Feature Y.

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

Final Response

Feature.

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

Information gain

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

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

Mutual information

Information gain

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

DTC Feature Selection

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

Information gain for TYPE

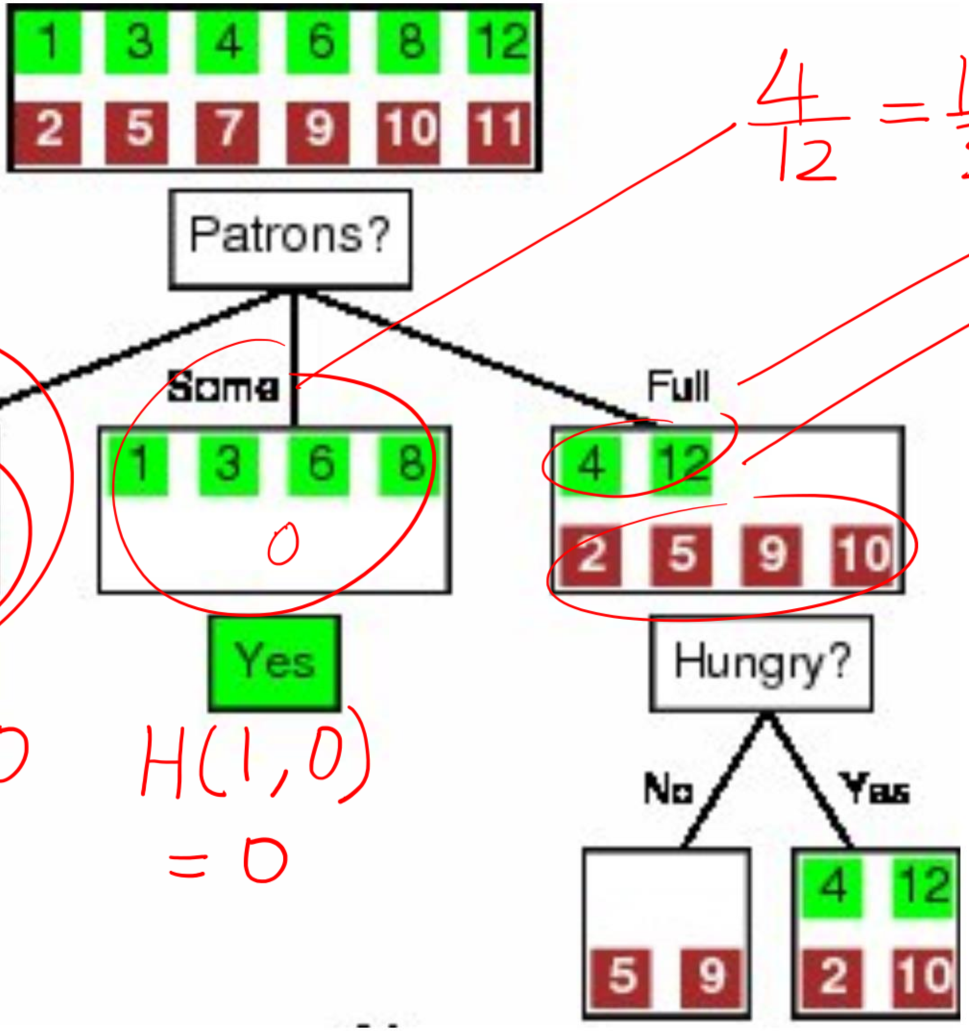
- The **information gain (IG)** for type feature is

Type

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

Information gain of Type = 0

IG for Patrons



$= 2$

$\frac{2}{12} = \frac{1}{6}.$

0

$H(0,1) = 0$

$H(1,0) = 0$

$\frac{4}{12} = \frac{1}{3}.$

$\frac{6}{12} = \frac{1}{2}$

$H\left(\frac{2}{6}, \frac{4}{6}\right) = H\left(\frac{1}{3}, \frac{2}{3}\right).$

IG for Patrons

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

Appendix

Pr(Full)

$H(X|Full)$

$= H(\frac{1}{3}, \frac{2}{3})$

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

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

$$= 1 - 0.46 = 0.54$$

Information gain Patrons.

Pr(None)

$H(0,1) = 0$

Pr(Some)

$H(X|Some)$.

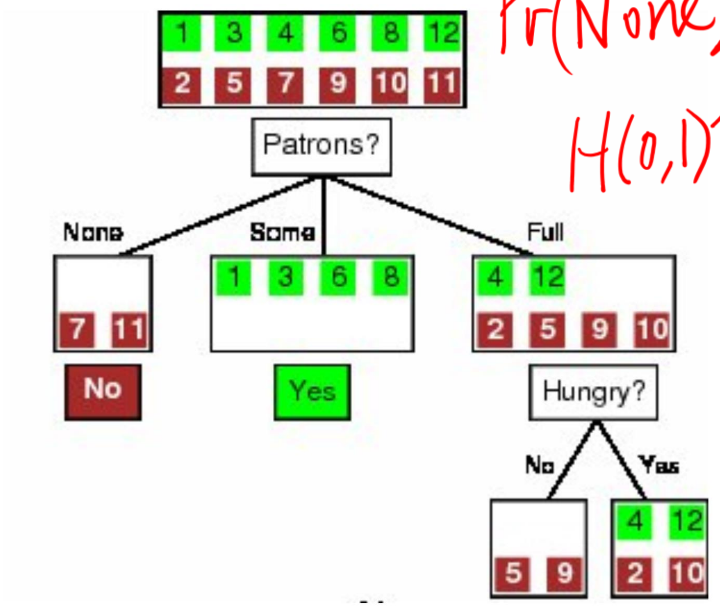
$H(1,0) = 0$.

$H(X|Patrons)$.

$H(X|None) . Pr(None)$

$+ H(X|Some) . Pr(Some)$

$+ H(X|Full) \times Pr(Full)$.



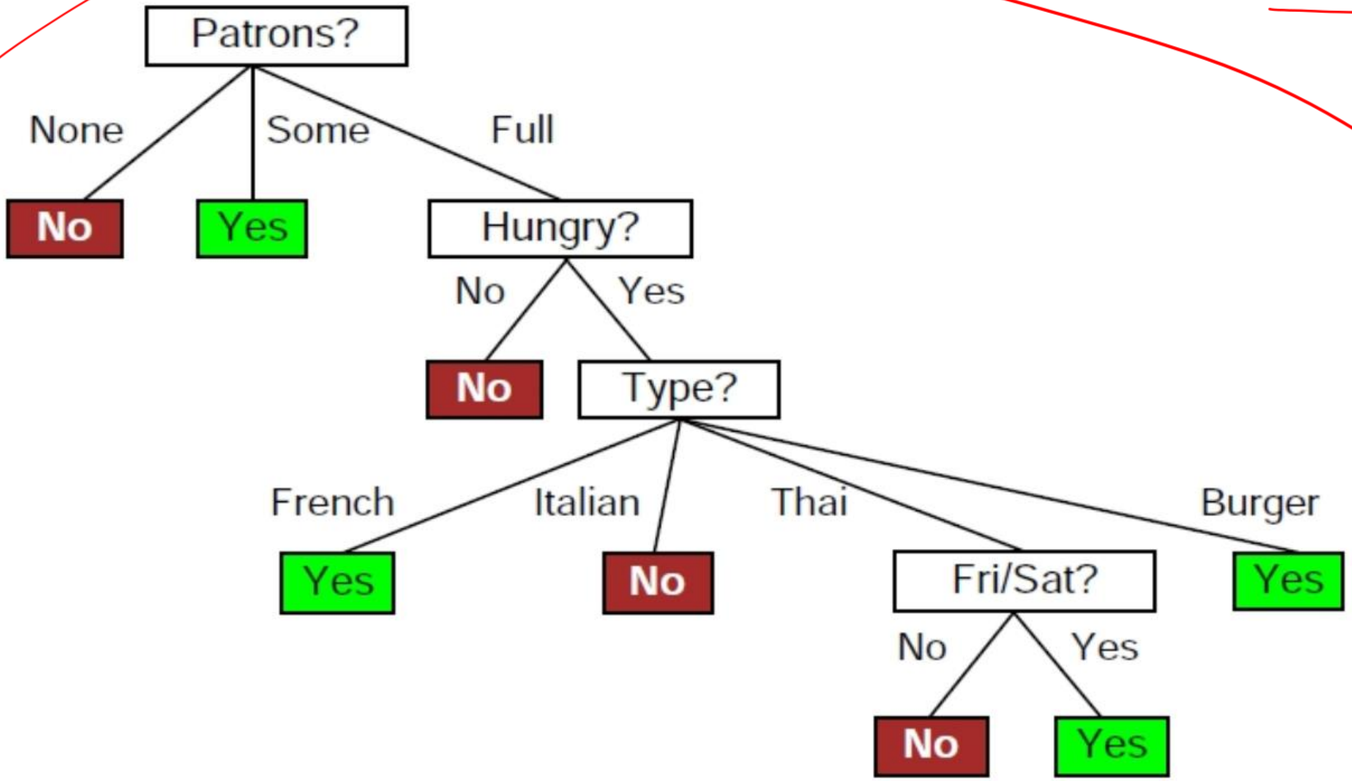
IG for Type

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

information gain
For patrons is higher

Final DTC

Restaurant Dataset



PYTHON DTC

```python
from sklearn.tree import DecisionTreeClassifier
from sklearn.tree import plot_tree
from sklearn import tree
from sklearn import datasets
```

Datasets

Plotting DTC

DTC module

PYTHON DTC

```
irisset = datasets.load_iris()
X = irisset.data
Y = irisset.target
```

Load IRIS Dataset

IRIS Data
IRIS Features

Response
Type of IRIS Flower

SVD and PCA

- This data sets consists of 3 different types of irises'
  - VERSICOLOR
  - SETOSA
  - VIRGINICA.

Iris Versicolor Iris Setosa Iris Virginica

SVD and PCA

- This data sets consists of 3 different types of irises’
  - Setosa
  - Versicolour
  - Virginica

Iris Versicolor Iris Setosa Iris Virginica

SVD and PCA

- The features are
  - SEPAL LENGTH
  - SEPAL WIDTH
  - PETAL LENGTH
  - PETAL WIDTH.
- Stored in a 150x4 numpy.ndarray

Iris Versicolor Iris Setosa Iris Virginica

SVD and PCA

- The features are
  - **Sepal Length**
  - **Sepal Width**
  - **Petal Length**
  - **Petal Width.**
- Stored in a 150x4 numpy.ndarray

150 Rows.

4 Features for each point

$$\bar{x} = \begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{bmatrix} \qquad y \begin{cases} 0 \\ 1 \\ 2 \end{cases}$$

Iris Versicolor | Iris Setosa | Iris Virginica

PYTHON DTC

```
cf=DecisionTreeClassifier(random_state=1234);
cf.fit(X,Y);

```

Random State

Fit DTC to IRIS Data

PYTHON DTC

```
decPlot = plot_tree(decision_tree=cf,
                    feature_names = ["sepal_length",
                    "sepal_width","petal_length","petal_width"],
                     class_names =["setosa", "vercicolor", "verginica"] ,
                     filled = True , precision = 4, rounded = True)
```

Plot DTC.

Color Filled Boxes

Precision of values.

PYTHON DTC

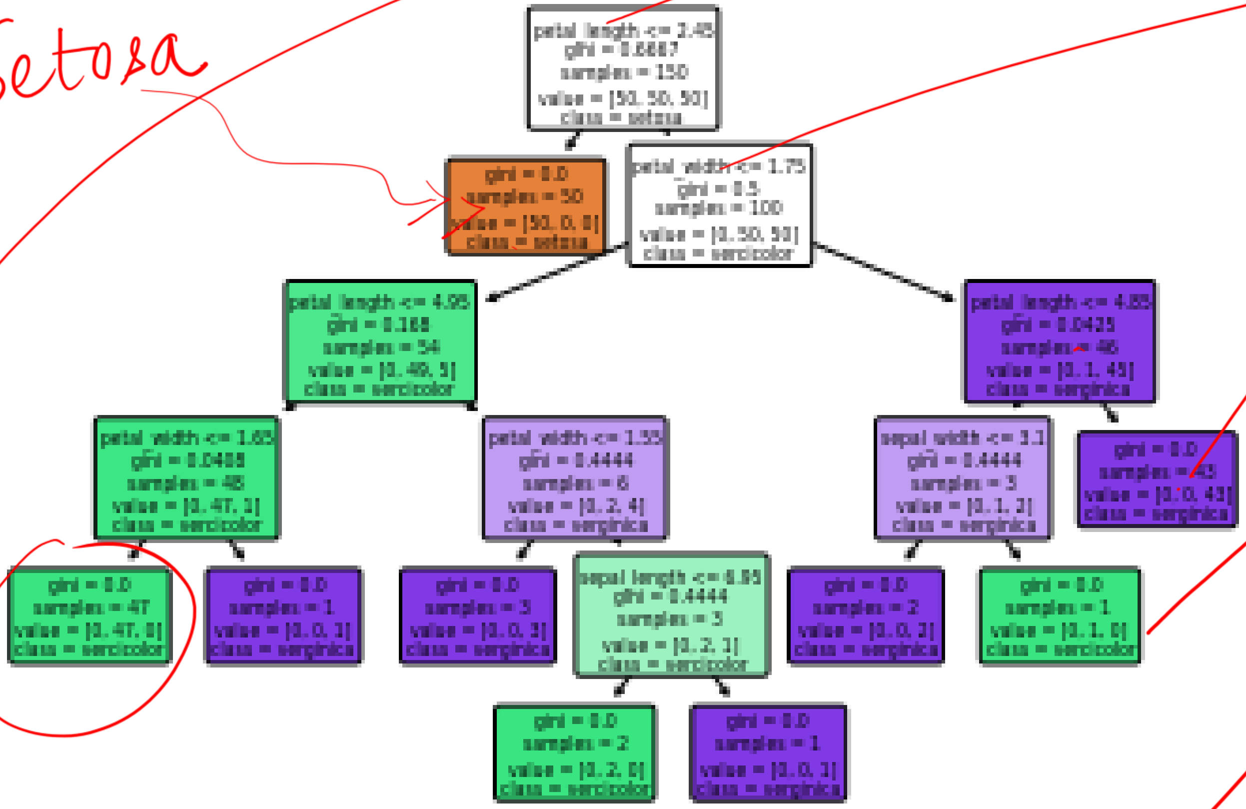
Petal Length

Setosa

Petal Width

Virginica

Versicolor



PYTHON DTC

DTC module

Feature names

```
text_representation = tree.export_text(cf,
                                        feature_names = ["sepal_length",
                                         "sepal_width","petal_length","petal_width"])

print(text_representation)
```

Print Text Representation

Text Representation of DTC

Generate Text Representation

PYTHON DTC

Text Representation of DTC.

```
|--- petal_length <= 2.45
|   |--- class: 0
|--- petal_length >  2.45
|   |--- petal_width <= 1.75
|   |   |--- petal_length <= 4.95
|   |   |   |--- petal_width <= 1.65
|   |   |   |   |--- class: 1
|   |   |   |--- petal_width >  1.65
|   |   |   |   |--- class: 2
|   |   |--- petal_length >  4.95
|   |   |   |--- petal_width <= 1.55
|   |   |   |   |--- class: 2
|   |   |   |--- petal_width >  1.55
|   |   |   |   |--- sepal_length <= 6.95
|   |   |   |   |   |--- class: 1
|   |   |   |   |--- sepal_length >  6.95
|   |   |   |   |   |--- class: 2
|   |--- petal_width >  1.75
|   |   |--- petal_length <= 4.85
|   |   |   |--- sepal_width <= 3.10
|   |   |   |   |--- class: 2
|   |   |   |--- sepal_width >  3.10
|   |   |   |   |--- class: 1
|   |   |--- petal_length >  4.85
|   |   |   |--- class: 2
```

Setosa

class 1

class 2

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

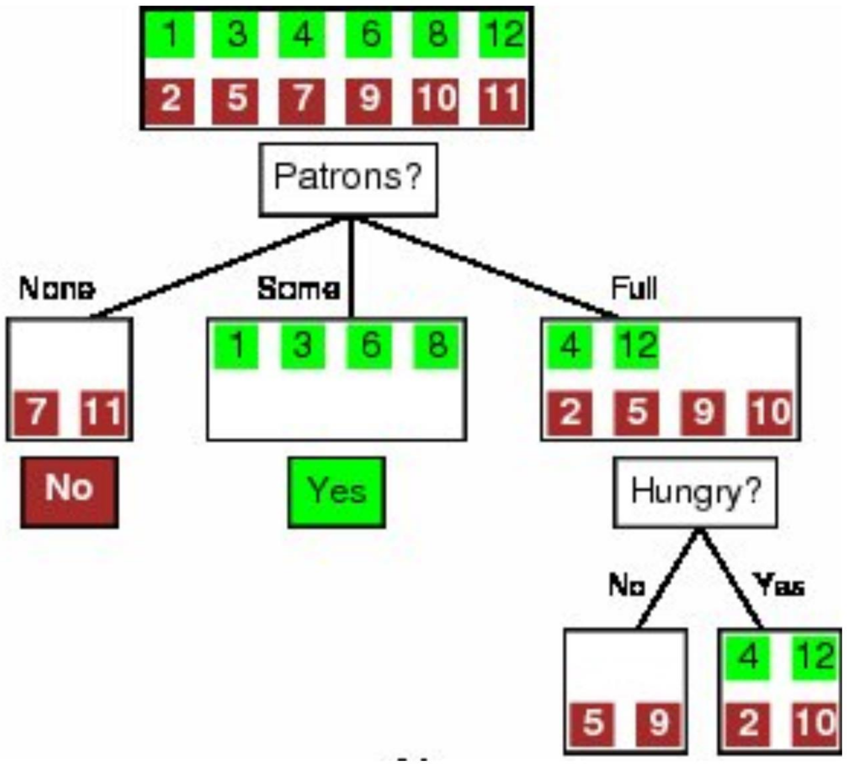
Font: Avenir (Book), Size: 24, Colour: Dark Grey

Do not use the space below.

Appendix: IG for Patrons

IG for Patrons

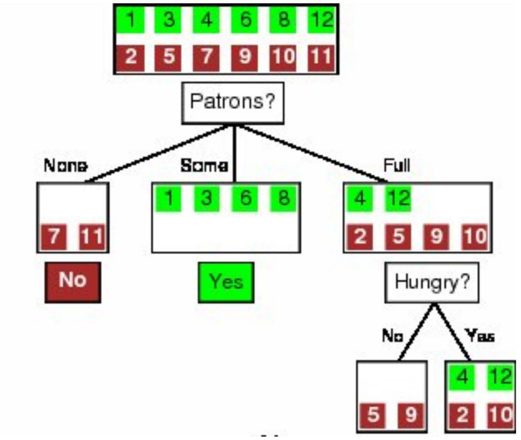
$$H(X) = H\left(\frac{1}{2}, \frac{1}{2}\right) = 1$$



IG for PATRONS

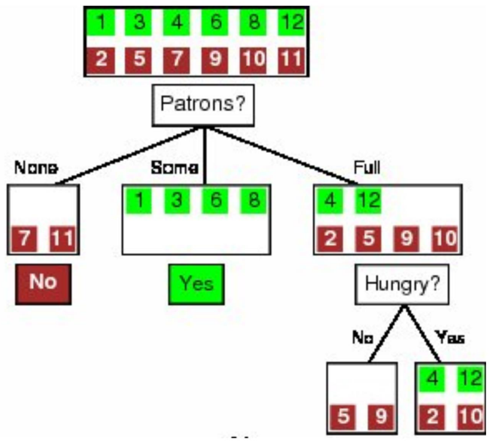
$$H(X|PATRONS)$$
$$= P(\text{None}) \times H(X|\text{None}) + P(\text{Some}) \times H(X|\text{Some}) + P(\text{Full}) \times H(X|\text{Full})$$

$$=$$



IG for PATRONS

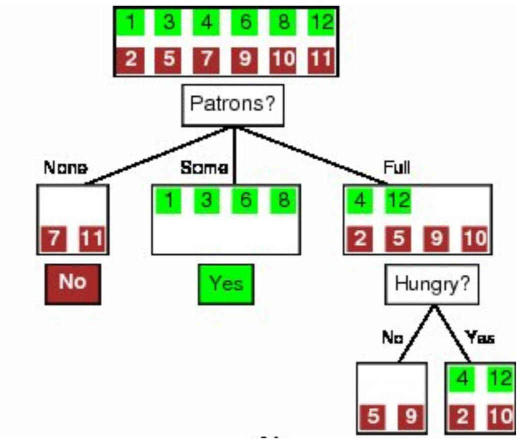
$$H(X|Y = PATRONS)$$
$$= P(\text{None}) \times H(X|\text{None}) + P(\text{Some})$$
$$\times H(X|\text{Some}) + P(\text{Full}) \times H(X|\text{Full})$$
$$= \frac{2}{12} \times 0 + \frac{4}{12} \times 0 + \frac{1}{2} \times H\left(\frac{1}{3}, \frac{2}{3}\right)$$
$$= 0.46$$



IG for Type

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

$$=$$

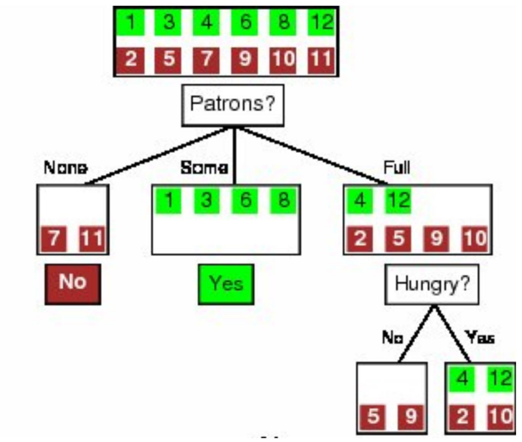


IG for Type

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

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

$$= 1 - 0.46 = 0.54$$



IG for Type

- Since

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