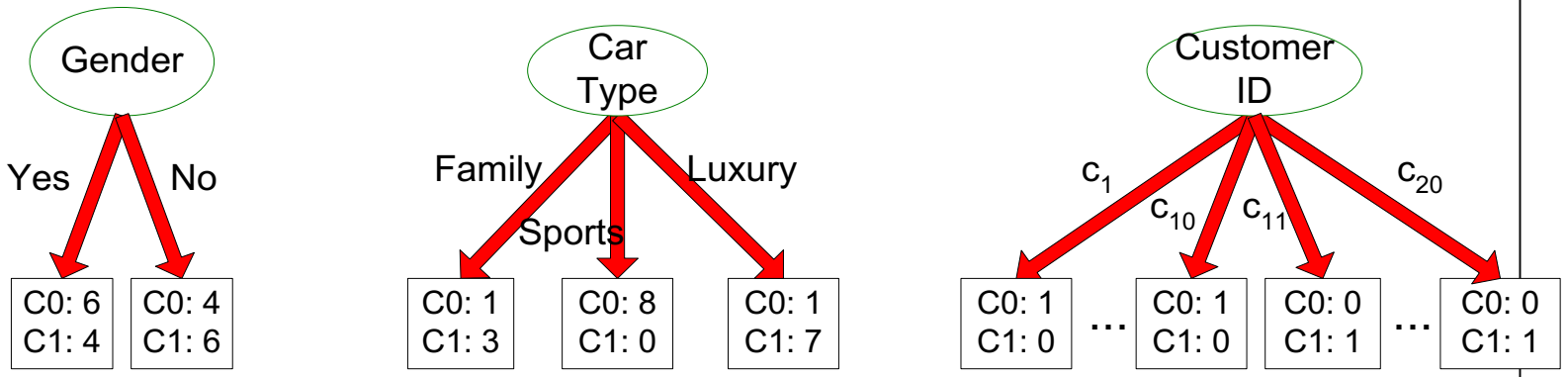
CLASSIFICATION

HOW TO DETERMINE THE BEST SPLIT

Before Splitting: 10 records of class 0 (c0),
10 records of class 1 (c1)

What are the values of the label for this data? How many cases / records for each label.

Learn the type of each attribute / feature, their values.



Which test condition is the best?

Customer Id	Gender	Car Type	Shirt Size	Class
1	M	Family	Small	C0
2	M	Sports	Medium	C0
3	M	Sports	Medium	C0
4	M	Sports	Large	C0
5	M	Sports	Extra Large	C0
6	M	Sports	Extra Large	C0
7	F	Sports	Small	C0
8	F	Sports	Small	C0
9	F	Sports	Medium	C0
10	F	Luxury	Large	C0
11	M	Family	Large	C1
12	M	Family	Extra Large	C1
13	M	Family	Medium	C1
14	M	Luxury	Extra Large	C1
15	F	Luxury	Small	C1
16	F	Luxury	Small	C1
17	F	Luxury	Medium	C1
18	F	Luxury	Medium	C1
19	F	Luxury	Medium	C1
20	F	Luxury	Large	C1

MEASURES OF NODE IMPURITY

- Gini Index

$$Gini\ Index = 1 - \sum_{i=0}^{c-1} p_i(t)^2$$

Where $\boldsymbol{p_i(t)}$ is the frequency of class $\boldsymbol{i}$ at node t, and $\boldsymbol{c}$ is the total number of classes

- Entropy

$$Entropy = -\sum_{i=0}^{c-1} p_i(t) log_2 p_i(t)$$

- Misclassification error (confusing matrix for decision tree)

$$Classification\ error = 1 - \max[p_i(t)]$$

FINDING THE BEST SPLIT

1. Compute impurity measure (P) before splitting
2. Compute impurity measure (M) after splitting
   - Compute impurity measure of each child node
   - M is the weighted impurity of child nodes
3. Choose the attribute test condition that produces the highest gain

Gain = P - M

or equivalently, lowest impurity measure after splitting (M)

GINI INDEX

- **What is Gini Index?**
- Gini index or Gini impurity measures the degree or probability of a particular variable being wrongly classified when it is randomly chosen.

$$Gini\ Index = 1 - \sum_{i=0}^{c-1} p_i(t)^2$$

- **But what is actually meant by 'impurity'?**
- If all the elements belong to a single class, then it can be called pure.
- The degree of Gini index varies between 0 and 1,
  0: all elements belong to a certain class or if there exists only one class, and
  1: the elements are randomly distributed across various classes.
- A Gini Index of 0.5 denotes equally distributed elements into some classes.

EXAMPLE OF GINI INDEX

attributes

label

Past Trend	Open Interest	Trading Volume	Return
Positive	Low	High	Up
Negative	High	Low	Down
Positive	Low	High	Up
Positive	High	High	Up
Negative	Low	High	Down
Positive	Low	Low	Down
Negative	High	High	Down
Negative	Low	High	Down
Positive	Low	Low	Down
Positive	High	High	Up

{P, N}

Goal: calculate the Gini index for feature / attribute "past trend"

Step1: find the values of this features (positive, negative)
Step2: For each value of this feature, we look at the corresponding labels and calculate the GI for this value.

positive < up / down

P(up | positive) = #(up ∪ p) / #(p) = 4/6

P(up|postive) = 4/6; p(down|positive) = 2/6
p(up|negavite) = 0/4; p(down|negative) = 4/4;
GI (past trend = positive) = 1 – [(4/6)^2 + (2/6)^2] = 0.45
GI (past trend = negative) = 1 – [(0)^2 +1^2] = 0

GI (past trend) = p(postive) * GI(positive) + p(negative) *GI(negative)
=(6/10)*0.45 + (4/10) * 0 = 0.27

EXAMPLE OF GINI INDEX

$GI(OI) = P(L) \cdot GI(L) + P(H) \cdot GI(H)$

$V = \{L, H\}$

Past Trend	Open Interest	Trading Volume	Return
Positive	Low	High	Up
Negative	High	Low	Down
Positive	Low	High	Up
Positive	High	High	Up
Negative	Low	High	Down
Positive	Low	Low	Down
Negative	High	High	Down
Negative	Low	High	Down
Positive	Low	Low	Down
Positive	High	High	Up

Goal: calculate the GI for the attribute / feature: open interest

1. Find all the values of OI (L, H)
2. For each value (L or H), we look at all the labels of each values.

L: U, D

$P(U|L) = \frac{\#U\&L}{\#L} = \frac{2}{6} = \frac{1}{3}$

$P(D|L) = \frac{4}{6} = \frac{2}{3}$

H: U, D

$P(U|H) = \frac{2}{4} = \frac{1}{2}$

$P(D|H) = \frac{2}{4} = \frac{1}{2}$

3. Calculate the GI for each value

$GI(L) = 1 - [P(U|L)^2 + P(D|L)^2] = 1 - [(\frac{1}{3})^2 + (\frac{2}{3})^2]$

$GI(H) = 1 - [P(U|H)^2 + P(D|H)^2] =$

MEASURE OF IMPURITY: ENTROPY

- Entropy at a given node t

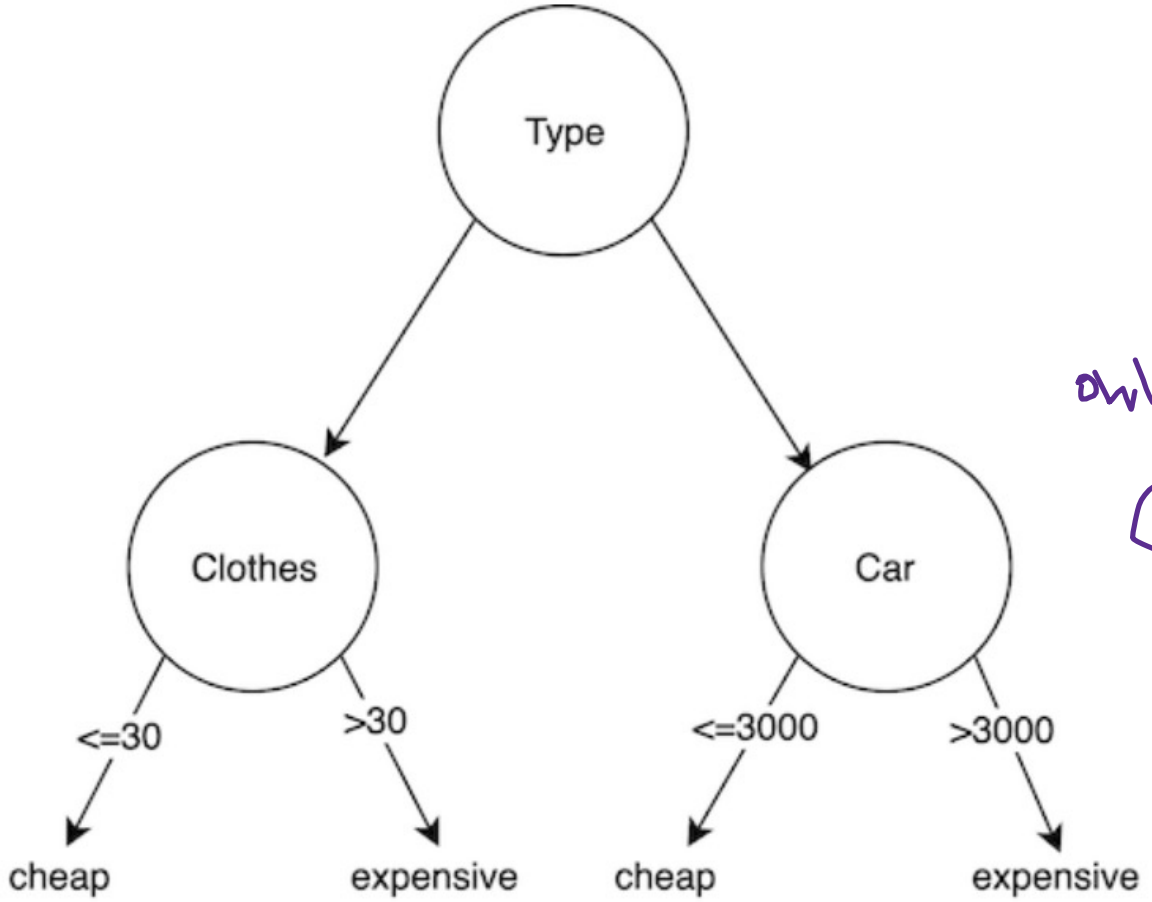
$$Entropy = -\sum_{i=0}^{c-1} p_i(t) log_2 p_i(t)$$

Where $p_i(t)$ is the frequency of class i at node t, and c is the total number of classes

  - Maximum of $\log_2 c$ when records are equally distributed among all classes, implying the least beneficial situation for classification
  - Minimum of 0 when all records belong to one class, implying most beneficial situation for classification

  - Entropy based computations are quite similar to the GINI index computations

ENTROPY

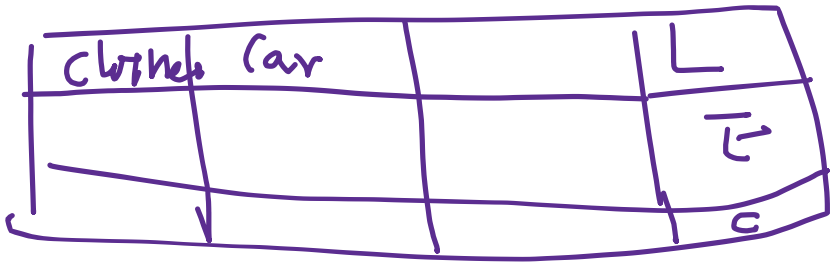


Entropy known as the controller for decision tree to decide where to split the data.

only 2 label

2-class Entropy: $(S) = -(p_1 * \log_2 p_1 + p_2 * \log_2 p_2)$

n-class Entropy -> $E(S) = \sum -(p_i * \log_2 p_i)$



ENTROPY

2-class Entropy: (S) =-(p1 * log2p1 + p2 * log2p2)

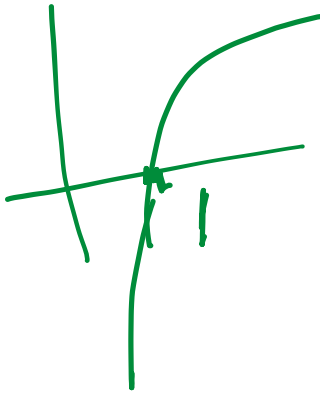
Sample Count	Sample Label
9	A
5	B

$$S = -\{ p_a \times \log_2 p_a + p_b \times \log_2 p_b \}$$

$$p_a = \frac{\#a}{\#T} = \frac{9}{14}$$

$$p_b = \frac{\#b}{\#T} = \frac{5}{14}$$

$$S = -\left[\frac{9}{14} * \log_2 \frac{9}{14} + \frac{5}{14} \cdot \log_2 \frac{5}{14} \right] = -$$



INFORMATION GAIN

Gain(S.A) = E(before) - G(after splitting)

Attribute

Weekend	Weather	Parental Availability	Wealthy	Decision (Class)
H1	Sunny	Yes	Rich	Cinema
H2	Sunny	No	Rich	Tennis
H3	Windy	Yes	Rich	Cinema
H4	Rainy	Yes	Poor	Cinema
H5	Rainy	No	Rich	Home
H6	Rainy	Yes	Poor	Cinema
H7	Windy	No	Poor	Cinema
H8	Windy	No	Rich	Shopping
H9	Windy	Yes	Rich	Cinema
H10	Sunny	No	Rich	Tennis

Goal: (1). Calculate the entropy for the dataset

- How many labels do you have (more than two)

L: Cinema, Tennis, H, S.

$$S = -\left[P_C * \log_2 P_C + P_T * \log_2 P_T + P_H * \log_2 P_H + P_S * \log_2 P_S \right]$$

$P_C = \frac{\#C}{\#T} = \frac{6}{10}$ $\quad P_T = \frac{\#T}{10} = \frac{2}{10}$

INFORMATION GAIN

$$Gain(S.A) = E(before) - G(after\ splitting)$$

Weekend	Weather	Parental Availability	Wealthy	Decision (Class)
H1	Sunny	Yes	Rich	Cinema
H2	Sunny	No	Rich	Tennis
H3	Windy	Yes	Rich	Cinema
H4	Rainy	Yes	Poor	Cinema
H5	Rainy	No	Rich	Home
H6	Rainy	Yes	Poor	Cinema
H7	Windy	No	Poor	Cinema
H8	Windy	No	Rich	Shopping
H9	Windy	Yes	Rich	Cinema
H10	Sunny	No	Rich	Tennis

Goal: (1). Calculate the entropy for the dataset

- How many labels do you have (more than two)

total Dataset

Gain (S. Weather)

① Values: { S, W, R }

② S: { Cinema (1/3), Tennis (2/3) }

R: { Cinema (2/3), Home (1/3) }

W: { Cinema (3/4), Shopping (1/4) }

INFORMATION GAIN

$$\text{Gain}(S.A) = E(\text{before}) - G(\text{after splitting})$$

Weekend	Weather	Parental Availability	Wealthy	Decision (Class)
H1	Sunny	Yes	Rich	Cinema
H2	Sunny	No	Rich	Tennis
H3	Windy	Yes	Rich	Cinema
H4	Rainy	Yes	Poor	Cinema
H5	Rainy	No	Rich	Home
H6	Rainy	Yes	Poor	Cinema
H7	Windy	No	Poor	Cinema
H8	Windy	No	Rich	Shopping
H9	Windy	Yes	Rich	Cinema
H10	Sunny	No	Rich	Tennis

Goal: (1). Calculate the entropy for the dataset

- How many labels do you have (more than two)

Calculate the entropy of each value for one attribute

$S = -[\ \ ]$

$S[\text{Sunny}] = -[P_C * \log_2 P_C + P_T * \log_2 P_T]$

$= -\left[\frac{1}{3}\log_2\frac{1}{3} + \frac{2}{3}\log\frac{2}{3}\right]$

INFORMATION GAIN

Gain(S.A) = E(before) - G(after splitting)

Weekend	Weather	Parental Availability	Wealthy	Decision (Class)
H1	Sunny	Yes	Rich	Cinema
H2	Sunny	No	Rich	Tennis
H3	Windy	Yes	Rich	Cinema
H4	Rainy	Yes	Poor	Cinema
H5	Rainy	No	Rich	Home
H6	Rainy	Yes	Poor	Cinema
H7	Windy	No	Poor	Cinema
H8	Windy	No	Rich	Shopping
H9	Windy	Yes	Rich	Cinema
H10	Sunny	No	Rich	Tennis

Goal: (1). Calculate the entropy for the dataset

- How many labels do you have (more than two)

Calculate the entropy of each value for one attribute

E(sunny) = 0.918, E(Windy) = 0.811 E(Rainy) = 0.918
E(S) = 1.571

Gain(S, weather) = E(S) – [p(sunny)*E(sunny) + p(windy)*E(windy) + p(rainy)*E(rainy)]

$$=1.571 - \left[\left(\frac{3}{10}\right) * 0.918 + \left(\frac{4}{10}\right) * 0.811 + \left(\frac{3}{10}\right) * 0.918\right]$$

= 0.7