DMML, 23 Jan 2019

Decision Trees

Items: Attributes $A_1, A_2, \ldots, A_N$,
Category C

Training data $\longrightarrow$ Model

Given $(a_1, a_2, \ldots, a_N)$, predict $c \in C$

Building a model – uniform algorithm

Specific training data $\longmapsto$ Specific model

Decision trees

- Ask questions about attribute values
  - Adaptive – next question depends on previous answers
- Each question prune available possibilities
- Stop when
  - Current set has a uniform class
  - Exhausted set of attributes to query
    - Answer is majority value

Which question to ask?

- Prefer small trees
- Exactly computing smallest tree is NP complete
- Instead, greedy heuristic

Measure: "Impurity" ~ uncertainty

Naive measure: Error Rate Minority %

More sophisticated measures of impurity/uncertainty

Information theory

$\{a, b, c, d, -, z\}$

Send a string of chars in binary

Uniform size encoding = 5 bits/char

Message of N chars → 5N bits

In fact - letters do not occur uniformly

Use variable length encoding - shorter seq for freq letters

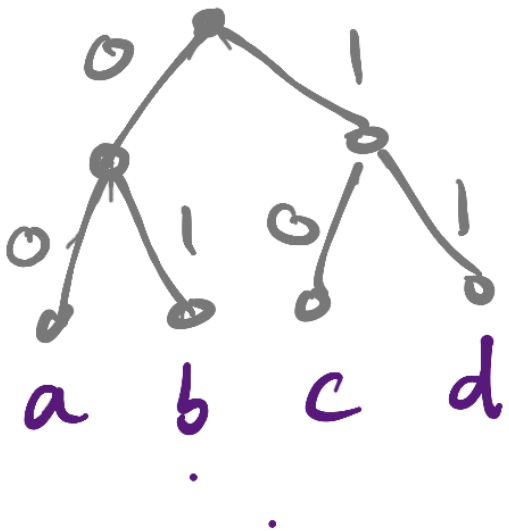
Claude Shannon

$\{a, b, c, d\}$ — uniform 2 bit code

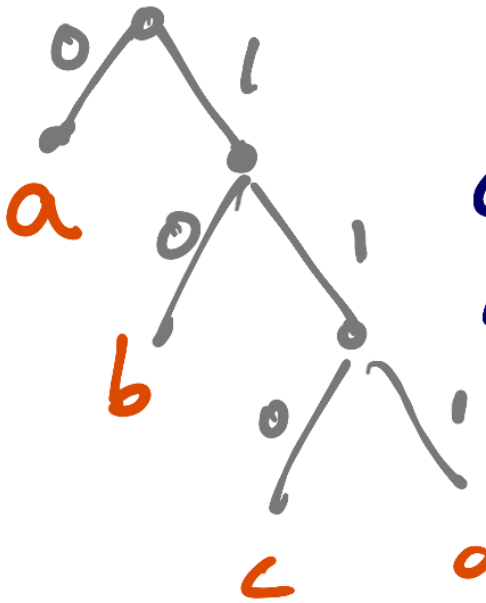
N char → 2N bits

$\frac{1}{2}$ $\frac{1}{4}$ $\frac{1}{8}$ $\frac{1}{8}$

Uniform encoding



Non uniform



a ↦ 0

b ↦ 10

c ↦ 110

d ↦ 111

200 characters $\{a, b, c, d\}$

$\frac{1}{2}$ $\frac{1}{4}$ $\frac{1}{8}$ $\frac{1}{8}$

Uniform = 400 bits

Nonuniform = 100 a's 50 b's. 25 c,d

100 + 100 + 75 + 75

= 350

Reverse analysis ⇒ 1.75 bits/char

Is this optimal?

Distribution of characters reflects uncertainty in their values

Shannon entropy

Letters/outcomes a_1 a_2 ... a_k

Probabilities p_1 p_2 ... p_k

$$\sum p_i = 1$$

$$\text{Entropy} \overset{\text{defn}}{=} -\sum_{i=1}^{k} p_i \log_2 p_i$$

Entropy measures uncertainty

a_1 a_2

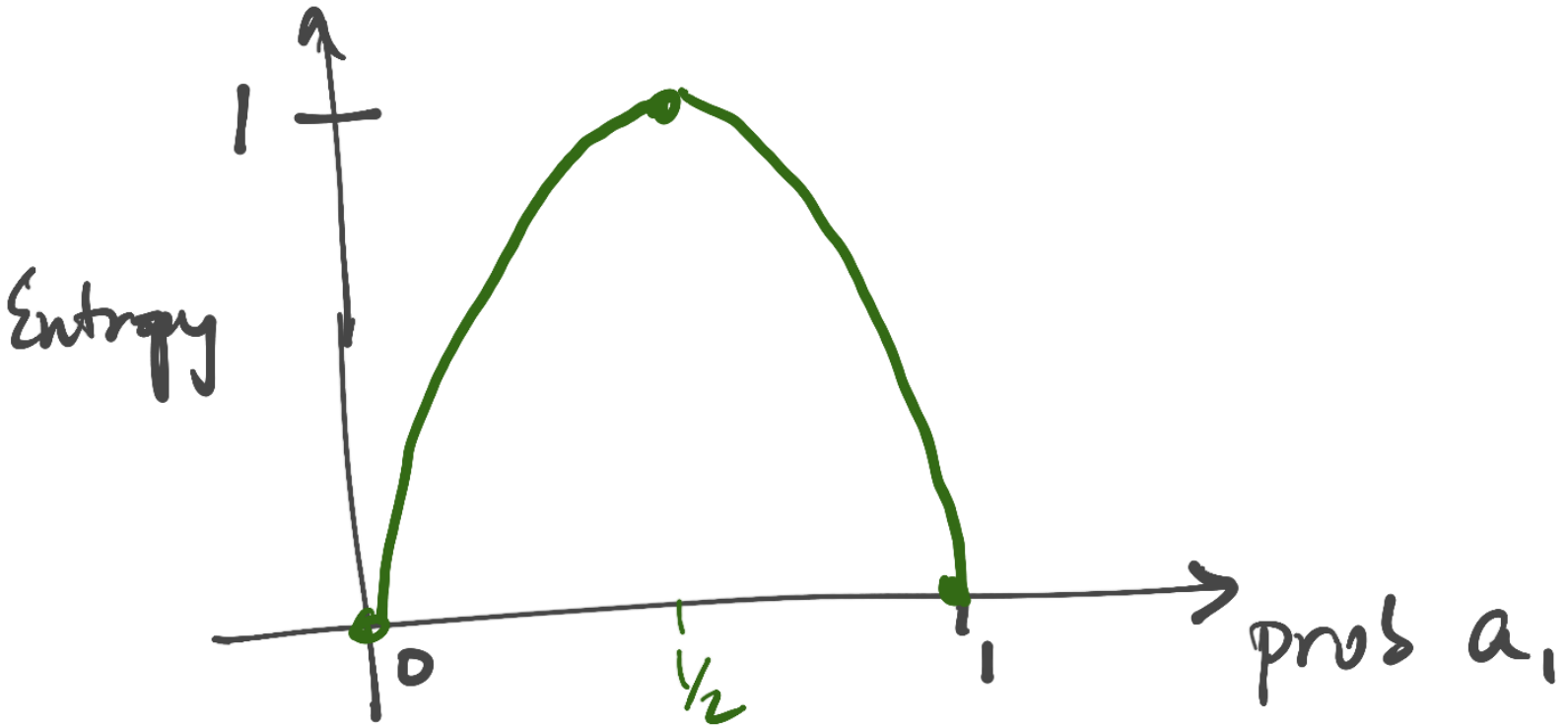
$\frac{1}{2}$ $\frac{1}{2}$ $\quad - \left(\underbrace{\frac{1}{2} \log \frac{1}{2}}_{-\frac{1}{2}} + \underbrace{\frac{1}{2} \log \frac{1}{2}}_{-\frac{1}{2}} \right)$

$= 1$

0 1 $\rightarrow$ $= 0$

1 0

$\left(\underbrace{0 \log 0}_{\text{assume } 0} + \underbrace{1 \log 1}_{0} \right)$



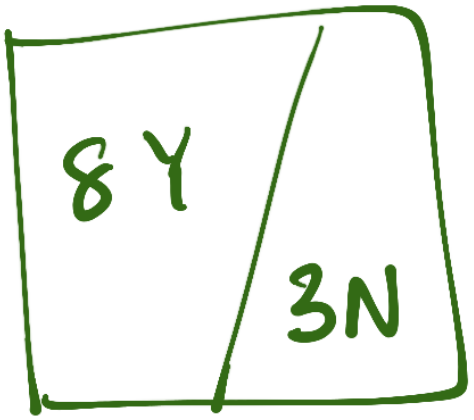
For $a_1, \ldots, a_k$ max entropy is $p_i = \frac{1}{k}$

a	b	c	d
1/2	1/4	1/8	1/8

$$\frac{1}{2}\log\frac{1}{2}^{-1/2} + \frac{1}{4}\log\frac{1}{4}^{-1/2} + 2\cdot\frac{1}{8}\log\frac{1}{8}^{-3/4}$$

$$-(-1.75) = 1.75$$

Entropy for impurity



Error Rate: $\frac{3}{11}$

Entropy: $-\left(\frac{8}{11}\log\frac{8}{11}+\frac{3}{11}\log\frac{3}{11}\right)$

Ross Quinlan – Entropy works better than other notions on many benchmarks

C4.5

Gini Index (Economics)

$$1 - \sum p_i^2$$

$$p_1 = p_2 = \frac{1}{2}$$

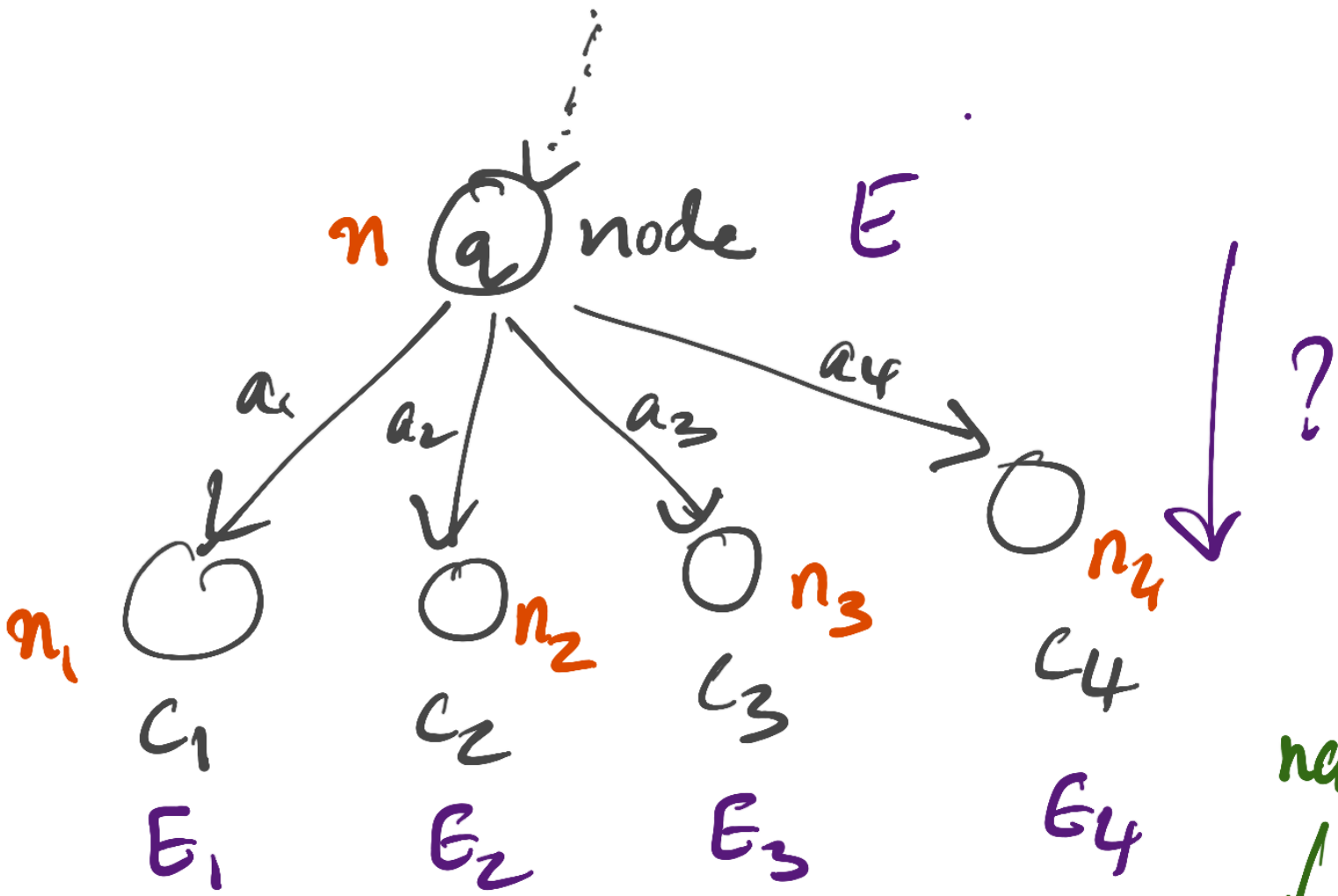
$$G.I = 1 - \frac{1}{2} = \frac{1}{2}$$

$$p_1 = p_2 \cdots = p_k = \frac{1}{k}$$

$$1 - k \cdot \frac{1}{k^2} = \frac{k-1}{k}$$

CART — Classification And Regression Tree implemented in R

Reduction in Impurity = Information Gain



$n_1 + n_2 + n_3 + n_4 = n$

Weighted Avg

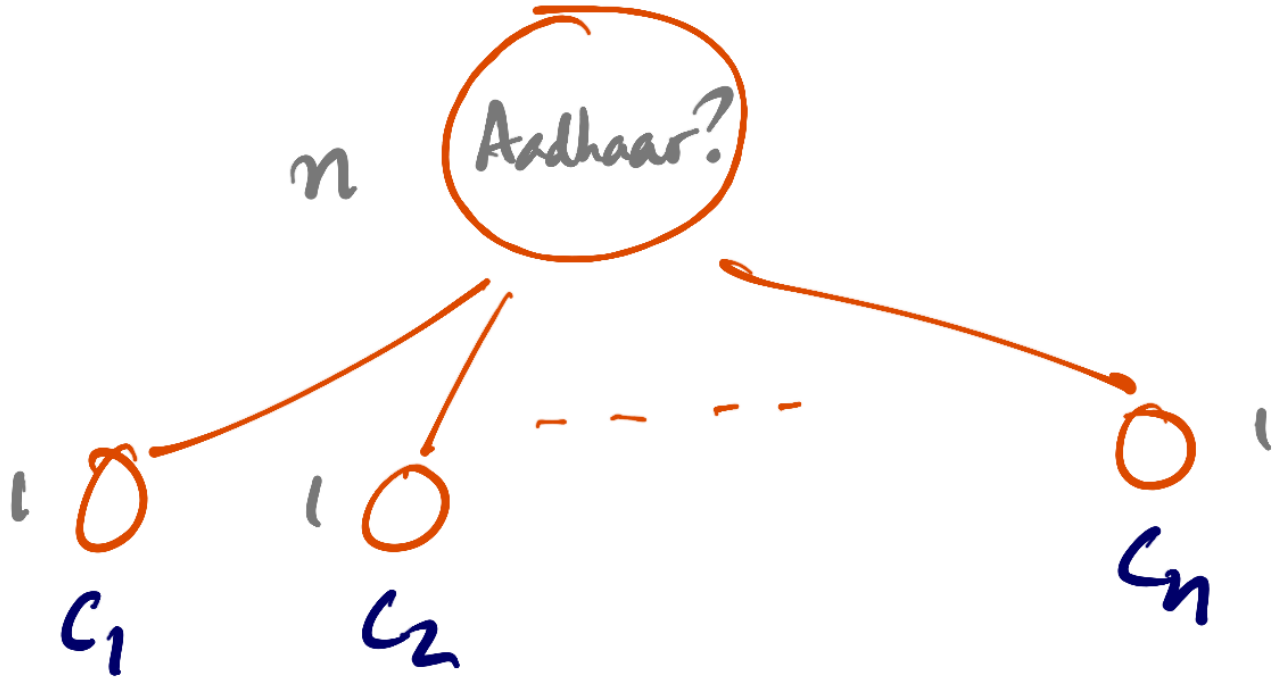
$\sum_i \frac{n_i}{n} \cdot E_i$

new E' after q

Special Case

What if A_i = Aadhaar No ?

Split on Aadhaar



$E_i = 0$ for all c_i $\rightarrow$ wt avg = 0

Measure entropy of an attribute

Aadhaar = k rows $\quad -\sum \frac{1}{k} \log \frac{1}{k}$

Moderate Information Gain (Absolute)

by entropy of attribute

Information Gain Ratio $= \dfrac{\text{Absolute Entropy Gain}}{\text{Entropy of Attribute}}$

Borrowed Entropy from Info Theory

Why does it make sense?

Because it works!