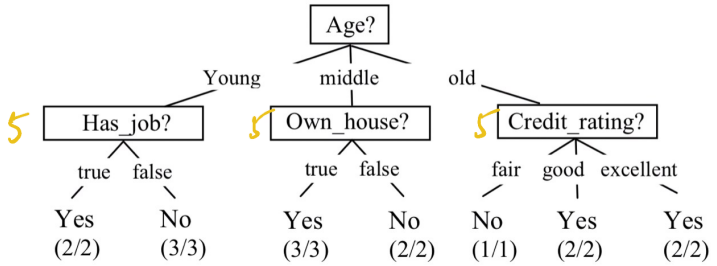
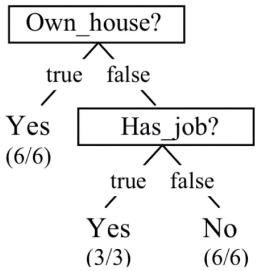
Supervised learning — Classification

Decision Trees

9 Yes
6 No

ID	Age	Has_job	Own_house	Credit_rating	Class
1	young	false	false	fair	**No**
2	young	false	false	good	**No**
3	young	true	false	good	**Yes**
4	young	true	true	fair	**Yes**
5	young	false	false	fair	**No**
6	middle	false	false	fair	**No**
7	middle	false	false	good	**No**
8	middle	true	true	good	**Yes**
9	middle	false	true	excellent	**Yes**
10	middle	false	true	excellent	**Yes**
11	old	false	true	excellent	**Yes**
12	old	false	true	good	**Yes**
13	old	true	false	good	**Yes**
14	old	true	false	excellent	**Yes**
15	old	false	false	fair	**No**





Prefer small trees

- Occam's Razor
- Minimize overfitting

Many definitions of "smallest"

Bad News For any reasonable definition, NP complete

Heuristic ("greedy")

Which attribute to query next?

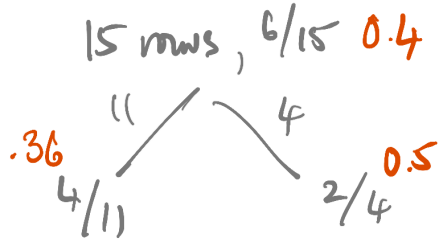
Impurity of data set

Minority # / Total

Reduce impurity

↓

Weighted avg



Better ways to measure impurity

Information theory [Shannon]

Messages over 4 letter alphabet {a, b, c, d}

Encode in binary

Natural choice is 2 bits / letter

a 00
b 01
c 10
d 11

10 11 00 10 11 00

N chars → 2N bits

26 letters a – z 5 bits / letter

Variable length coding

More frequent – shorter code

Characters $c_1, c_2, \ldots c_k$

Frequency / prob $p_1, p_2, \ldots p_k$

$$\sum p_i = 1$$

$$\sum_i -p_i \log_2 p_i = \text{Shannon entropy}$$

$\sum -p_i \log_2 p_i$ $\{a, b, c, d\}$

$\frac{1}{4}\ \frac{1}{4}\ \frac{1}{4}\ \frac{1}{4}$

$$\sum_{i=1}^{4} -\frac{1}{4} \log_2 \frac{1}{4}$$

$$\sum_{i=1}^{4} -\frac{1}{4} \cdot -2$$

$$= \sum_{i=1}^{4} \frac{1}{2} = 2$$

$\log_2 \frac{1}{4}$

$= \log_2 (4)^{-1}$

$= -1 \cdot \log_2 4$

$= -2$

a b c d

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

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

$$\frac{1}{2} + \frac{1}{2} + \frac{3}{4}$$

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

Entropy is maximized when each $p_i = \frac{1}{k}$

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

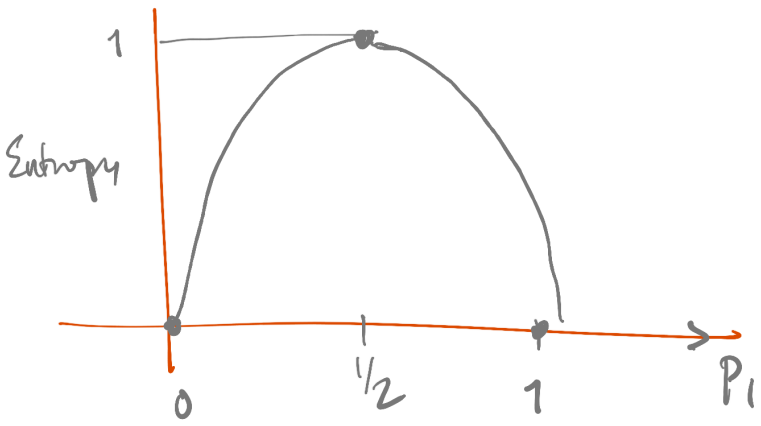
$$\sum_{i=1}^{2} -\frac{1}{2} \log \frac{1}{2}$$

$= 1$

$p_1 = 0$ $p_2 = 1$

$0 \log 0$ $1 \log 1$

$= 0$ $= 0$



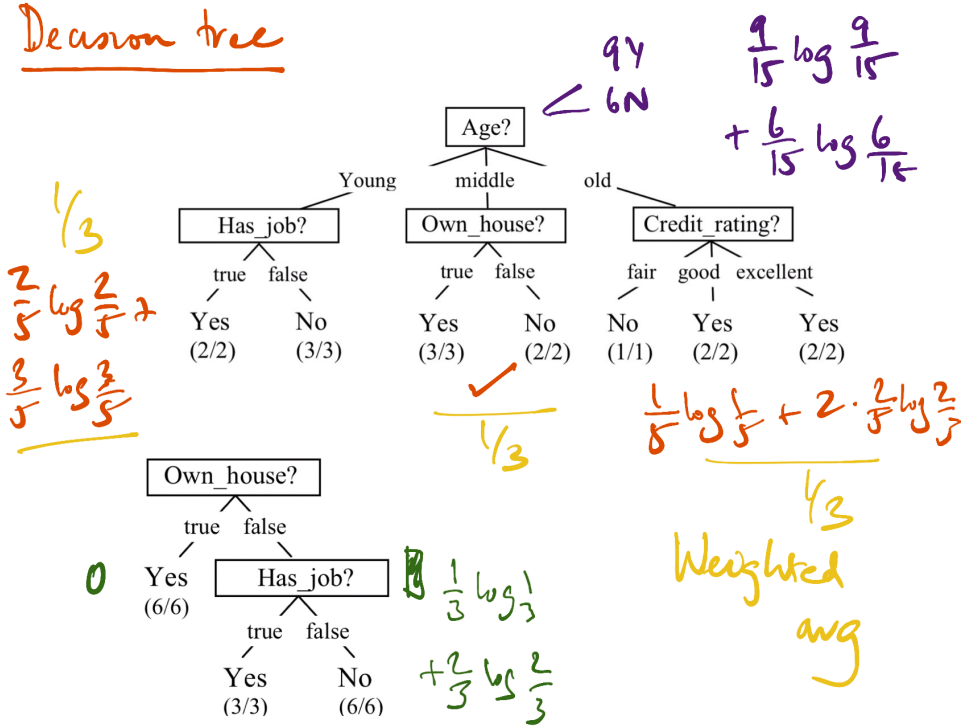
$$\sum_{i=1}^{k} \frac{1}{k} \log \frac{1}{k} \qquad \log_2 k \;,\; k>2$$

Choose attribute that reduces entropy the most — Information Gain

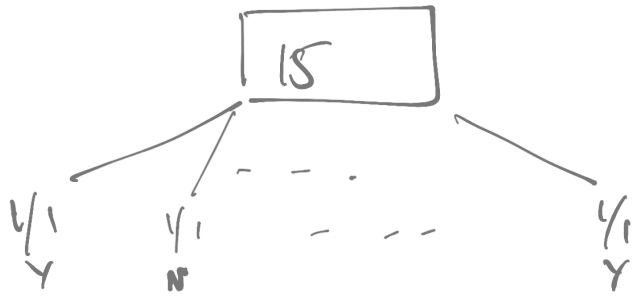
Ross Quinlan

Now, this description does violence to statistical notions of sampling and confidence limits, so the reasoning should be taken with a large grain of salt. Like many heuristics with questionable underpinnings, however, the estimates that it produces seem frequently to yield acceptable results.

Decision tree



What if Aadhaar is an attribute?



Entropy of Aadhaar $= \sum_{i=1}^{k} \frac{1}{k} \log \frac{1}{k}$

Normalize information gain by entropy of attribute

Information Gain = Reduction in Entropy

Entropy of Attribute

Algorithm

BuildTree $(A_1, \ldots, A_k)$

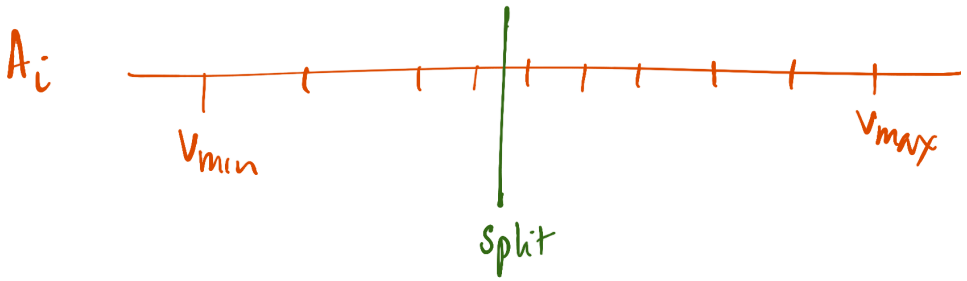
Choose A_j that maximizes Info Gain Ratio

For each child created

BuildTree $(A_1, \ldots, A_{j-1}, A_{j+1}, \ldots A_k)$

Stop if < homogeneous / no attributes left — use majority

What about numeric attributes



$A_i < split$?

Testing a classifier

Set aside some training data for validation

Randomly split as → Training / Test

Cross validation

10 x remove 10%

What is correctness?

Percentage of correct answers

Accuracy

Most classifiers are looking at skewed samples

Vast majority → "Normal"

Small minority → "Abnormal"

Refined calculation

Yes is the Abnormal (minority) case

Actual data \ Classifier says	Y	N
Y	0	5
N	0	95

Usually Precision ↑ Recall ↓

Selection of candidates

Screening Exam — High Recall

+

Interview — High Precision

Recall \ Precision		
	25	25
	50	900

$\frac{25}{50}$

$\frac{50}{75}$ mistake

$\frac{25}{75}$