Lecture 4: 15 April, 2021

Madhavan Mukund

https://www.cmi.ac.in/~madhavan

Data Mining and Machine Learning
April–July 2021

Example: Loan application data set

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

9Y
6N
15

Decision tree algorithm

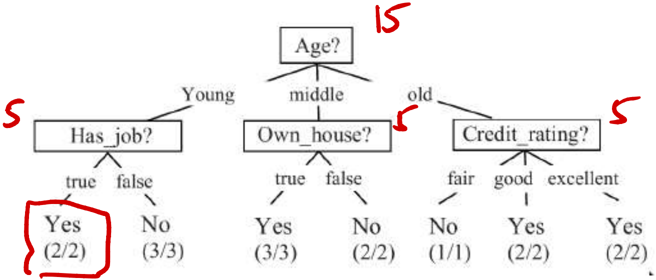
A : current set of attributes

Pick $a \in A$, create children corresponding to resulting partition with attributes $A \setminus \{a\}$

Stopping criterion:

- Current node has uniform class label
- A is empty — no more attributes to query

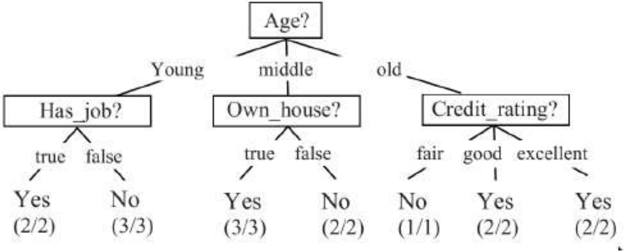
If a leaf node is not uniform, use majority class as prediction



- Non-uniform leaf node — identical combination of attributes, but different classes
- Attributes do not capture all criteria used for classification

Decision trees

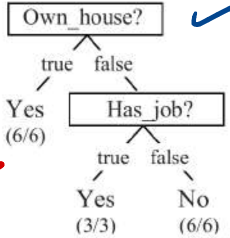
- Tree is not unique
- Which tree is better?
- Prefer small trees
  - Explainability
  - Generalize better (see later)



Unfortunately

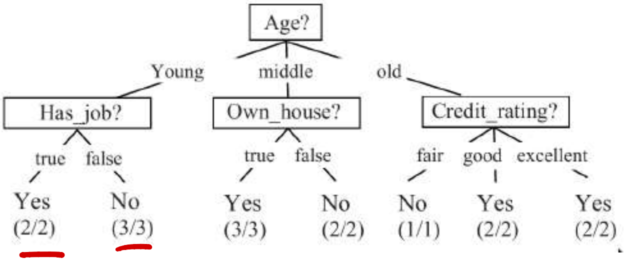
- Finding smallest tree is NP-complete — for any definition of "smallest"
- Instead, greedy heuristic

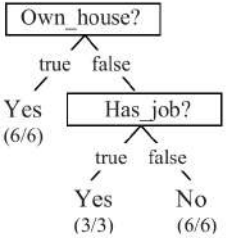
Locally best choice



Greedy heuristic

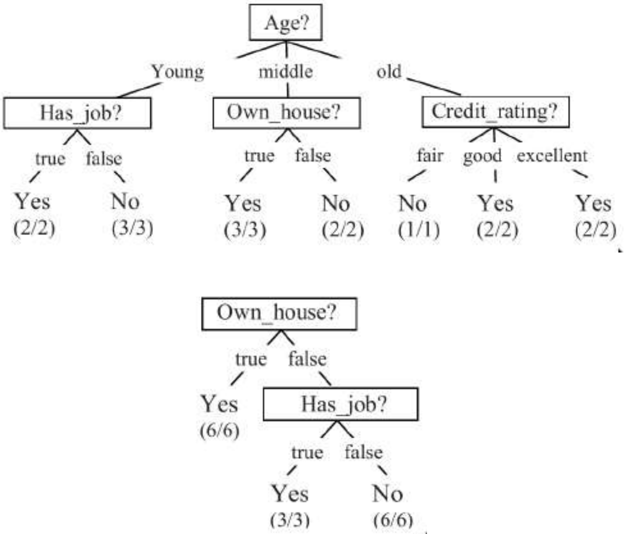
- Goal: partition with uniform category — pure leaf





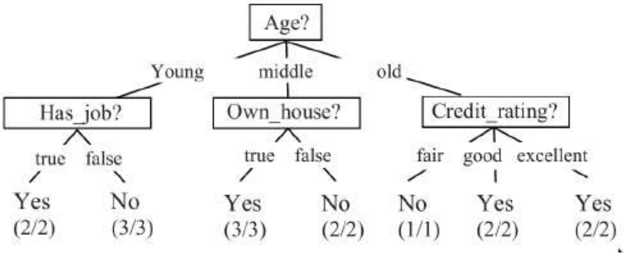
Greedy heuristic

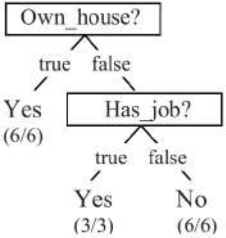
- Goal: partition with uniform category — pure leaf
- Impure node — best prediction is majority value



Greedy heuristic

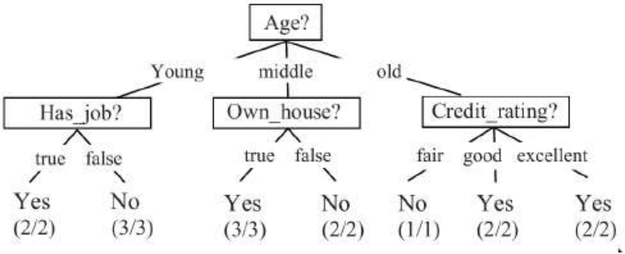
- Goal: partition with uniform category — pure leaf
- Impure node — best prediction is majority value
- Minority ratio is impurity

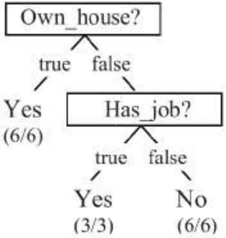




Greedy heuristic

- Goal: partition with uniform category — pure leaf
- Impure node — best prediction is majority value
- Minority ratio is impurity
- Heuristic: reduce impurity as much as possible



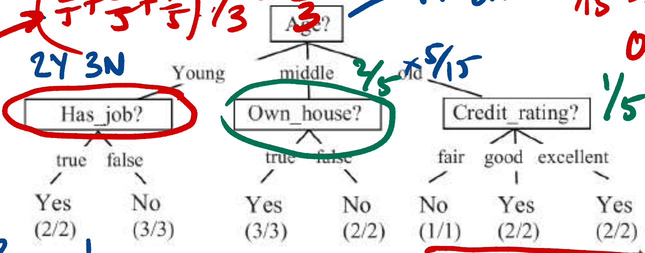


Greedy heuristic

- Goal: partition with uniform category — pure leaf
- Impure node — best prediction is majority value
- Minority ratio is impurity
- Heuristic: reduce impurity as much as possible
- For each attribute, compute weighted average impurity of children

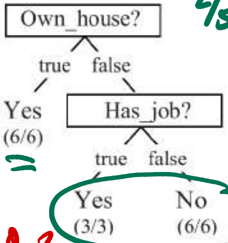
$\left(\frac{2}{5}+\frac{2}{5}+\frac{1}{5}\right)\frac{1}{3} = \frac{1}{3}$ 0.33 9Y 6N $6/15 = \frac{2}{5}$ 0.4

$\frac{5}{15} \times \frac{2}{5}$ 2Y 3N $2/5 \times 5/15$ $1/5 \times 5/15$



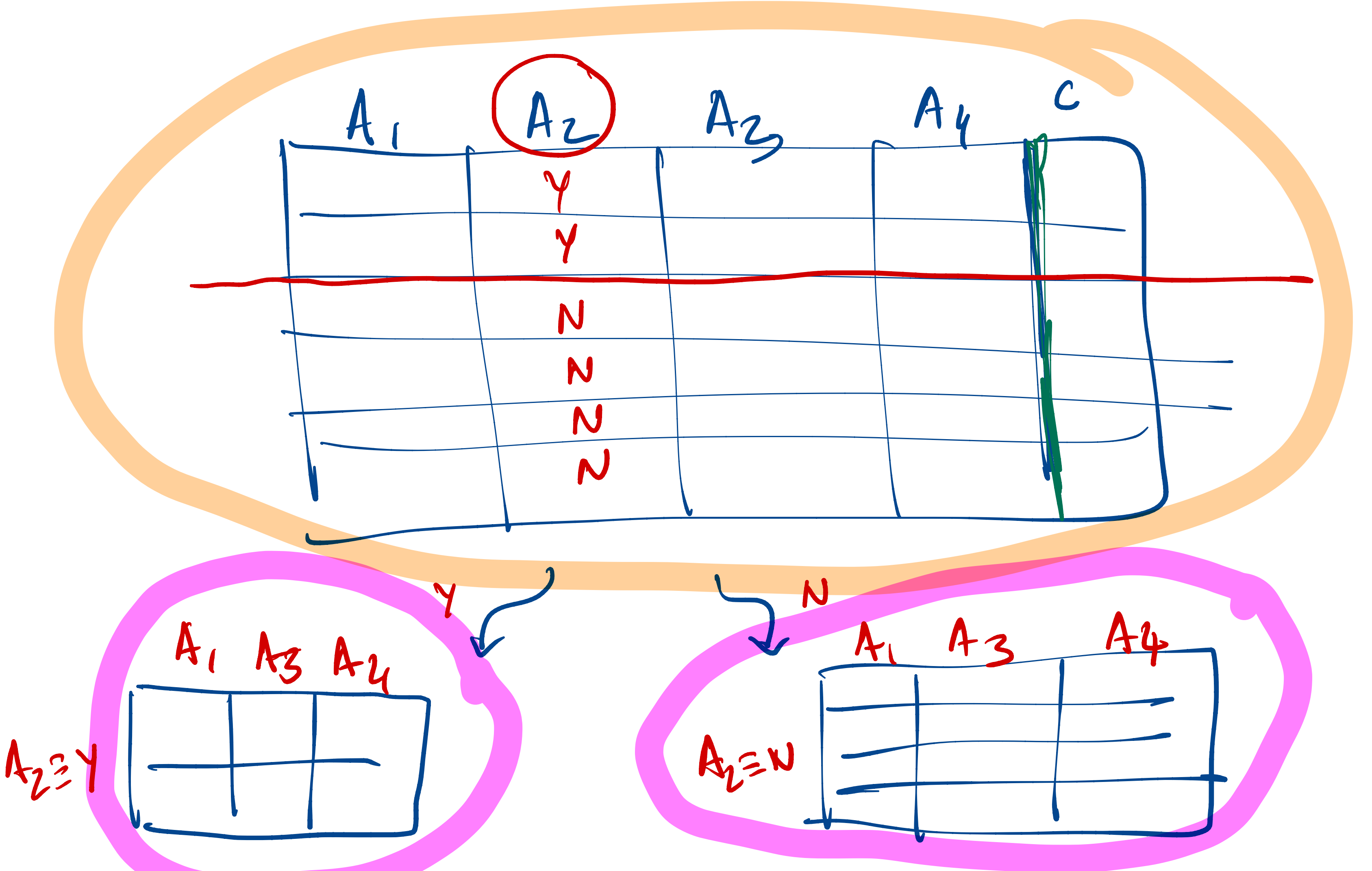
$\frac{2}{5} \to \frac{1}{3}$ 0.07

$2/5$



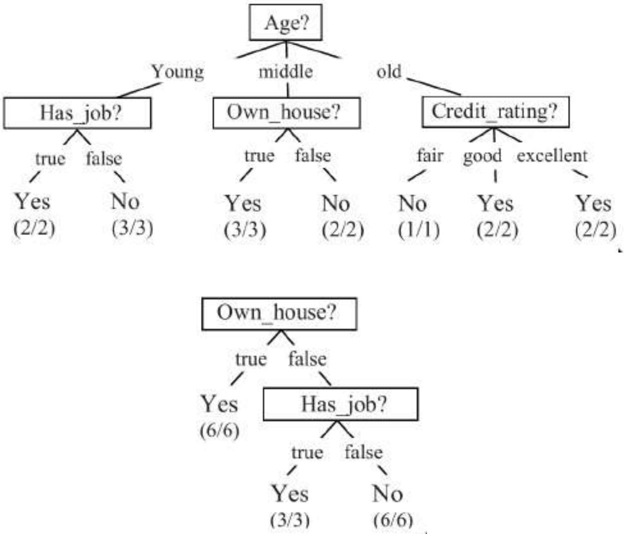
$\frac{6}{15} \cdot \frac{0}{6}$ $\frac{9}{15} \cdot \frac{3}{9}$

$\frac{6}{15} \cdot \frac{0}{6} + \frac{9}{15} \cdot \frac{3}{9} = \frac{3}{15} = 1/5$ $\frac{2}{5} \to \frac{1}{5}$ 0.2


A1
A2
A3
A4
C
Y
Y
N
N
N
N
Y
N
A1 A3 A4
A2=Y
A1 A3 A4
A2=N

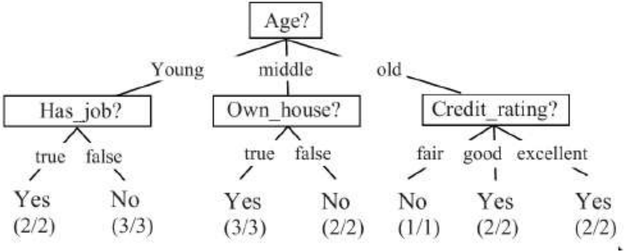
Greedy heuristic

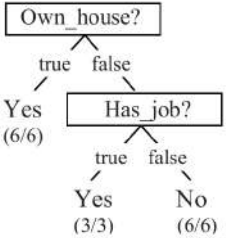
- Goal: partition with uniform category — pure leaf
- Impure node — best prediction is majority value
- Minority ratio is impurity
- Heuristic: reduce impurity as much as possible
- For each attribute, compute weighted average impurity of children
- Choose the minimum



Greedy heuristic

- Goal: partition with uniform category — pure leaf
- Impure node — best prediction is majority value
- Minority ratio is impurity
- Heuristic: reduce impurity as much as possible
- For each attribute, compute weighted average impurity of children
- Choose the minimum
- Will see better heuristics





Algorithm

Current impurity I, Set of attributes $A_1, A_2, \ldots, A_k$

For each A_j

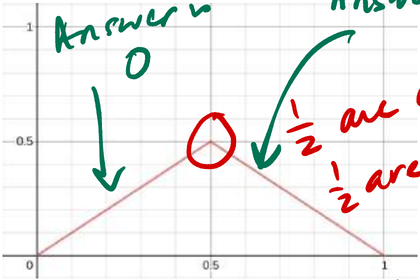
Compute weighted avg impurity I_j if we split on A_j [ask question A_j?]

Choose A_j for which $I - I_j$ is maximum

Split table into subtable with $A_1, A_2, \ldots, A_{j-1}, A_{j+1}, \ldots A_k$

A better impurity function

- Misclassification rate is linear



Answer is 0

Answer is 1

$\frac{1}{2}$ are $c=1$

$\frac{1}{2}$ are $c=0$

pure

$c = 1$

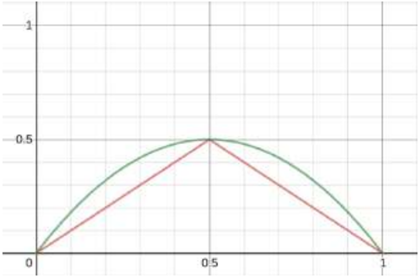
pure

fraction with $c=1$

$c \in \{0,1\}$

A better impurity function

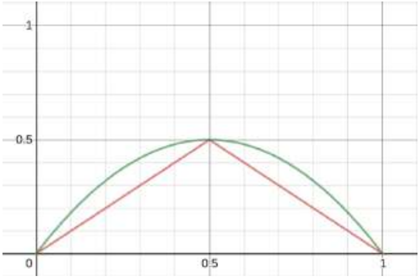
- Misclassification rate is linear
- Impurity measure that increases more sharply performs better, empirically



$c = 1$

A better impurity function

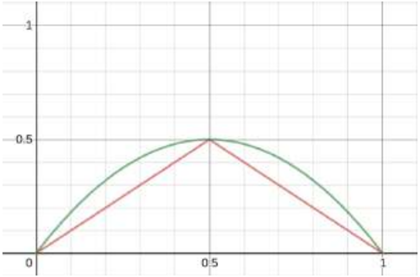
- Misclassification rate is linear
- Impurity measure that increases more sharply performs better, empirically
- Entropy — [Quinlan]



$c = 1$

A better impurity function

- Misclassification rate is linear
- Impurity measure that increases more sharply performs better, empirically
- Entropy — [Quinlan]
- Gini index — [Breiman]



$c = 1$

Entropy

- Information theoretic measure of randomness
- Minimum number of bits to transmit a message — [Shannon]

Entropy

- Information theoretic measure of randomness
- Minimum number of bits to transmit a message — [Shannon]
- n data items
  - n_0 with $c = 0$, $p_0 = n_0/n$
  - n_1 with $c = 1$, $p_1 = n_1/n$

$n_0 + n_1 = n$

Entropy

- Information theoretic measure of randomness
- Minimum number of bits to transmit a message — [Shannon]
- n data items
  - n_0 with $c = 0$, $p_0 = n_0/n$
  - n_1 with $c = 1$, $p_1 = n_1/n$
- Entropy

$$E = -\underbrace{(p_0 \log_2 p_0 + p_1 \log_2 p_1)}_{\text{negative}}$$

$\log p_0 < 0$

$\downarrow$

$0 - 1$

- Information theoretic measure of randomness
- Minimum number of bits to transmit a message — [Shannon]
- n data items
  - n_0 with $c = 0$, $p_0 = n_0/n$
  - n_1 with $c = 1$, $p_1 = n_1/n$
- Entropy
  $E = -(p_0 \log_2 p_0 + p_1 \log_2 p_1)$
- Minimum when $p_0 = 1, p_1 = 0$ or vice versa — note, declare $0 \log_2 0$ to be 0
- Maximum when $p_0 = p_1 = 0.5$

$- (\boxed{0 \log 0} + \underline{1 \log 1})$ 0

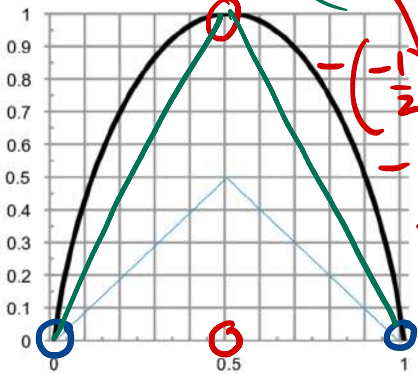
$p_0 = p_1 = 1/2$

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

$-\left(-\frac{1}{2} + -\frac{1}{2}\right)$

$-(-1)$

$= +1$



$p_0 = 0$
$p_1 = 1$

$c = 1$

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

$\log_2 \frac{1}{2} = -1$

Gini Index

- Measure of unequal distribution of wealth
- Economics — [Corrado Gini]
- As before, n data items
  - n_0 with $c = 0$, $p_0 = n_0/n$
  - n_1 with $c = 1$, $p_1 = n_1/n$

Gini Index

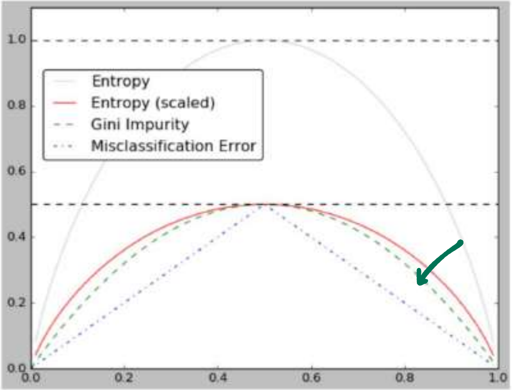
- Measure of unequal distribution of wealth
- Economics — [Corrado Gini]
- As before, n data items
  - n_0 with $c = 0$, $p_0 = n_0/n$
  - n_1 with $c = 1$, $p_1 = n_1/n$
- Gini Index $G = 1 - (p_0^2 + p_1^2)$

Gini Index

- Measure of unequal distribution of wealth
- Economics — [Corrado Gini]
- As before, n data items
  - n_0 with $c = 0$, $p_0 = n_0/n$
  - n_1 with $c = 1$, $p_1 = n_1/n$
- Gini Index $G = 1 - (p_0^2 + p_1^2)$
- $G = 0$ when $p_0 = 0$, $p_1 = 0$ or v.v.
  $G = 0.5$ when $p_0 = p_1 = 0.5$

$p_0 = p_1 = \frac{1}{2}$ $\quad 1 - \left(\frac{1}{4} + \frac{1}{4}\right) = \frac{1}{2}$

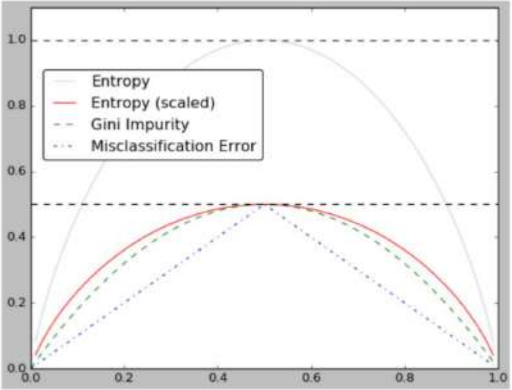
$p_0 = 0, p_1 = 1$ $\quad 1 - (0 + 1) = 0$



$c = 1$

Gini Index

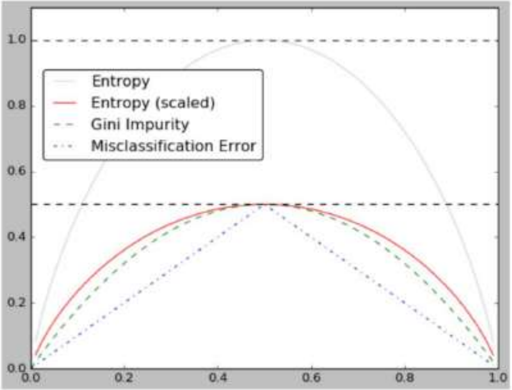
- Measure of unequal distribution of wealth
- Economics — [Corrado Gini]
- As before, n data items
  - n_0 with $c = 0$, $p_0 = n_0/n$
  - n_1 with $c = 1$, $p_1 = n_1/n$
- Gini Index $G = 1 - (p_0^2 + p_1^2)$
- $G = 0$ when $p_0 = 0$, $p_1 = 0$ or v.v.
  $G = 0.5$ when $p_0 = p_1 = 0.5$
- Entropy curve is slightly steeper, but Gini index is easier to compute



$c = 1$

Gini Index

- Measure of unequal distribution of wealth
- Economics — [Corrado Gini]
- As before, n data items
  - n_0 with $c = 0$, $p_0 = n_0/n$
  - n_1 with $c = 1$, $p_1 = n_1/n$
- Gini Index $G = 1 - (p_0^2 + p_1^2)$
- $G = 0$ when $p_0 = 0$, $p_1 = 0$ or v.v.
  $G = 0.5$ when $p_0 = p_1 = 0.5$
- Entropy curve is slightly steeper, but Gini index is easier to compute
- Decision tree libraries usually use Gini index



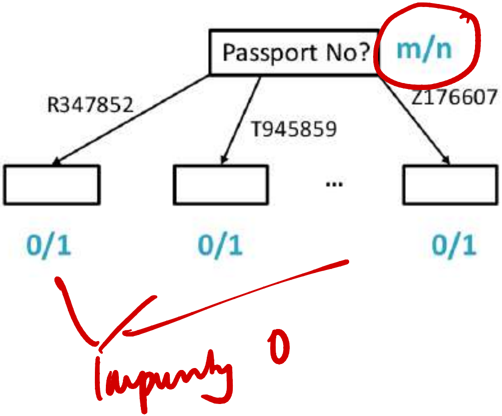
$c = 1$

Information gain

- Greedy strategy: choose attribute to maximize reduction in impurity — maximize information gain

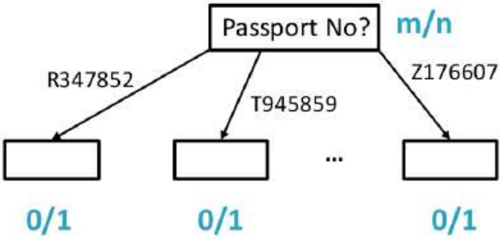
Information gain

- Greedy strategy: choose attribute to maximize reduction in impurity — maximize information gain
- Suppose an attribute is a unique identifier
  - Roll number, passport number, Aadhaar ...



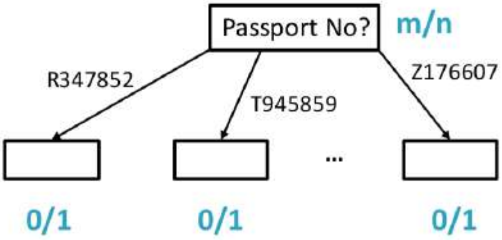
Information gain

- Greedy strategy: choose attribute to maximize reduction in impurity — maximize information gain
- Suppose an attribute is a unique identifier
  - Roll number, passport number, Aadhaar ...
- Querying this attribute produces partitions of size 1
  - Each partition guaranteed to be pure
  - New impurity is zero



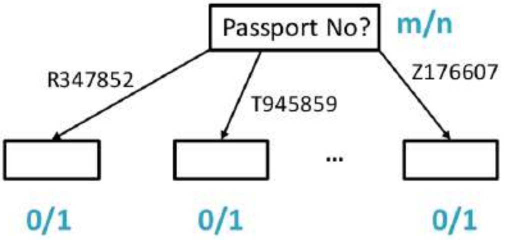
Information gain

- Greedy strategy: choose attribute to maximize reduction in impurity — maximize information gain
- Suppose an attribute is a unique identifier
  - Roll number, passport number, Aadhaar . . .
- Querying this attribute produces partitions of size 1
  - Each partition guaranteed to be pure
  - New impurity is zero
- Maximum possible impurity reduction, but useless!



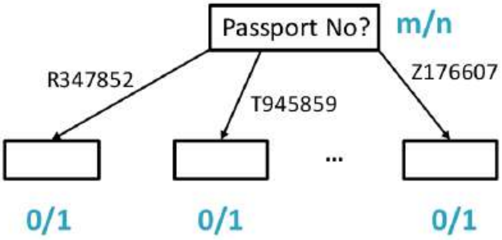
Information gain

- Tree building algorithm blindly picks attribute that maximizes information gain



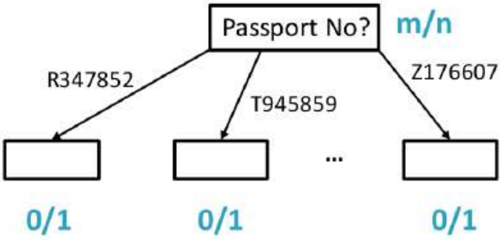
Information gain

- Tree building algorithm blindly picks attribute that maximizes information gain
- Need a correction to penalize attributes with highly scattered attributes



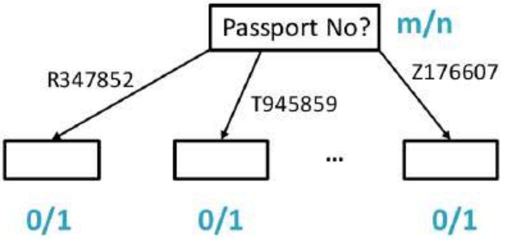
Information gain

- Tree building algorithm blindly picks attribute that maximizes information gain
- Need a correction to penalize attributes with highly scattered attributes
- Extend the notion of impurity to attributes



Attribute Impurity

- Attribute takes values $\{v_1, v_2, \ldots, v_k\}$
- v_i appears n_i times across n rows
- $p_i = n_i/n$



Attribute Impurity

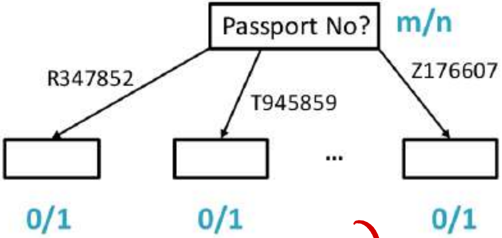
- Attribute takes values $\{v_1, v_2, \ldots, v_k\}$
- v_i appears n_i times across n rows
- $p_i = n_i / n$
- Entropy across k values

$$-\sum_{i=1}^{k} p_i \log_2 p_i$$

$P_0 + P_1 = 1$

$P_1 + P_2 + \ldots + P_k = 1$

$-(p_1 \log p_1 + p_2 \log p_2 - + p_k \log p_k)$



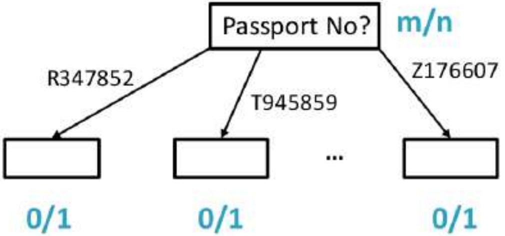
Attribute Impurity

- Attribute takes values $\{v_1, v_2, \ldots, v_k\}$
- v_i appears n_i times across n rows
- $p_i = n_i / n$
- Entropy across k values

  $$-\sum_{i=1}^{k} p_i \log_2 p_i$$

- Gini index across k values

  $$1 - \sum_{i=1}^{k} p_i^2$$



$1 - (p_0^2 + p_1^2)$

$1 - (p_1^2 + p_2^2 + \ldots p_k^2)$

Attribute Impurity

- Extreme case, each $p_i = 1/n$

Attribute Impurity

- Extreme case, each $p_i = 1/n$
- Entropy

$$-\sum_{i=1}^{n} \underline{\frac{1}{n}\log_2} \frac{1}{n} = -\cancel{n \cdot \frac{1}{n}}(-\log_2 n) = \log_2 n$$

Attribute Impurity

- Extreme case, each $p_i = 1/n$
- Entropy

$$-\sum_{i=1}^{n} \frac{1}{n} \log_2 \frac{1}{n} = -n \cdot \frac{1}{n}(-\log_2 n) = \log_2 n$$

- Gini index

$$1 - \sum_{i=1}^{n} \left(\frac{1}{n}\right)^2 = 1 - \frac{n}{n^2} = \frac{n-1}{n}$$

Attribute Impurity

- Extreme case, each $p_i = 1/n$
- Entropy

$$-\sum_{i=1}^{n} \frac{1}{n} \log_2 \frac{1}{n} = -n \cdot \frac{1}{n}(-\log_2 n) = \log_2 n$$

- Gini index

$$1 - \sum_{i=1}^{n} \left(\frac{1}{n}\right)^2 = 1 - \frac{n}{n^2} = \frac{n-1}{n}$$

- Both increase as n increases

n increases

$v_1 \; v_2 \; \ldots \; v_n$

Attribute Impurity

- Extreme case, each $p_i = 1/n$
- Entropy

$$-\sum_{i=1}^{n} \frac{1}{n} \log_2 \frac{1}{n} = -n \cdot \frac{1}{n}(-\log_2 n) = \log_2 n$$

- Gini index

$$1 - \sum_{i=1}^{n} \left(\frac{1}{n}\right)^2 = 1 - \frac{n}{n^2} = \frac{n-1}{n}$$

- Both increase as n increases

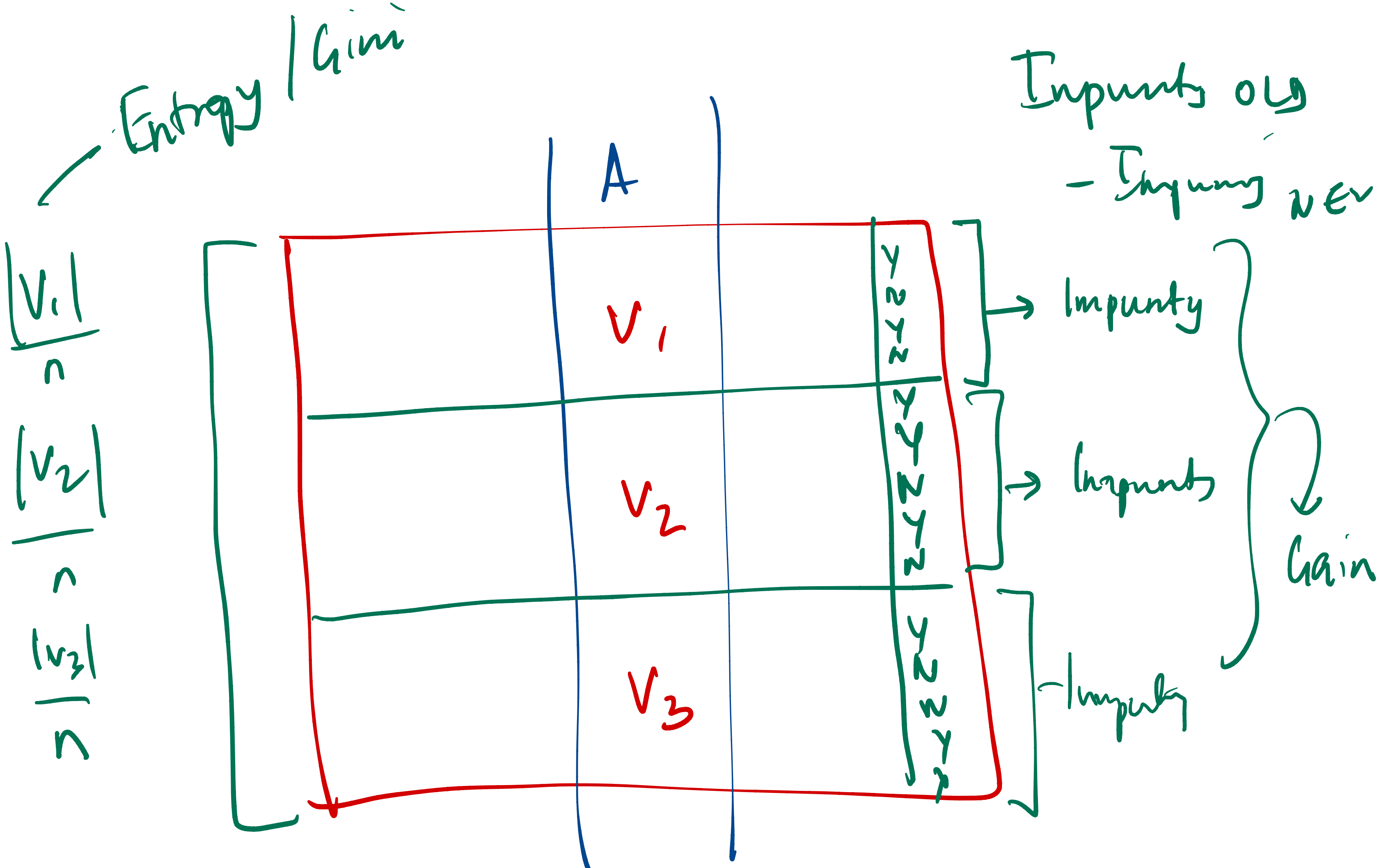
Penalizing scattered attributes

- Divide information gain by attribute impurity
- Information gain ratio(A)

$$\frac{\text{Information-Gain(A)}}{\text{Impurity}(A)}$$

Random

- Scattered attributes have high denominator, counteracting high numerator



Entropy / Gini
Impurity OLD
− Impurity NEW
A
V_1
V_2
V_3
$\frac{|V_1|}{n}$
$\frac{|V_2|}{n}$
$\frac{|V_3|}{n}$
Y N Y N
Y Y N Y N
Y N N Y Y
Impurity
Impurity
Impurity
Gain