香港城市大學
City University of Hong Kong
專業 創新 胸懷全球
Professional · Creative
For The World

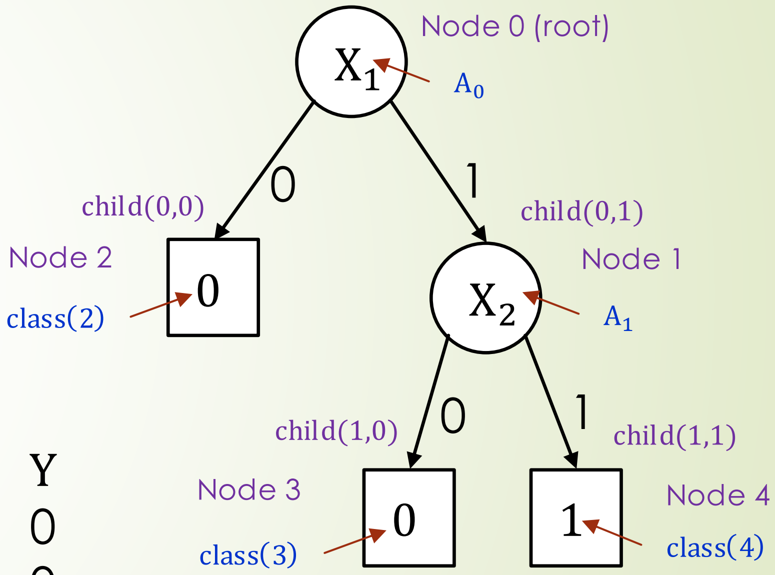
Classification:
Decision Tree Induction

C5483 Data Warehousing and Data Mining

What is a decision tree?

- Internal nodes t (circle)
  - Label A_t (splitting criterion)
  - For each $\mathrm{A}_t = j$ (outcome), an edge to $\texttt{child}(t,j)$ (child node)
- Leave nodes (square)
  - label $\texttt{class}(t)$ (decision)

	X_1	X_2	Y
1.	0	0	0
2.	0	1	0
3.	1	0	0
4.	1	1	1
	x_1	x_2	y



How to classify?
Trace from root to leaves

Input: feature vector $\boldsymbol{x}$

Output: predicted class $\hat{y}$

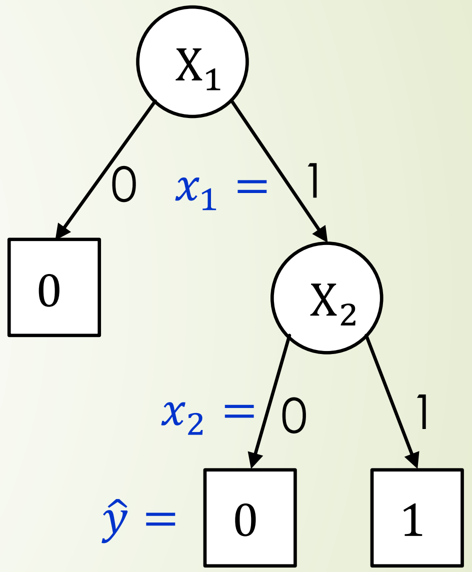
$t \leftarrow$ root

while t is not a leaf

    $t \leftarrow$ child(t, j) where $\mathrm{A}_t = j$ for $\boldsymbol{x}$

$\hat{y} \leftarrow$ class(t)

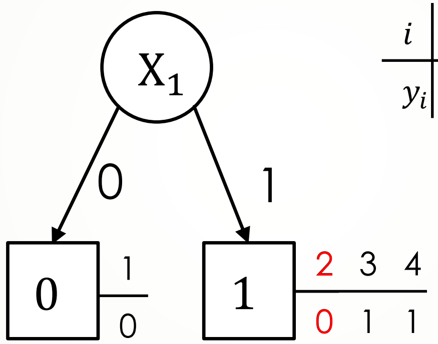
$\boldsymbol{x} = (x_1, x_2) = (1,0)$



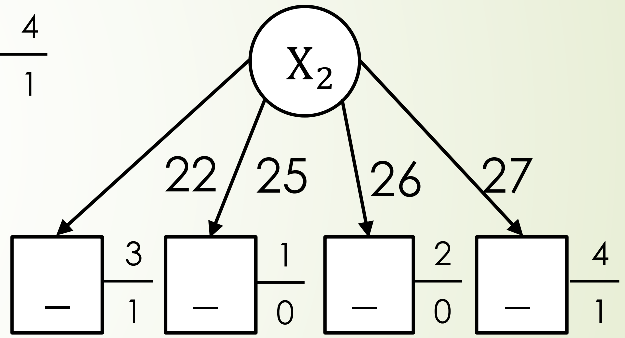
How to build a decision stump?

- A decision stump is a decision tree with depth ≤1.

	X_1	X_2	Y
1.	0	25	0
2.	1	26	0
3.	1	22	1
4.	1	27	1



i	1	2	3	4
y_i	0	0	1	1



error rate = ______

error rate = ______

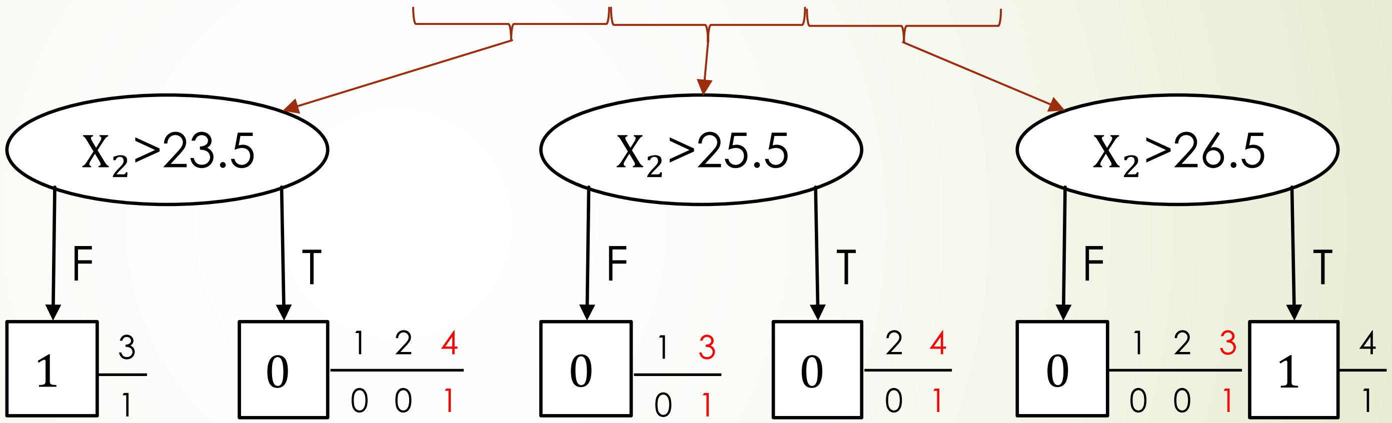
- Choose a splitting attribute.
- Use majority voting to determine $\text{class}(t)$.
- Which decision stump is better? <u>Left/right</u> because of ______________________________.

Binary splits for numeric attributes

Observed values of X_2 in ascending order:

$22 < 25 < 26 < 27$

	X_1	X_2	Y
1.	0	25	0
2.	1	26	0
3.	1	22	1
4.	1	27	1



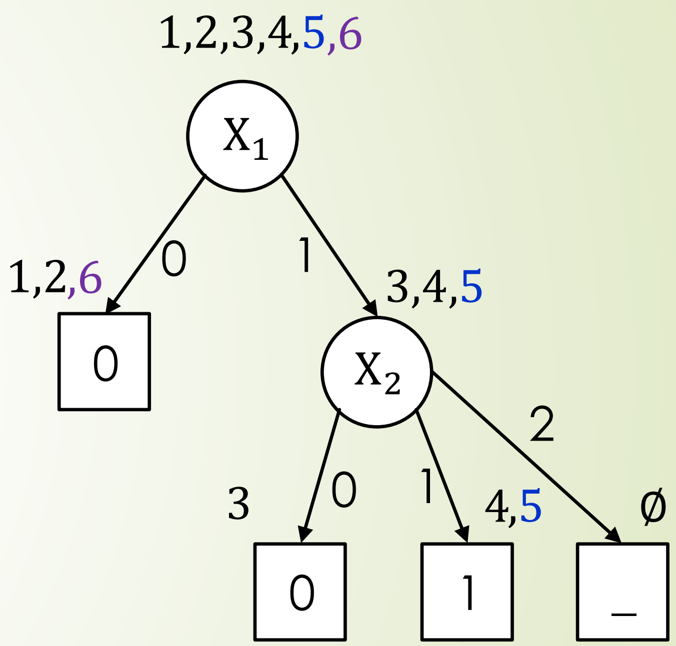
- **c**____________ **m**___-points as **s**_______________ points.
- Which is/are the best split(s)? _______________
- How to build a tree instead of a stump? R__________ly split (d_______ and c_____).

How to build a decision tree?

Greedy algorithm (See Han11 Fig 8.3 for the full version)

```
Input: training data D
Output: root node of the decision tree
function Split(D)
    create node t
    A_t ← a good criterion A to split D
    if A_t is not null
        for each outcome j of A_t
            D_j ← {(x, y) ∈ D | x satisfies A_t = j}
            child(t, j) ← _____ if D_j ≠ ∅
                          else _____
    else  # further splitting is not useful when _____
        class(t) ← _____
    return t
```

	X_1	X_2	Y
1.	0	0	0
2.	0	1	0
3.	1	0	0
4.	1	1	1
5.	1	1	0
6.	0	2	0



(a) $\arg\max_k |\{(\boldsymbol{x}, y) \in D \mid y = k\}|$

(b) new node t' with $\text{class}(t') \leftarrow \arg\max_k |\{(\boldsymbol{x}, y) \in D \mid y = k\}|$

(c) $\text{Split}(D_j)$

How to find good splitting attribute?

- Given the data D to split, choose the splitting attribute A that minimizes e________________ of decision stump by A.
- What is the precise formula?

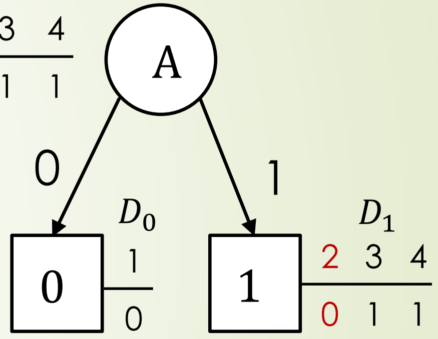
$$\mathrm{Misclass}_{\mathrm{A}}(D) \coloneqq \sum_j \frac{|D_j|}{|D|}\left(1 - \max_k p_{k|j}\right)$$

$$= 1 - \underbrace{\sum_j \frac{|D_j|}{|D|} \max_k p_{k|j}}_{\text{accuracy}}$$

- D_j: set $\{(\boldsymbol{x}, y) \in D | \mathrm{A} = j \text{ for } \boldsymbol{x}\}$ of tuples in D satisfying $\mathrm{A} = j$.
- $p_{k|j}$: fraction $\frac{|\{(\boldsymbol{x},y) \in D_j | y = k\}|}{|D_j|}$ of tuples in D_j belonging to class k.

D

i	1	2	3	4
y_i	0	0	1	1



$\frac{|D_0|}{|D|} = \frac{1}{4}$ $\qquad$ $\frac{|D_1|}{|D|} =$ ____

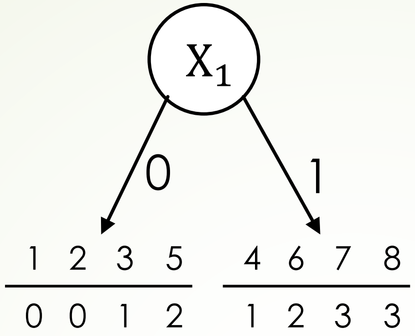
$p_{0|0} = 1$ $\qquad$ $p_{0|1} = \frac{1}{3}$

$p_{1|0} = 0$ $\qquad$ $p_{1|1} =$ ____

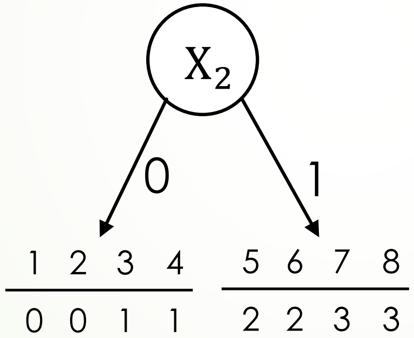
$\max_k p_{k|0}$ $\qquad$ $\max_k p_{k|1}$

$\mathrm{Misclass}_A(D) =$ ____________

	X_1	X_2	X_3	Y
1.	0	0	0	0
2.	0	0	0	0
3.	0	0	1	1
4.	1	0	1	1
5.	0	1	0	2
6.	1	1	0	2
7.	1	1	1	3
8.	1	1	1	3



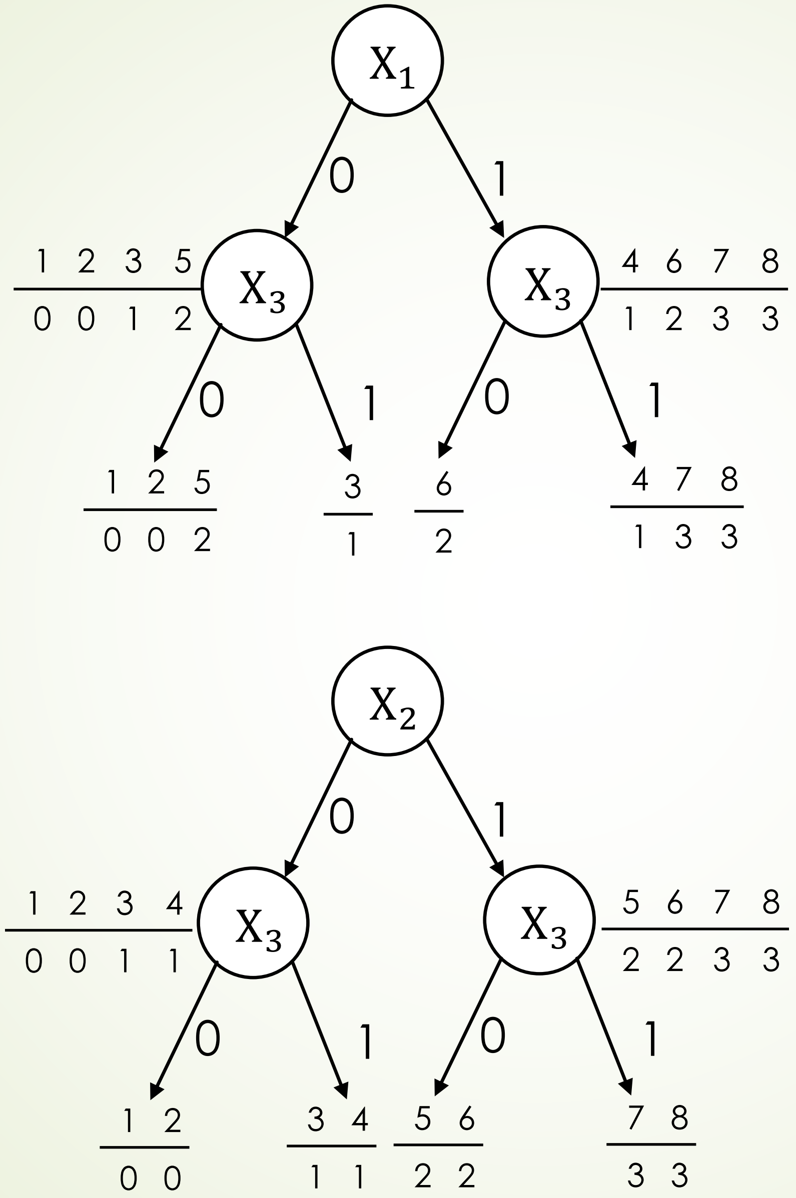
$$\text{Misclass}_{X_1}(D) = \_\_\_\_$$



$$\text{Misclass}_{X_2}(D) = \_\_\_\_$$

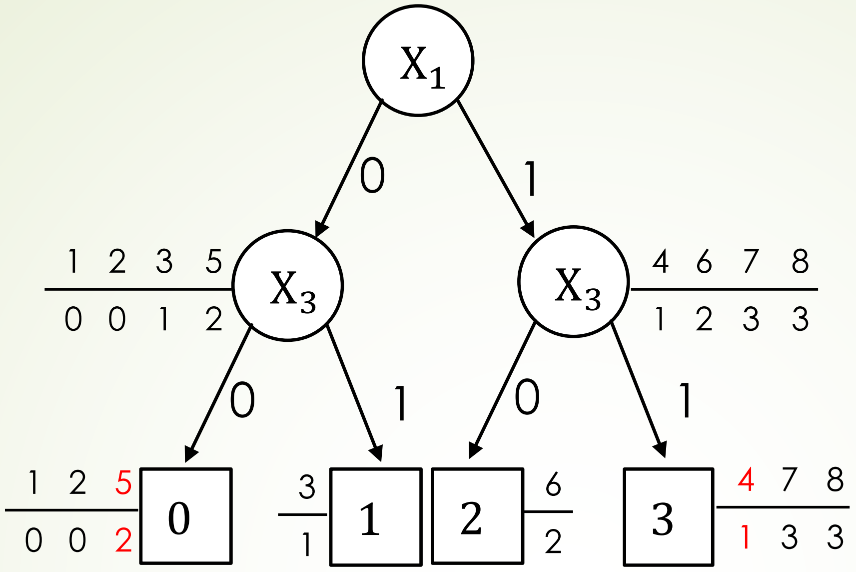
- What is the best splitting attribute? X_1 / X_2 / same

	X_1	X_2	X_3	Y
1.	0	0	0	0
2.	0	0	0	0
3.	0	0	1	1
4.	1	0	1	1
5.	0	1	0	2
6.	1	1	0	2
7.	1	1	1	3
8.	1	1	1	3

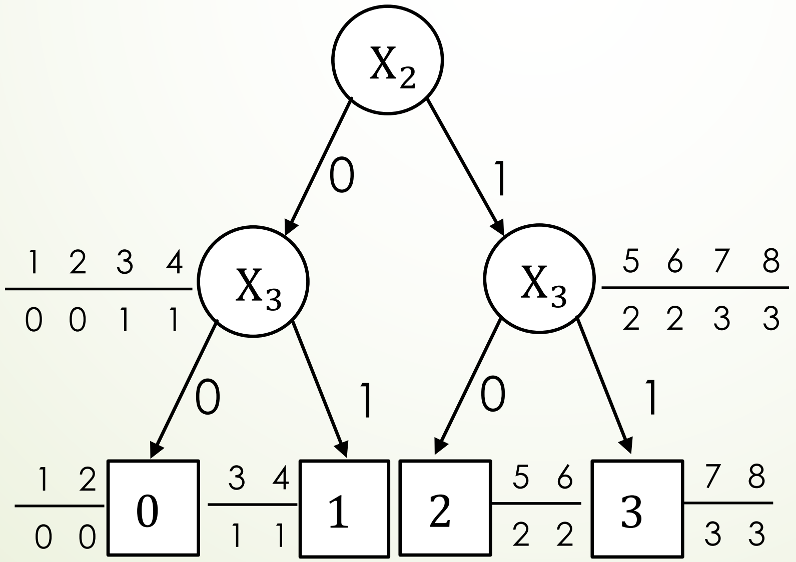


Further split on X_3

	X_1	X_2	X_3	Y
1.	0	0	0	0
2.	0	0	0	0
3.	0	0	1	1
4.	1	0	1	1
5.	0	1	0	2
6.	1	1	0	2
7.	1	1	1	3
8.	1	1	1	3



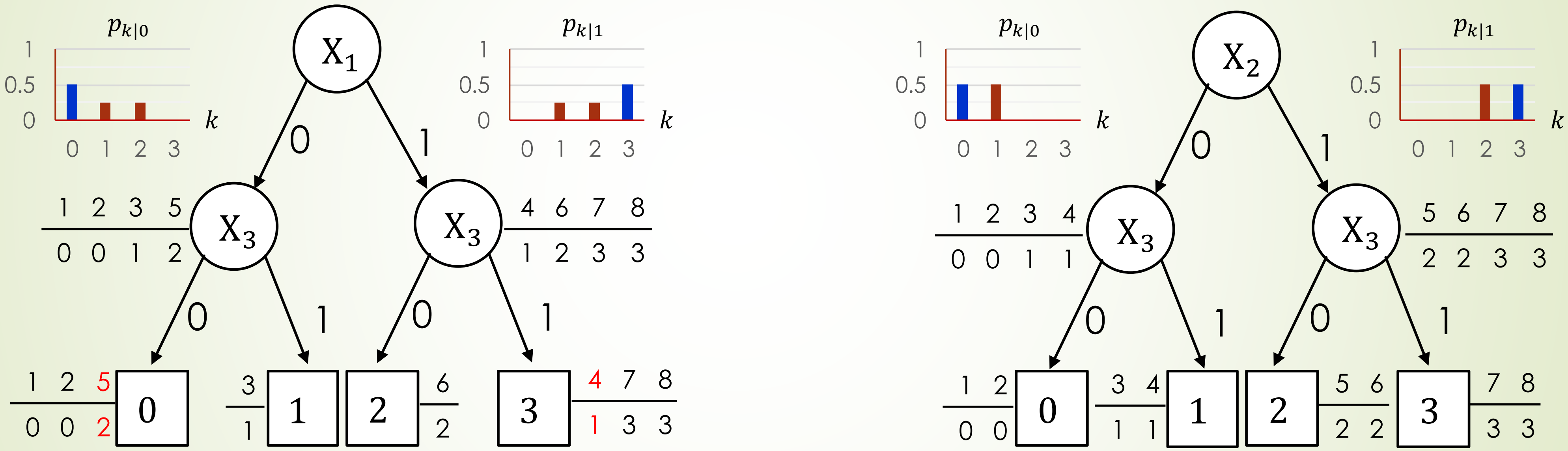
Error rate = ________



Error rate = ________

Issue of greedy algorithm

- **Locally optimal** split may not be **g________________ optimal**.



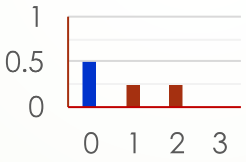
- Why splitting on X_1 is not good? Child nodes of X_1 are **less p________**.
- Why misclassification rate fails?
  It neglects the **distribution of the class values of m___________________ instances**.

How to remain greedy but not myopic

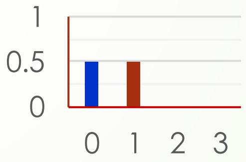
- Find better **i___________ measures** than misclassification rate.
- How to measure impurity?
- E.g., order the following distributions in ascending order of impurities:

___ (purest) < ___< ___< ___

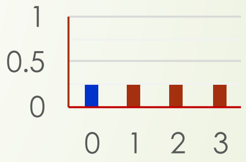
(i)



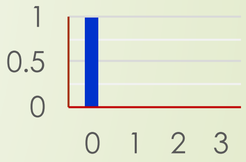
(ii)



(iii)



(iv)



- Given a distribution p_k of the class values of D, how to define a non-negative function of p_k's that respect the above ordering?

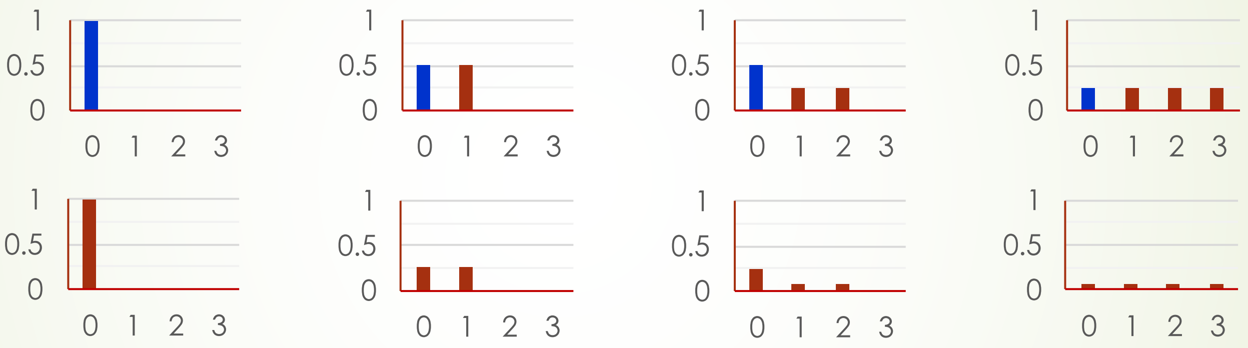
$1 - \max_k p_k$ works? $\underline{\text{Yes/No}}$ $\qquad$ $1 - \sum_k p_k$ works? $\underline{\text{Yes/No}}$

Gini impurity index

$$\text{Gini}(D) \coloneqq g(p_0, p_1, \ldots) \coloneqq \sum_k p_k(1 - p_k) = 1 - \sum_k p_k^2$$

p_k

p_k^2



$\sum_k p_k^2$	1	0.5	______	______
Gini(D)	0	0.5	______	______

Why it works?

- $g(p_0, p_1, \dots) \geq 0$. Equality iff $\forall k, p_k \in \{0,1\}$. Why?

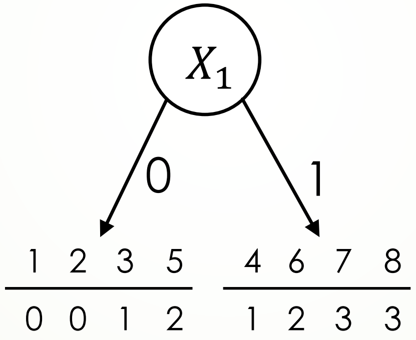
- $g(p_0, p_1, \dots, p_n) \leq 1 - \frac{1}{n}$. Equality iff $p_k =$__________. Why?

Finding the best split using Gini impurity

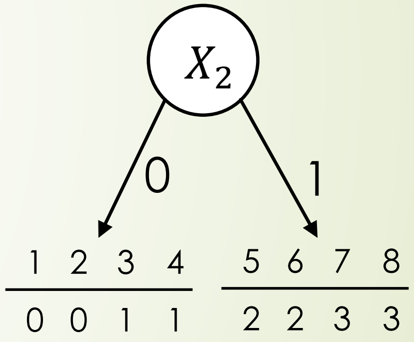
- Minimize the Gini impurity given A:

$$\mathrm{Gini}_A(D) := \sum_j \frac{|D_j|}{|D|}\mathrm{Gini}(D_j)$$

	X_1	X_2	X_3	Y
1.	0	0	0	0
2.	0	0	0	0
3.	0	0	1	1
4.	1	0	1	1
5.	0	1	0	2
6.	1	1	0	2
7.	1	1	1	3
8.	1	1	1	3



$$\mathrm{Gini}(D_0) = \mathrm{Gini}(D_1) = 0.625$$
$$= \mathrm{Gini}_A(D)$$



$$\mathrm{Gini}(D_0) = \mathrm{Gini}(D_1) = \_\_\_\_\_\_$$
$$= \mathrm{Gini}_A(D)$$

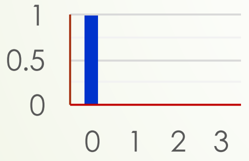
- What is the best splitting attribute? X_1 / X_2 / same

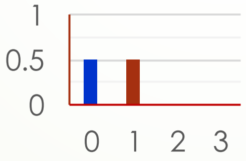
An impurity measure from information theory
Shannon's entropy

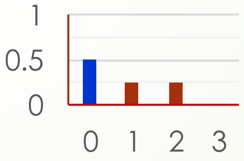
$$\text{Info}(D) := h(p_1, p_2, \ldots)$$
$$:= \sum_k p_k \log \frac{1}{p_k} = -\sum_{k:p_k>0} p_k \log p_k$$

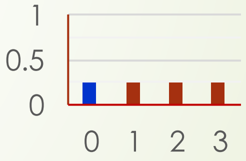
- Measured in bits with base-2 logarithm. Why?
- $0 \log 0$ is regarded as $\lim_{p \to 0} p \log p$ even though $\log 0$ is undefined.

p_k









$p_k \log_2 \frac{1}{p_k}$	$0,0,0,0$	$a, a, 0, 0$	$a, b, b, 0$	b, b, b, b
$\text{Info}(D)$	0	___	___	___

$a =$ ___
$b =$ ___

Why it works?

- $h(p_0, p_1, \ldots) \geq 0$. Equality iff $\forall k, p_k \in \{0,1\}$. Why?

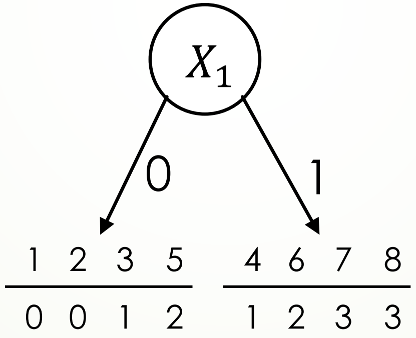
- $h(p_0, p_1, \ldots, p_n) \leq \log_2 n$. Equality iff $p_k =$__________. Why?

Finding the best split by conditional entropy

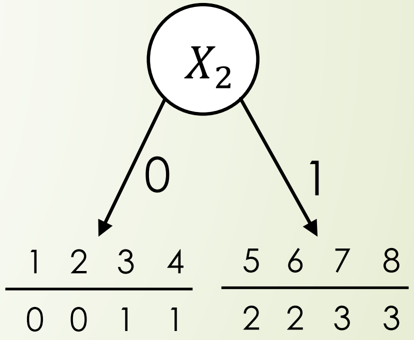
- Minimize the entropy given A,

$$\mathrm{Info}_A(D) := \sum_j \frac{|D_j|}{|D|}\mathrm{Info}(D_j)$$

	X_1	X_2	X_3	Y
1.	0	0	0	0
2.	0	0	0	0
3.	0	0	1	1
4.	1	0	1	1
5.	0	1	0	2
6.	1	1	0	2
7.	1	1	1	3
8.	1	1	1	3



$$\mathrm{Info}(D_0) = \mathrm{Info}(D_1) = 1.5$$
$$= \mathrm{Info}_A(D)$$



$$\mathrm{Info}(D_0) = \mathrm{Info}(D_1) = \underline{\qquad\qquad}$$
$$= \mathrm{Info}_A(D)$$

- What is the best splitting attribute? <u>X_1 / X_2 / same</u>

Which impurity measure is used?

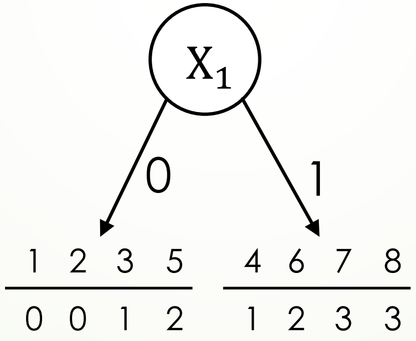
- **ID3** (Iteractive Dichotomiser 3) maximizes

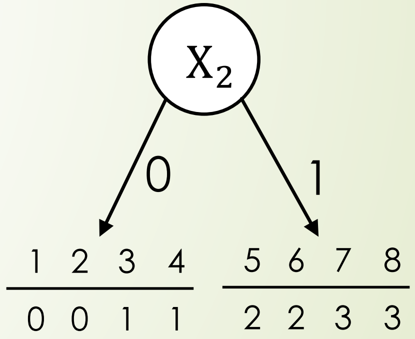
  $\mathrm{Gain}_{\mathrm{A}}(D) \coloneqq \mathrm{Info}(D) - \mathrm{Info}_{\mathrm{A}}(D)$ (information gain or mutual information)

- **CART** (Classification and Regression Tree)

  $\Delta\mathrm{Gini}_{\mathrm{A}}(D) \coloneqq \mathrm{Gini}(D) - \mathrm{Gini}_{\mathrm{A}}(D)$ (Drop in Gini impurity)

	X_1	X_2	X_3	Y
1.	0	0	0	0
2.	0	0	0	0
3.	0	0	1	1
4.	1	0	1	1
5.	0	1	0	2
6.	1	1	0	2
7.	1	1	1	3
8.	1	1	1	3





$\mathrm{Info}(D) = 2$
$\mathrm{Gini}(D) = 0.75$

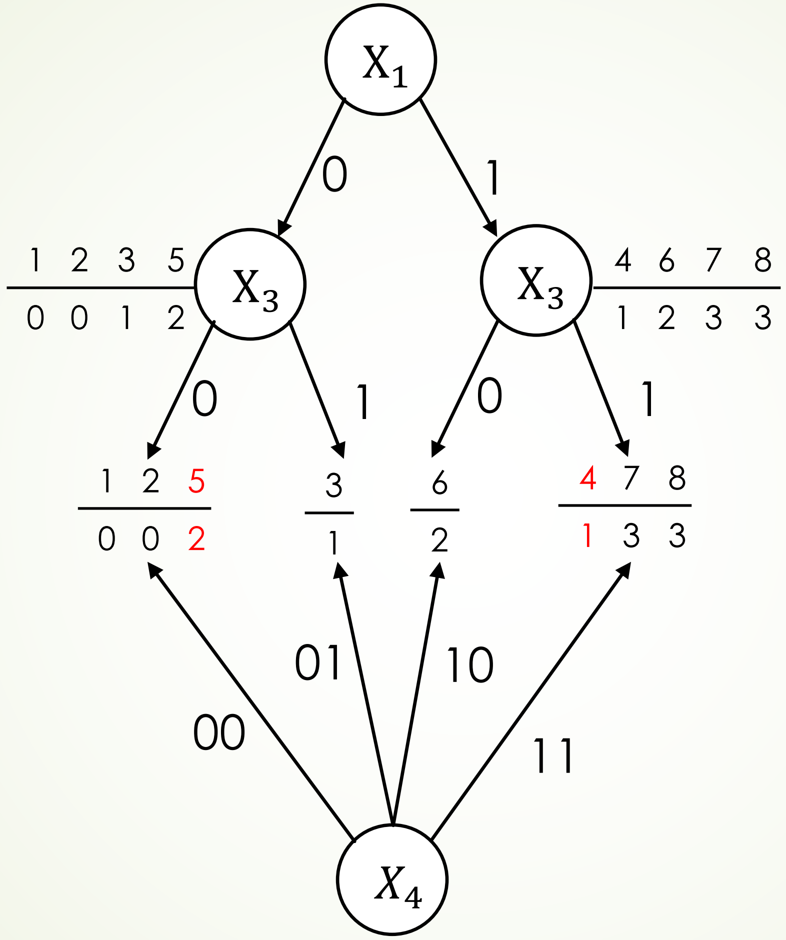
$\mathrm{Info}_{\mathrm{X}_1}(D) = 1.5$ $\quad$ $\mathrm{Gain}_{\mathrm{X}_1}(D) = 0.5$
$\mathrm{Gini}_{\mathrm{X}_1}(D) = 0.625$ $\quad$ $\Delta\mathrm{Gini}_{\mathrm{X}_1}(D) = 0.125$

$\mathrm{Info}_{\mathrm{X}_2}(D) = 1$ $\quad$ $\mathrm{Gain}_{\mathrm{X}_2}(D) =$ ______
$\mathrm{Gini}_{\mathrm{X}_2}(D) = 0.5$ $\quad$ $\Delta\mathrm{Gini}_{\mathrm{X}_2}(D) =$ ______

- What is the best splitting attribute? $\underline{\mathrm{X}_1 \ / \ \mathrm{X}_2 \ / \ \text{same}}$

	X_1	X_2	X_3	X_4	Y
1.	0	0	0	00	0
2.	0	0	0	00	0
3.	0	0	1	01	1
4.	1	0	1	11	1
5.	0	1	0	00	2
6.	1	1	0	10	2
7.	1	1	1	11	3
8.	1	1	1	11	3

$X_4 \coloneqq X_1 \circ X_3$



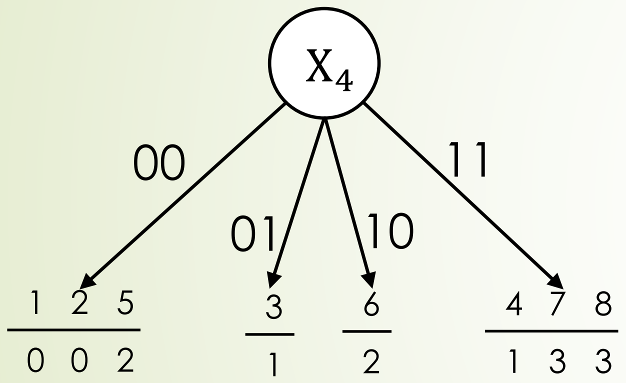
- Is X_4 a good splitting attribute? Yes/No.

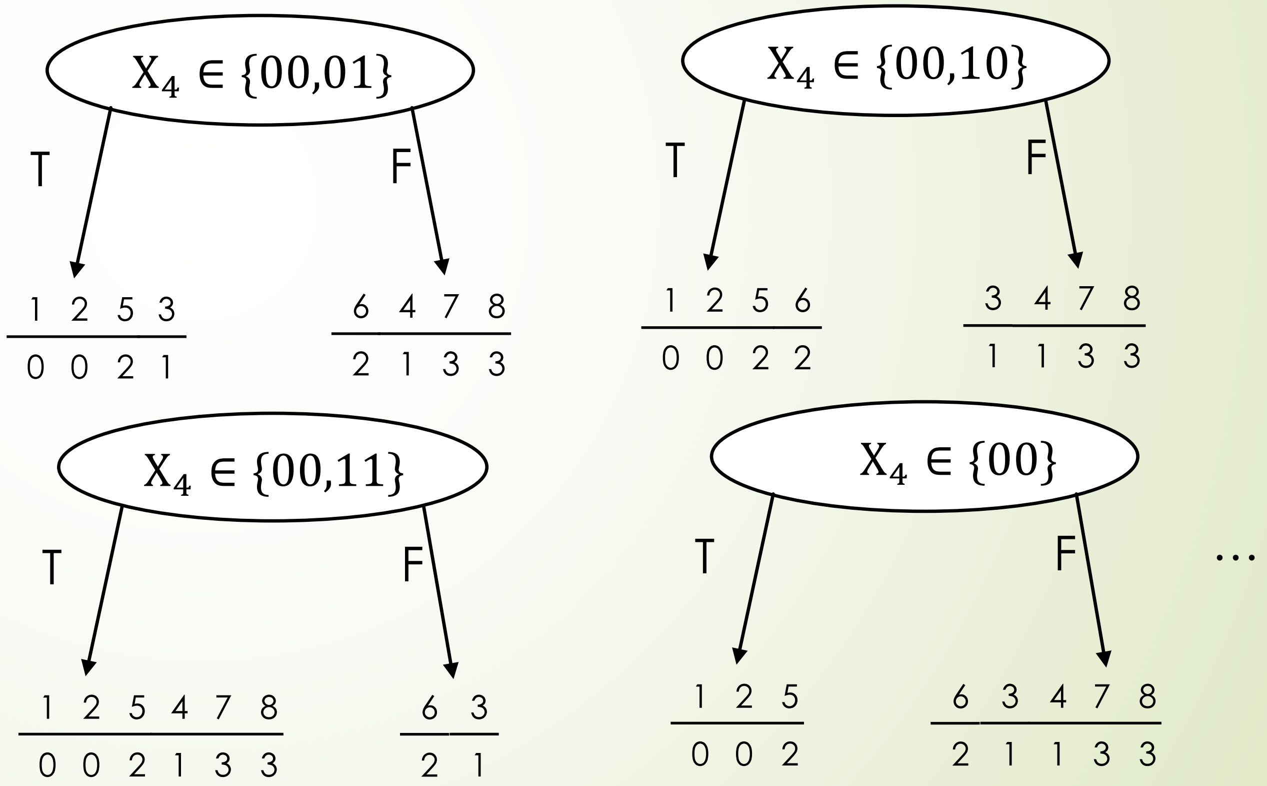
Bias towards attributes with many outcomes

- An attribute with more outcomes tends to
  - reduce impurity more but
  - result in more comparisons.
- Issues: Such attribute may not *minimize impurity per comparison*.
- Remedies?

Binary split also for nominal attributes

- CART uses a **s**____________________ S to generate a **binary split** (whether $A \in S$).
- The number of outcomes is therefore limited to ______.





$\max_{S} \Delta \mathrm{Gini}_{X_4 \in S}(D)$ = ______________
achieved by S = ________________

Normalization by split information

- C4.5/J48 allows **m**____________ split but uses **information gain ratio**

$$\frac{\mathrm{Gain}_{\mathrm{A}}(D)}{\mathrm{SplitInfo}_{\mathrm{A}}(D)} \quad \text{where} \quad \mathrm{SplitInfo}_{\mathrm{A}}(D) = \sum_j \frac{|D_j|}{|D|} \log_2 \frac{1}{|D_j|/|D|}.$$

  - $\mathrm{SplitInfo}_{\mathrm{A}}(D)$is the entropy of ______________________ because __________________.
  - Attributes with many outcomes tend to have smaller/larger $\mathrm{SplitInfo}_{\mathrm{A}}(D)$.

How to avoid overfitting

- **P__-pruning**: Limit the size of the tree as we build it. E.g.,
  - Ensure each node is supported by enough examples.
    (C4.5: minimum number of objects.)
  - Split only if we are confident enough about the improvement.
    (C4.5: confidence factor.)
- **P__-pruning**: Reduce the size of the tree after we build it. E.g.,
  - Contract leaf nodes if complexity outweighs the risk.
    (CART: **cost-complexity pruning**)

References

- 8.1 Basic Concepts
- 8.2 Decision Tree Induction
- Optional readings
  - https://en.wikipedia.org/wiki/C4.5_algorithm
  - Cover, T., & Thomas, J. (2006). *Elements of information theory* (2nd ed.). Hoboken, N.J.: Wiley-Interscience. Chapter 1 and 2.