Decision Tree Induction: Algorithm

- Basic algorithm → อัลกอริทึมพื้นฐาน
  - ต้นไม้ถูกสร้างขึ้น บนลงล่าง แบ่ง
  - Tree is constructed in a **top-down, recursive, divide-and-conquer manner**
  - At start, all the training examples are at the root ตัวอย่างการฝึกทั้งหมดอยู่ที่รูท
  - Examples are partitioned recursively based on selected attributes จะถูกแบ่งพาร์ติชันแบบเรียกซ้ำตามแอตทริบิวต์ที่เลือก
  - ในแต่ละโหนด จะถูกเลือกตามตัวอย่าง
  - On each node, attributes are selected based on the training examples on that node, and a heuristic or statistical measure (e.g., **information gain**)
- Conditions for stopping partitioning เงื่อนไขในการหยุดแบ่ง
  - All samples for a given node belong to the same class ตัวอย่างทั้งหมดสำหรับ node ที่กำหนดอยู่ในคลาสเดียวกัน
  - There are no remaining attributes for further partitioning ไม่มี attributes เหลือ ก็ไม่ทำต่อ
  - There are no samples left ไม่มีตัวอย่างเหลือ
- Prediction
  - **Majority voting** is employed for classifying the leaf

How to Handle Continuous-Valued Attributes?

- Method 1: Discretize continuous values and treat them as categorical values
  - E.g., age: < 20, 20..30, 30..40, 40..50, > 50
- Method 2: Determine the ***best split point*** for continuous-valued attribute A
  - Sort the value A in increasing order:, e.g. 15, 18, 21, 22, 24, 25, 29, 31, …
  - *Possible split point:* the midpoint between *each pair of adjacent values*
    - $(a_i+a_{i+1})/2$ is the midpoint between the values of a_i and a_{i+1}
    - e.g., (15+18/2 = 16.5, 19.5, 21.5, 23, 24.5, 27, 30, …
  - The point with the *maximum information gain* for A is selected as the **split-point** for A
- Split: Based on split point P
  - The set of tuples in D satisfying A ≤ P vs. those with A > P

Method 1 Categorical

15, 18, 21, 22, 24, 25, 29, 31,

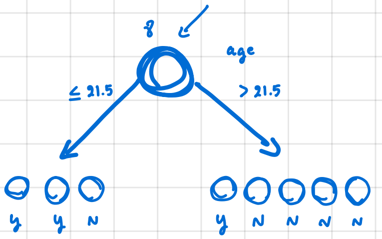
จัดกลุ่ม < 18, $18-22$, $22-30$, ≥ 31

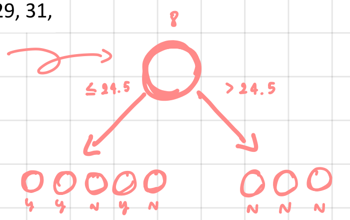
Method 2 best split point

1. เรียงข้อมูลก่อน

y y N | y N | N N N

15, 18, 21, | 22, 24, | 25, 29, 31,





$\frac{5}{8} \overset{\le}{I}(3,2) + \frac{3}{8} \overset{>}{I}(0,3)$

$age_{21.5} = \frac{3}{8} I(2,1) + \frac{5}{8} I(1,4)$

Method 3 Random

1 | 2 3 4 5 6 7 8

15, | 18, 21, 22, 24, 25, 29, 31,

สุ่มเลขมาถ้าได้ 1 ก็แบ่งตรง 15

Gain Ratio: A Refined Measure for Attribute Selection

- Information gain measure is biased towards attributes with a large number of values
- Gain ratio: Overcomes the problem (as a normalization to information gain)

$$SplitInfo_A(D) = -\sum_{j=1}^{v} \frac{|D_j|}{|D|} \times \log_2\left(\frac{|D_j|}{|D|}\right)$$

  - GainRatio(A) = Gain(A)/SplitInfo(A)
  - The attribute with the maximum gain ratio is selected as the splitting attribute
  - Gain ratio is used in a popular algorithm C4.5 (a successor of ID3) by R. Quinlan
- Example
  - $\mathrm{SplitInfo}_{\mathrm{income}}(\mathrm{D}) = -\frac{4}{14}\log_2\frac{4}{14} - \frac{6}{14}\log_2\frac{6}{14} - \frac{4}{14}\log_2\frac{4}{14} = 1.557$
  - GainRatio(income) = 0.029/1.557 = 0.019

Another Measure: Gini Index

- Gini index: Used in CART, and also in IBM IntelligentMiner
- If a data set D contains examples from n classes, gini index, $gini(D)$ is defined as
  - $gini(D) = 1 - \sum_{j=1}^{n} p_j^2$ $\longrightarrow 1 - \left\{ \left(\frac{3}{8}\right)^2 + \left(\frac{5}{8}\right)^2 \right\}$
    - p_j is the relative frequency of class j in D
- If a data set D is split on A into two subsets D_1 and D_2, the $gini$ index $gini(D)$ is defined as

  g(y,N

  - $gini_A(D) = \frac{|D_1|}{|D|} gini(D_1) + \frac{|D_2|}{|D|} gini(D_2)$
- Reduction in Impurity:
  - $\Delta gini(A) = gini(D) - gini_A(D)$
- The attribute provides the smallest $gini_{split}(D)$ (or the largest reduction in impurity) is chosen to split the node (*need to enumerate all the possible splitting points for each attribute*)

Computation of Gini Index

- Example: D has 9 tuples in buys_computer = "yes" and 5 in "no"

$$gini(D) = 1 - \left(\frac{9}{14}\right)^2 - \left(\frac{5}{14}\right)^2 = 0.459$$

- Suppose the attribute income partitions D into 10 in D_1: {low, medium} and 4 in D_2
  - $gini_{income\in\{low,medium\}}(D) = \frac{10}{14}gini(D_1) + \frac{4}{14}gini(D_2)$
    $$= \frac{10}{14}\left(1 - \left(\frac{7}{10}\right)^2 - \left(\frac{3}{10}\right)^2\right) + \frac{4}{14}\left(1 - \left(\frac{2}{4}\right)^2 - \left(\frac{2}{4}\right)^2\right) = 0.443$$
    $$= Gini_{income\in\{high\}}(D)$$
  - $Gini_{\{low,high\}}$ is 0.458; $Gini_{\{medium,high\}}$ is 0.450
  - Thus, split on the {low,medium} (and {high}) since it has the lowest Gini index
- All attributes are assumed continuous-valued
- May need other tools, e.g., clustering, to get the possible split values
- Can be modified for categorical attributes

Comparing Three Attribute Selection Measures

- The three measures, in general, return good results but
  - **Information gain**:
    - biased towards multivalued attributes
  - **Gain ratio**:
    - tends to prefer unbalanced splits in which one partition is much smaller than the others
  - **Gini index**:
    - biased to multivalued attributes
    - has difficulty when # of classes is large
    - tends to favor tests that result in equal-sized partitions and purity in both partitions

Other Attribute Selection Measures

- Minimal Description Length (MDL) principle
  - Philosophy: The simplest solution is preferred
  - The best tree as the one that requires the fewest # of bits to both (1) encode the tree, and (2) encode the exceptions to the tree
- CHAID: a popular decision tree algorithm, measure based on χ^2 test for independence
- Multivariate splits (partition based on multiple variable combinations)
  - CART: finds multivariate splits based on a linear combination of attributes
- There are many other measures proposed in research and applications
  - E.g., G-statistics, C-SEP
- Which attribute selection measure is the best?
  - Most give good results, none is significantly superior than others

test data → การจะมีมากถ้าจำนวน
data มาก

training data | real world data

all data

แยกก้อนไว้ทดสอบ

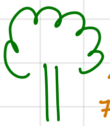
training data | test data

training data

training data | Validation data

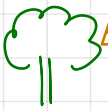
นำมา train ต้นไม้ 3 ต้น

นำ Validation มา test



70 %

สูง 2 ชั้น



80 %

มีใบ 100 ใบ

90 %

→ เลือกต้นนี้เพราะคะแนนสูงสุด

split แบบ random

ได้

split แบบ random

↘

นำมา train อีกครั้ง จากนั้นจะได้

Spit random เรียนด้วย data ที่เยอะขึ้น

สอนจบ...

เมื่อสอนจบ นำ test data

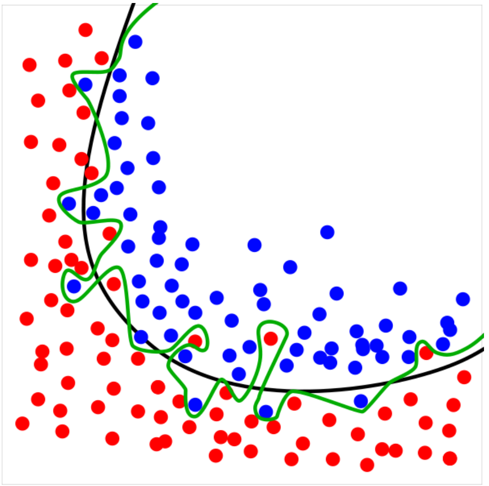
ได้ 73 %

เราจะได้ Tree randomsplit มีความแม่นยำ = 73 %

แปลว่า โลมอกตัวนี้การทำนายได้อยู่ที่ 73 %

ตัดแต่งกิ่ง

Overfitting and Tree Pruning



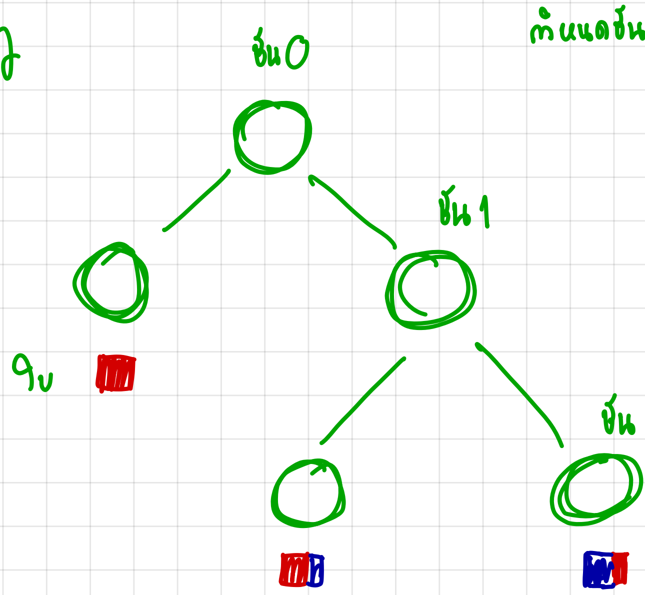
- Overfitting: An induced tree may overfit the training data
  - Too many branches, some may reflect anomalies due to noise or outliers
  - Poor accuracy for unseen samples
- Two approaches to avoid overfitting

  หยุดการโตของต้นไม้ก่อน (กำหนดชั้นของต้นไม้)
  - Prepruning: *Halt tree construction early*-do not split a node if this would result in the goodness measure falling below a threshold
    - Difficult to choose an appropriate threshold

  ตัดกิ่งที่ต้นไม้โตสมบูรณ์แล้ว
  - Postpruning: *Remove branches* from a “fully grown” tree—get a sequence of progressively pruned trees
    - Use a set of data different from the training data to decide which is the “best pruned tree”

Prepruning

กำหนดชั้นต้นไม้ เท่ากับ 2 ชั้น



ดูว่า % สีไหนมากตอบอันนั้น

Postpruning