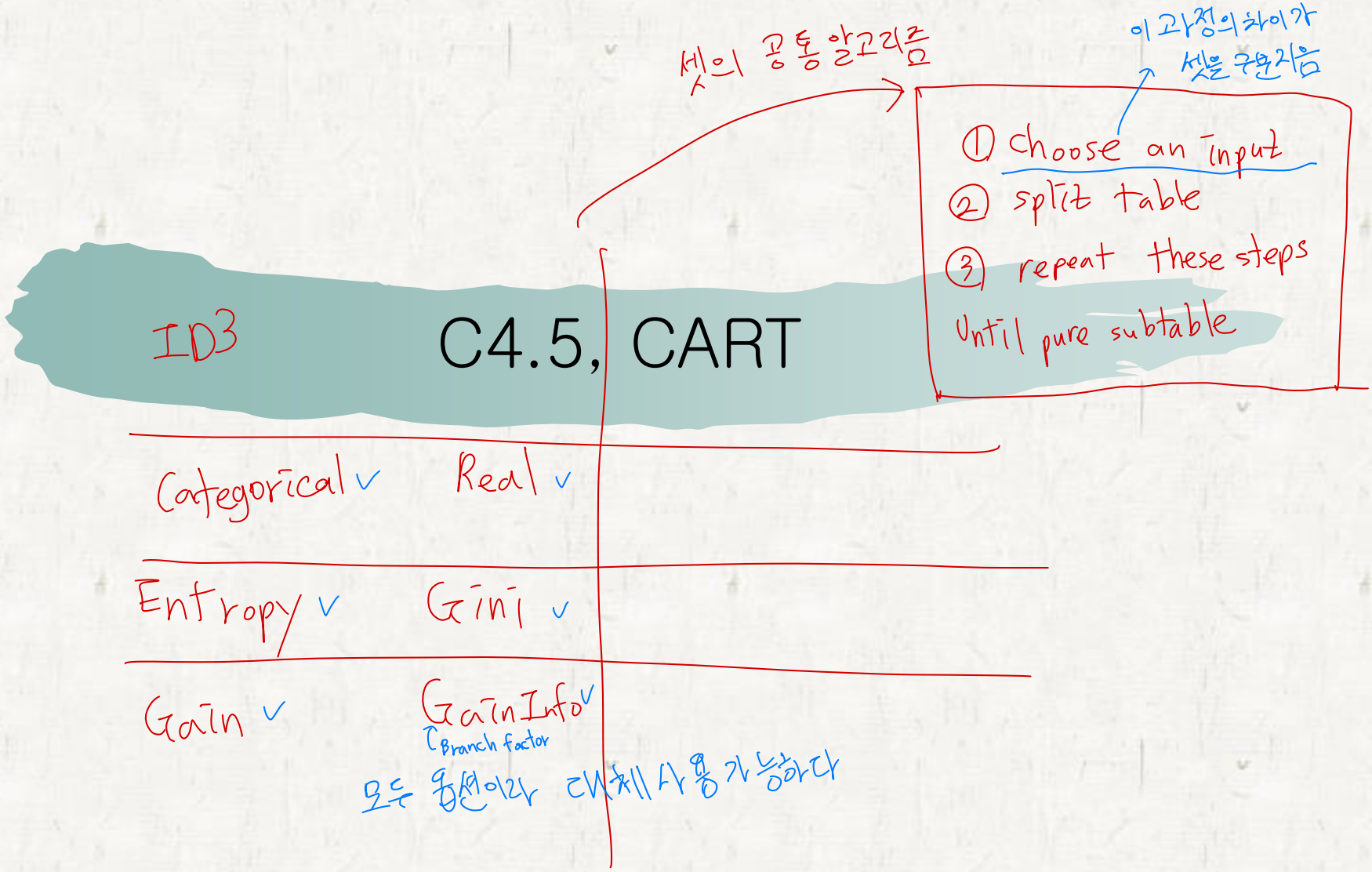

셋의 공통 알고리즘
이 과정의 차이가 셋을 구분지음
① choose an input
② split table
③ repeat these steps
Until pure subtable
ID3
C4.5, CART
Categorical ∨
Real ∨
Entropy ∨
Gini ∨
Gain ∨
GainInfo ∨
Branch factor
모두 옵션이나 대체사용가능하다

C4.5

- Dataset with Continuous-Valued Attributes

OUTLOOK	TEMP.	HUMIDITY	WINDY	PLAY
sunny	85	85	false	No
sunny	80	90	true	No
overcast	83	78	false	Yes
rain	70	96	false	Yes
rain	68	80	false	Yes
rain	65	70	true	No
overcast	64	65	true	Yes
sunny	72	95	false	No
sunny	69	70	false	Yes
rain	75	80	false	Yes
sunny	75	70	true	Yes
overcast	72	90	true	Yes
overcast	81	75	false	Yes
rain	71	80	true	No

C4.5

Categorical 처리방법

1. Evaluate the entropy of the given table
2. Split the table with an input
3. Evaluate the entropy of each sub-table
4. Evaluate the average of the entropies
5. Evaluate the Gain: (1)-(4)

* Repeat this for all inputs and choose the best

- Categorical Attribute: Split as we did

① Choose an input
- try all inputs → 속성은 모든 input 다 볼수 있음
① split with an input
② Entropy ③ Gain
- choose the best

② split table

S: [9+,5-]
E=0.940

Outlook

Sunny | Overcast | Rain

$S_{outlook=Sunny}$: [2+,3-]
E=0.971

$S_{outlook=Overcast}$: [4+,0-]
E=0.00

$S_{outlook=Rain}$: [3+,2-]
E=0.971

Gain(S, Outlook)
= 0.940-(5/14)*0.971-(5/14)*0.971
= 0.246

S: [9+,5-]
E=0.940

Wind

Weak | Strong

$S_{wind=Weak}$: [6+,2-]
E=0.811

$S_{wind=Strong}$: [3+,3-]
E=1.00

Gain(S, Wind)
= 0.940-(8/14)*0.811-(6/14)*1.00
= 0.048

C4.5

1. Evaluate the entropy of the given table
2. Split the table with an input
3. Evaluate the entropy of each sub-table
4. Evaluate the average of the entropies
5. Evaluate the Gain: (1)-(4)

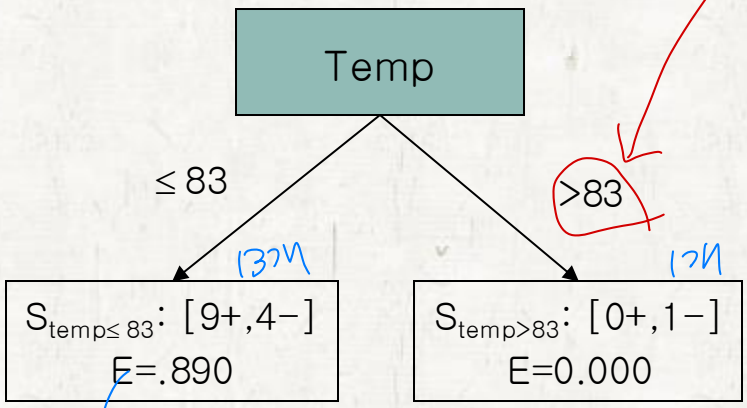
* Repeat this for all inputs and choose the best

- Continuous-Valued Attributes
  - Binary Split

실수형의 attribute

* ① choose the 기준점 split point → a real value
② split the table using the split point
③ Entropy ④ Gain

OUTLOOK	TEMP.	HUMIDITY	WINDY	PLAY
sunny	85	85	false	No
sunny	80	90	true	No
overcast	83	78	false	Yes
rain	70	96	false	Yes
rain	68	80	false	Yes
rain	65	70	true	No
overcast	64	65	true	Yes
sunny	72	95	false	No
sunny	69	70	false	Yes
rain	75	80	false	Yes
sunny	75	70	true	Yes
overcast	72	90	true	Yes
overcast	81	75	false	Yes
rain	71	80	true	No



$Gain = 0.940 - \frac{13}{14} \times 0.890$

How to choose the split point?

Temp → 13 13 13

$SplitInfo = -\frac{13}{14}\log\frac{13}{14} - \frac{1}{14}\log\frac{1}{14}$

Temp 가 몇에서 split 되었지? → play가 몇에서 split 되었을지 ≠ Entropy → E는 두 집합의 Entropy의 평균으로 계산

C4.5

Continuous input 처리 방법

1. Evaluate the entropy of the given table
2-1. Choose the split point
2-2. Split the table with the split point
3. Evaluate the entropy of each sub-table
4. Evaluate the average of the entropies
5. Evaluate the Gain: (1)-(4)
* Repeat this for all inputs and choose the best

- Continuous-Valued Attributes
  - How to choose the split point?

OUTLOOK	TEMP.	HUMIDITY	WINDY	PLAY
sunny	85	85	false	No
sunny	80	90	true	No
overcast	83	78	false	Yes
rain	70	96	false	Yes
rain	68	80	false	Yes
rain	65	70	true	No
overcast	64	65	true	Yes
sunny	72	95	false	No
sunny	69	70	false	Yes
rain	75	80	false	Yes
sunny	75	70	true	Yes
overcast	72	90	true	Yes
overcast	81	75	false	Yes
rain	71	80	true	No

All values in the dataset is the candidates of split point

① Evaluate all split point
② choose the best one

C4.5

1. Evaluate the entropy of the given table
2-1. Choose the split point
2-2. Split the table with the split point
3. Evaluate the entropy of each sub-table
4. Evaluate the average of the entropies
5. Evaluate the Gain: (1)-(4)
* Repeat this for all inputs and choose the best

- How to choose the split point
  - Step1: Try all the candidate split points

OUTLOOK	TEMP.	HUMIDITY	WINDY	PLAY
overcast	64	65	true	Yes
rain	65	70	true	No
rain	68	80	false	Yes
sunny	69	70	false	Yes
rain	70	96	false	Yes
rain	71	80	true	No
sunny	72	95	false	No
overcast	72	90	true	Yes
rain	75	80	false	Yes
sunny	75	70	true	Yes
sunny	80	90	true	No
overcast	81	75	false	Yes
overcast	83	78	false	Yes
sunny	85	85	false	No

(1) ≤64 : [1+,0−], >64: [8+,5−]
(2) ≤65 : [1+,1−], >65: [8+,4−]
(3) ≤68 : [2+,1−], >68: [7+,4−]
(4) ≤69 : [3+,1−], >69: [6+,4−]
(5) ≤70 : [4+,1−], >70: [5+,4−]
(6) ≤71 : [4+,2−], >71: [5+,3−]
(7) ≤72 : [5+,3−], >72: [4+,2−]
(8) ≤75 : [7+,3−], >75: [2+,2−]
(9) ≤80 : [7+,4−], >80: [2+,1−]
(10) ≤81 : [8+,4−], >81: [1+,1−]
(11) ≤83 : [9+,4−], >83: [0+,1−]
(12) ≤85 : [9+,5−], >85: [0+,0−]

C4.5

1. Evaluate the entropy of the given table
2-1. Choose the split point
2-2. Split the table with the split point
3. Evaluate the entropy of each sub-table
4. Evaluate the average of the entropies
5. Evaluate the Gain: (1)-(4)
* Repeat this for all inputs and choose the best

- How to choose the split point
  - Step2: Evaluate split qualities of them

avg Entropy

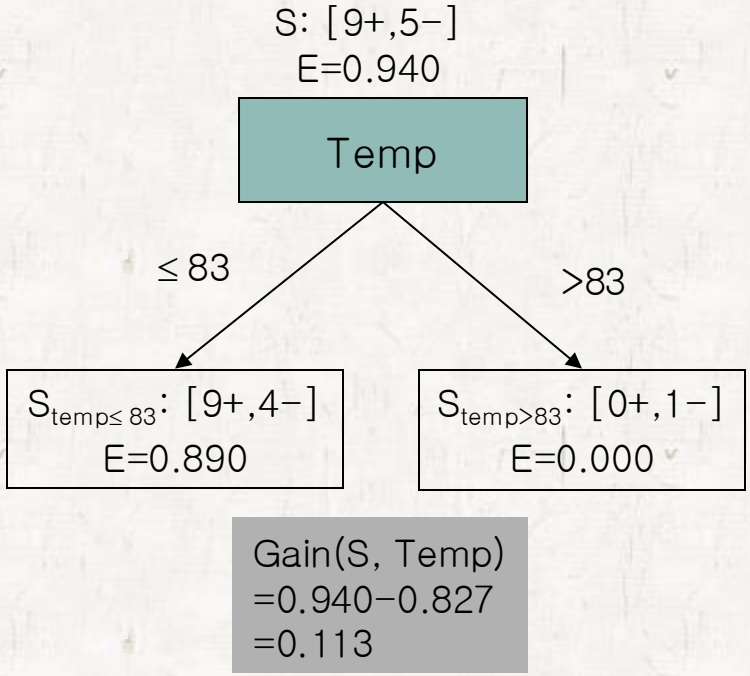
(1) ≤ 64 : [1+,0−], >64: [8+,5−] -> Entropy = 0.893
(2) ≤ 65 : [1+,1−], >65: [8+,4−] -> Entropy = 0.930
(3) ≤ 68 : [2+,1−], >68: [7+,4−] -> Entropy = 0.940
(4) ≤ 69 : [3+,1−], >69: [6+,4−] -> Entropy = 0.925
(5) ≤ 70 : [4+,1−], >70: [5+,4−] -> Entropy = 0.895
(6) ≤ 71 : [4+,2−], >71: [5+,3−] -> Entropy = 0.939
(7) ≤ 72 : [5+,3−], >72: [4+,2−] -> Entropy = 0.939
(8) ≤ 75 : [7+,3−], >75: [2+,2−] -> Entropy = 0.915
(9) ≤ 80 : [7+,4−], >80: [2+,1−] -> Entropy = 0.940
(10) ≤ 81 : [8+,4−], >81: [1+,1−] -> Entropy = 0.930
(11) ≤ 83 : [9+,4−], >83: [0+,1−] -> Entropy = 0.827
(12) ≤ 85 : [9+,5−], >85: [0+,0−] -> Entropy = 0.940

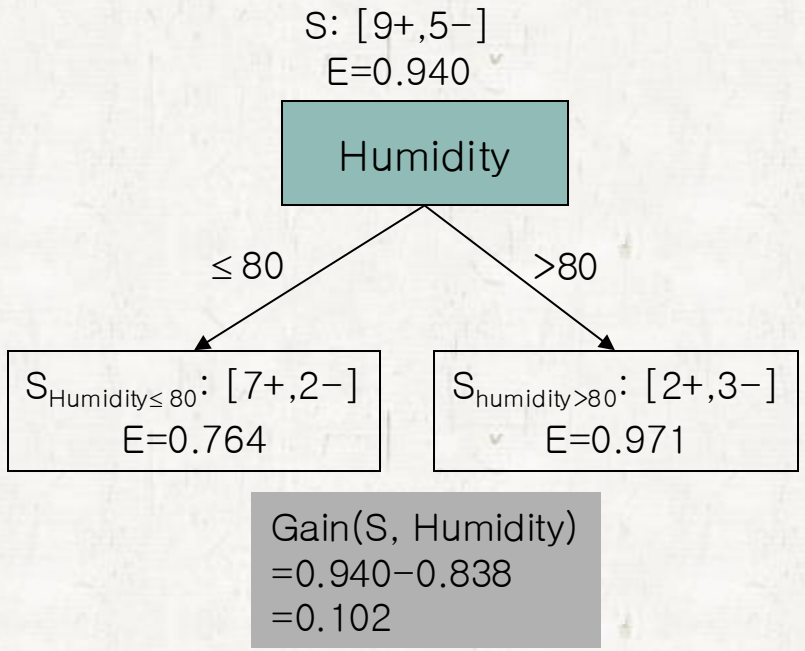
Entropy = confuseness

min → 0.827

C4.5

- Dataset with Continuous-Valued Attributes

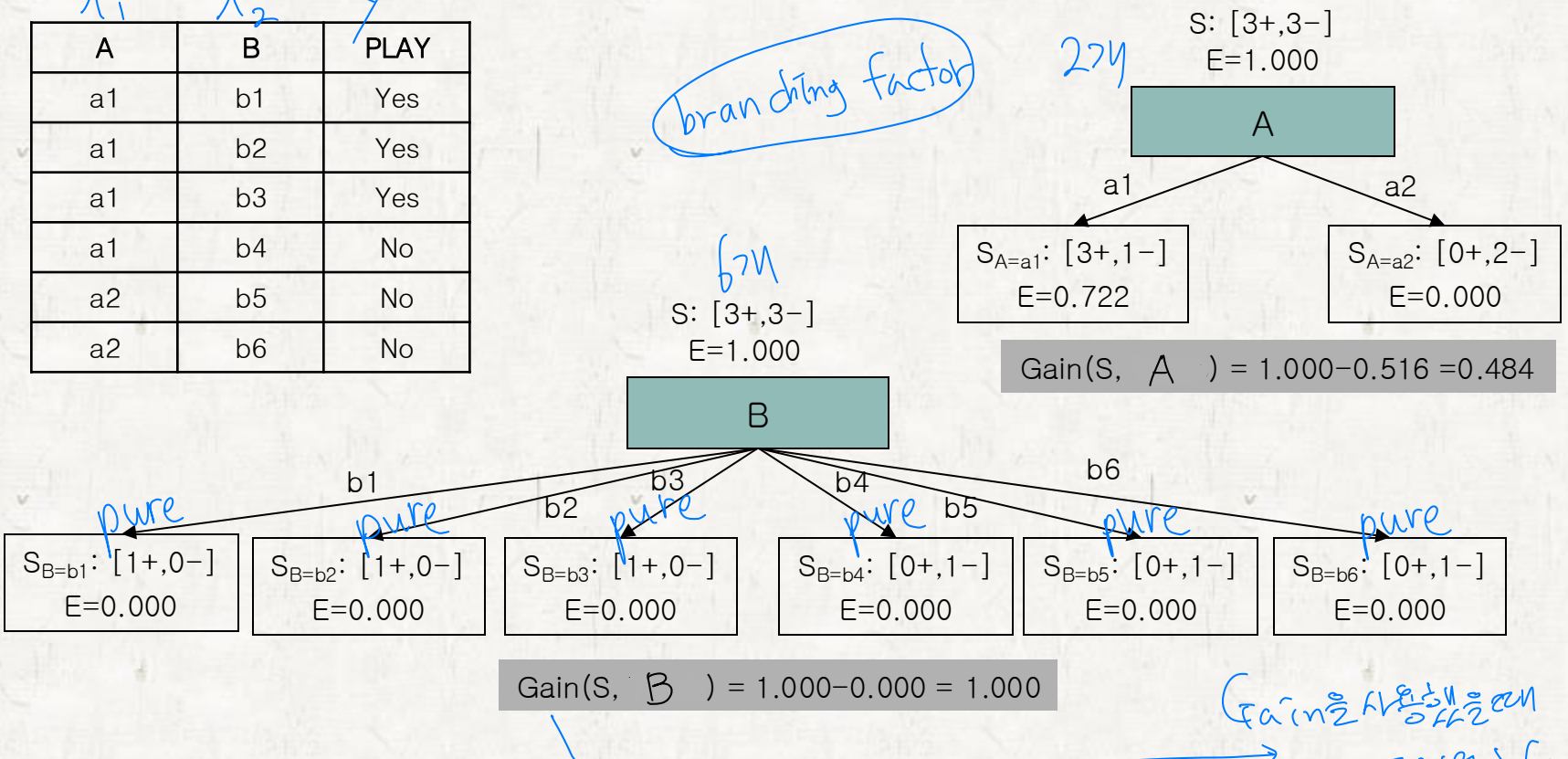




C4.5

- Alternative Measures for Selecting Attributes (1)

A	B	PLAY
a1	b1	Yes
a1	b2	Yes
a1	b3	Yes
a1	b4	No
a2	b5	No
a2	b6	No



Which one is better? Gain does not consider the branching factor

C4.5

- Alternative Measures for Selecting Attributes (2)
  - Considering not only Gain but also branching factor

- SplitInfo: Measuring branching factor
  - Entropy of the attribute

A	B	PLAY
a1	b1	Yes
a1	b2	Yes
a1	b3	Yes
a1	b4	No
a2	b5	No
a2	b6	No

$$A = \{a1, a1, a1, a1, a2, a2\}$$

$$SplitInfo(set, A) = -\frac{4}{6}\log_2\frac{4}{6} - \frac{2}{6}\log_2\frac{2}{6} = 0.918$$

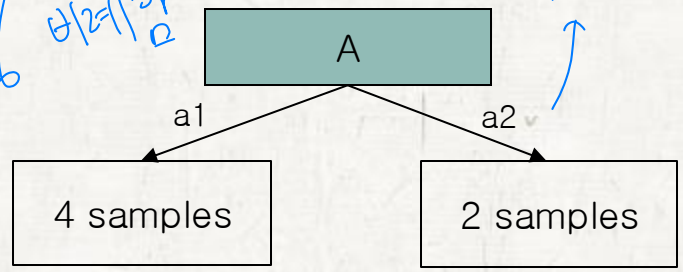
$$B = \{b1, b2, b3, b4, b5, b6\}$$

$$SplitInfo(set, B) = 6\left(-\frac{1}{6}\log_2\frac{1}{6}\right) = 2.585$$

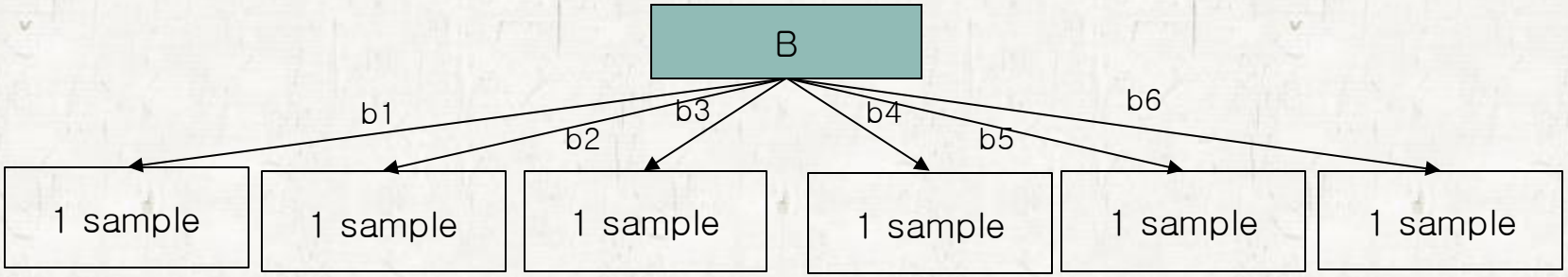
C4.5

- SplitInfo: Measuring branching factor

A	B	PLAY
a1	b1	Yes
a1	b2	Yes
a1	b3	Yes
a1	b4	No
a2	b5	No
a2	b6	No



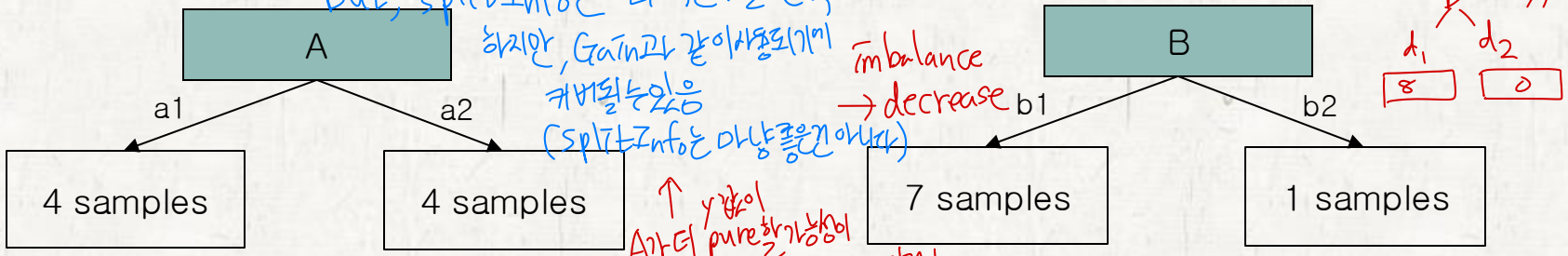
$$SplitInfo(set, A) = -\frac{4}{6}\log_2\frac{4}{6} - \frac{2}{6}\log_2\frac{2}{6} = 0.918$$



$$SplitInfo(set, B) = 6\left(-\frac{1}{6}\log_2\frac{1}{6}\right) = 2.585$$

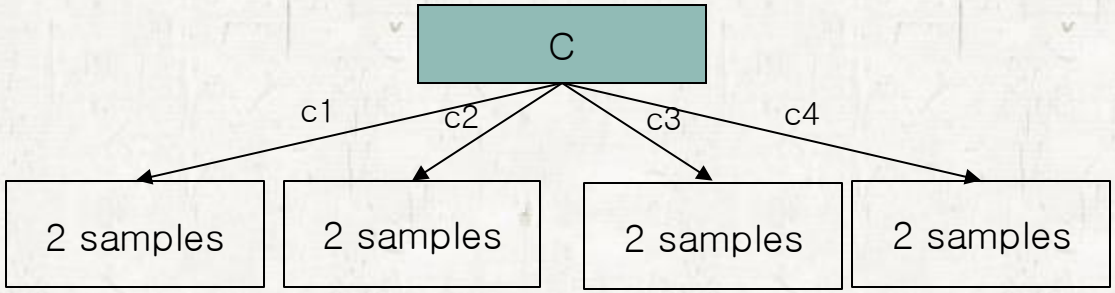
C4.5

- SplitInfo: Measuring branching factor



$$SplitInfo(set, A) = -\frac{4}{8}\log_2\frac{4}{8} - \frac{4}{8}\log_2\frac{4}{8} = 1.0$$

$$SplitInfo(set, B) = -\frac{7}{8}\log_2\frac{7}{8} - \frac{1}{8}\log_2\frac{1}{8} = 0.544$$



$$SplitInfo(set, C) = -\frac{2}{8}\log_2\frac{2}{8} - \frac{2}{8}\log_2\frac{2}{8} - \frac{2}{8}\log_2\frac{2}{8} - \frac{2}{8}\log_2\frac{2}{8} = 2$$

C4.5

usefulness of an input $\propto$ Gain, $\frac{1}{B.F}$

- Alternative Measures for Selecting ~~Attributes~~ (3)
  - Considering not only Gain but also branching factor

$$GainInfo(set, A) = \frac{Gain(set, A)}{SplitInfo(set, A)}$$

$\approx$ measure of branching factor

나뉘어지는 가짓수

A	B	PLAY
a1	b1	Yes
a1	b2	Yes
a1	b3	Yes
a1	b4	No
a2	b5	No
a2	b6	No

$$GainInfo(set, A) = \frac{0.484}{0.918} = 0.527$$

b.f = 2

$$GainInfo(set, B) = \frac{1.000}{2.585} = 0.387$$

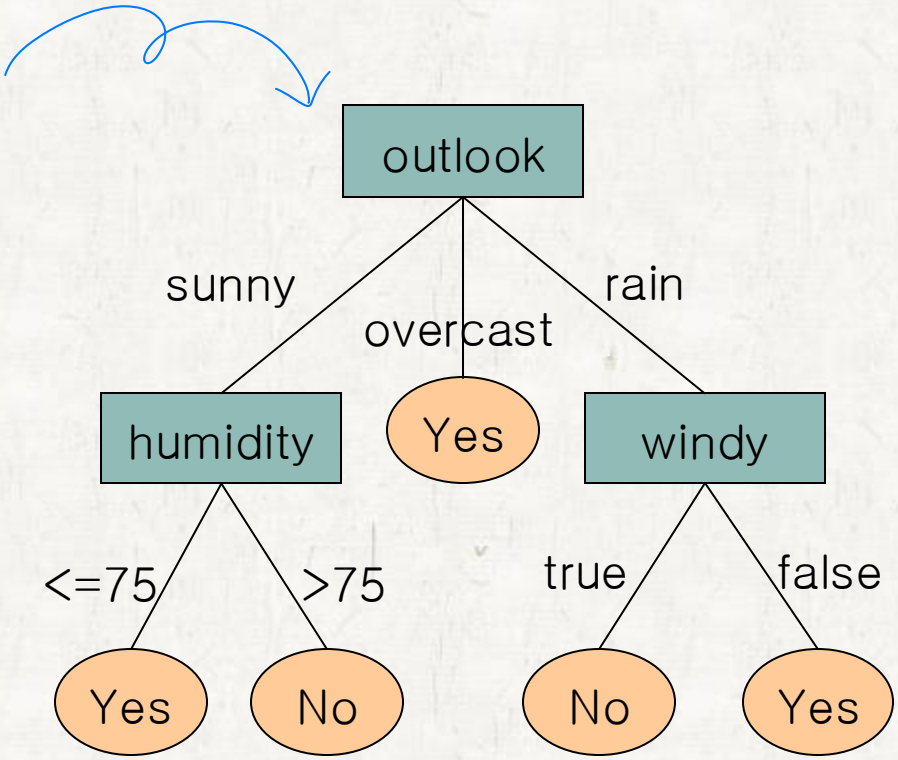
b.f = 6

A가 더 나은 듯

C4.5

- Incorporating Continuous-Valued Attributes (4)

OUTLOOK	TEMP.	HUMIDITY	WINDY	PLAY
sunny	85	85	false	No
sunny	80	90	true	No
overcast	83	78	false	Yes
rain	70	96	false	Yes
rain	68	80	false	Yes
rain	65	70	true	No
overcast	64	65	true	Yes
sunny	72	95	false	No
sunny	69	70	false	Yes
rain	75	80	false	Yes
sunny	75	70	true	Yes
overcast	72	90	true	Yes
overcast	81	75	false	Yes
rain	71	80	true	No

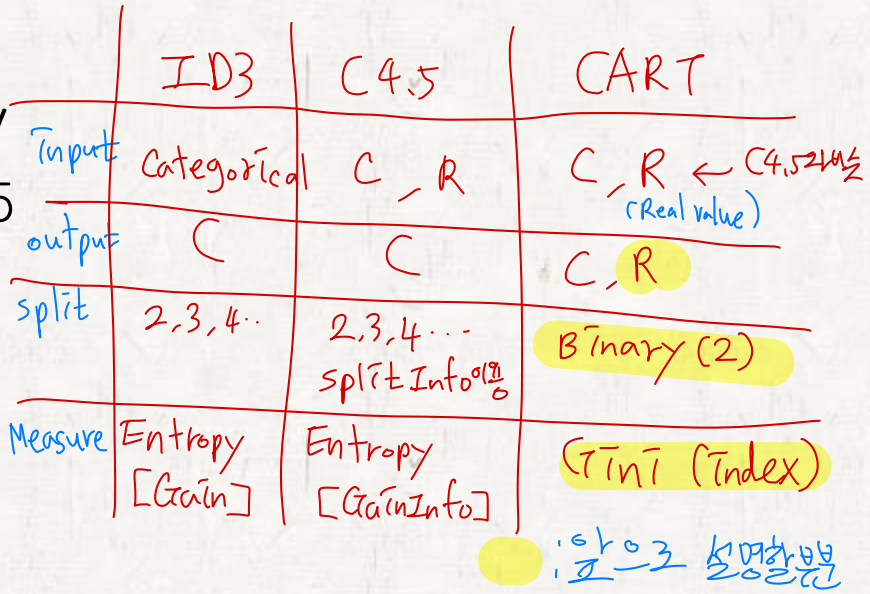


CART

- CART developed by Breiman Friedman Olsen and Stone: "Classification and Regression Trees"

- Gini Index instead of Entropy
  - All steps are very similar to C4.5

- Binary Split
  - Eg: Tuesday vs. Not Tuesday



- Real-Valued Output (Regression Tree)

CART

- Gini Index

$$i(p) = \sum_{i \neq j} p_i p_j = 1 - \sum_j p_j^2$$

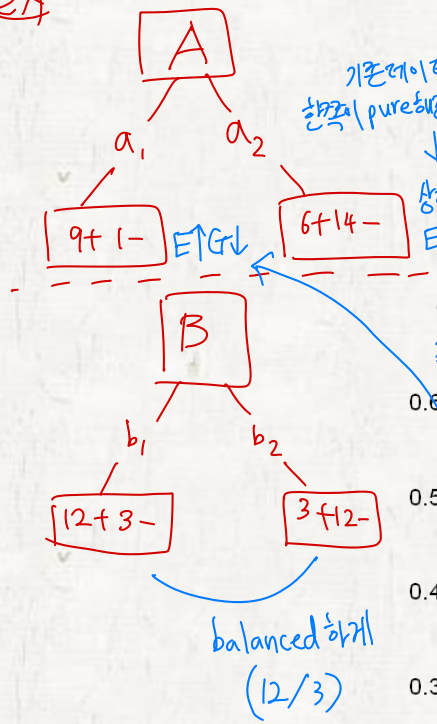
$$i(11111000) = \frac{5}{9}\cdot\frac{4}{9} + \frac{4}{9}\cdot\frac{5}{9} = 1 - \left[\left(\frac{5}{9}\right)^2 + \left(\frac{4}{9}\right)^2\right]$$

Entropy 의 종류이다.

$$E(T) = -\frac{5}{9}\log_2\frac{5}{9} - \frac{4}{9}\log_2\frac{4}{9}$$

$$Gini(T) = 1 - (p_1^2 + p_0^2) = 1 - \left(\left(\frac{5}{9}\right)^2 + \left(\frac{4}{9}\right)^2\right)$$

ex)



기존데이터고정하고 한쪽이 pure해진다면 다른한쪽 상대적으로 덜 pure 해짐

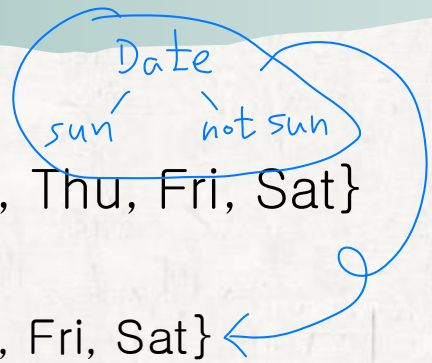
상대적으로 E↓ G↑

⇒ 균형잡힌 분할(B)를 선택할 것이다.

Max Value 보지말고 모양만보자!

similar하다

CART → 무조건 Binary split



- Binary Split
  - Date = {Sun, Mon, Tue, Wed, Thu, Fri, Sat}
  - Test all the split points
    - {Sun}, {Mon, Tue, Wed, Thu, Fri, Sat}
    - {Mon}, {Sun, Tue, Wed, Thu, Fri, Sat}
    - {Tue}, {Sun, Mon, Wed, Thu, Fri, Sat}
    - {Wed}, {Sun, Mon, Tue, Thu, Fri, Sat}
    - {Thu}, {Sun, Mon, Tue, Wed, Fri, Sat}
    - {Fri}, {Sun, Mon, Tue, Wed, Thu, Sat}
    - {Sat}, {Sun, Mon, Tue, Wed, Thu, Fri}

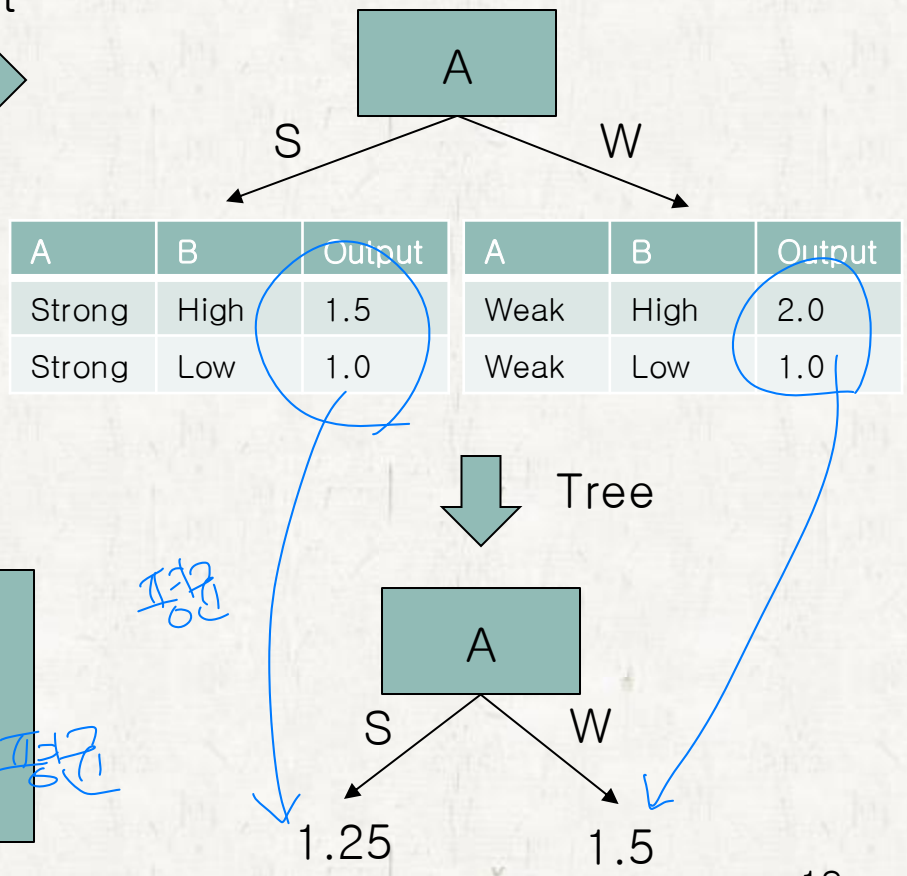
① Try all the split points
② choose the best split point

Regression Tree

- How to Build a Regression Tree

A	B	Output
Strong	High	1.5
Strong	Low	1.0
Weak	High	2.0
Weak	Low	1.0

Split



The output of node n

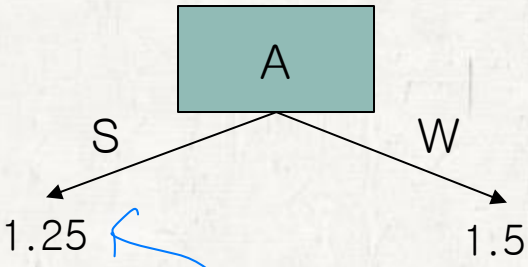
$$y_n = \frac{1}{|n|}\sum_{d \in n} t_d$$

Regression Tree

1. Evaluate the entropy of the given table
2. Split the table with an input
3. Evaluate the entropy of each sub-table
4. Evaluate the average of the entropies
5. Evaluate the Gain: (1)-(4)

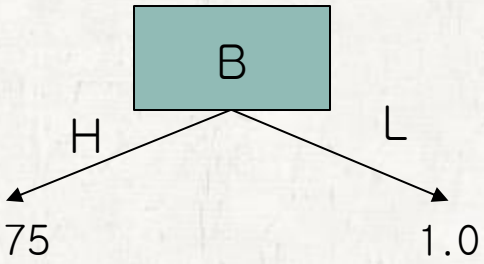
* Repeat this for all inputs and choose the best

- How to Build a Tree
  - Step1: Split the dataset with respect to all the attributes



A	B	Output
Strong	High	1.5
Strong	Low	1.0

A	B	Output
Weak	High	2.0
Weak	Low	1.0



A	B	Output
Strong	High	1.5
Weak	High	2.0

A	B	Output
Strong	Low	1.0
Weak	Low	1.0

target: 1.5
prediction: 1.25
$\rightarrow (1.5-1.25)^2$

target: 1.0
prediction: 1.25
$\rightarrow (1.0-1.25)^2$

Which split is better?

Instead of Entropy, we use MSE

$\Rightarrow MSE : \sum_{i=1}^{n} (\hat{y}-y)^2$

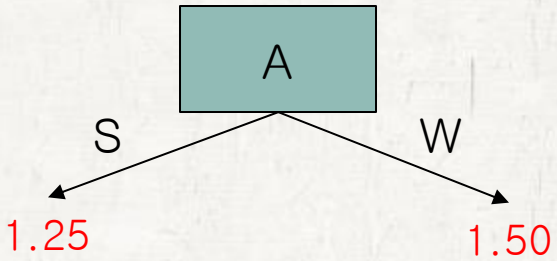
Regression Tree

1. Split the table with an input
2. Evaluate the MSEs of each sub-table
3. Evaluate the sum of MSEs

* Repeat this for all inputs and choose the best

Entropy대신에 사용

- How to Build a Tree
  - Step2: Evaluate the error of trees with respect to training data



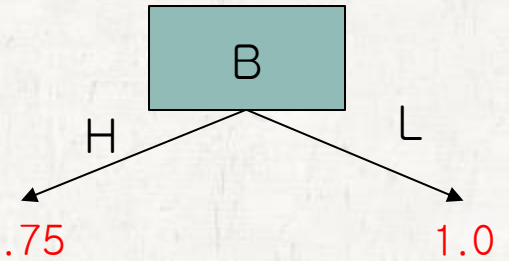
A	B	Output
Strong	High	1.5
Strong	Low	1.0

$E=(1.5-1.25)^2+(1.0-1.25)^2 = 0.125$

A	B	Output
Weak	High	2.0
Weak	Low	1.0

$E=(2.0-1.5)^2+(1.0-1.5)^2 = 0.5$

$$MSE = 0.125 + 0.5 = 0.625$$



A	B	Output
Strong	High	1.5
Weak	High	2.0

$E=(1.5-1.75)^2+(2.0-1.75)^2 = 0.125$

A	B	Output
Strong	Low	1.0
Weak	Low	1.0

$E=(1.0-1.0)^2+(1.0-1.0)^2 = 0.0$

둘을 비교

$$MSE = 0.125 + 0 = 0.125$$

less Error!

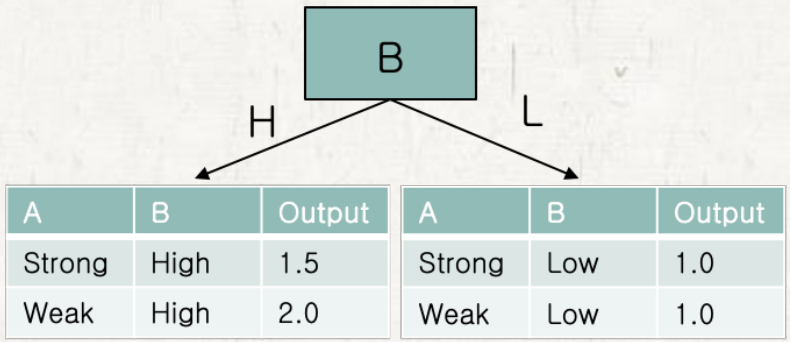
$$MSE = \sum_{i \in L} (y_i - \bar{y}_L)^2 + \sum_{i \in R} (y_i - \bar{y}_R)^2$$

=분산

Regression Tree

- How to Build a Tree
  - Step3: Choose the attr. with the smallest MSE, and repeat

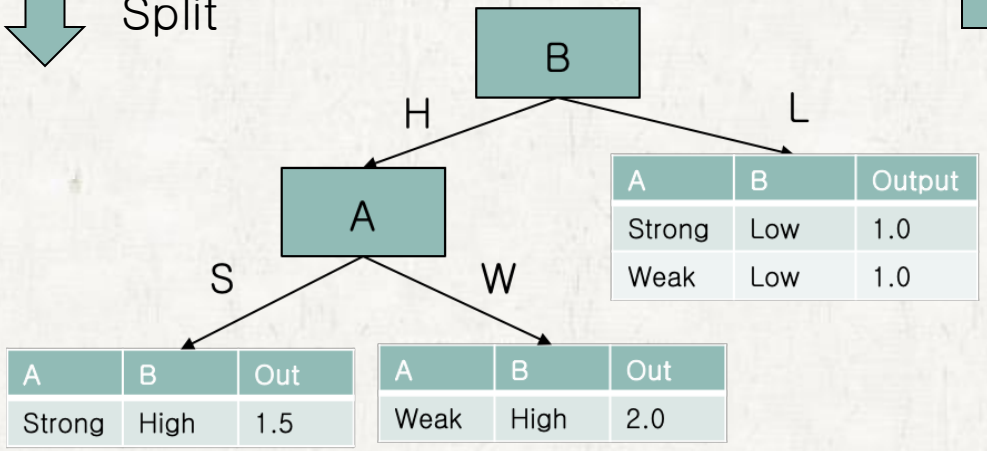
MSE가 0이 될 때까지 반복



A	B	Output
Strong	High	1.5
Weak	High	2.0

A	B	Output
Strong	Low	1.0
Weak	Low	1.0

Split

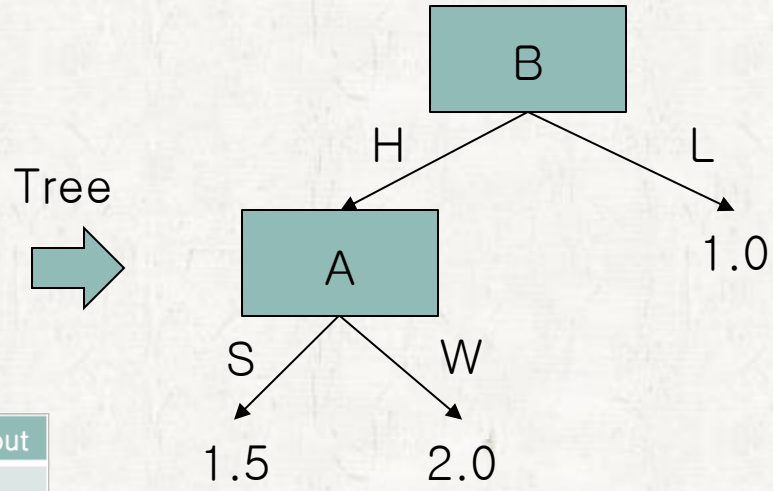


A	B	Output
Strong	Low	1.0
Weak	Low	1.0

A	B	Out
Strong	High	1.5

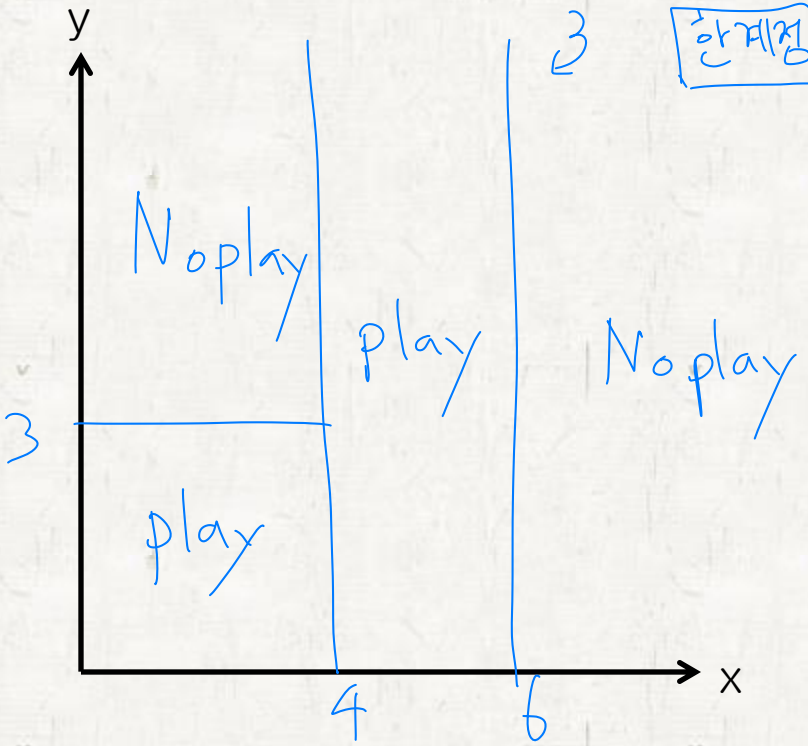
A	B	Out
Weak	High	2.0

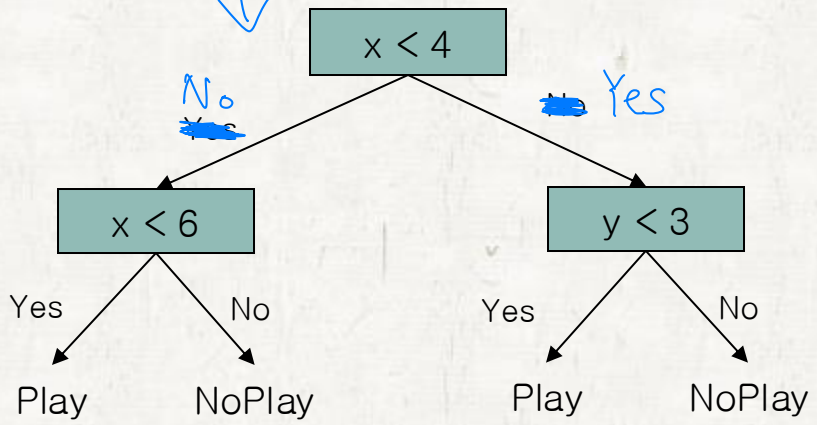
Tree



Discussion

- How a Decision Splits Input Space
  - Continuous attribute case





Input1	Input2	output
21	y	play
10	5	No
5	17	No
6	10	yes