Decision Tree

중간 범위X

Regression

이전에 배운 Logistic Regression

Naïve Beyesian

그와는 달리 수학적 개념이 사용X

Tennis Playing

- Tennis playing record

	Outlook	Temp	Humidity	Wind	Play
1	Sunny	Hot	High	Weak	No
2	Sunny	Hot	High	Strong	No
3	Overcast	Hot	High	Weak	Yes
4	Rain	Mild	High	Weak	Yes
5	Rain	Cool	Normal	Weak	Yes
6	Rain	Cool	Normal	Strong	No
7	Overcast	Cool	Normal	Strong	Yes
8	Sunny	Mild	High	Weak	No
9	Sunny	Cool	Normal	Weak	Yes
10	Rain	Mild	Normal	Weak	Yes
11	Sunny	Mild	Normal	Strong	Yes
12	Overcast	Mild	High	Strong	Yes
13	Overcast	Hot	Normal	Weak	Yes
14	Rain	Mild	High	Strong	No

Tennis Playing

- Today is Sunny, Mild, High and Strong
  - Will he play tennis today?
- How can you predict his tennis playing?
- What is the rule for playing tennis?
  - How to extract patterns from the data?

Basic Idea

- We have a simple data

Day	A	B	C	Play
1	T	F	T	T
2	T	F	F	T
3	F	T	F	T
4	F	F	T	F
5	F	F	F	F

- Will he play tennis if A=F, B=T, C=T ?
- Will he play tennis if A=T, B=T, C=F ?

Basic Idea

① choose
②

- A simple idea: See one of inputs and answer based on the majority

How evaluate to How Important?

Day	A	B	C	Play
1	T	F	T	T
2	T	F	F	T
3	F	T	F	T
4	F	F	T	F
5	F	F	F	F

A=T

Day	A	B	C	Play
1	T	F	T	T
2	T	F	F	T

2T 0F
100%

A=F

Day	A	B	C	Play
3	F	T	F	T
4	F	F	T	F
5	F	F	F	F

1T 2F
67%
not bad

Basic Idea

- A simple idea: See one of inputs and answer based on the majority

Day	A	B	C	Play
1	T	F	T	T
2	T	F	F	T
3	F	T	F	T
4	F	F	T	F
5	F	F	F	F

B=T

Day	A	B	C	Play
3	F	T	F	T

100%

B=F

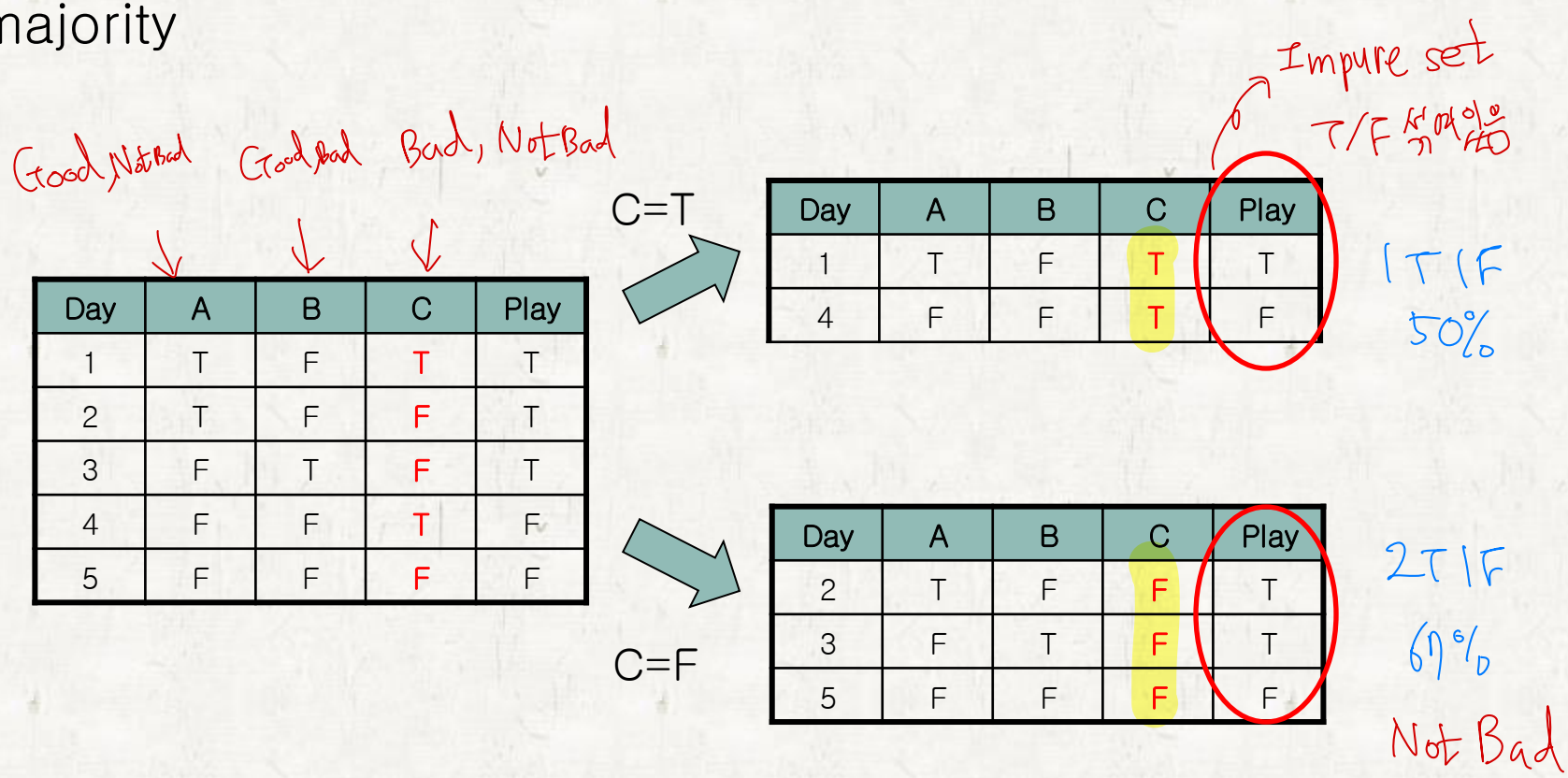
Day	A	B	C	Play
1	T	F	T	T
2	T	F	F	T
4	F	F	T	F
5	F	F	F	F

2T2F
50%

Not useful
A가 다냈다!

Basic Idea

- A simple idea: See one of inputs and answer based on the majority



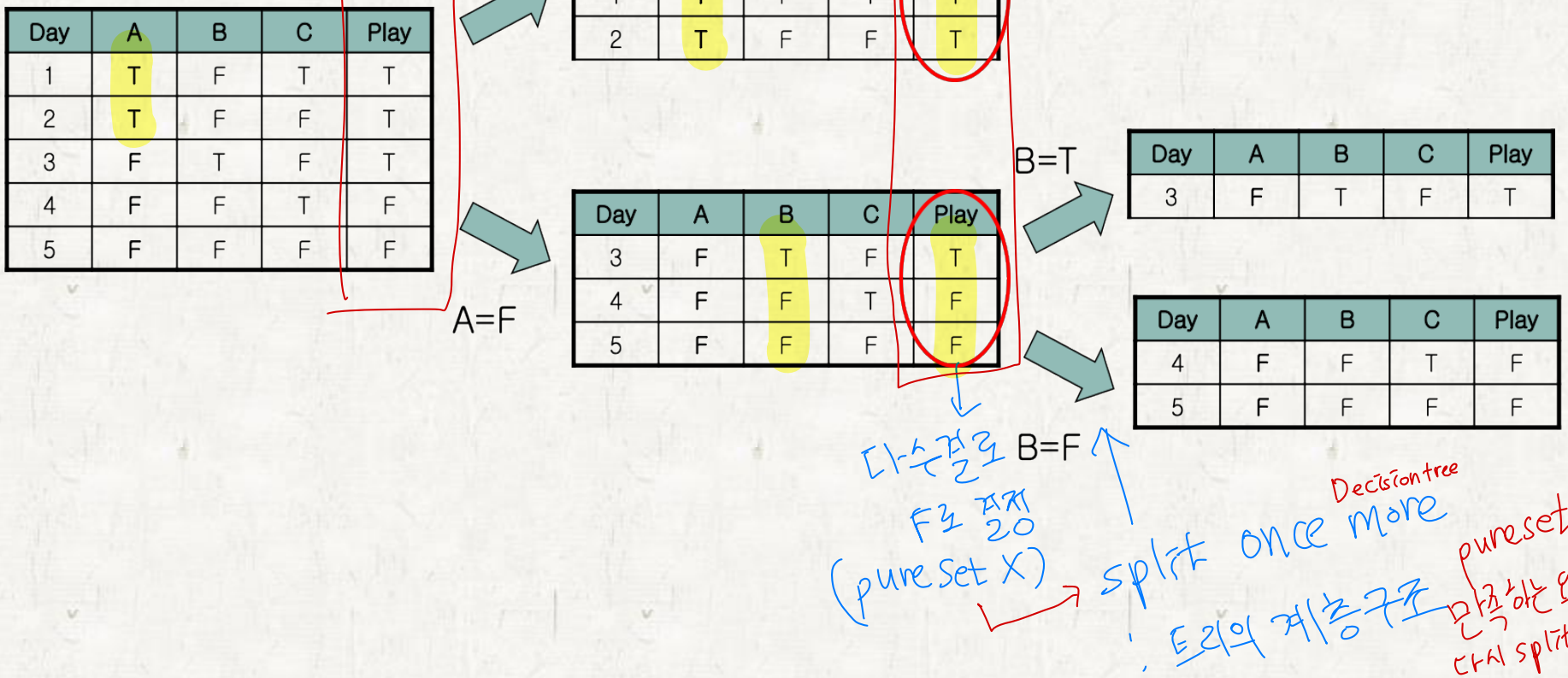
Day	A	B	C	Play
1	T	F	T	T
2	T	F	F	T
3	F	T	F	T
4	F	F	T	F
5	F	F	F	F

Day	A	B	C	Play
1	T	F	T	T
4	F	F	T	F

Day	A	B	C	Play
2	T	F	F	T
3	F	T	F	T
5	F	F	F	F

Basic Idea

- A simple idea: See one of inputs and answer based on the majority



Day	A	B	C	Play
1	T	F	T	T
2	T	F	F	T
3	F	T	F	T
4	F	F	T	F
5	F	F	F	F

Day	A	B	C	Play
1	T	F	T	T
2	T	F	F	T

Day	A	B	C	Play
3	F	T	F	T
4	F	F	T	F
5	F	F	F	F

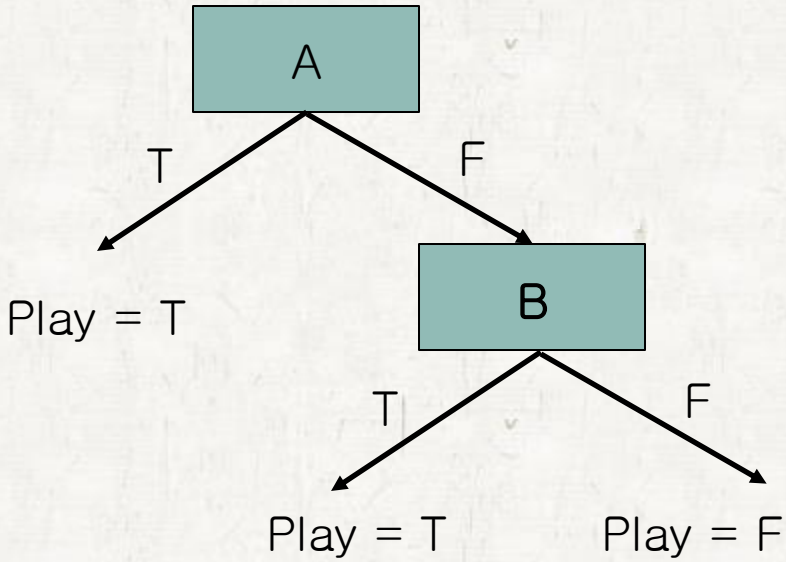
Day	A	B	C	Play
3	F	T	F	T

Day	A	B	C	Play
4	F	F	T	F
5	F	F	F	F

Basic Idea

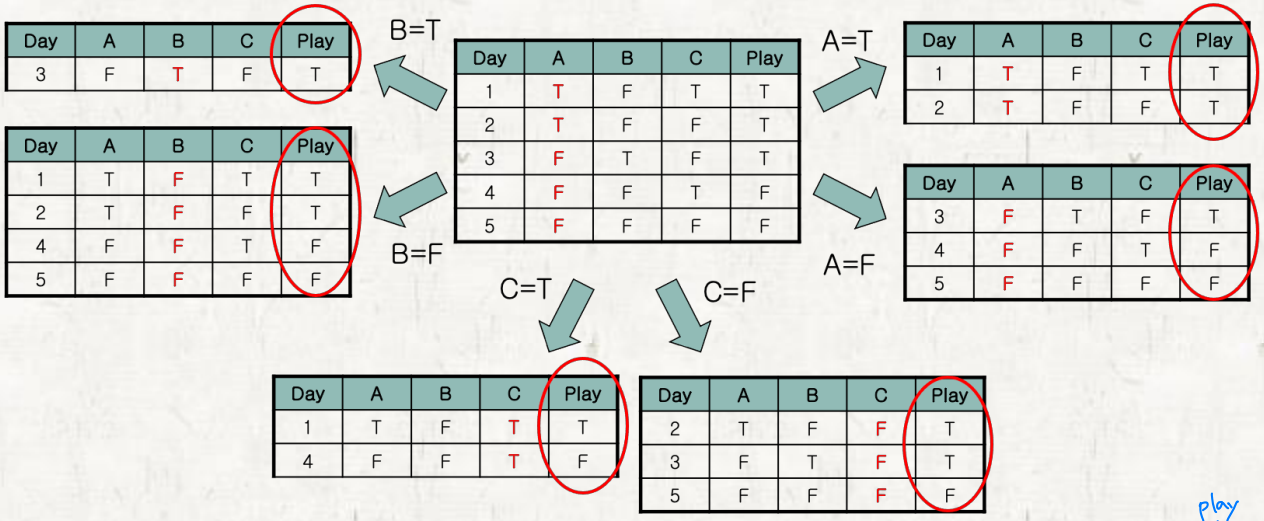
- A simple idea: See one of inputs and answer based on the majority

Day	A	B	C	Play
1	T	F	T	T
2	T	F	F	T
3	F	T	F	T
4	F	F	T	F
5	F	F	F	F



Attribute Selection for Split

- Which Attribute is the Best?



Day	A	B	C	Play
3	F	T	F	T

Day	A	B	C	Play
1	T	F	T	T
2	T	F	F	T
3	F	T	F	T
4	F	F	T	F
5	F	F	F	F

Day	A	B	C	Play
1	T	F	T	T
2	T	F	F	T
4	F	F	T	F
5	F	F	F	F

Day	A	B	C	Play
1	T	F	T	T
2	T	F	F	T

Day	A	B	C	Play
3	F	T	F	T
4	F	F	T	F
5	F	F	F	F

Day	A	B	C	Play
1	T	F	T	T
4	F	F	T	F

Day	A	B	C	Play
2	T	F	F	T
3	F	T	F	T
5	F	F	F	F

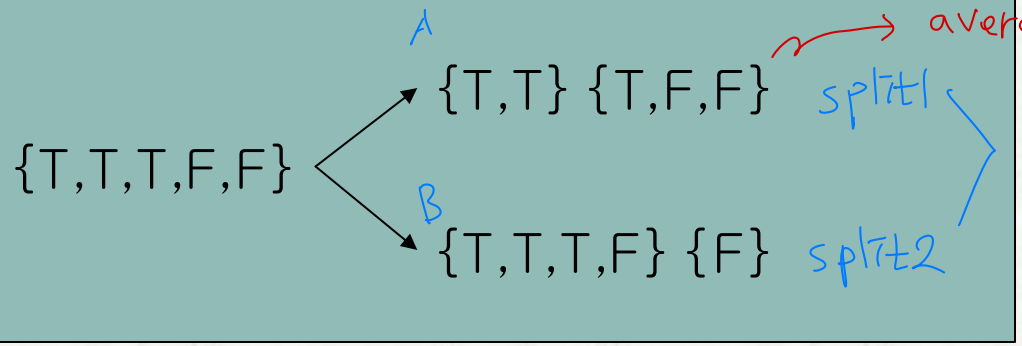
- We will define
  - Impurity of single sets: Entropy
  - Impurity of splits: Average of Entropy
- Then, present attribute selection measure: Gain

Impurity Measure

데이터가 많아지면 Impurity 측정 힘듦 → 측정방법을 공식화하자

- We need the impurity of splits

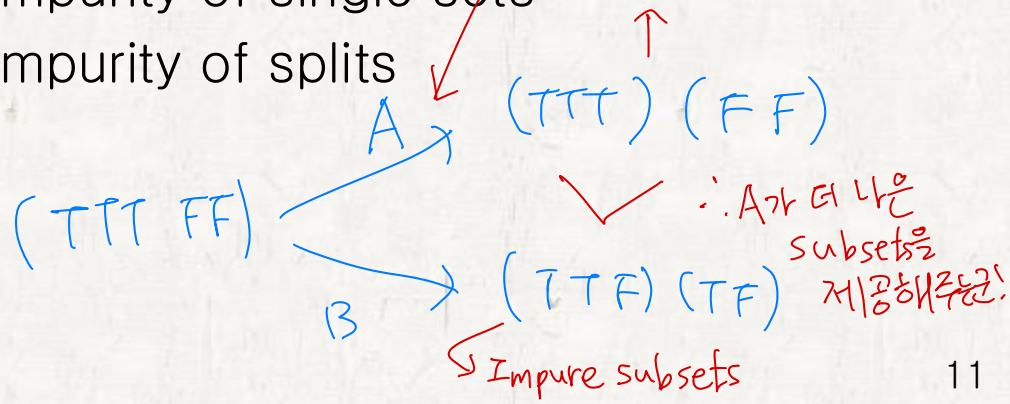
Which split is compose of purer subsets?



average of impurity of subsets 를 측정하여 A, B 비교

- Step1: We will define the impurity of single sets
- Step2: We will define the impurity of splits

much pure subsets



4step 중에 ①

Impurity Measure for Single Sets

- Let's define impurity measure for single sets

$$F(Set) = \text{Impurity of Set in } [0,1]$$

single element → pure (하나의 요소만 있으면) (portion of element)

불순성을 따지는 함수

Set_1: 1, 1, 1, 1, 1, 1, 1, 1	pure set (purest 가장) → 순서는 중요하지 X 같은 데이터가 있는지?
Set_2: 1, 1, 1, 1, 1, 1, 1, 0	
Set_3: 1, 1, 1, 1, 1, 1, 0, 0	
Set_4: 1, 1, 1, 1, 1, 0, 0, 0	
Set_5: 1, 1, 1, 1, 0, 0, 0, 0	Impure set (Impurest 가장)
Set_6: 1, 1, 1, 0, 0, 0, 0, 0	
Set_7: 1, 1, 0, 0, 0, 0, 0, 0	
Set_8: 1, 0, 0, 0, 0, 0, 0, 0	
Set_9: 0, 0, 0, 0, 0, 0, 0, 0	pure set (purest)

$F(Set_1) = 0$ All elements are 1's

$F(Set_5) = 1$ Half and half. Totally, mixed up

$F(Set_9) = 0$ All elements are 0's

→ 함수로 만들어보자 (찾아보자)

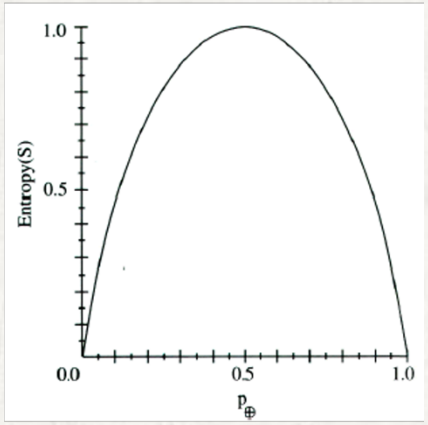
Impurity Measure for Single Sets

- Impurity of single sets

$$Entropy(set) = -P_1 \log_2 P_1 - P_0 \log_2 P_0$$

P_1 = the probability that 1 appears in *set*

P_0 = the probability that 0 appears in *set*



$P_1 = 0$

A={0,0,0,0,0,0,0,0} $Entropy(A) = -\frac{0}{8} \cdot \log_2 \frac{0}{8} - \frac{8}{8} \cdot \log_2 \frac{8}{8} = 0$

A={1,1,0,0,0,0,0,0} $Entropy(A) = -\frac{2}{8} \cdot \log_2 \frac{2}{8} - \frac{6}{8} \cdot \log_2 \frac{6}{8} = 0.81$

$P_1 = 1/2, P_0 = 1/2$

A={1,1,1,1,0,0,0,0} $Entropy(A) = -\frac{4}{8} \cdot \log_2 \frac{4}{8} - \frac{4}{8} \cdot \log_2 \frac{4}{8} = 1$

A={1,1,1,1,1,1,1,1} $Entropy(A) = -\frac{8}{8} \cdot \log_2 \frac{8}{8} - \frac{0}{8} \cdot \log_2 \frac{0}{8} = 0$

Impurity Measure for Single Sets

{aa bbb ccccc}
→ Impurity: $-\frac{2}{10}\log\frac{2}{10} - \frac{3}{10}\log\frac{3}{10} - \frac{5}{10}\log\frac{5}{10}$

- Impurity of single sets
  - If *set* is composed of $c_1, c_2, \ldots, c_n$

→ $-P_0\log P_0 - P_1\log P_1 - P_2\log P_2$

$$Entropy(set) = -\sum_{i=1}^{n} P_{c_i} \log_2 P_{c_i}$$

$$P_{c_i} = \text{the probability that } c_i \text{ appears in } set$$

서로다른 요소가 모두 같은 개수로 존재할 때 Maximum 가지만 이해!

more than two element

A={0,0,0,1,1,1,3,3,3} $Entropy(A) = -\frac{3}{9}\cdot\log_2\frac{3}{9} - \frac{3}{9}\cdot\log_2\frac{3}{9} - \frac{3}{9}\cdot\log_2\frac{3}{9}$

A={0,0,1,1,1,1,1,3,3} $Entropy(A) = -\frac{2}{9}\cdot\log_2\frac{2}{9} - \frac{5}{9}\cdot\log_2\frac{5}{9} - \frac{2}{9}\cdot\log_2\frac{2}{9}$

4step 중에 (2)

Impurity Measure for Splits

- Impurity of splits

{T,T,T,F,F} → {T,T} {T,F,F}

5 elements

How can we measure the impurity of the split?

$$Impurity(Split) = \text{Average of impurites of its subsets}$$

subset1 subset2

$$Impurity(TT, TFF) = \frac{2}{5} \times Impurity(TT) + \frac{3}{5} \times Impurity(TFF)$$

↙ Entropy

$$= \frac{2}{5} \times \left(-\frac{2}{2} \cdot \log_2 \frac{2}{2} - \frac{0}{2} \cdot \log_2 \frac{0}{2} \right) + \frac{3}{5} \times \left(-\frac{1}{3} \cdot \log_2 \frac{1}{3} - \frac{2}{3} \cdot \log_2 \frac{2}{3} \right)$$

Impurity Measure for Splits

- Entropy of Splits
  - Example:

orginal set

1, 1, 1, 1, 0, 0, 0, 0

Split 1

1, 1, 0, 0, 0

1, 1, 0

$-P_1 \log_2 P_1 - P_0 \log_2 P_0 = -2/5 \log_2 2/5 - 3/5 \log_2 3/5 = 0.97$

$-P_1 \log_2 P_1 - P_0 \log_2 P_0 = -2/3 \log_2 2/3 - 1/3 \log_2 1/3 = 0.92$

$\frac{5}{8} \times 0.97 + \frac{3}{8} \times 0.92 = 0.95$

Impurity: 0.95 > 0.51 → B가 더 낫다!

Split 2

1, 1, 1, 1, 0

0, 0, 0

$-P_1 \log_2 P_1 - P_0 \log_2 P_0 = -4/5 \log_2 4/5 - 1/5 \log_2 1/5 = 0.72$

$-P_1 \log_2 P_1 - P_0 \log_2 P_0 = -3/3 \log_2 3/3 - 0/3 \log_2 0/3 = 0.0$

$\frac{5}{8} \times 0.72 + \frac{3}{8} \times 0 = 0.51$

Attribute Selection

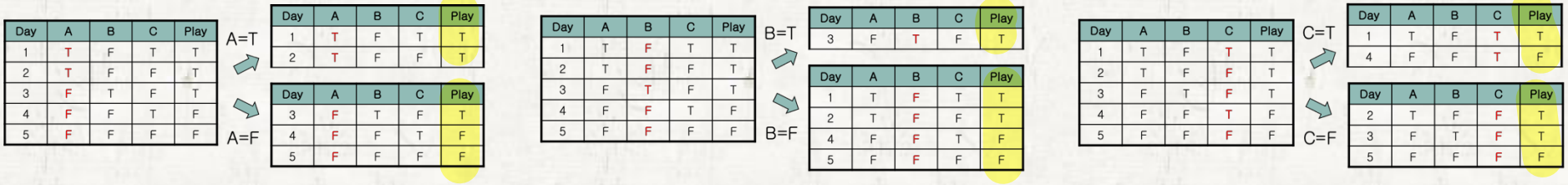
- Let's return back to a simple dataset

Day	A	B	C	Play
1	T	F	T	T
2	T	F	F	T
3	F	T	F	T
4	F	F	T	F
5	F	F	F	F

Play:{3T, 2F}

- Which attribute is good for "Split"

예전 슬라이드 참고해서 비교후 제일 Impurity 낮은거 choose



Selecting Attribute

- Gain: Measure how much an attribute reduce the entropy

original table

$$Gain(S, a) = Entropy(S) - Entopy\ of\ splits\ by\ a$$

chosen input to split

Day	A	B	C	Play
1	T	F	T	T
2	T	F	F	T
3	F	T	F	T
4	F	F	T	F
5	F	F	F	F

A=T

Day	A	B	C	Play
1	T	F	T	T
2	T	F	F	T

A=F

Day	A	B	C	Play
3	F	T	F	T
4	F	F	T	F
5	F	F	F	F

$$Gain(Play, A) = E(TTTFF) - E\ (TT, TFF)$$

Entropy

Day	A	B	C	Play
1	T	F	T	T
2	T	F	F	T
3	F	T	F	T
4	F	F	T	F
5	F	F	F	F

B=T

Day	A	B	C	Play
3	F	T	F	T

B=F

Day	A	B	C	Play
1	T	F	T	T
2	T	F	F	T
4	F	F	T	F
5	F	F	F	F

$$Gain(Play, B) = E(TTTFF) - E(T, TTFF)$$

Day	A	B	C	Play
1	T	F	T	T
2	T	F	F	T
3	F	T	F	T
4	F	F	T	F
5	F	F	F	F

C=T

Day	A	B	C	Play
1	T	F	T	T
4	F	F	T	F

C=F

Day	A	B	C	Play
2	T	F	F	T
3	F	T	F	T
5	F	F	F	F

$$Gain(Play, C) = E(TTTFF) - E(TF, TTF)$$

Selecting Attribute

- Measure the entropy of the original set

$$E(TTTFF) = -P_T \log_2 P_T - P_F \log_2 P_F = -3/5 \cdot \log_2 3/5 - 2/5 \cdot \log_2 2/5 = 0.97$$

- Measure the entropy of each split

$$E(TT) = -P_T \log_2 P_T - P_F \log_2 P_F = -2/2 \cdot \log_2 2/2 - 0/2 \cdot \log_2 0/2 = 0.0$$
$$E(TFF) = -P_T \log_2 P_T - P_F \log_2 P_F = -1/3 \cdot \log_2 1/3 - 2/3 \cdot \log_2 2/3 = 0.92$$

$$E(TT, TFF) = \frac{2}{5} \times 0 + \frac{3}{5} \times 0.92 = 0.55$$

$$E(T) = -P_T \log_2 P_T - P_F \log_2 P_F = -1/1 \cdot \log_2 1/1 - 0/1 \cdot \log_2 0/1 = 0.0$$
$$E(TTFF) = -P_T \log_2 P_T - P_F \log_2 P_F = -2/4 \cdot \log_2 2/4 - 2/4 \cdot \log_2 2/4 = 1.0$$

$$E(T, TTFF) = \frac{1}{5} \times 0 + \frac{4}{5} \times 1.0 = 0.80$$

$$E(TF) = -P_T \log_2 P_T - P_F \log_2 P_F = -1/2 \cdot \log_2 1/2 - 1/2 \cdot \log_2 1/2 = 1.0$$
$$E(TTF) = -P_T \log_2 P_T - P_F \log_2 P_F = -2/3 \cdot \log_2 2/3 - 1/3 \cdot \log_2 1/3 = 0.92$$

$$E(TF, TTF) = \frac{2}{5} \times 1.0 + \frac{3}{5} \times 0.92 = 0.952$$

Selecting Attribute

- Gain: Measure how much an attribute reduce the entropy

$$Gain(S, a) = Entropy(S) - Entopy\ of\ splits\ by\ a$$

Confuseness SS Decrease of Impurity

Day	A	B	C	Play
1	T	F	T	T
2	T	F	F	T
3	F	T	F	T
4	F	F	T	F
5	F	F	F	F

A=T

Day	A	B	C	Play
1	T	F	T	T
2	T	F	F	T

A=F

Day	A	B	C	Play
3	F	T	F	T
4	F	F	T	F
5	F	F	F	F

$$Gain(Play, A) = 0.97 - 0.55 = 0.42$$

Day	A	B	C	Play
1	T	F	T	T
2	T	F	F	T
3	F	T	F	T
4	F	F	T	F
5	F	F	F	F

B=T

Day	A	B	C	Play
3	F	T	F	T

B=F

Day	A	B	C	Play
1	T	F	T	T
2	T	F	F	T
4	F	F	T	F
5	F	F	F	F

$$Gain(Play, B) = 0.97 - 0.80 = 0.17$$

Day	A	B	C	Play
1	T	F	T	T
2	T	F	F	T
3	F	T	F	T
4	F	F	T	F
5	F	F	F	F

C=T

Day	A	B	C	Play
1	T	F	T	T
4	F	F	T	F

C=F

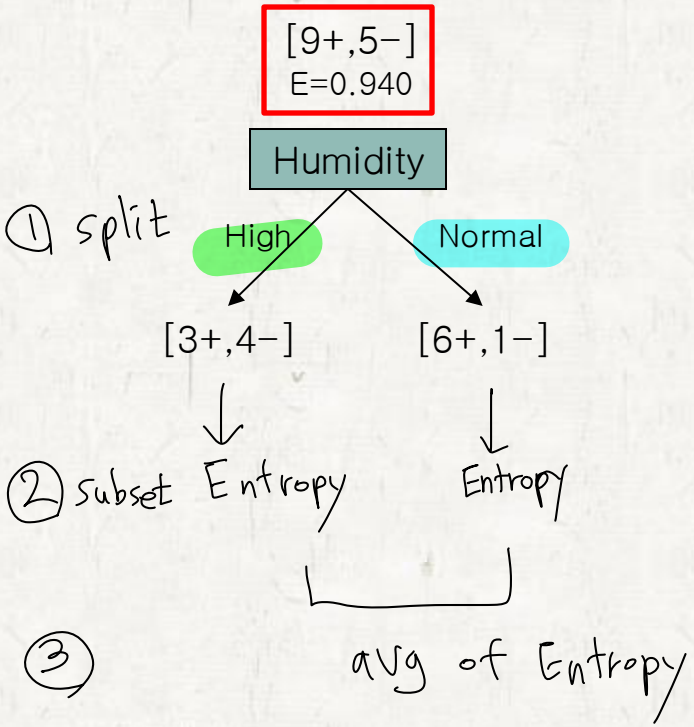
Day	A	B	C	Play
2	T	F	F	T
3	F	T	F	T
5	F	F	F	F

$$Gain(Play, C) = 0.97 - 0.952 = 0.018$$

Example

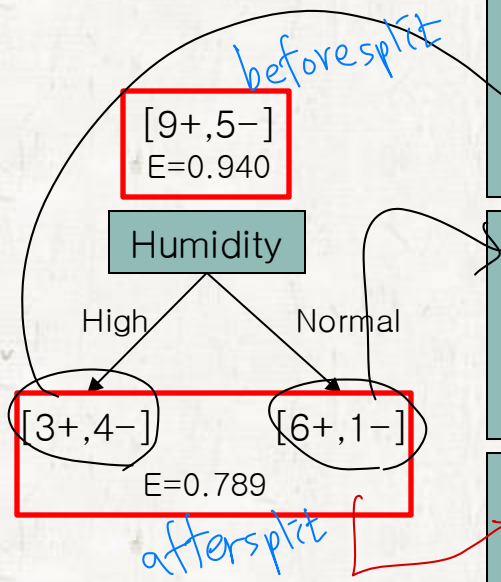
- Tennis playing record again

	Outlook	Temp	Humidity	Wind	Play
1	Sunny	Hot	High	Weak	No
2	Sunny	Hot	High	Strong	No
3	Overcast	Hot	High	Weak	Yes
4	Rain	Mild	High	Weak	Yes
5	Rain	Cool	Normal	Weak	Yes
6	Rain	Cool	Normal	Strong	No
7	Overcast	Cool	Normal	Strong	Yes
8	Sunny	Mild	High	Weak	No
9	Sunny	Cool	Normal	Weak	Yes
10	Rain	Mild	Normal	Weak	Yes
11	Sunny	Mild	Normal	Strong	Yes
12	Overcast	Mild	High	Strong	Yes
13	Overcast	Hot	Normal	Weak	Yes
14	Rain	Mild	High	Strong	No



Example

- Reduction of disorder after dividing
  - S is divided into two subsets
  - What is the average entropy of the two subsets ?



Entropy of the subset with Humidity = High

$$Entropy(S_{Humidity=High}) = -\frac{3}{7}\log_2\frac{3}{7} - \frac{4}{7}\log_2\frac{4}{7} = 0.985$$

Entropy of the subset with Humidity = Normal

$$Entropy(S_{Humidity=Normal}) = -\frac{6}{7}\log_2\frac{6}{7} - \frac{1}{7}\log_2\frac{1}{7} = 0.592$$

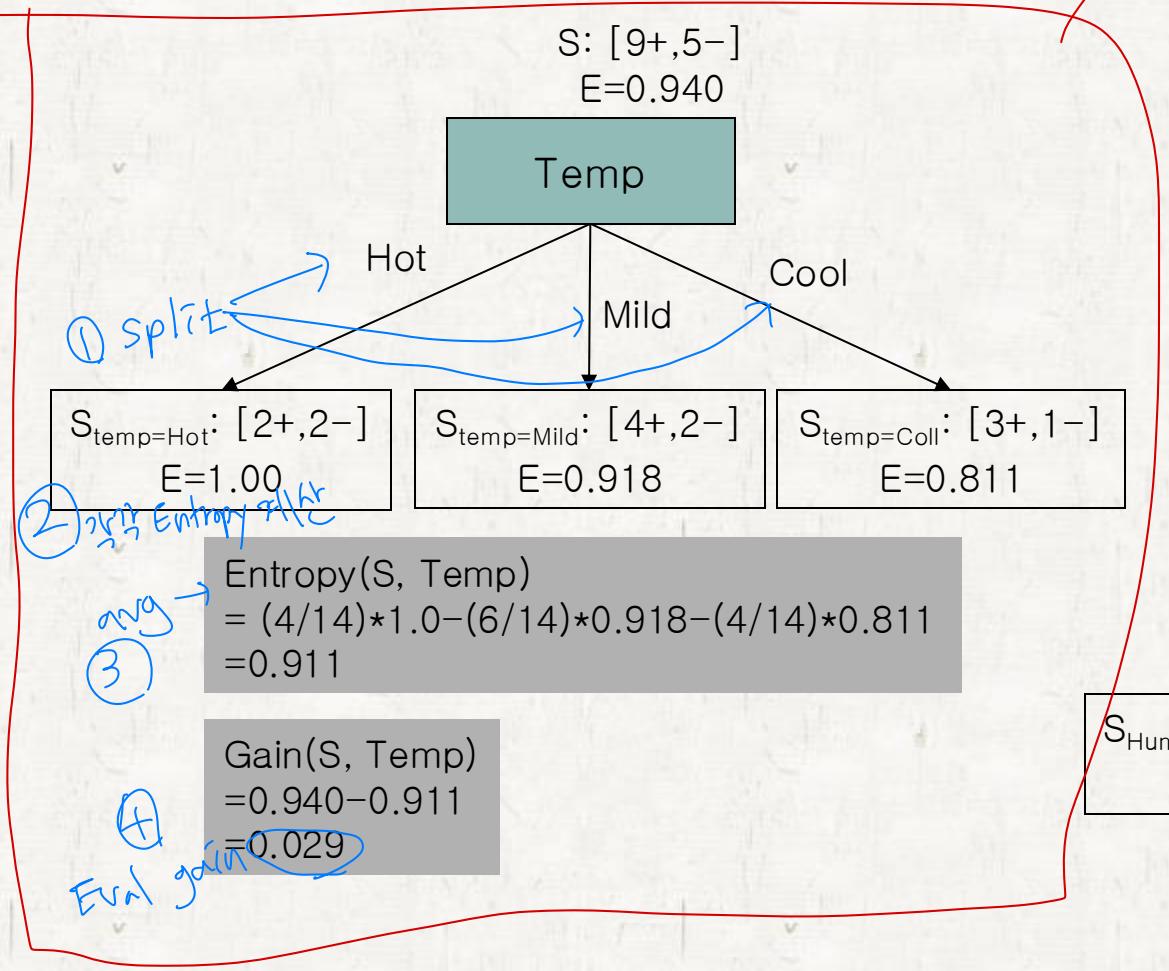
Average entropy of the two subset

$$\frac{7}{14}\times 0.985 + \frac{7}{14}\times 0.592 = 0.789$$

heterogeneity
E : impurity
(= confuseness)

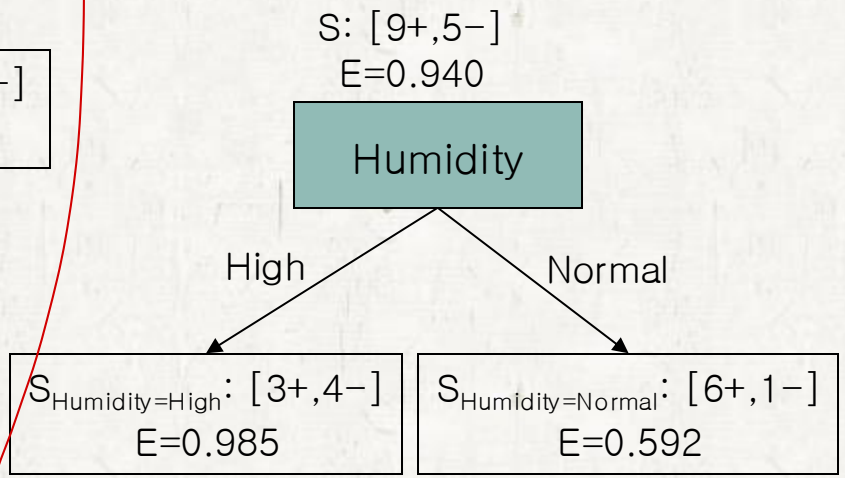
Example

- Which attribute is the best classifier?



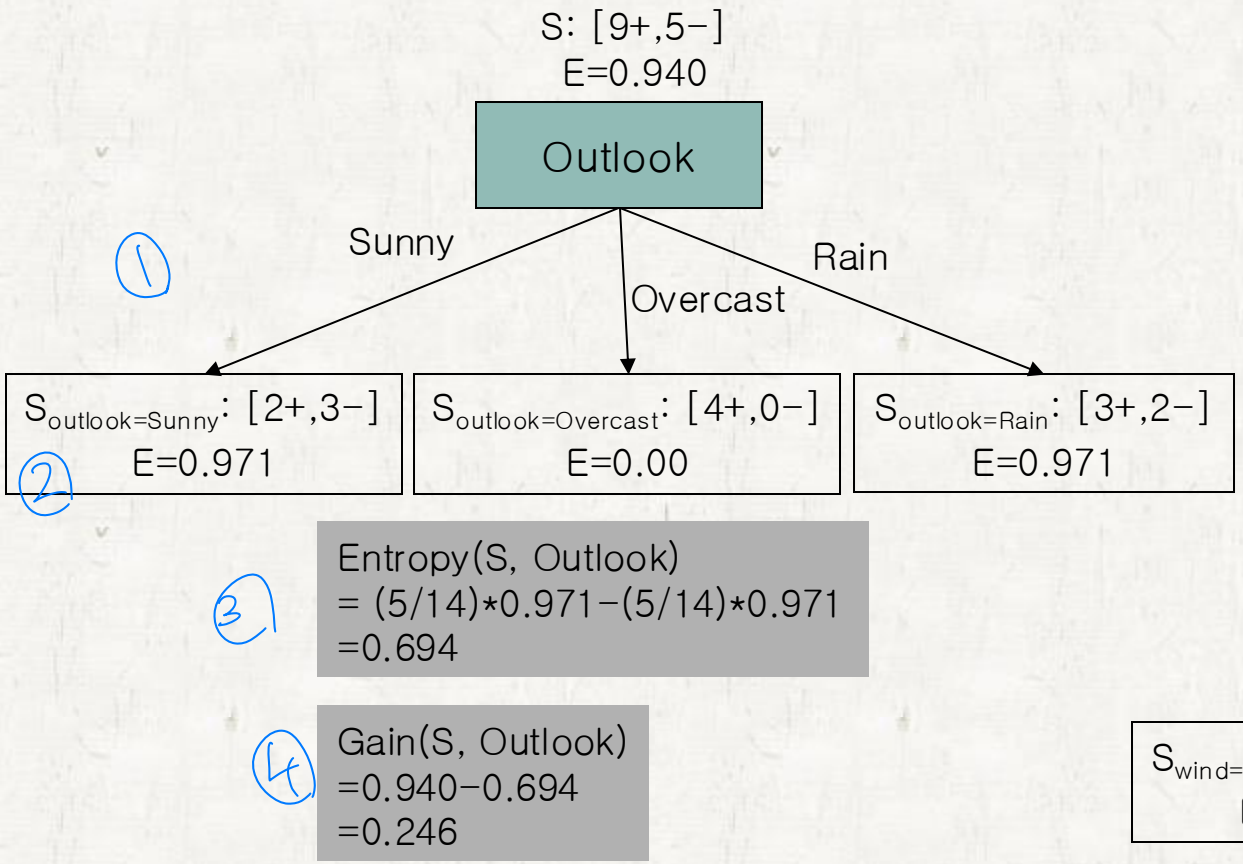
Entropy(S, Humidity)
= (7/14)*0.985−(7/14)*0.592
= 0.798

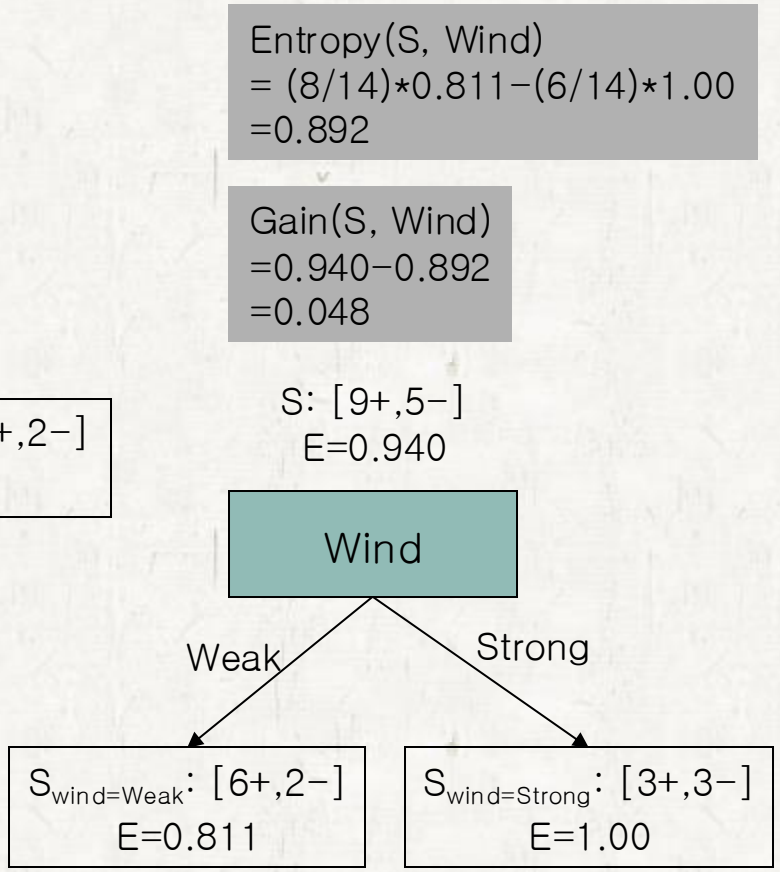
Gain(S, Humidity)
=0.940−0.798
=0.151



Example

- Which attribute is the best classifier?





Example

- Which attribute is the best classifier?
  - At root node, we will classify the data with “Outlook”

Gain(S, Outlook) = 0.246
Gain(S, Humidity) = 0.151
Gain(S, Wind) = 0.048
Gain(S, Temp) = 0.029

Example

- Which attribute is the best classifier?
  - At root node, we will classify the data with “Outlook”

	Outlook	Temp	Humidity	Wind	Play
4	Rain	Mild	High	Weak	Yes
5	Rain	Cool	Normal	Weak	Yes
6	Rain	Cool	Normal	Strong	No
10	Rain	Mild	Normal	Weak	Yes
14	Rain	Mild	High	Strong	No

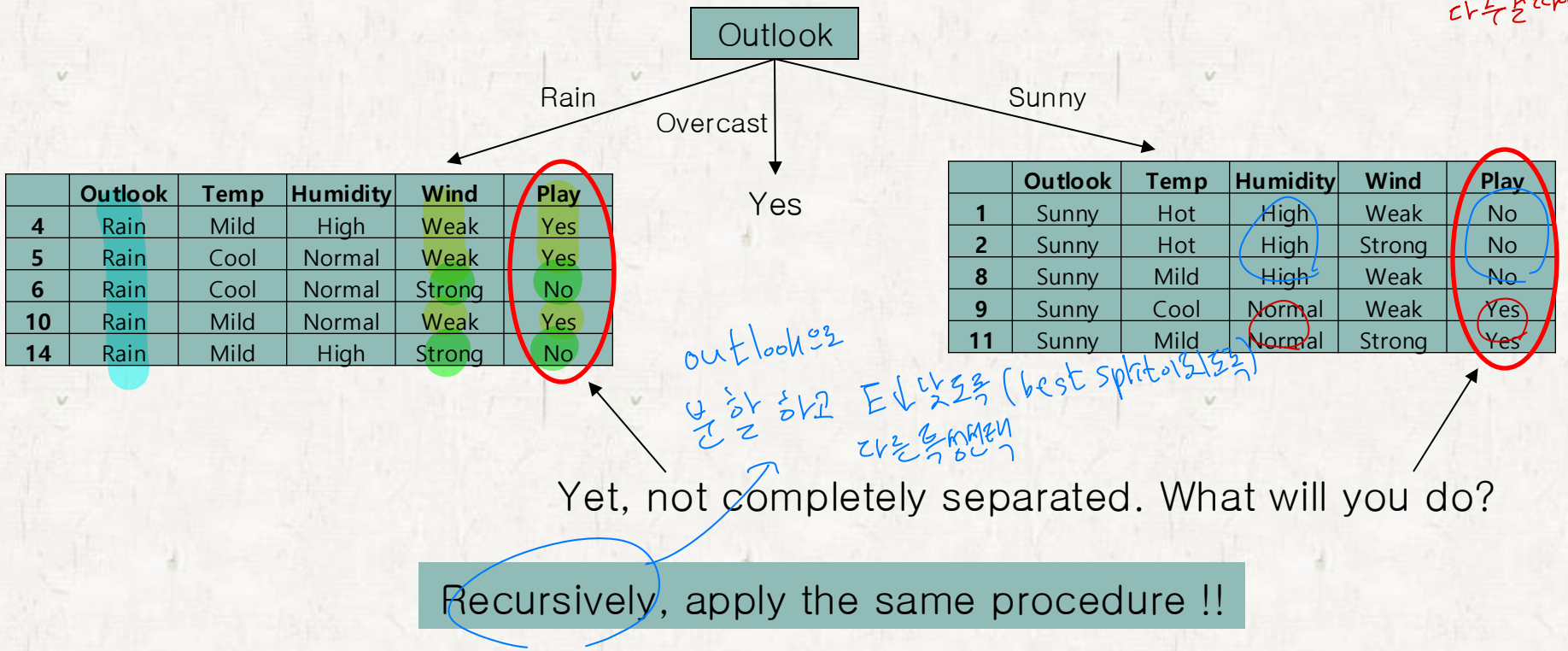
	Outlook	Temp	Humidity	Wind	Play
1	Sunny	Hot	High	Weak	No
2	Sunny	Hot	High	Strong	No
8	Sunny	Mild	High	Weak	No
9	Sunny	Cool	Normal	Weak	Yes
11	Sunny	Mild	Normal	Strong	Yes

	Outlook	Temp	Humidity	Wind	Play
3	Overcast	Hot	High	Weak	Yes
7	Overcast	Cool	Normal	Strong	Yes
12	Overcast	Mild	High	Strong	Yes
13	Overcast	Hot	Normal	Weak	Yes

all yes → E=0

Example

- Which attribute is the best classifier?
  - At root node, we will classify the data with “Outlook”

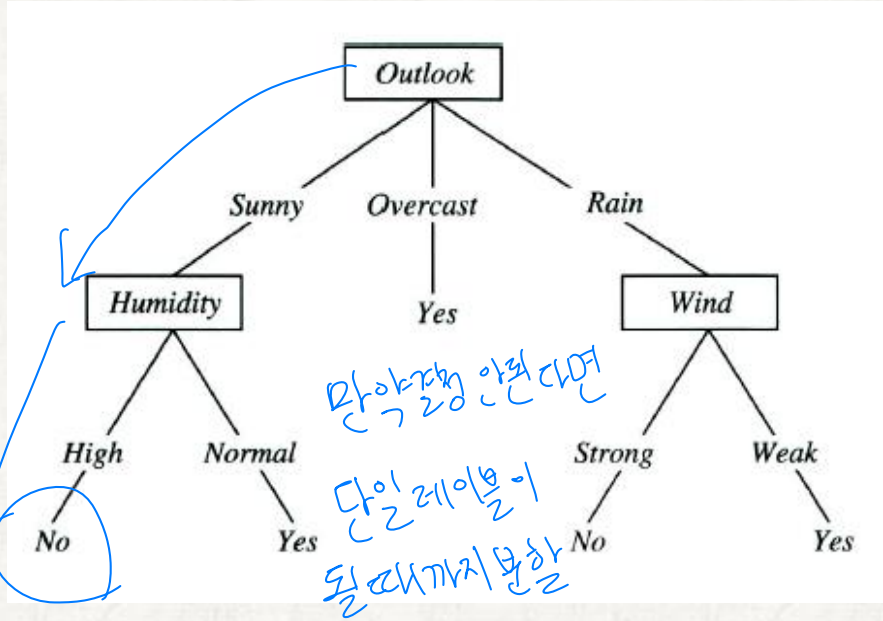


	Outlook	Temp	Humidity	Wind	Play
4	Rain	Mild	High	Weak	Yes
5	Rain	Cool	Normal	Weak	Yes
6	Rain	Cool	Normal	Strong	No
10	Rain	Mild	Normal	Weak	Yes
14	Rain	Mild	High	Strong	No

	Outlook	Temp	Humidity	Wind	Play
1	Sunny	Hot	High	Weak	No
2	Sunny	Hot	High	Strong	No
8	Sunny	Mild	High	Weak	No
9	Sunny	Cool	Normal	Weak	Yes
11	Sunny	Mild	Normal	Strong	Yes

Example

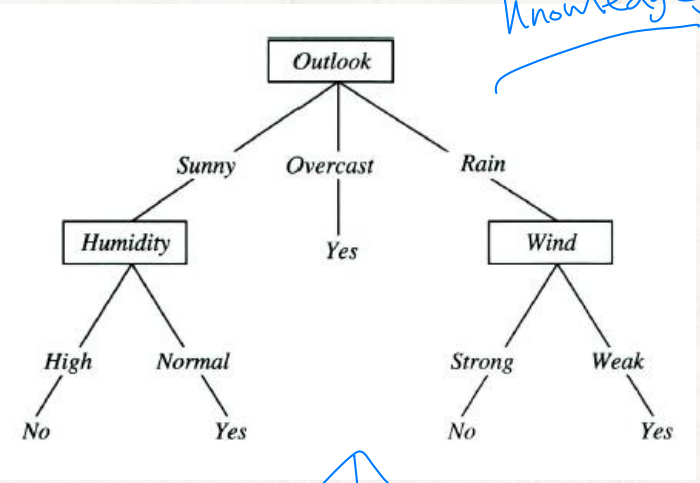
- Created Decision Tree



- Prediction
  - Sunny, Mild, High, Strong -> No

Example

- Rule Generation from Decision Tree



If Outlook = Sunny & Humidity = High
then Tennis = No

If Outlook = Sunny & Humidity = Normal
then Tennis = Yes

If Outlook = Overcast
then Tennis = Yes

If Outlook = Rain & Wind = Strong
then Tennis = No

If Outlook = Sunny & Wind = Weak
then Tennis = Yes

3개 Questions

- 2. There is the unique Decision Tree for a given dataset "No"
- Which tree is the best?
- 2. ID3 generates the best Decision Tree "No" ID3는 Best가 아님

3. ID3 guarantees "the best" tree "No"

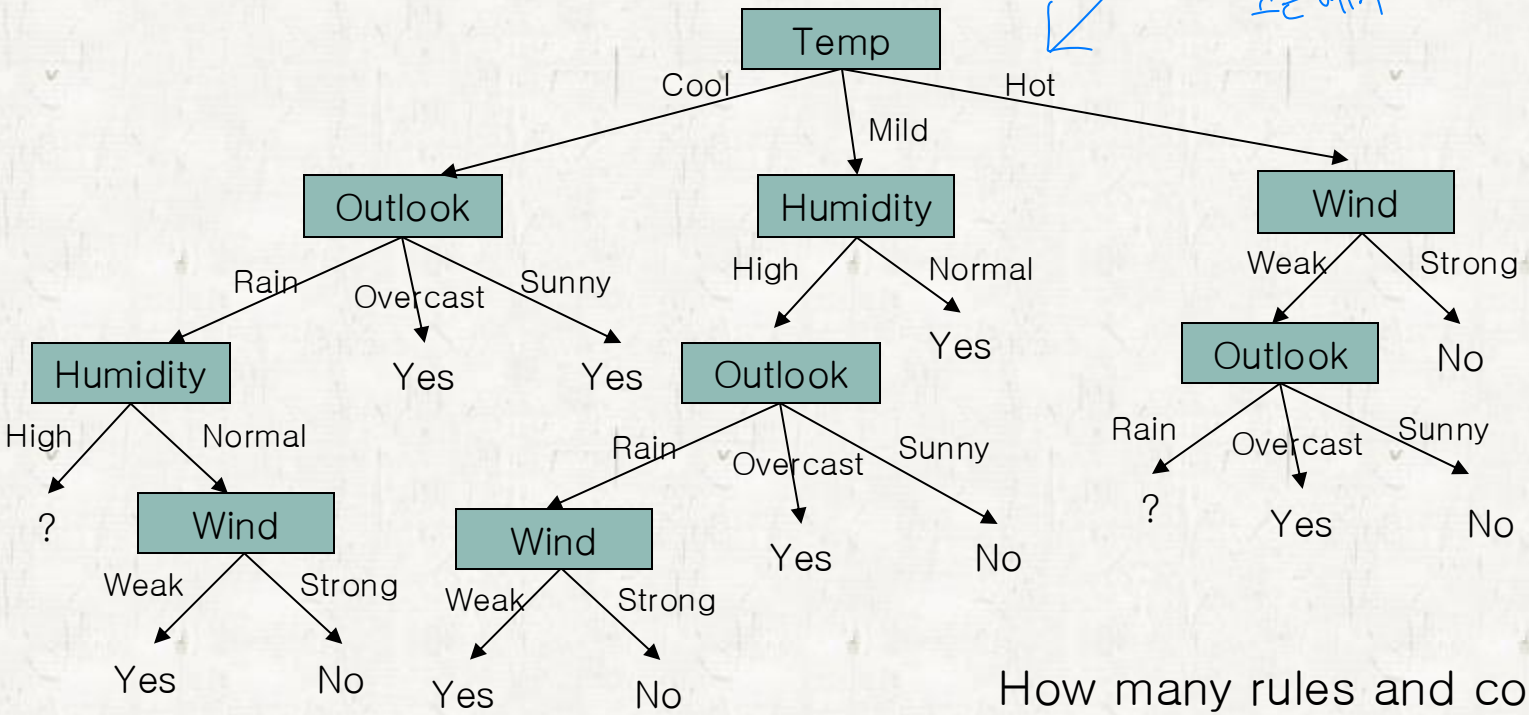
Questions

① choose an input
② split the table (pure table 될때까지)
③ Repeat this

Entropy, GINI 인덱스, Randomly (best아닌 가능성)

- Decision Tree by ID3 is not the unique decision tree for given data

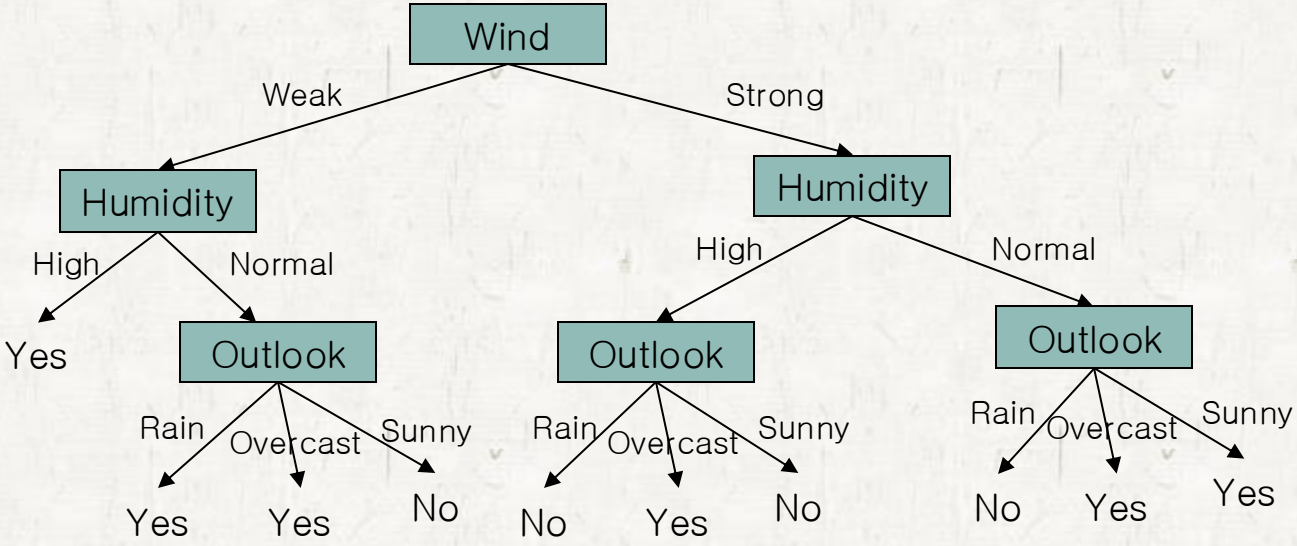
D.T(Randomly) 고른 예시



How many rules and conditions?

Questions

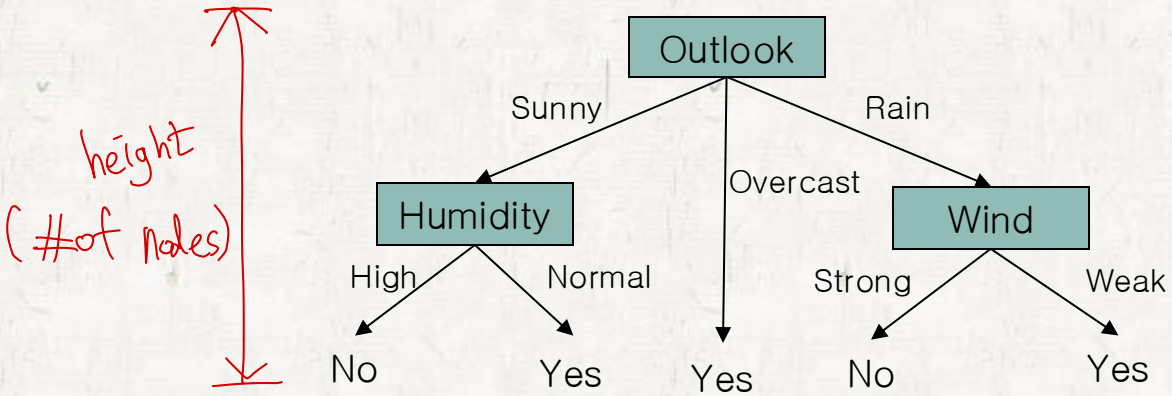
- Decision Tree by ID3 is not unique decision tree for given data



How many rules and conditions?

Questions

- Decision Tree by ID3 is not unique decision tree for given data



D.T (Entropy)

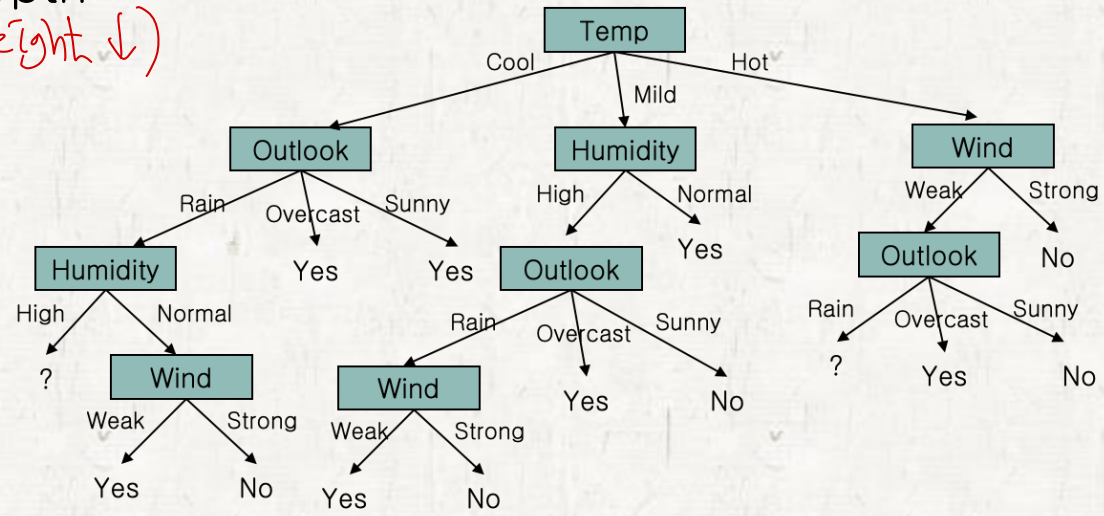
"The simpler is the better"

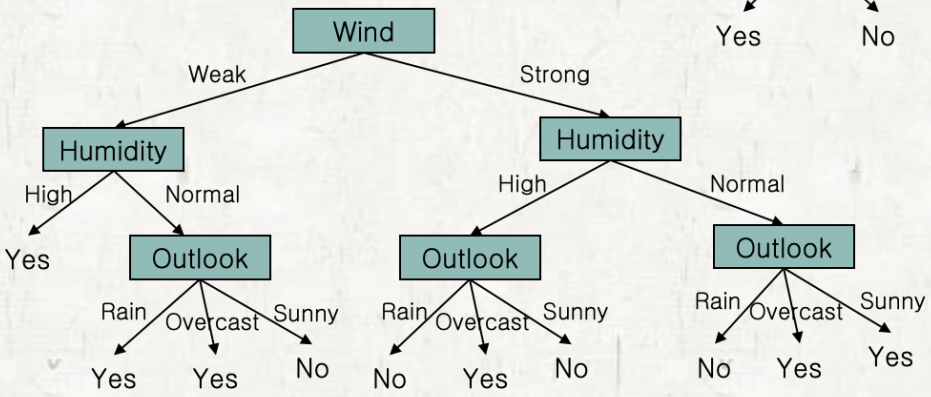
왜? 복잡한 것은 overfitting되기 쉽기때문에 (주어진 데이터에)

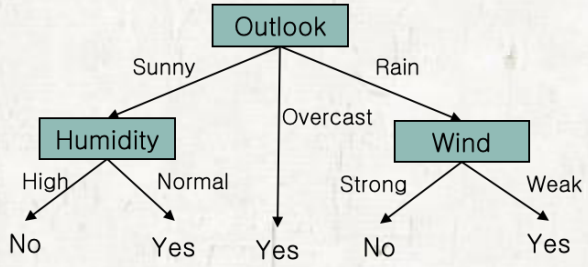
How many rules and conditions?

Questions

- Which tree is the best? = simple
  - With fewer nodes
  - With shallower depth (= short = height ↓)





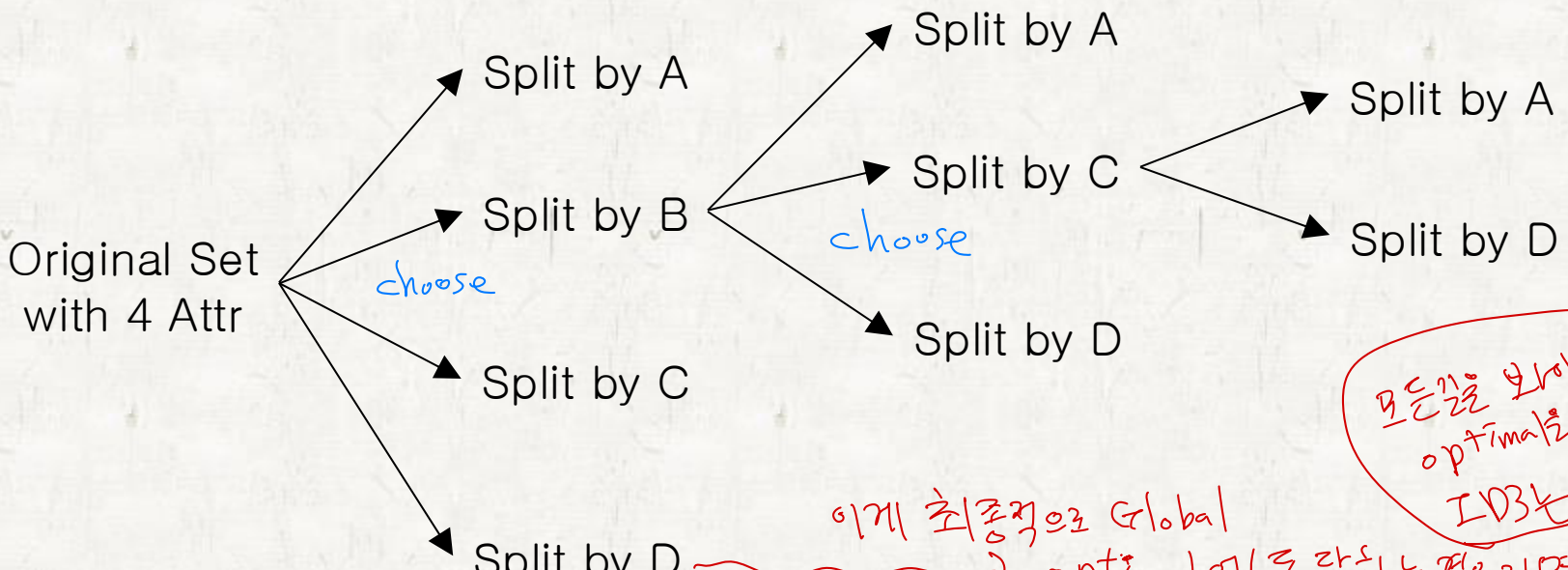


Questions

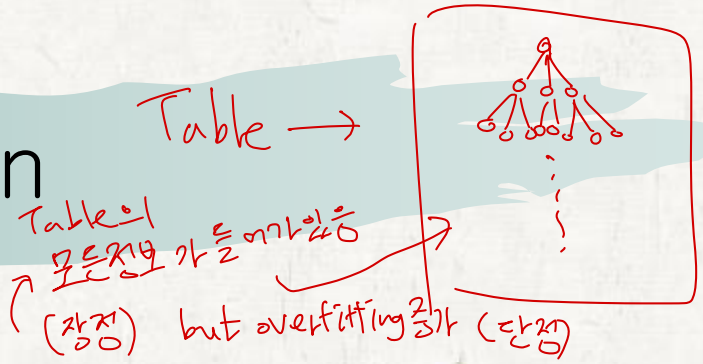
- Does ID3 generate the best Decision Tree?

Greedy Algorithm!!

이기때문에 optimal을 보장할수X
(back tracking 없기에)



Discussion

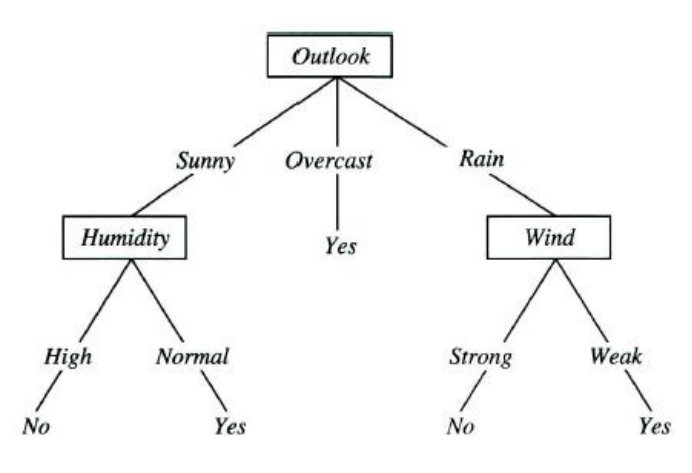


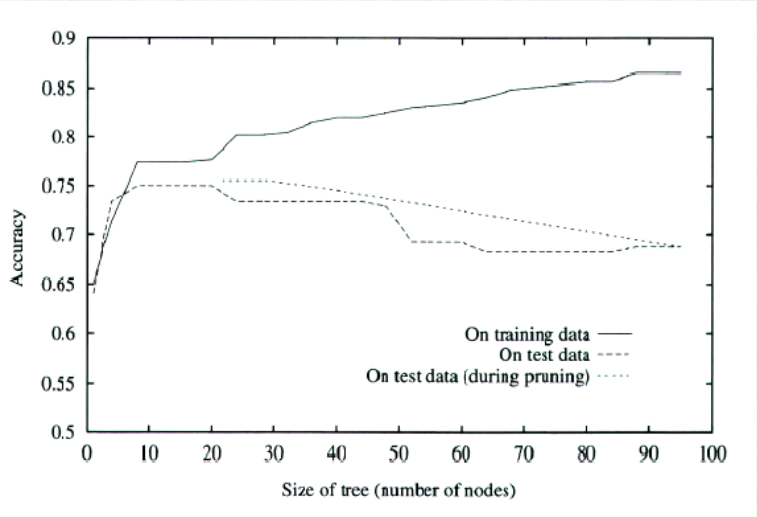
- Overfitting
  - As # of nodes is increasing, the Decision Tree fits to training data more and more, but possibly overfitted

- How to avoid overfitting
  - When generating: limit the maximum depth, or do not split nodes with small samples
  - After generating: Pruning

Discussion

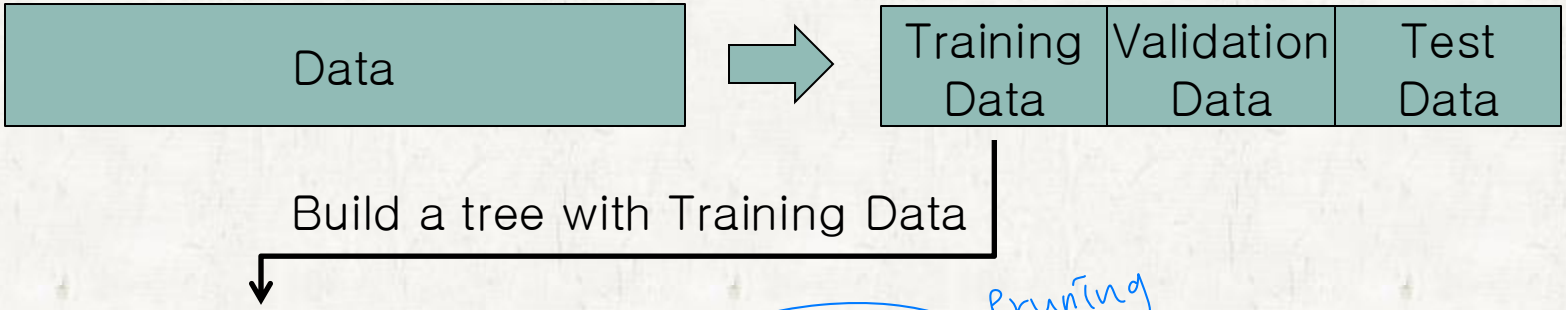
- Pruning
  - Reduced Error Pruning
    - Divide data into three set : training set, validation set, testing set
    - Create a decision tree with training set (allow overfit)
    - Prune a leaf node if removing it does not affect the classification of validation set
    - Do this until there is no more such nodes



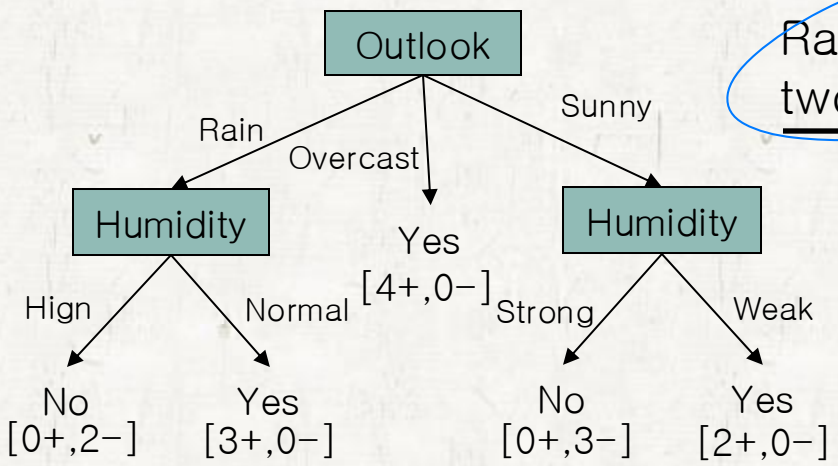


Discussion

- Pruning
  - Reduced Error Pruning

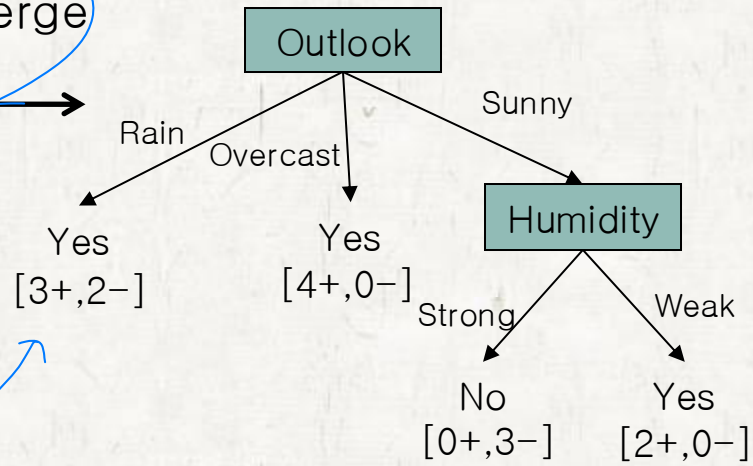


Build a tree with Training Data



Pruning

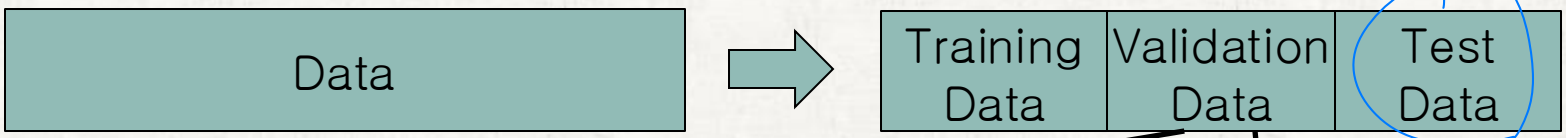
Randomly merge two nodes



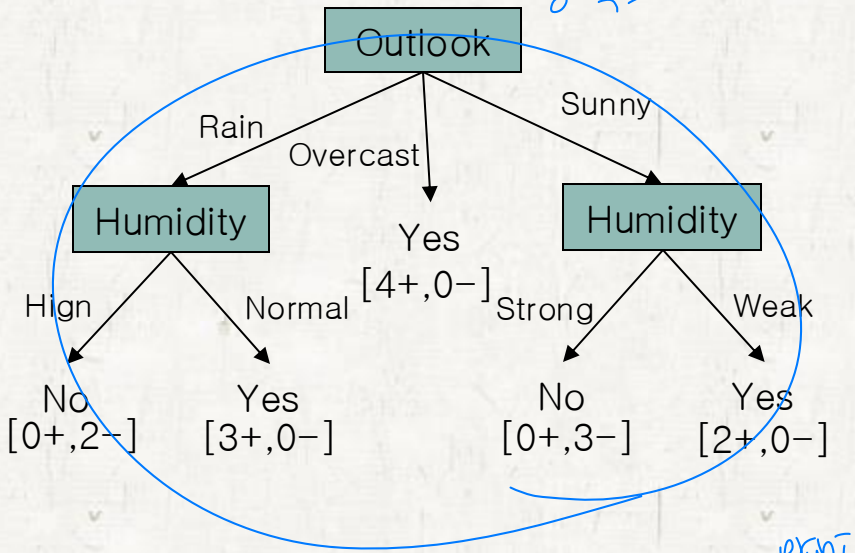
merge ⇒ [3+,2−] 다수결 yes

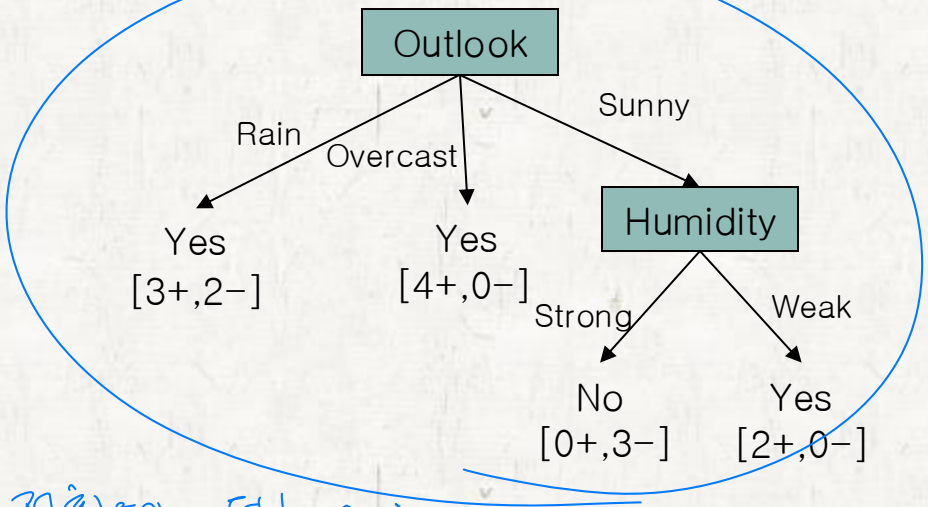
Discussion

- Pruning
  - Reduced Error Pruning



Evaluate trees with validation data





Discussion

optimal tree는 불가능하다

- Pruning
  - Post Rule Pruning (다른 방법)
    - Split data into training set and validation set
    - Build a consistent decision tree (allow overfit) (=full, complete) (리프노드 pure한 데이터 존재할때 (conflict 없을때))
    - Convert the tree into a set of "if-then" rules
    - For each rule, remove any attributes if estimated accuracy of validation set is not reduced (조건)

If (Outlook = Sunny) & (Humidity = High) then Tennis = No
If (Outlook = Sunny) & (Humidity = Normal) then Tennis = Yes
If (Outlook = Overcast) then Tennis = Yes
If (Outlook = Rain) & (Wind = Strong) then Tennis = No
If (Outlook = Sunny) & (Wind = Weak) then Tennis = Yes

⇒ 9개 조건 중에 1개를 지움 그후 정확도를 측정함

지우기전/후 정확도 중 높은것으로