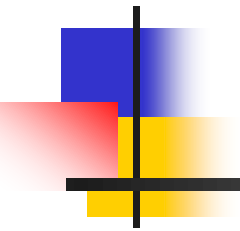


# 분류 및 예측 (1)
# : 의사결정나무(Decision Tree)

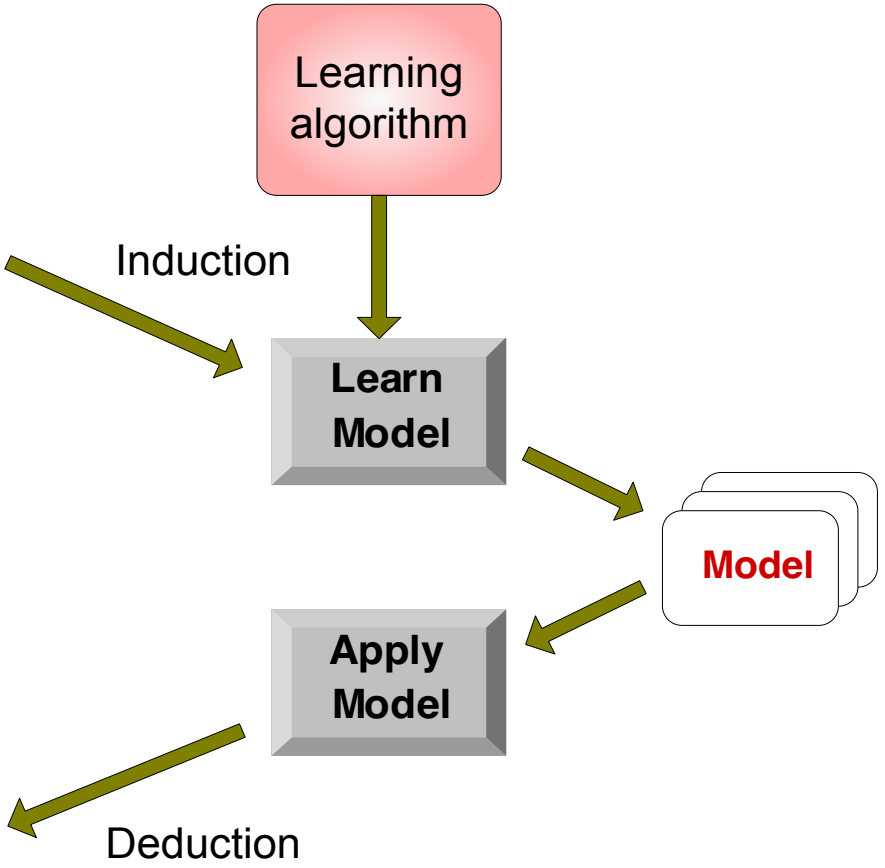
# 분류/예측 분석의 개념

| Tid | Attrib1 | Attrib2 | Attrib3 | Class |
|---|---|---|---|---|
| 1 | Yes | Large | 125K | No |
| 2 | No | Medium | 100K | No |
| 3 | No | Small | 70K | No |
| 4 | Yes | Medium | 120K | No |
| 5 | No | Large | 95K | Yes |
| 6 | No | Medium | 60K | No |
| 7 | Yes | Large | 220K | No |
| 8 | No | Small | 85K | Yes |
| 9 | No | Medium | 75K | No |
| 10 | No | Small | 90K | Yes |

Training Set

| Tid | Attrib1 | Attrib2 | Attrib3 | Class |
|---|---|---|---|---|
| 11 | No | Small | 55K | ? |
| 12 | Yes | Medium | 80K | ? |
| 13 | Yes | Large | 110K | ? |
| 14 | No | Small | 95K | ? |
| 15 | No | Large | 67K | ? |

Test Set




*Source: Introduction to Data Mining by Tan, Steinbach, Kumar*

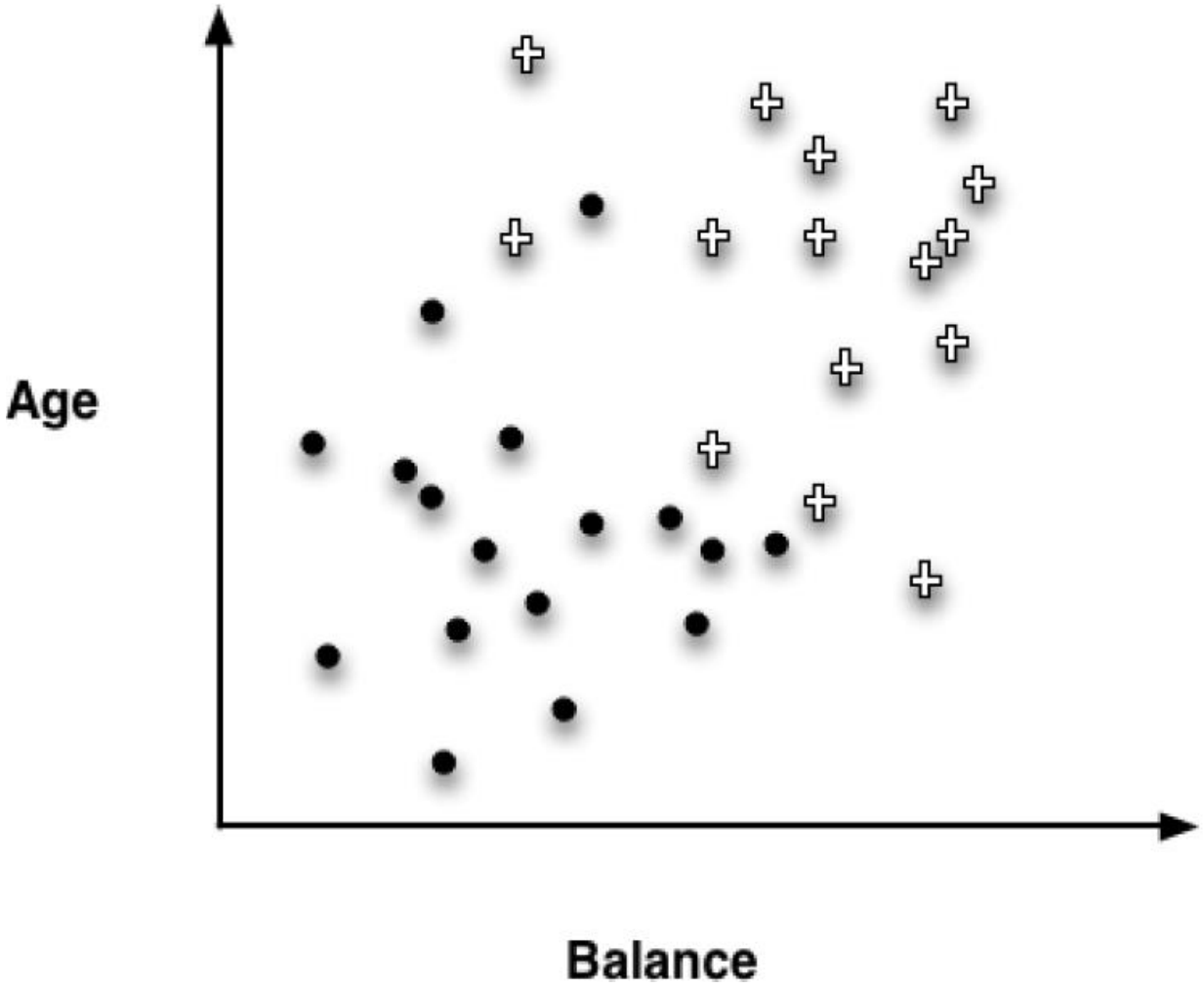
# 분류/예측 분석의 응용

- 특정 상품(예: 화장품)의 구매 가능성이 높은 고객은 누구인가?
- 향후 6개월 안에 이탈할 가능성이 높은 고객들은 누구이며 그들의 특징은?
- 향후 1개월 안에 내점하여 최소 10만원 이상 구매할 가능성이 높은 고객은 누구인가?
- H백화점의 고객별 LTV를 계산해 주는 모델은?
  - 예: LTV = 현재 고객의 기여가치 + 추가구매확률에 의한 기여가치 – 이탈확률에 의한 손실 가치
- 최고의 고객들은  어떤 특성을 갖는가?
- 최고 등급의 고객이 될 가능성이 높은 고객들은?

# 분류/예측 분석기법

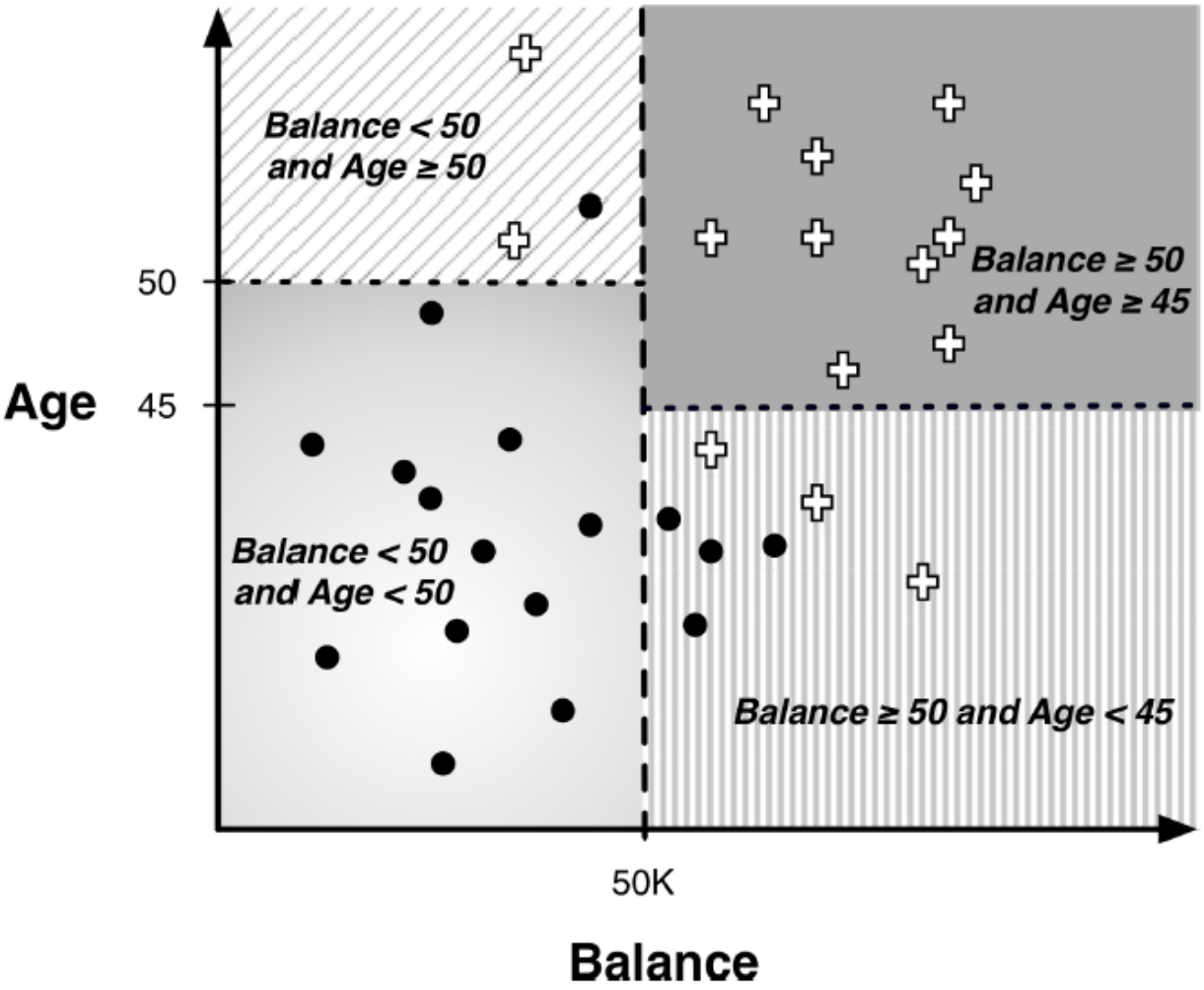
- **Decision Tree based Methods**
- **Neural Networks**
- **Support Vector Machines**
- kNN(k Nearest Neighbor)
- Naïve Bayes & Bayesian Belief Networks
- Rule-based Methods
- Logistic Regression
- Discriminant Analysis
- **Ensemble Methods**

# 분류/예측 분석기법 (시각적 해석)



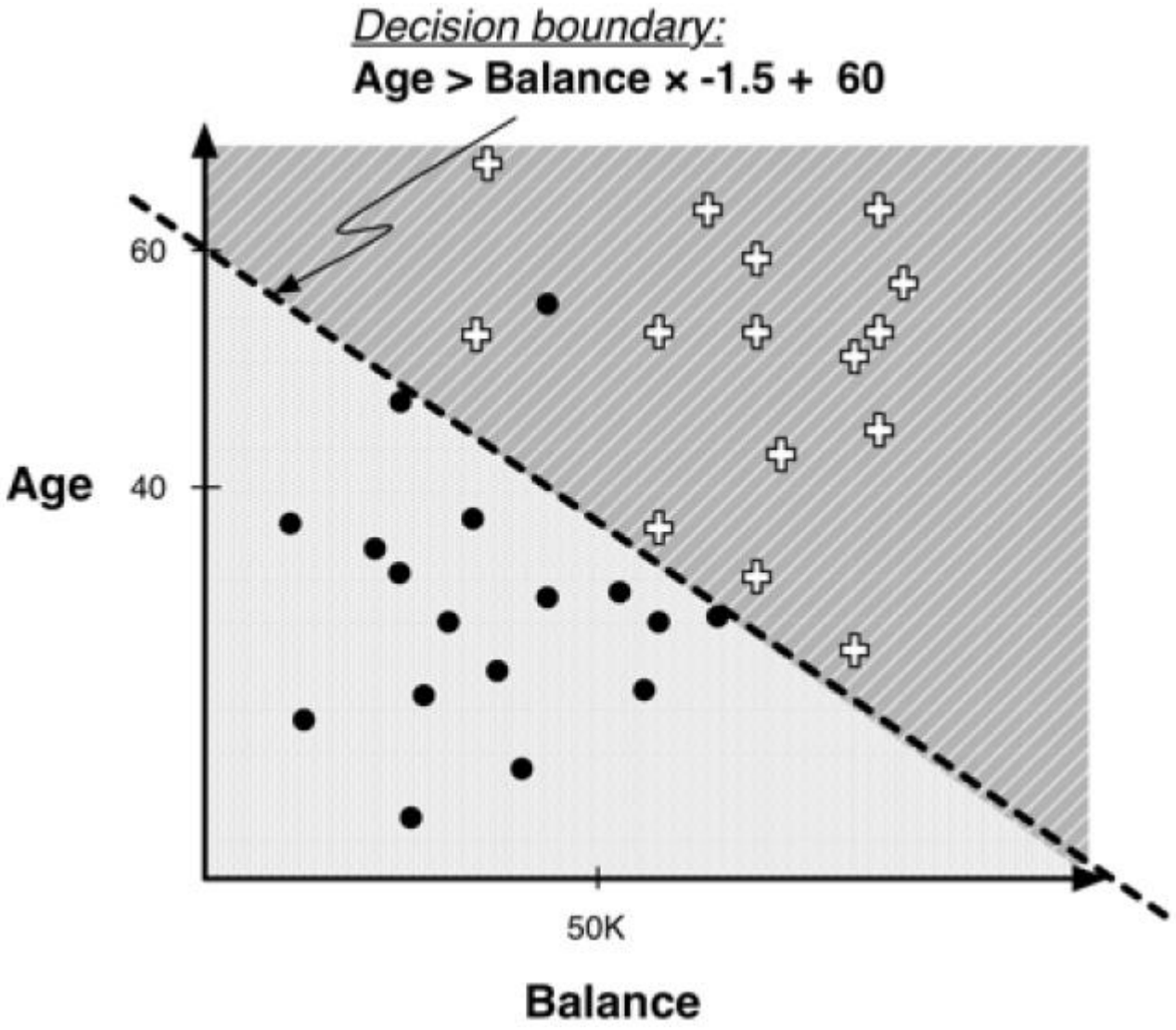

*Source: Data Science for Business by Provost & Fawcett*

# 분류/예측 분석기법
# – Decision Tree based Methods



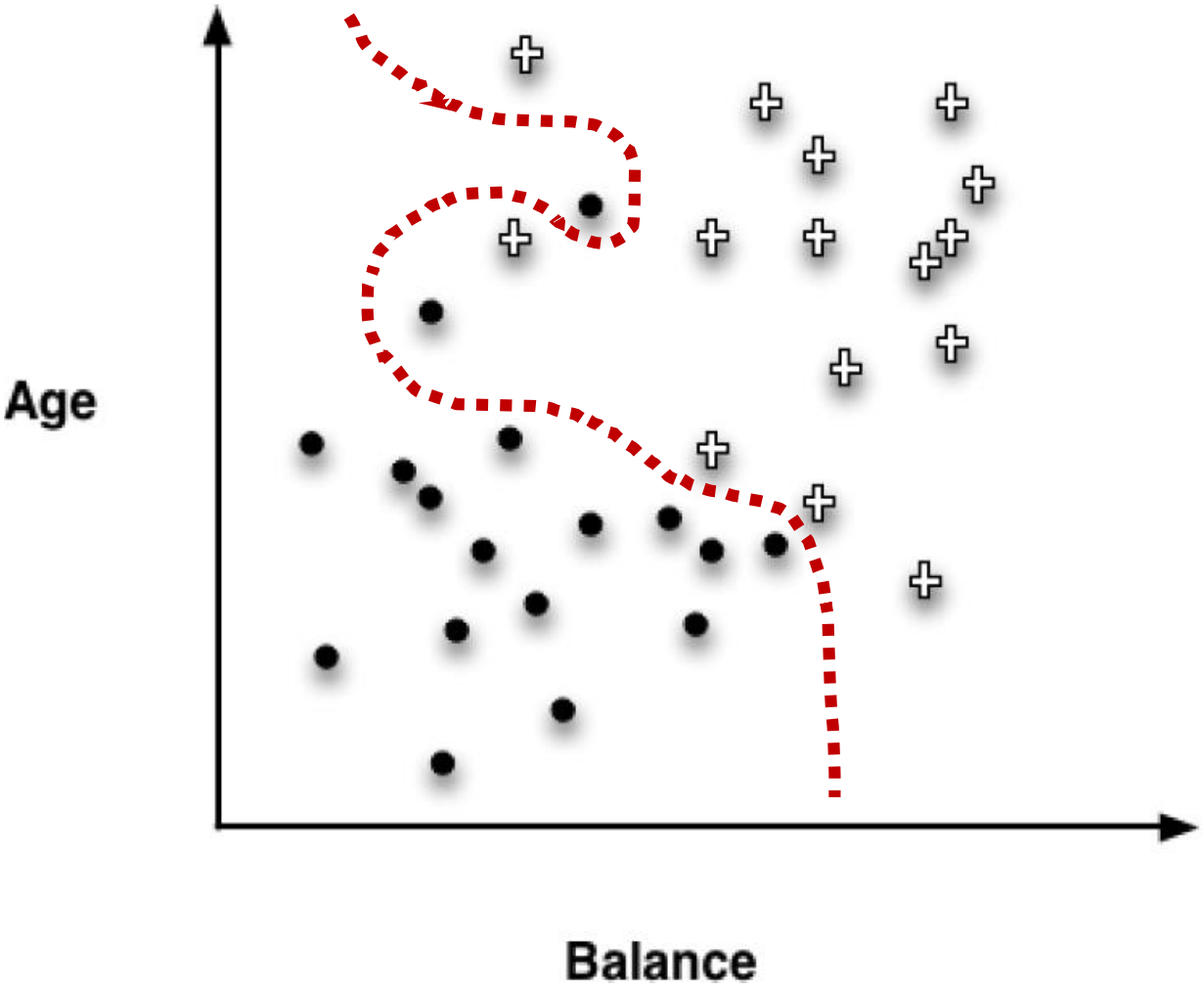

Source: Data Science for Business by Provost & Fawcett

# 분류/예측 분석기법
# – Support Vector Machines




Source: Data Science for Business by Provost & Fawcett

# 분류/예측 분석기법
# – Neural Networks



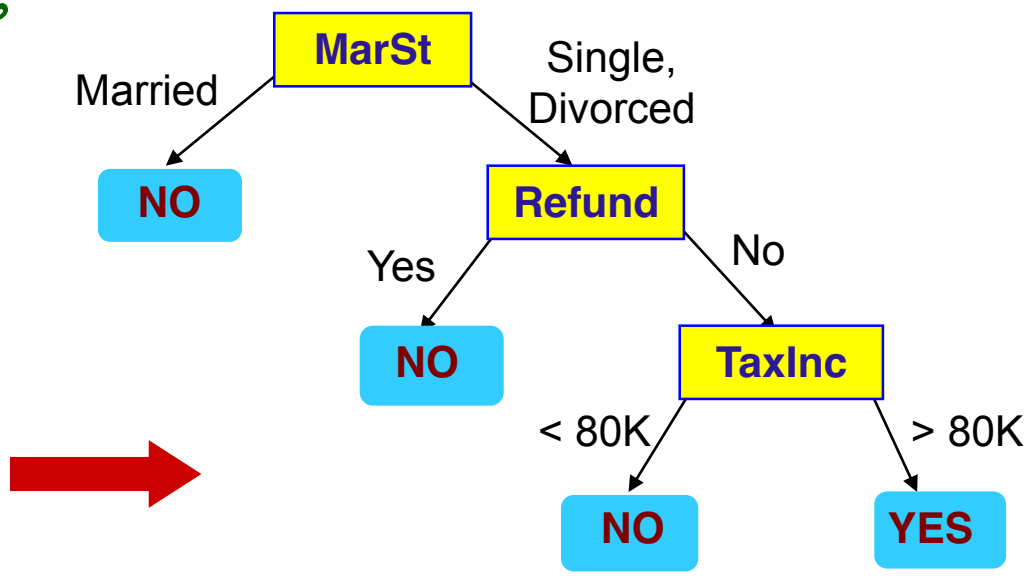

*Source: Data Science for Business by Provost & Fawcett*

# 의사결정나무(Decision Tree)의 개요

- 데이터 분류와 예측에서 강력하고 인기 있는 도구
- 장점
  - 인간이 쉽게 이해할 수 있는 언어로 표현할 수 있는 규칙을 기반으로 판단
- 적용 예
  - 메일 마케팅 회사의 마케팅 모델
    - 회원들에 잘 반응하는 마케팅 모델을 예측하는데 사용
  - 은행에서 대출 상담의 경우
    - 대출 상담자에게 대출을 해 줄 경우, 정확한 근거를 제시하여 대출을 거부할 자료를 제시할 수 있음
    - If 연간 수입 $20,000이상 and 관련 계좌가 3개 이상 then 대출
- 의사결정나무 기법
  - 명목형 목표변수 : C5.0, QUEST(Quick Unbiased Efficient Statistical Tree)
  - 연속형 목표변수 : CART(Classification & Regression Tree),
    CHAID(Chisquared Automatic Interaction Detection)

# Example of Decision Tree

| Tid | Refund (categorical) | Marital Status (categorical) | Taxable Income (continuous) | Cheat (class) |
|---|---|---|---|---|
| 1 | Yes | Single | 125K | No |
| 2 | No | Married | 100K | No |
| 3 | No | Single | 70K | No |
| 4 | Yes | Married | 120K | No |
| 5 | No | Divorced | 95K | Yes |
| 6 | No | Married | 60K | No |
| 7 | Yes | Divorced | 220K | No |
| 8 | No | Single | 85K | Yes |
| 9 | No | Married | 75K | No |
| 10 | No | Single | 90K | Yes |

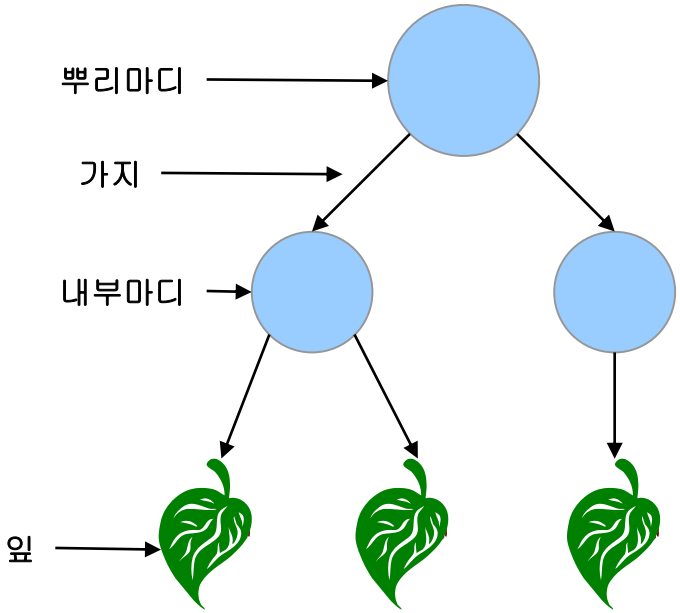



**There could be more than one tree that fits the same data!**

*Source: Introduction to Data Mining by Tan, Steinbach, Kumar*

# 의사결정나무의 구성

- 뿌리마디(root node): 최상단에 위치
- 내부마디(internal node): 속성의 분리 기준을 포함
- 가지(link): 마디와 마디를 이어줌
- 잎(leaf): 최종 분류



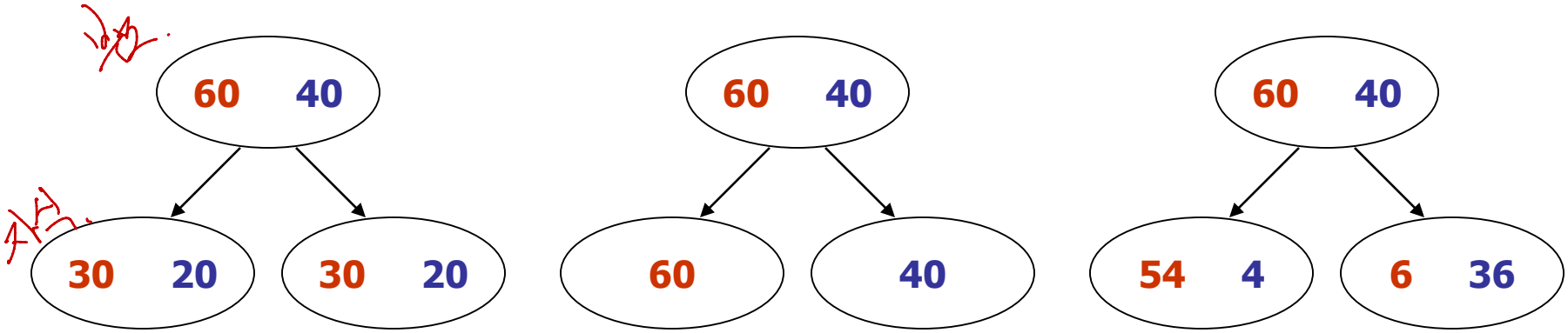
# 의사결정나무의 형성과정

• 첫분류가 가장 큰분류기준이 된다.

| | |
|---|---|
| **Step 1**<br>의사결정나무 형성 | ✓ 분석의 목적과 자료구조에 따라, 적절한 최적의 분리기준(split criterion)을 찾아서 나무를 성장 시킨다. 정지규칙(stop rule) 을 만족하면 성장을 중단한다. |
| **Step 2**<br>**가지치기** | ✓ 분류오류(classification error)를 크게 할 위험(risk)이 높거나 부적절한 추론규칙 (induction rule)을 가지고 있는 가지(branch)를 제거한다. 또한, 불필요한 가지를 제거한다. |
| **Step 3**<br>**타당성 평가** | ✓ 이익도표(gain chart)나 위험도표(risk chart) 또는 검증용 자료 (test sample)의 사용, 또는 교차타당성 (cross validation) 등을 이용하여 의사결정나무를 평가한다. |
| **Step 4**<br>**해석 및 예측** | ✓ 구축된 나무모형을 해석하고 예측모형을 설정한다 |

# 순수도 & 분리기준

- 순수도
  - 목표변수의 특정 범주에 개체들이 포함되는 정도
- 분리기준
  - 하나의 부모마디로부터 자식마디들이 형성될 때 입력변수의 선택과 범주의 병합이 이루어 질 기준을 의미
  - 어떤 입력변수를 이용하여 분리하는 것이 목표변수의 분포를 가장 잘 구별해 주는지를 파악하여 자식마디가 생성되는데, 부모마디의 순수도에 비해서 자식마디들의 순수도가 증가하도록 자식마디를 형성



# C5.0 알고리즘 (1/10)

C5.0은 퀼란(Quilan, 1988)의 C4.5를 개량한 알고리즘으로 의사결정 나무 모델의 하나이며, 범주형(flag/set type)인 목표변수를 이용하여 다른 입력변수들의 분류를 통해, 세분화 모델이나 목표변수를 예측할 수 있는 분석 방법이다.

## C5.0의 규칙

✓ 예측변수의 선택기준

- 목표변수(종속변수)를 예측하는 데 있어 $X_1$, ..., $X_p$를 예측변수로 활용하고자 한다고 할때, C5.0처럼 나무구조의 결정규칙을 생성하기 위해서는 각 단계에서 p개의 예측 변수 중 어느 것에 의하여 가지분리를 할 것인가를 선택해야 한다. 이 때 결정규칙들은 각기 다른 변수선택 기준을 쓴다.

✓ 정지규칙/가지치기(Pruning) 규칙

- 의사결정 나무 구조에서 사용되는 가지치기 규칙은 일반적으로 2 가지가 있다. 하나는 일정한 임계 치를 기준으로 그보다 높은 경우에만 새로운 가지가 나오도록 하는 정지규칙이고, 다른 하나는 의사결정 나무가 다 자란 후 중요도가 비교적 작은 가지 분리들을 취소함으로써 나무의 크기를 감축시키는 가지치기규칙이다. C5.0는 가지치기 방법을 택하고 있으며, 이는 계산시간이 비교적 긴 반면 가지의 중요도를 확인한다는 점에서 큰 의미가 있다.

✓ 최소 레코드 수

- 가지 분리를 허용하는 조건으로 가지에 배속되는 레코드 수가 특정 값보다 커야 한다는 조건. 이 최소수에 대한 디폴트 값은 2인데, 이것을 큰 값으로 세팅할 수록 나무 규모가 줄어들게 된다. C5.0 에서는 분리될 가지 중 2개 이상이 최소 레코드 수보다 커야 가지 분리가 허용된다.

✓ 엔트로피 지수 (Entropy Index) : C5.0에서 결정규칙을 분리할 수 있는 기준으로 데이터의 무질서 정도를 측정할 수 있는 방법

- $Entropy(T) = -\sum_{i=1}^{k} p_i \log p_i$ ($p$ = 각 범주의 비율)

# C5.0 알고리즘 (2/10)

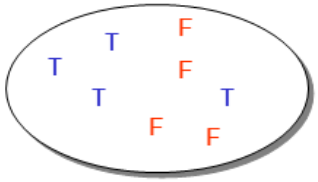
## C5.0 –엔트로피(Entropy)

- ✓ C5.0 / C4.5의 기본 개념은 불순도를 측정하는(C&RT의 Gini와 유사)하나의 측정 지표가 엔트로피라는 것이다.
  원래 엔트로피는 물리학의 용어로서 자연 현상의 변화는 일정한 방향으로만 진행하는, 즉, 자연현상은 물질계의 엔트로피가 증가하는 방향으로 진행한다는 엔트로피 증가의 법칙에서 나왔으며, 분자운동이 확률이 적은 질서 있는 상태로부터 확률이 큰 무질서한 상태로 이동한다는 것을 의미한다.
- ✓ 이런 무질서와 확률을 이용한 개념을 데이터에서 구분을 잘 짓고, 못 짓는 속성 값을 찾아내는 지수로 사용하는 것으로 의사결정나무에서 활용을 한다.
- ✓ 의사결정 나무에서는 엔트로피가 높을 수록 Target 구분을 잘 못해주는 속성 필드가 되며, 낮을 수록 구분을 잘 해주는 유의한 속성 필드가 된다.

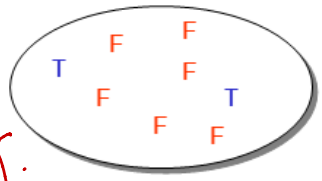
엔트로피 계산법

(Target 범주가 2개)

▶ **Entropy(S) = - (T 범주의 수 / 전체건수) * Log2(T 범주의 수 / 전체건수)**
**- (F 범주의 수 / 전체건수) * Log2(F 범주의 수 / 전체건수)**




-T범주수 : 4개, F범주수: 4개, 전체건수 8개인 경우
-실계산 : -4/8*log2(4/8) - 4/8*log2(4/8) = 1
※ 밑이 2인 로그임.




-T범주수 : 2개, F범주수:6 개, 전체건수 8개인 경우
-실계산 : -2/8*log2(2/8) - 6/8*log2(6/8) = 0.5936
※ 밑이 2인 로그임.

# C5.0 알고리즘 (3/10)

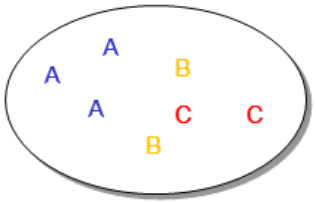
## C5.0 – 엔트로피(Entropy)

✓ 엔트로피의 성질 : 집합에 범주가 고르게 분포할 수록 엔트로피 값은 높다. (불확실성 상태가 크다.)

엔트로피 계산법

(Target 범주가 3개)

▶ **Entropy(S) = – (1st 범주의 수 / 전체건수) * Log(1st 범주의 수 / 전체건수)**
**– (2nd 범주의 수 / 전체건수) * Log(2nd 범주의 수 / 전체건수)**
**…**
**– (Nth 범주의 수 / 전체건수) * Log(Nth 범주의 수 / 전체건수)**




**A범주수: 3개, B범주수: 2개, C범주수: 2개, 전체건수 8개인 경우**

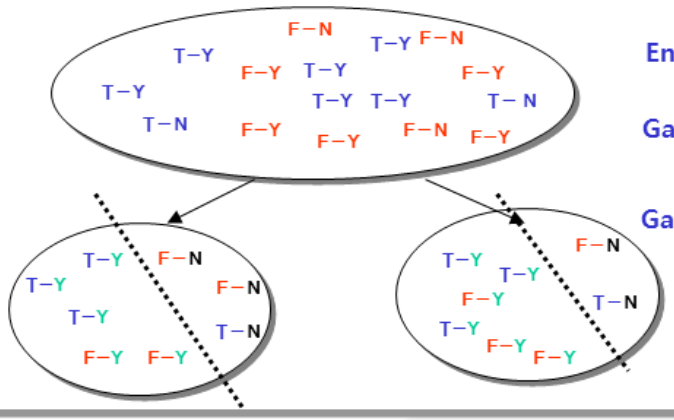
**-실계산 : -3/8*log(4/8) - 2/8*log(2/8) -2/8*log(2/8)**
**= 0.4607**

**※ 밑이 10인 상용 로그임(주의 할 것).**

# C5.0 알고리즘 (4/10)

## C5.0 – 정보획득함수(Information Gains)

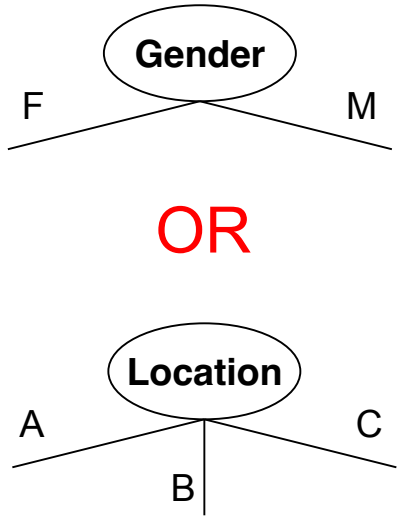
- ✓ Information Gains라는 것은 C5.0에서 분류를 하는 지표가 되는 함수로서, CHAID의 Chi-square 통계량, C&RT, Gini Index와 같은 역할을 하게 된다.
- ✓ Information Gains라는 의미는 특정한 속성(attribute: 필드, 변수라고 생각하면 된다.)에 대한 정보를 알았을 때 얻어지는 정보량이 얼마만큼 되는지를 측정하는 지수가 되는 것이다.
- ✓ Entropy는 이 Information Gains를 계산하는데 필요한 측정식의 일부가 된다.

| | |
|---|---|
| Information Gains<br>함수식 | $$Gain(S, A) = Entropy(S) - \sum_{v \in Values(A)} \frac{\lvert S_v \rvert}{\lvert S \rvert} Entropy(S_v)$$ |
| Information Gains<br>함수의 예제(개요) |  |

# C5.0 알고리즘 (5/10)

**C5.0 – 알고리즘 사례 ①**

| 성별(Gender) | 거주지역(Location) | 응답여부(Respond) |
|---|---|---|
| M | A | Y |
| M | B | Y |
| M | A | Y |
| M | C | Y |
| F | B | N |
| F | A | N |
| F | B | N |
| M | C | N |
| M | A | N |
| M | A | Y |




| | | Respond | | Total |
|---|---|---|---|---|
| | | Y | N | |
| Gender | F | 0 | 3 | 3 |
| | M | 5 | 2 | 7 |
| Total | | 5 | 5 | 10 |

| | | Respond | | Total |
|---|---|---|---|---|
| | | Y | N | |
| Location | A | 3 | 2 | 5 |
| | B | 1 | 2 | 3 |
| | C | 1 | 1 | 2 |
| Total | | 5 | 5 | 10 |

# C5.0 알고리즘 (6/10)

## Information Gain

- Gender

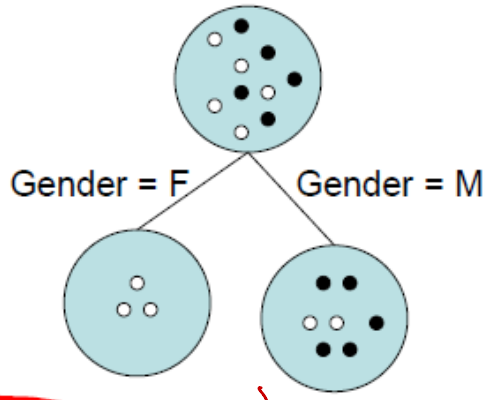
| | | Respond | | Total |
|---|---|---|---|---|
| | | Y | N | |
| Gender | F | 0 | 3 | 3 |
| | M | 5 | 2 | 7 |
| Total | | 5 | 5 | 10 |

$$E_{before} = -\frac{5}{10}\log_2\frac{5}{10} - \frac{5}{10}\log_2\frac{5}{10} = 1$$

$$E_{left} = -0 - \frac{3}{3}\log_2\frac{3}{3} = 0$$

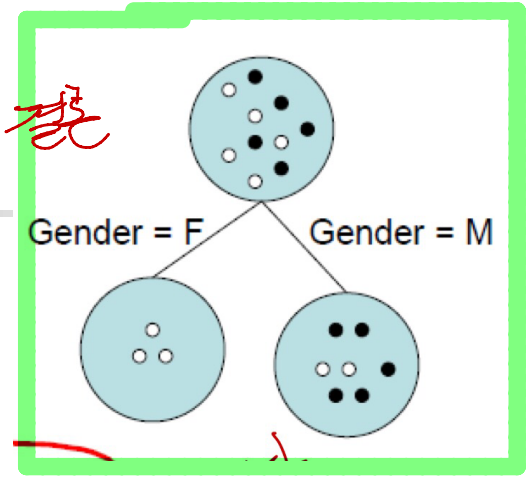
$$E_{right} = -\frac{5}{7}\log_2\frac{5}{7} - \frac{2}{7}\log_2\frac{2}{7} = 0.863121$$

$$E_{after} = \frac{3}{10}\times 0 + \frac{7}{10}\times 0.863121 = 0.604185$$




$$IG_{Gender} = E_{before} - E_{after} = 1 - 0.604185 = 0.395815$$

# C5.0 알고리즘 (7/10)




## Information Gain

- Location

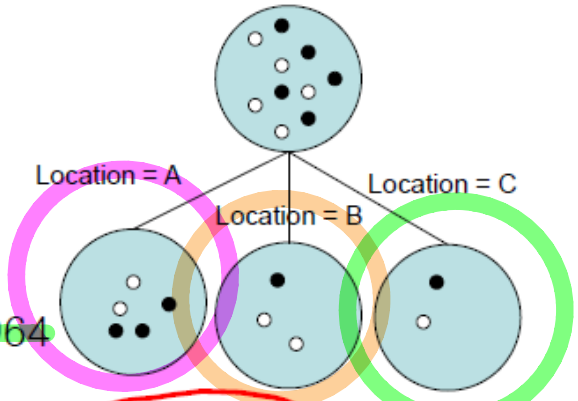
| | | Respond | | Total |
|---|---|---|---|---|
| | | Y | N | |
| Location | A | 3 | 2 | 5 |
| | B | 1 | 2 | 3 |
| | C | 1 | 1 | 2 |
| Total | | 5 | 5 | 10 |

$$E_{before} = -\frac{5}{10}\log_2\frac{5}{10} - \frac{5}{10}\log_2\frac{5}{10} = 1$$

$$E_{left} = -\frac{3}{5}\log_2\frac{3}{5} - \frac{2}{5}\log_2\frac{2}{5} = 0.970951$$

$$E_{middle} = -\frac{1}{3}\log_2\frac{1}{3} - \frac{2}{3}\log_2\frac{2}{3} = 0.918296$$

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




$$E_{after} = \frac{5}{10}\times 0.970951 + \frac{3}{10}\times 0.918296 + \frac{2}{10}\times 1 = 0.960964$$

$$IG_{location} = E_{before} - E_{after} = 1 - 0.960964 = 0.039036$$

IG값이 큰 Gender가 분기 기준으로 선택!

# C5.0 알고리즘 (8/10)

**C5.0 – 알고리즘 사례 ②**

| ID | STR | INCOME | SEX | HOUSE | TARGET |
|---|---|---|---|---|---|
| 1 | 서울 | 고소득 | 남자 | 아파트 | 구매 |
| 2 | 서울 | 고소득 | 남자 | 주택기타 | 구매 |
| 3 | 수도권 | 고소득 | 남자 | 아파트 | 비구매 |
| 4 | 지방 | 중간소득 | 남자 | 아파트 | 비구매 |
| 5 | 지방 | 저소득 | 여자 | 아파트 | 비구매 |
| 6 | 지방 | 저소득 | 여자 | 주택기타 | 구매 |
| 7 | 수도권 | 저소득 | 여자 | 주택기타 | 비구매 |
| 8 | 서울 | 중간소득 | 남자 | 아파트 | 구매 |
| 9 | 서울 | 저소득 | 여자 | 아파트 | 비구매 |
| 10 | 지방 | 중간소득 | 여자 | 아파트 | 비구매 |
| 11 | 서울 | 중간소득 | 여자 | 주택기타 | 비구매 |
| 12 | 수도권 | 중간소득 | 남자 | 주택기타 | 비구매 |
| 13 | 수도권 | 고소득 | 여자 | 아파트 | 비구매 |
| 14 | 지방 | 중간소득 | 남자 | 주택기타 | 구매 |

# C5.0 알고리즘 (9/10)

## C5.0 – 알고리즘 사례

**1단계 : Target에 대한 Entropy를 계산한다.**

→ Target 필드 (구매 5건, 비구매 9건) :

**Entropy = -(5/14) * log2(5/14) – (9/14) * log2(9/14) = 0.9402**

**2단계 : Target을 제외한 다른 나머지 Input 필드(설명변수)의 Information Gains 값을 계산한다.**

→ STR(지역)의 Information Gains 값 계산하기

1) "서울"범주의 Entropy(구매 3, 비구매 2) = -(3/5)*log2(3/5) – (2/5)*log2(2/5) = 0.9709

2) "수도권"범주의 Entropy(구매 0, 비구매 4) = -(0/4)*log2(0/4) – (4/4)*log2(4/4) = 0

3) "지방"범주의 Entropy(구매 2, 비구매 3) = -(2/5)*log2(2/5) – (3/5)*log2(3/5) = 0.9709

**Information Gains = 0.9402 – (5/14)*0.9709 – (4/14)*0 – (5/14)*0.9709 = 0.2467**

→ INCOME(소득)의 Information Gains 값 계산하기

1) "고소득"범주의 Entropy(구매 2, 비구매 2) = -(2/4)*log2(2/4) – (2/4)*log2(2/4) = 1

2) "중간소득"범주의 Entropy(구매 2, 비구매 4) = -(2/6)*log2(2/6) – (4/6)*log2(4/6) = 0.9183

3) "저소득"범주의 Entropy(구매 1, 비구매 3) = -(1/4)*log2(1/4) – (3/4)*log2(3/4) = 0.8112

**Information Gains = 0.9402 – (4/14)*1 – (5/14)*0.9183 – (4/14)*0.8112 = 0.0948**

→ SEX(성별)의 Information Gains 값 계산하기

1) "남자"범주의 Entropy(구매 4, 비구매 3) = -(4/7)*log2(4/7) – (3/7)*log2(3/7) = 0.9852

2) "여자"범주의 Entropy(구매 1, 비구매 6) = -(1/7)*log2(1/7) – (6/7)*log2(6/7) = 0.5916

**Information Gains = 0.9402 – (7/14)*0.9852 – (7/14)*0.5916 = 0.1518**

→ HOUSE(주거종류)의 Information Gains 값 계산하기

1) "아파트"범주의 Entropy(구매 2, 비구매 6) = -(2/8)*log2(2/8) – (6/8)*log2(6/8) = 0.8112

2) "주택기타"범주의 Entropy(구매 2, 비구매 4) = -(3/6)*log2(3/6) – (3/6)*log2(3/6) = 1

**Information Gains = 0.9402 – (8/14)*0.8112 – (6/14)*1 = 0.048**

→ 가장 Information Gains 값이 큰 **STR**이 첫 번째 분류 기준 필드로 선정이 된다.

**3단계 : 1단계 분류가 되면 자식노드(leaf)를 하나의 어미노드로 인식을 하여, 분류된 노드마다 2단계를 재 수행한다.**

→ 단, STR 필드의 수도권의 범주와 같이 0이 되면 더 이상 분류를 하지 않는다.(100% 만족된 상태)

# Practical Issues of Classification

- **Under-fitting and Over-fitting**
- **Missing Values**
- **Costs of Classification**
- **The Curse of Dimensionality**

Decision Tree는 일반적으로 고능률인다.

ex) 대출여부에서 대출해주는것이 High risk.

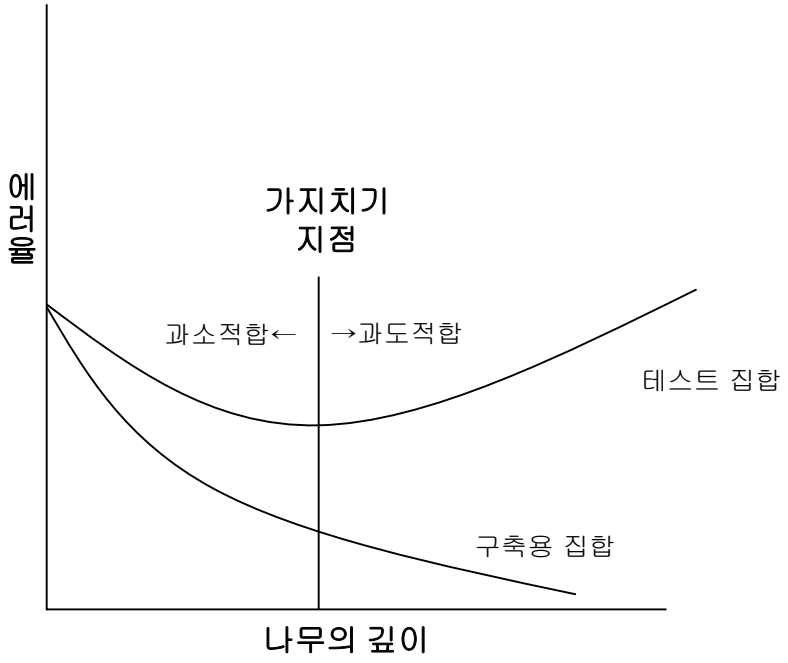
차원의저주

# 과도/과소 적합

- 과도적합(overfitting)
  - 모델이 데이터에 필요이상으로 적합한 모델
  - 데이터 내에 존재하는 규칙 뿐만 아니라 불완전한 레코드도 학습
- 과소적합(underfitting)
  - 모델이 데이터에 제대로 적합하지 못한 모델
  - 데이터 내에 존재하는 규칙도 제대로 학습하지 못함

# 가지치기 (1/3)

- 가지치기 규칙(pruning rule)
  - 최종 마디의 수가 너무 많으면 모형이 과적합 상태가 됨 -> 현실문제에 적용할 수 있는 규칙이 나오지 않는다.
  - 불필요하게 복잡해진 나무의 의미 없는 가지를 제거하는 작업



# 가지치기 (2/3)

- **Pre-Pruning**
  - **Stopping Rule**

**현재의 마디가 더 이상 분리가 일어나지 못하게 하는 규칙**

- 모든 자료가 한 그룹에 속할 때
- 마디에 속하는 자료가 일정한 수 이하일 때
- 불순도의 감소량이 아주 적을 때
- 뿌리마디로부터 깊이가 일정 수 이상일 때

# 가지치기 (3/3)

- **Post-pruning**
  - **Grow decision tree to its entirety**
  - **Trim the nodes of the decision tree in a bottom-up fashion**
  - **If generalization error improves after trimming, replace sub-tree by a leaf node.**
  - **Class label of leaf node is determined from majority class of instances in the sub-tree**

# C5.0 알고리즘 (10/10)

| | |
|---|---|
| Class = Yes | 20 |
| Class = No | 10 |
| Error = 10/30 | |

$$Error(t) = 1 - \max P(i \mid t)$$

**Training Error (Before splitting) = 10/30**

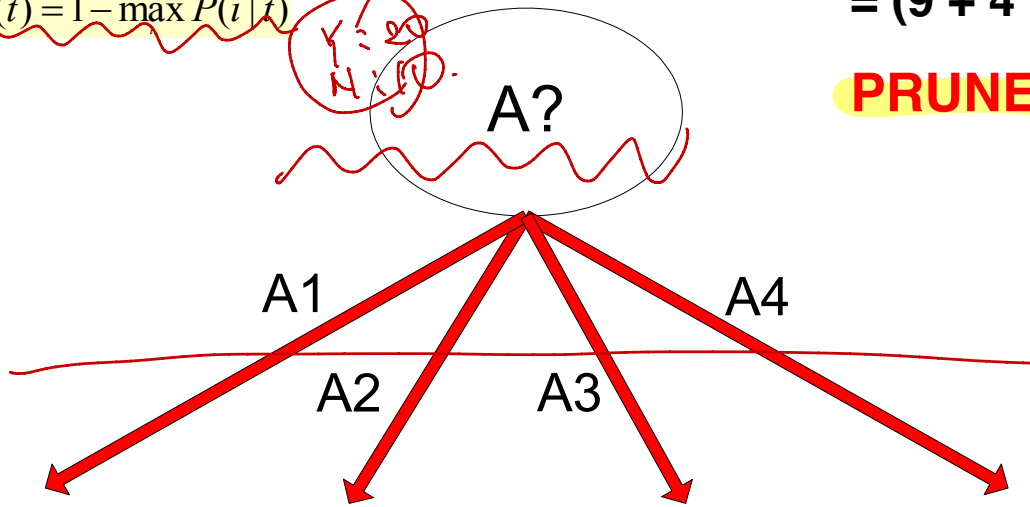
**Pessimistic error = (10 + 0.5)/30 = 10.5/30**

**Training Error (After splitting) = 9/30**

**Pessimistic error (After splitting)**

**= (9 + 4 × 0.5)/30 = 11/30**

**PRUNE!**




| | | | | | | | |
|---|---|---|---|---|---|---|---|
| Class = Yes | 8 | Class = Yes | 3 | Class = Yes | 4 | Class = Yes | 5 |
| Class = No | 4 | Class = No | 4 | Class = No | 1 | Class = No | 1 |