

Decision tree

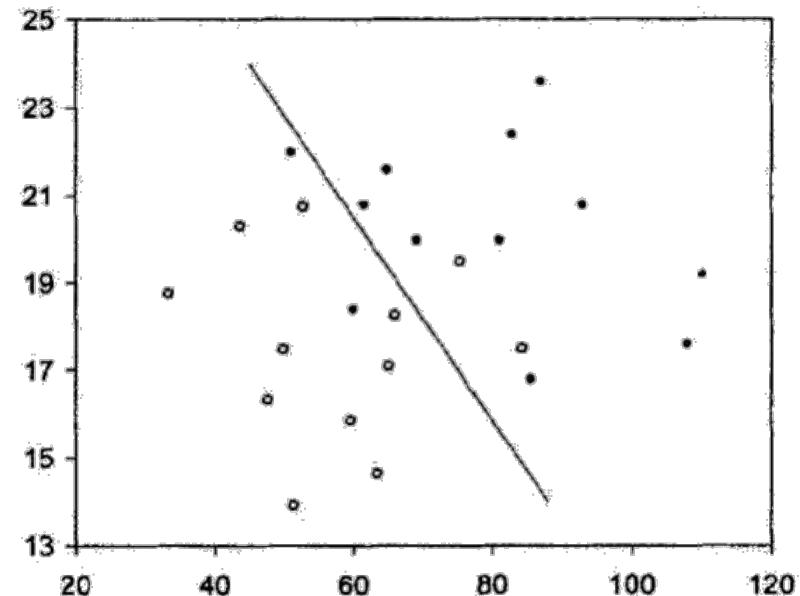
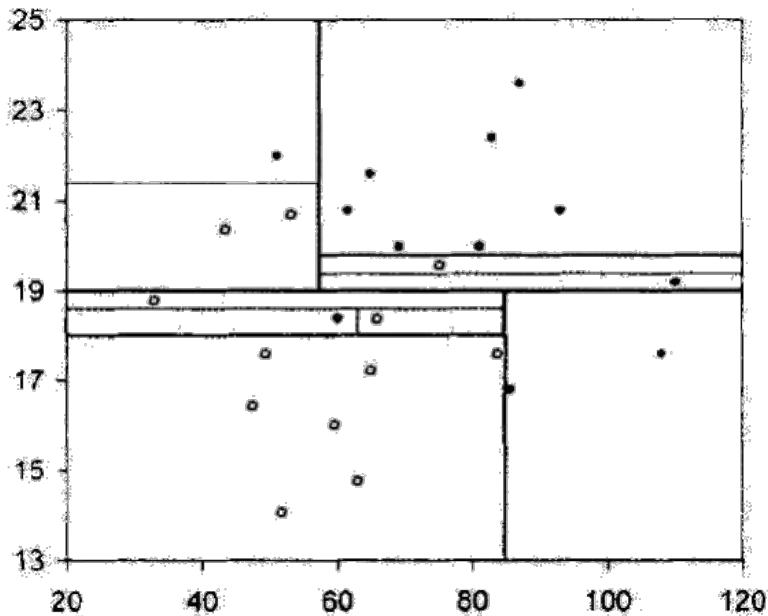
: Classification and regression tree (CART)

고태훈 (taehoonko@dm.snu.ac.kr)

Classification revisited

- ❖ Why are there many different classification algorithms?
 - ▶ So many ways to reach the same result.

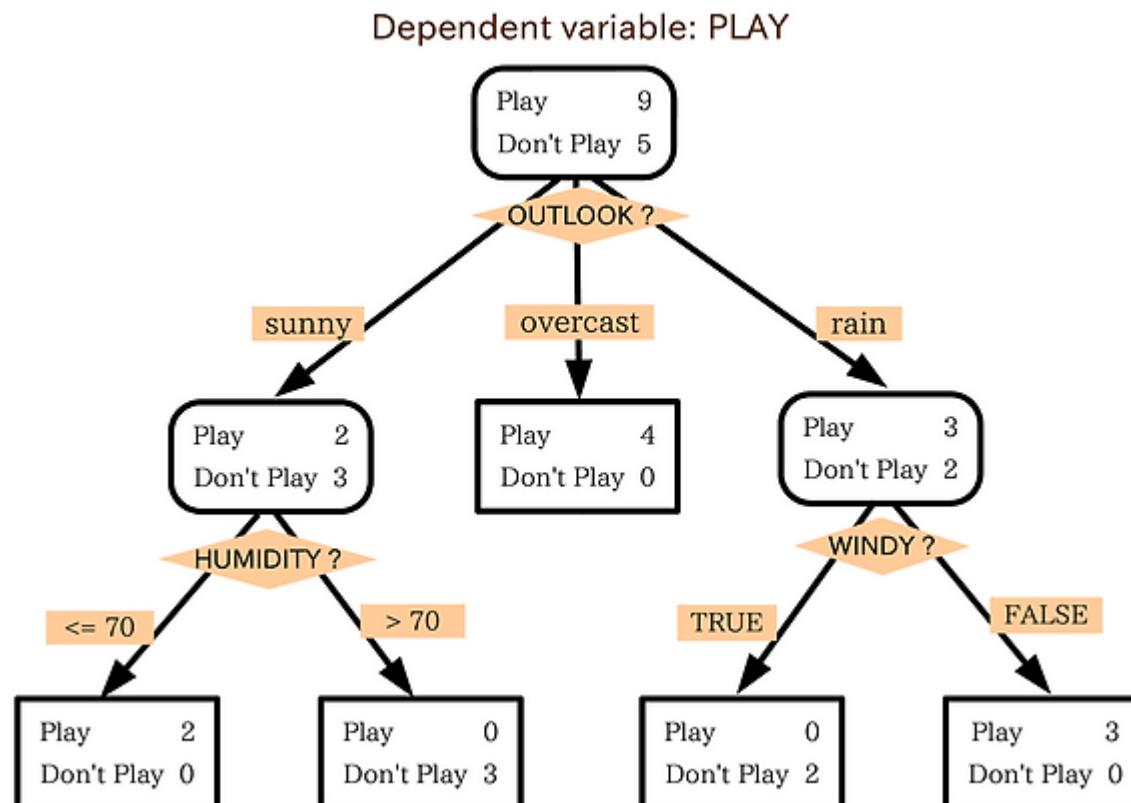
“Separate the riding mower buyers(●) from non-buyers(○)”



Decision Tree

❖ Goal

- ▶ Classify or predict an outcome based on a set of predictors.



Rule example

If outlook is sunny
and if humidity > 70
then he does not play

or

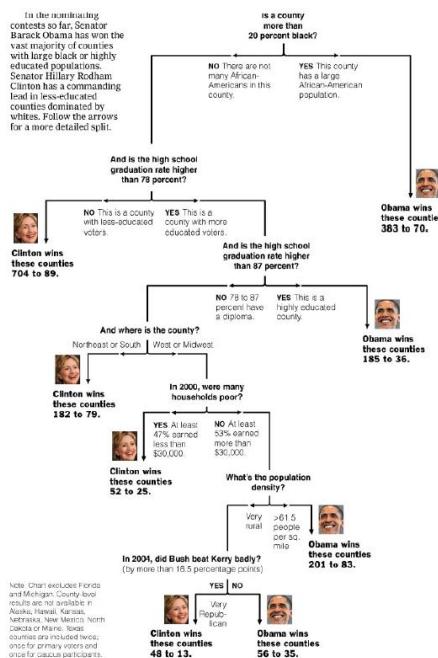
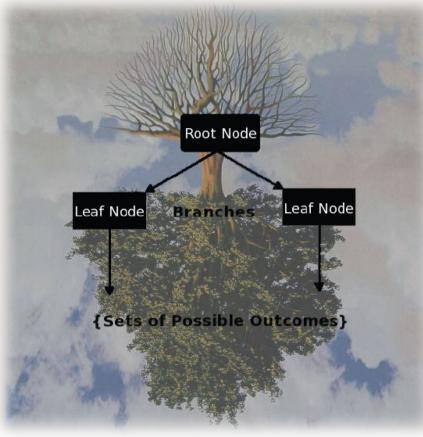
If outlook is rainy
and it is not windy
then he does play

Classification and Regression Trees (CART)

❖ Why CART?

- ▶ 분류 분석과 회귀 분석 모두 가능
- ▶ 입력 변수들 중 데이터를 가장 잘 분류하는 변수를 택하여 분기가 계속 일어남
- ▶ 해석력이 좋다는 장점
 - 분기되는 변수는 중요한 변수로 간주할 수 있음
 - 분류 규칙을 추출할 수 있음
- ▶ 변수 선택이 모델 자체에서 이루어짐
- ▶ 수치형, 명목형, 카테고리 변수도 처리 가능

Classification and Regression Trees (CART)



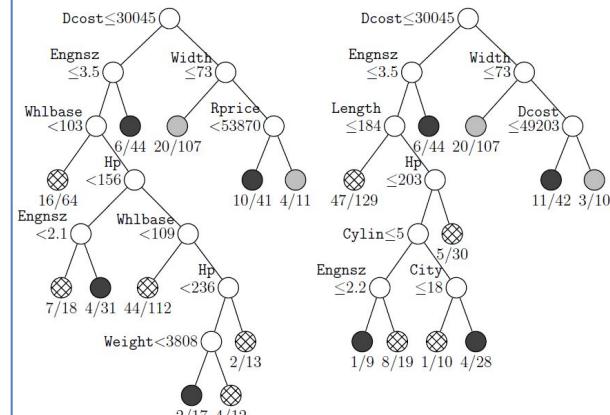
Classification and Regression Tree (CART)

- 개별 변수의 영역을 반복적으로 분할함으로써 전체 영역에서의 규칙을 생성하는 지도학습 기법 (Breiman, 1984)
- If-then 형식으로 표현되는 규칙(rules)을 생성함으로써, 결과에 대한 예측과 함께 그 이유를 설명할 수 있는 장점이 있음
- 수치형 변수와 범주형 변수에 대한 동시 처리 가능

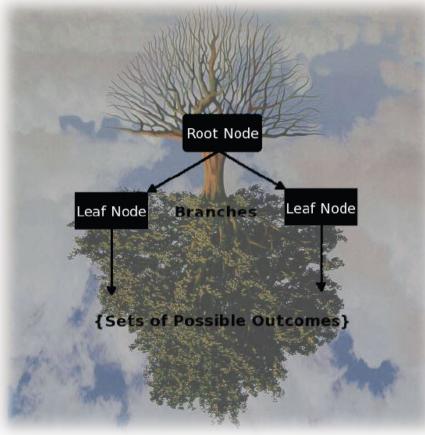
재귀적 분기 (Recursive Partitioning)

가지치기 (Pruning)

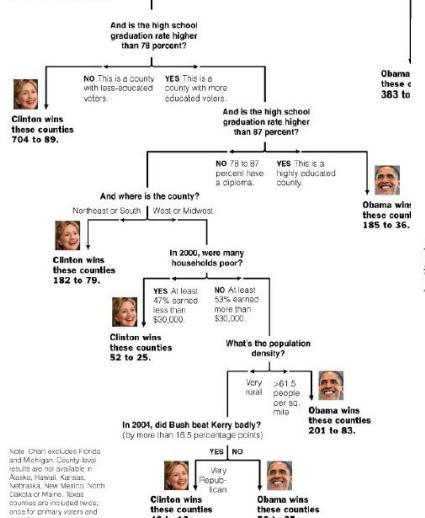
- 특정 영역(부모 노드)에 속하는 개체들을 하나의 기준 변수 값의 범위에 따라 분기
- 분기에 의해 새로 생성된 자식 노드의 동질성이 최대화 되도록 분기점 선택
- 불순도를 측정하는 기준으로는 범주형 변수에 대해서는 지니계수, 수치형 변수에 대해서는 분산을 이용



Classification and Regression Trees (CART)



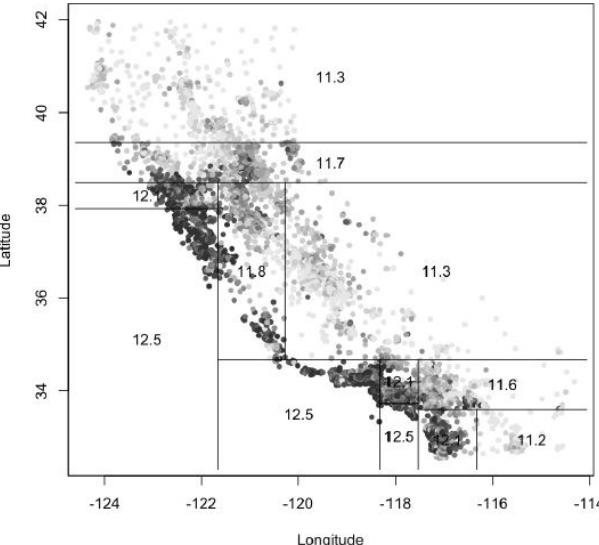
In the nominating contests so far, Senator Barack Obama has won the vast majority of counties with large black or highly educated populations. Senator Hillary Rodham Clinton has a commanding lead in less-educated counties dominated by whites. Follow the arrows for a more detailed split.



Classification and Regression Tree (CART)

- 개별 변수의 영역을 반복적으로 분할함으로써 전체 영역에서의 규칙을 생성하는
지도학습 기법 (Breiman, 1984)
- If-then 형식으로 표현되는 규칙(rules)을 생성함으로써, 결과에 대한 예측과 함께 그 이유를 설명할 수 있는 장점이 있음
- 수치형 변수와 범주형 변수에 대한 동시 처리 가능

재귀적 분기 (Recursive Partitioning)



가지치기 (Pruning)

- 과적합(Over-fitting)을 방지하기 위하여 하위 노드들을 상위 노드로 결합
- Pre-pruning: Tree를 생성하는 과정에서 최소 분기 기준을 이용하는 사전적 가지치기
- Post-pruning: Full-tree 생성 후, 검증 데이터의 오분류율과 Tree의 복잡도(말단 노드의 수) 등을 고려하는 사후적 가지치기

의사결정나무 예시 : 포유류 분류

- ❖ 여러 동물들에 대한 정보를 조사하여 다음과 같은 데이터가 있다고 가정해보자.

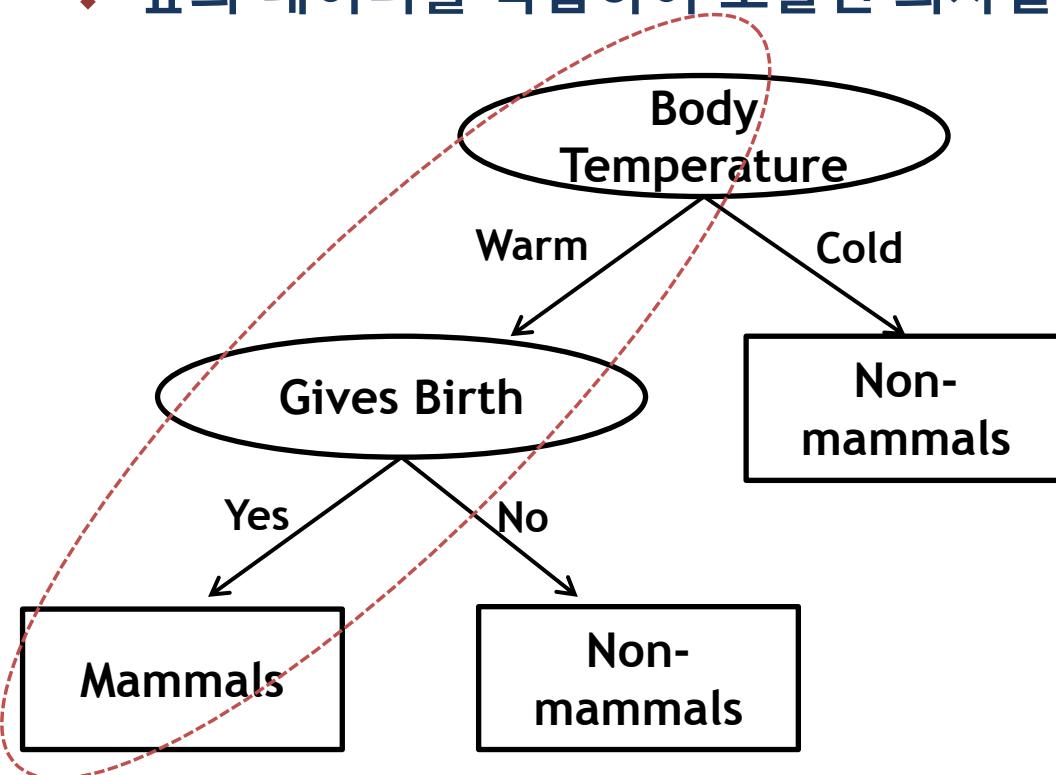
ID	Body Temperature	Gives Birth?	Weight(kg)	...	Mammal?
1	Warm	Yes	65	...	Mammal
2	Warm	No	0.32	...	Non-mammal
3	Cold	No	0.14	...	Non-mammal
4	Warm	Yes	3,000	...	Mammal
5	Cold	No	127	...	Non-mammal
6	Cold	No	0.35	...	Non-mammal
7	Warm	Yes	1.5	...	Mammal
...

Body Temperature = Warm → 온혈동물(혹은 정온동물)
Body Temperature = Cold → 냉혈동물(혹은 변온동물)

Gives Birth = Yes → 태생
Gives Birth = No → 난생

의사결정나무 예시 : 포유류 분류

- ❖ 앞의 데이터를 학습하여 도출된 의사결정나무 모델은 다음과 같다.

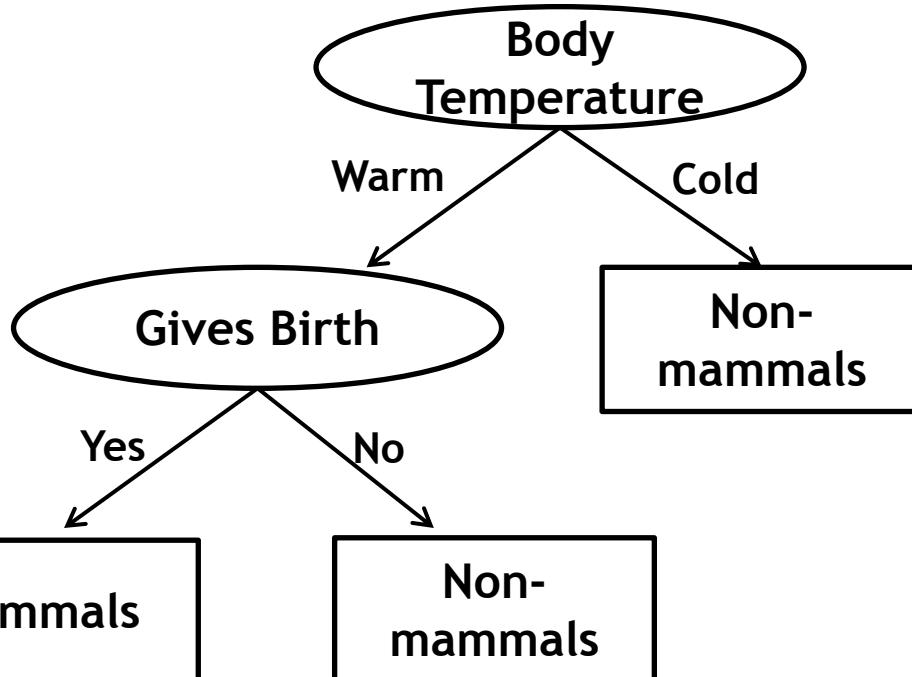


▶ “체온”과 “새끼를 낳는 방법”이 포유류를 분류하는 데에 있어 중요한 변수라는 것을 알 수 있다.

- ▶ 의사결정나무를 통해 포유류으로 분류하는 규칙을 구할 수 있다.
 - 온혈동물이고 태생이면 포유류이다.
 - IF Body Temperature = Warm and Gives Birth = Yes, then Mammals.

의사결정나무 예시 : 포유류 분류

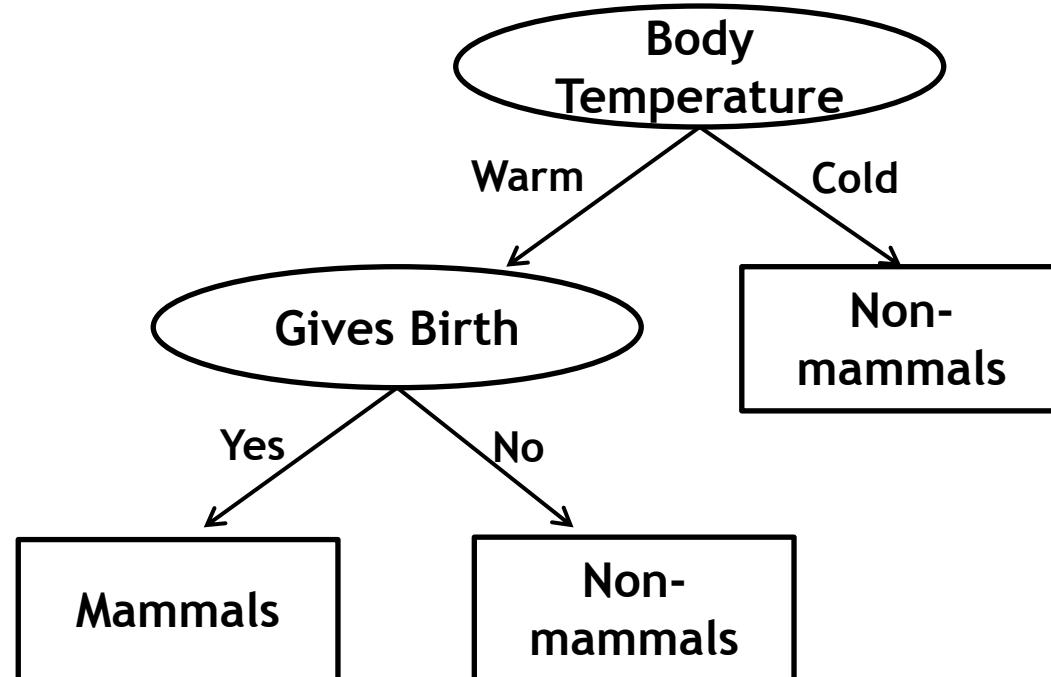
❖ 새로운 데이터를 의사결정나무 모델을 이용하여 분류해보자.



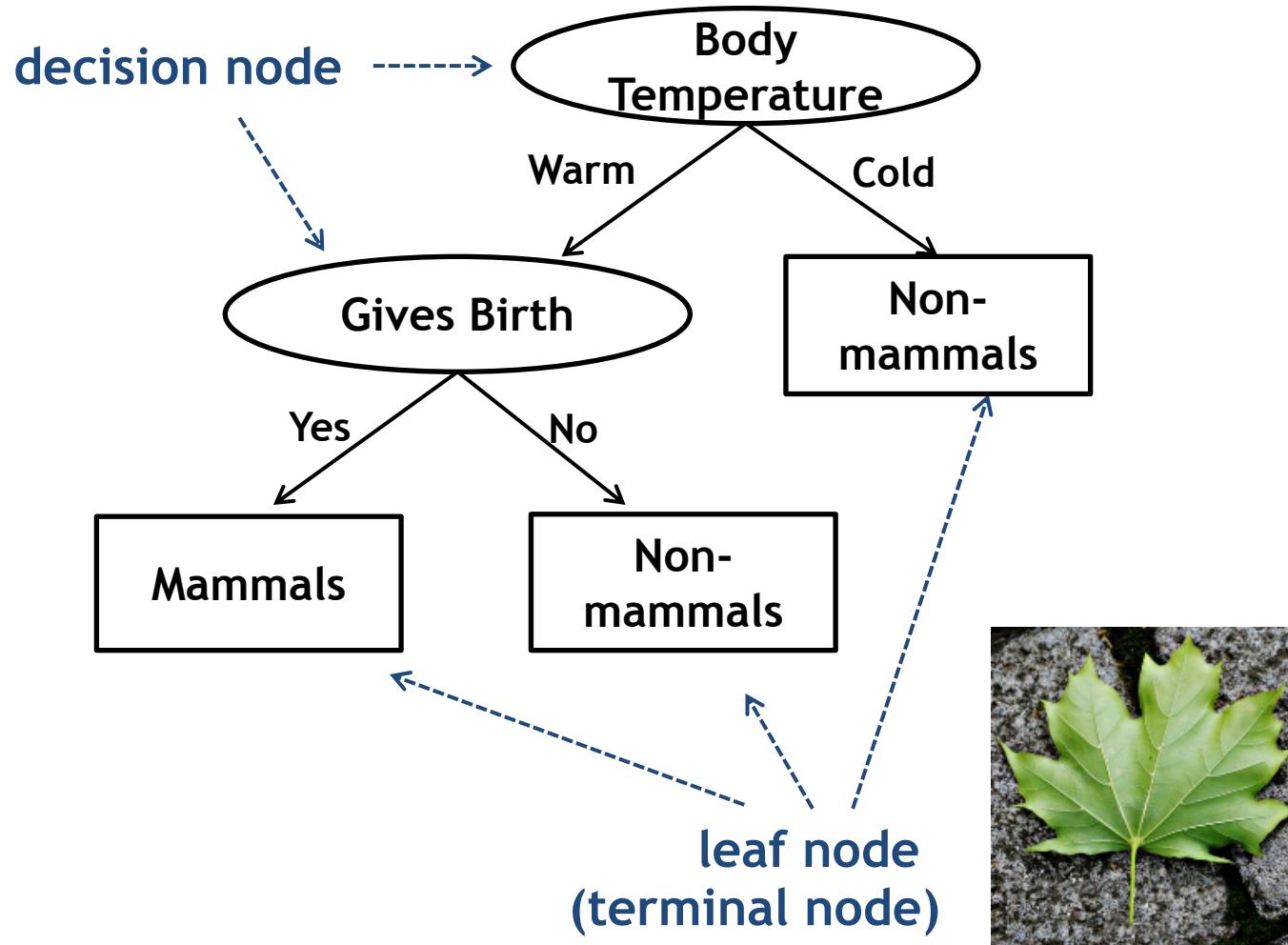
새로운 데이터

의사결정나무 예시 : 포유류 분류

❖ 분류 완료



의사결정나무의 구성



의사결정나무 예시 : 채무불이행자 분류

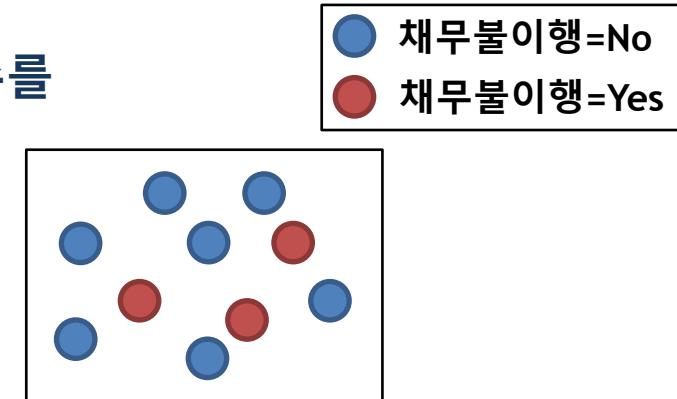
- ❖ 세 가지 입력변수와 범주형 출력변수로 구성된 다음 데이터를 이용하여, 의사결정나무 모델을 학습한 후 채무불이행자를 분류해보자.

ID	집 소유	결혼	연소득(K)	채무 불이행
1	Yes	미혼	125	No
2	No	기혼	100	No
3	No	미혼	70	No
4	Yes	기혼	120	No
5	No	이혼	95	Yes
6	No	기혼	60	No
7	Yes	이혼	220	No
8	No	미혼	85	Yes
9	No	기혼	75	No
10	No	미혼	90	Yes

레코드를 분기할 변수 찾기

- ❖ 10개의 레코드를 도식화하여, 의사결정나무가 변수를 선택하는 과정을 살펴보자.

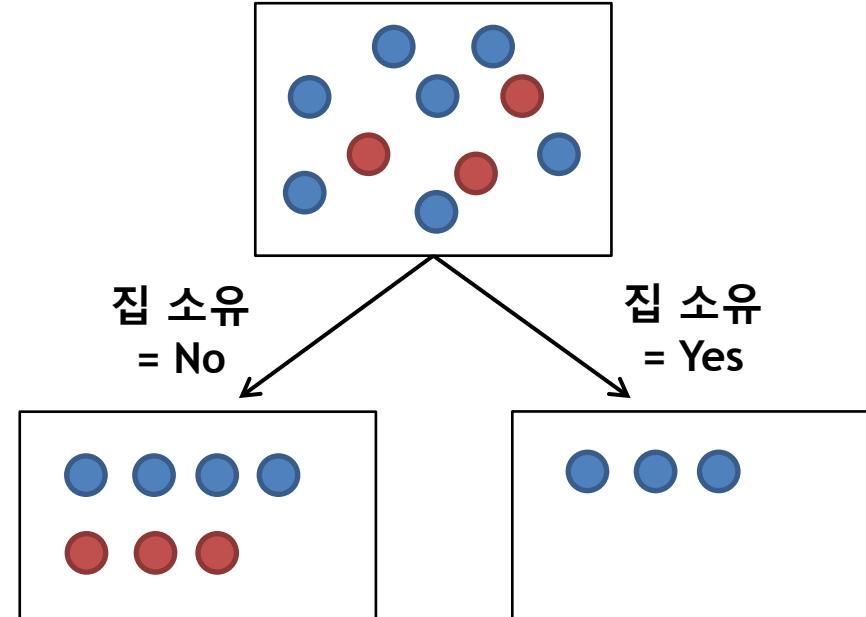
ID	집 소유	결혼	연소득 (K)	채무 불이행
1	Yes	미혼	125	No
2	No	기혼	100	No
3	No	미혼	70	No
4	Yes	기혼	120	No
5	No	이혼	95	Yes
6	No	기혼	60	No
7	Yes	이혼	220	No
8	No	미혼	85	Yes
9	No	기혼	75	No
10	No	미혼	90	Yes



레코드를 분기할 변수 찾기

❖ “집 소유” 변수로 분기 시, Gini index를 계산

ID	집 소유	결혼	연소득 (K)	채무 불이행
1	Yes	미혼	125	No
2	No	기혼	100	No
3	No	미혼	70	No
4	Yes	기혼	120	No
5	No	이혼	95	Yes
6	No	기혼	60	No
7	Yes	이혼	220	No
8	No	미혼	85	Yes
9	No	기혼	75	No
10	No	미혼	90	Yes



$$1 - \left(\frac{4}{7} \right)^2 - \left(\frac{3}{7} \right)^2 = \frac{24}{49}$$
$$1 - \left(\frac{3}{3} \right)^2 = 0$$

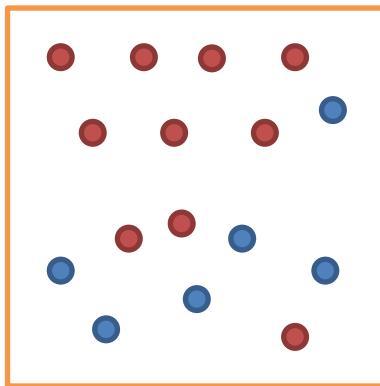
$$GINI_{집\ 소유} = \frac{24}{49} \times \frac{7}{10} + 0 \times \frac{3}{10} = 0.343$$

Measuring Impurity: Gini Index

❖ Gini Index for rectangle A containing m records

$$I(A) = 1 - \sum_{k=1}^m p_k^2$$

- ▶ p = proportion of cases in rectangle A that belong to class k.



$$\begin{aligned} I(A) &= 1 - \sum_{k=1}^m p_k^2 \\ &= 1 - \left(\frac{6}{16} \right)^2 - \left(\frac{10}{16} \right)^2 \\ &\approx 0.47 \end{aligned}$$

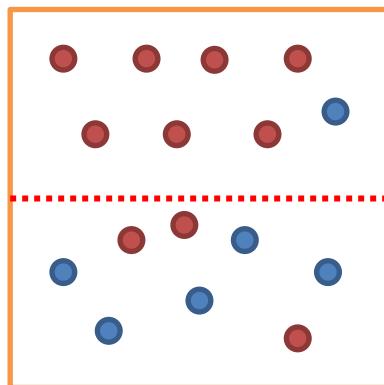
- ▶ $I(A) = 0$ when all cases belong to the same class.
- ▶ Max value when all classes are equal represented (=0.5 in binary case)

Measuring Impurity: Gini Index

- ❖ When there are more than two rectangles

$$I(A) = \sum_{i=1}^d \left(R_i \left(1 - \sum_{k=1}^m p_{ik}^2 \right) \right)$$

- ▶ R_i = proportion of cases in rectangle R_i among the training data.



$$\begin{aligned} I(A) &= 0.5 \times \left(1 - \left(\frac{7}{8} \right)^2 - \left(\frac{1}{8} \right)^2 \right) + 0.5 \times \left(1 - \left(\frac{3}{8} \right)^2 - \left(\frac{5}{8} \right)^2 \right) \\ &= 0.34 \end{aligned}$$

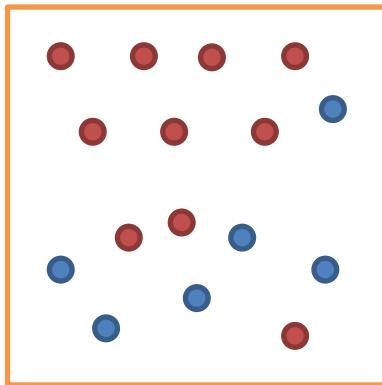
- ▶ “Information gain” after splitting: $0.47 - 0.34 = 0.13$

Measuring Impurity: Entropy

- ❖ Entropy for rectangle A containing m records

$$Entropy(A) = -\sum_{k=1}^m p_k \log_2(p_k)$$

- ▶ p = proportion of cases in rectangle A that belong to class k.



$$\begin{aligned} E(A) &= -\sum_{k=1}^m p_k \log_2(p_k) \\ &= -\frac{6}{16} \log_2\left(\frac{6}{16}\right) - \frac{10}{16} \log_2\left(\frac{10}{16}\right) \\ &\approx 0.95 \end{aligned}$$

- ▶ Entropy ranges between 0 (most pure) and $\log_2(m)$ (equal representation of classes)

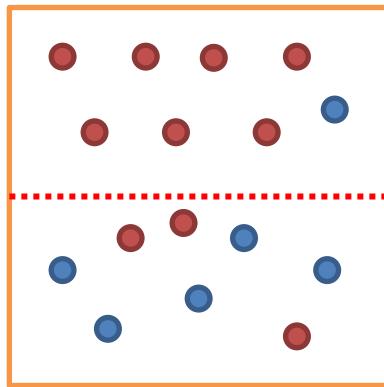
Measuring Impurity: Entropy

- ❖ When there are more than two rectangles

$$Entropy(A) = \sum_{i=1}^d R_i \times \left(- \sum_{k=1}^m p_{ik} \log_2(p_{ik}) \right)$$

- ▶ R_i = proportion of cases in rectangle R_i among the training data.

$$E(A)$$



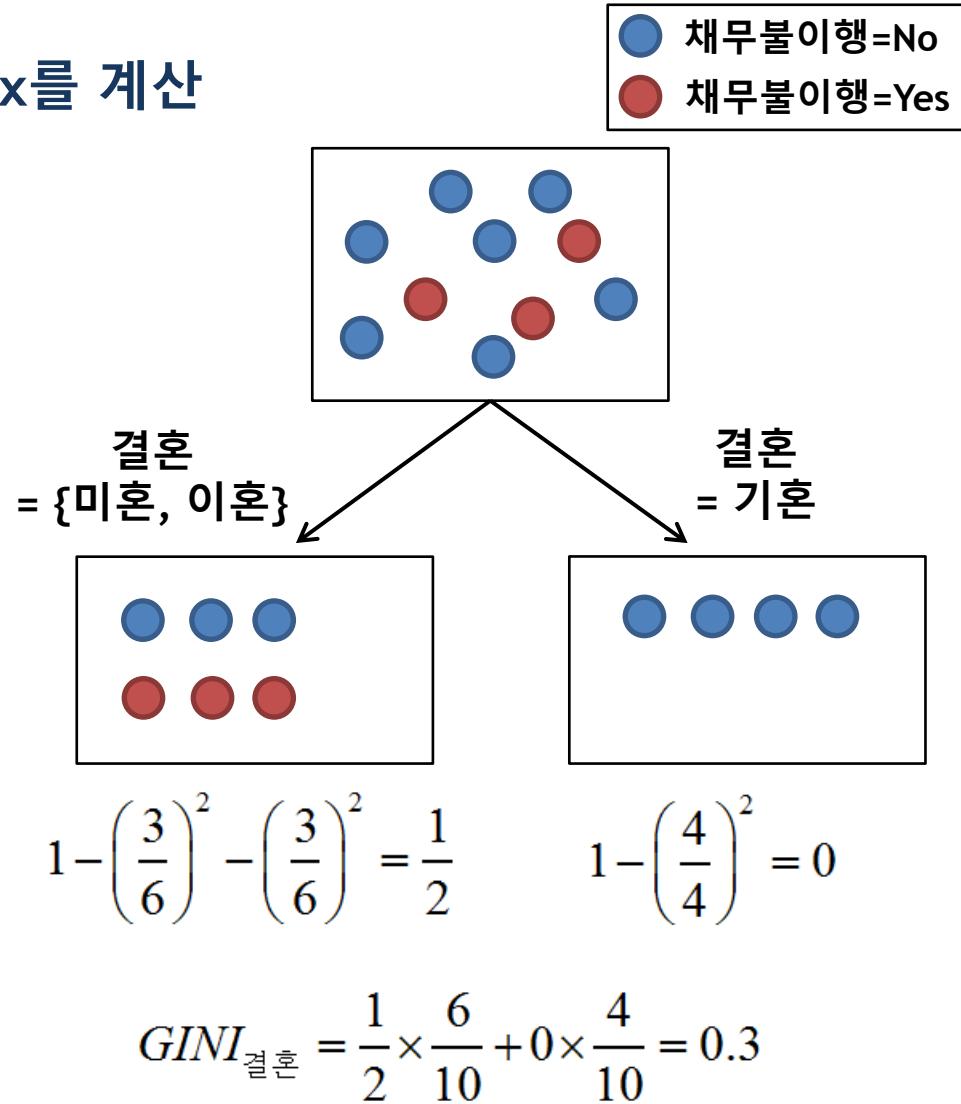
$$\begin{aligned} E(A) &= 0.5 \times \left(-\frac{1}{8} \log_2 \left(\frac{1}{8} \right) - \frac{7}{8} \log_2 \left(\frac{7}{8} \right) \right) \\ &\quad + 0.5 \times \left(-\frac{3}{8} \log_2 \left(\frac{3}{8} \right) - \frac{5}{8} \log_2 \left(\frac{5}{8} \right) \right) \\ &= 0.75 \end{aligned}$$

- ▶ “Information gain” after splitting: $0.95 - 0.75 = 0.20$

레코드를 분기할 변수 찾기

- ❖ “결혼” 변수로 분기 시, Gini index를 계산

ID	집 소유	결혼	연소득 (K)	채무 불이행
1	Yes	미혼	125	No
2	No	기혼	100	No
3	No	미혼	70	No
4	Yes	기혼	120	No
5	No	이혼	95	Yes
6	No	기혼	60	No
7	Yes	이혼	220	No
8	No	미혼	85	Yes
9	No	기혼	75	No
10	No	미혼	90	Yes



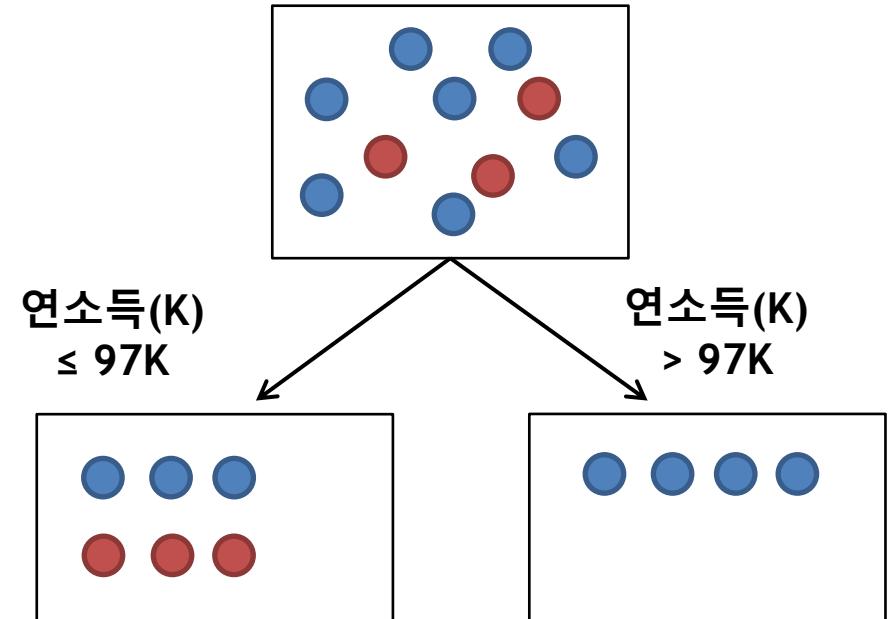
Note: How to split categorical variables?

- ❖ Examine all possible ways in which the categories can be split.
 - ▶ Example: 3 categories A, B, and C
 - {A} vs. {B,C}
 - {B} vs. {A,C}
 - {C} vs. {A,B}
- ❖ With many categories, the number of splits becomes huge.

레코드를 분기할 변수 찾기

- ❖ “연소득(K)” 변수로 분기 시, Gini index를 계산

ID	집 소유	결혼	연소득 (K)	채무 불이행
1	Yes	미혼	125	No
2	No	기혼	100	No
3	No	미혼	70	No
4	Yes	기혼	120	No
5	No	이혼	95	Yes
6	No	기혼	60	No
7	Yes	이혼	220	No
8	No	미혼	85	Yes
9	No	기혼	75	No
10	No	미혼	90	Yes



$$1 - \left(\frac{3}{6} \right)^2 - \left(\frac{3}{6} \right)^2 = \frac{1}{2}$$

$$1 - \left(\frac{4}{4} \right)^2 = 0$$

$$GINI_{연소득} = \frac{1}{2} \times \frac{6}{10} + 0 \times \frac{4}{10} = 0.3$$

Note: How to split numerical variables?

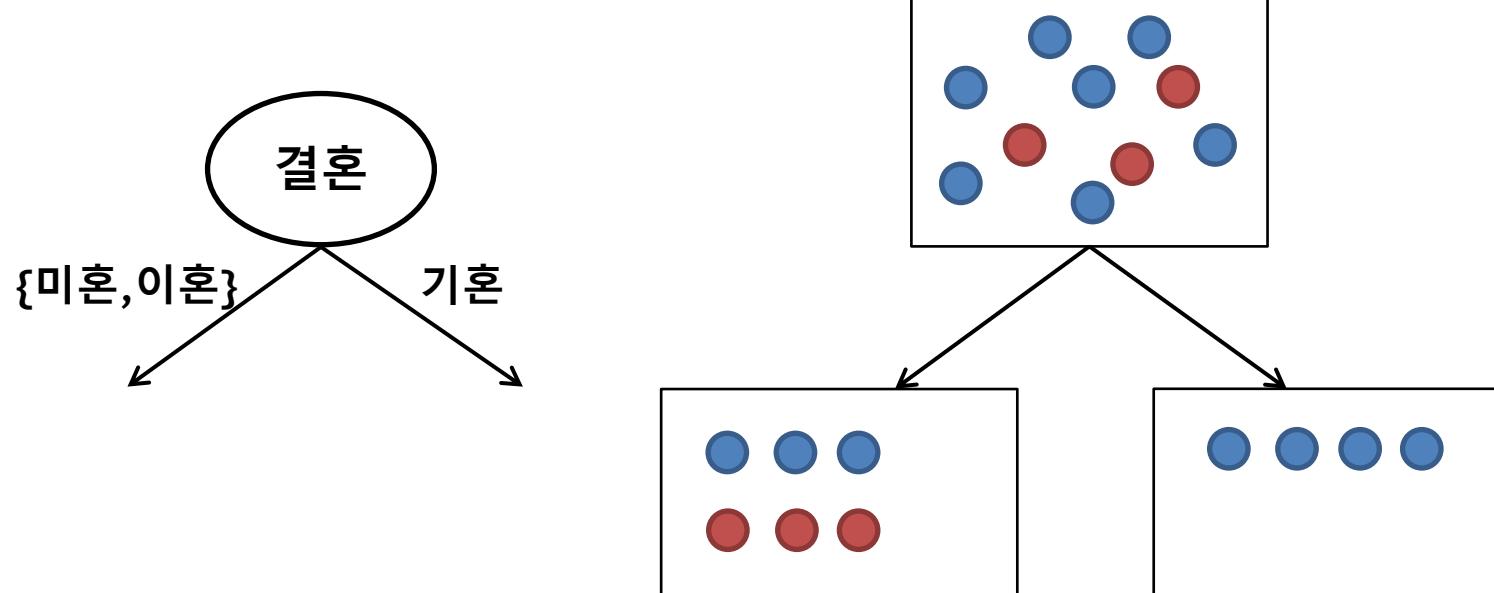
- ❖ 연속형 변수에 대해 정렬한 후, 일정한 간격으로 분기점 후보들을 선정
- ❖ 각 분기점에서의 Gini index (또는 Entropy)를 계산한 후, 최소인 때로 분기점을 정한다.

채무 불이행	No	No	No	Yes	Yes	Yes	No	No	No	No										
	연소득																			
	60	70	75	85	90	95	100	120	125	220										
분기점	55	65	72	80	87	92	97	110	122	172	230									
	<=	>	<=	>	<=	>	<=	>	<=	>	<=									
Yes	0	3	0	3	0	3	1	2	2	1	3	0	3	0	3	0				
No	0	7	1	6	2	5	3	4	3	4	3	4	4	3	5	2	6	1	7	0
Gini	0.420	0.400	0.375	0.343	0.417	0.400	0.300	0.343	0.375	0.400	0.420									

레코드를 분기할 변수 찾기

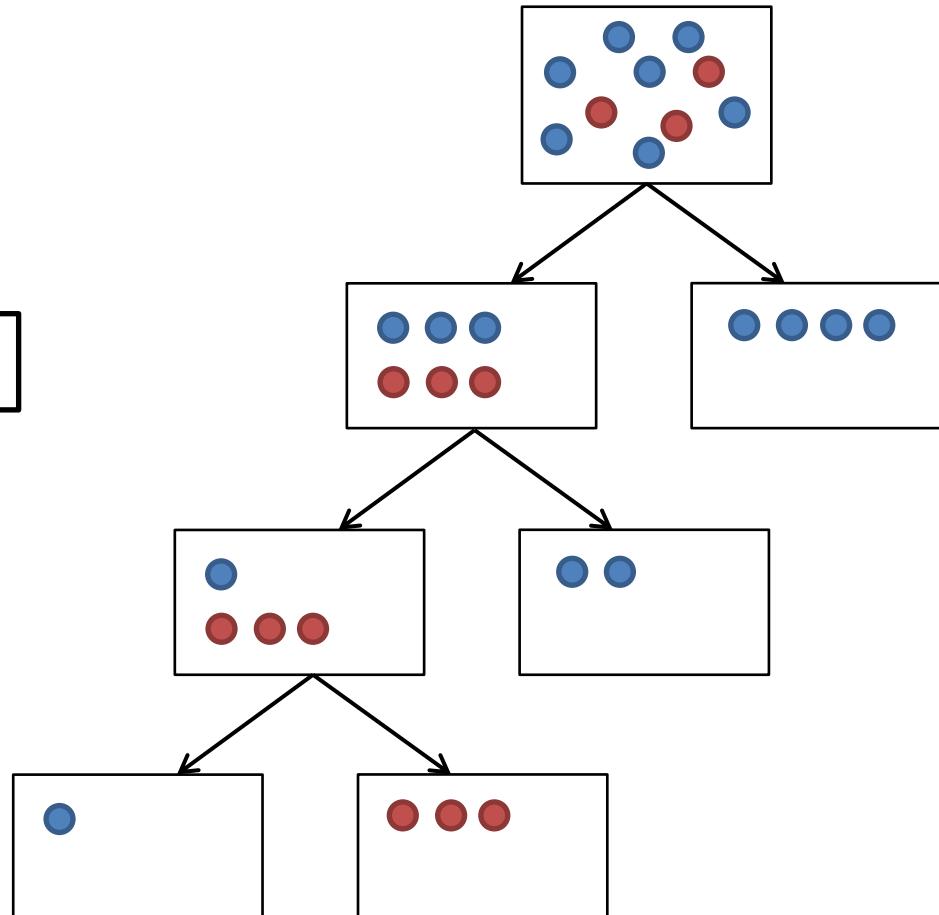
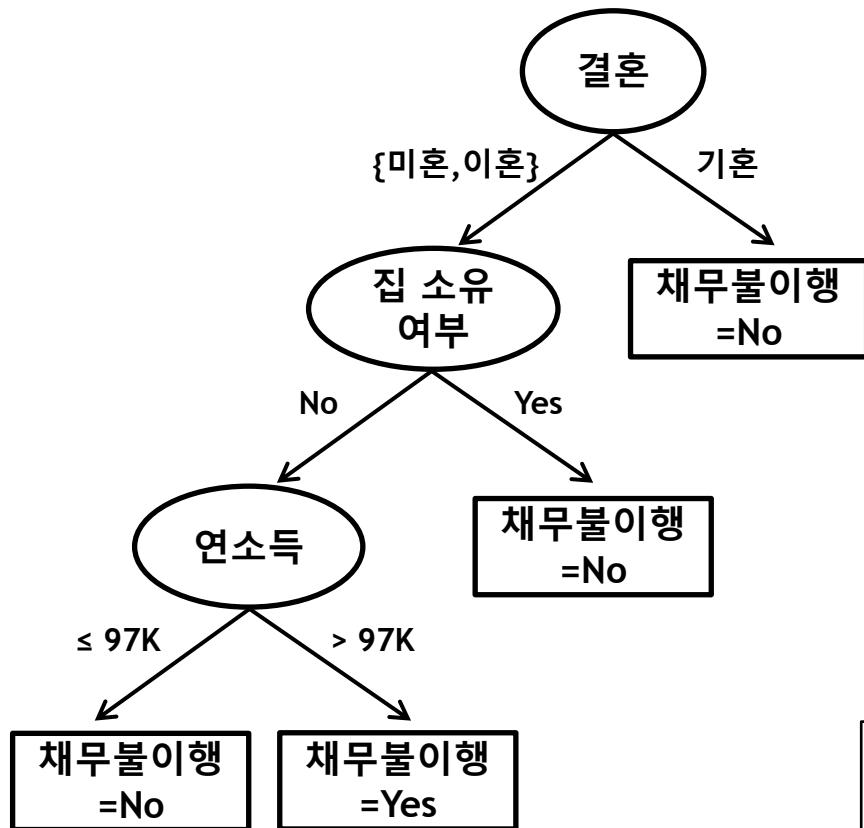
❖ 첫 번째로 분기할 변수 선정 : “결혼” 선정

입력변수	Gini index
집 소유	0.343
결혼	0.3
연소득	0.3



의사결정나무 모델의 학습

- ❖ 이와 같은 방법으로 분기를 계속 진행하면 다음과 같은 의사결정나무를 도출할 수 있다.

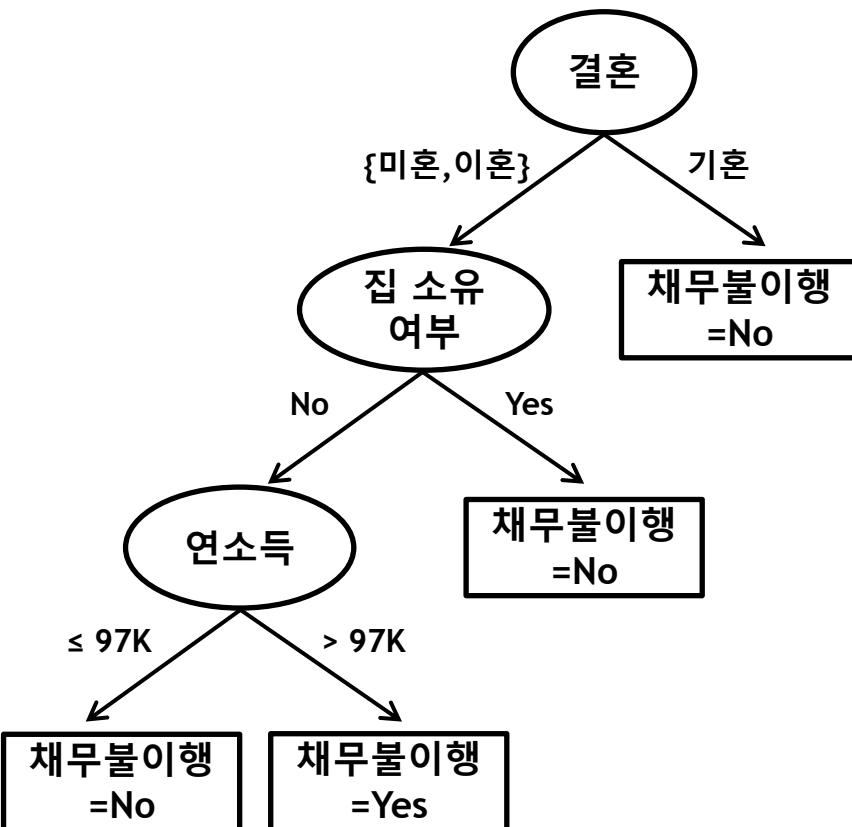


Recursive Partitioning (재귀적 분기)

- ❖ Pick one of the predictor variables, x_i
- ❖ Pick a value of x_i , e.g., s_i , that divides the training data into two (not necessarily equal) portions.
- ❖ Measure how “**pure**” or “**homogeneous**” each of resulting portions are
 - ❖ “Pure” = containing records of mostly one class.
 - ❖ Algorithm tries different values of x_i , and s_i , so as to maximize the purity in initial split.
 - ❖ After getting a “**maximum purity**” split, repeat the process for a second split, and so on.

새로운 데이터의 클래스를 예측하기

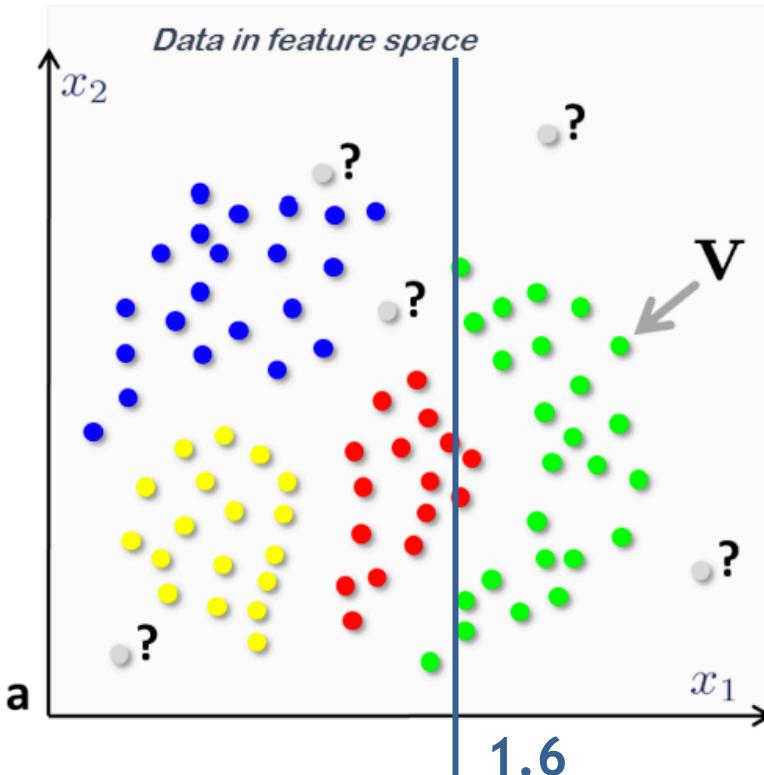
- ❖ 다음과 같이 새로운 데이터가 주어지면, 미리 학습 데이터를 이용하여 구축한 의사결정나무 모델을 이용하여 새로운 데이터의 채무불이행 여부를 예측할 수 있다.



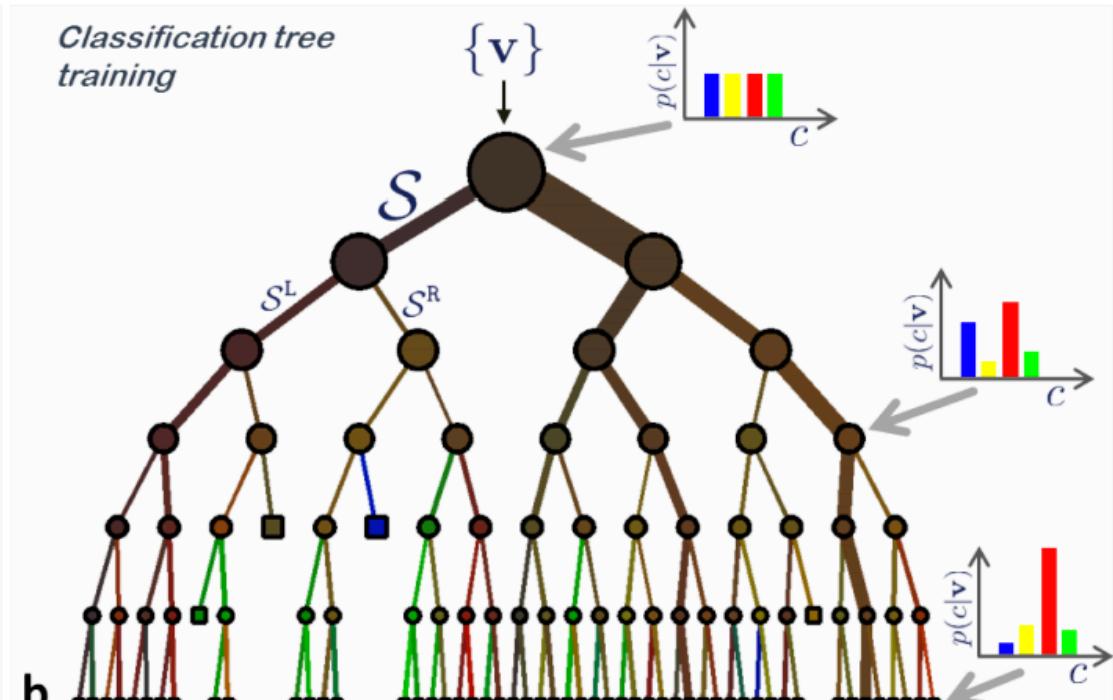
ID	집 소유	결혼	연소득 (K)	채무 불이행	예측 클래스
11	No	미혼	55	?	No
12	Yes	기혼	80	?	No
13	Yes	미혼	110	?	No
14	No	기혼	95	?	No
15	No	이혼	300	?	Yes

Single decision tree model

A decision tree choose the split that results in the purest daughter



(a) Input data
with 4 classes

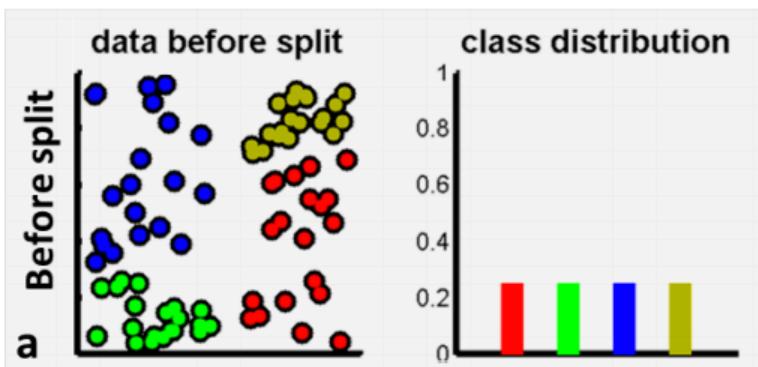


(b) A classification tree

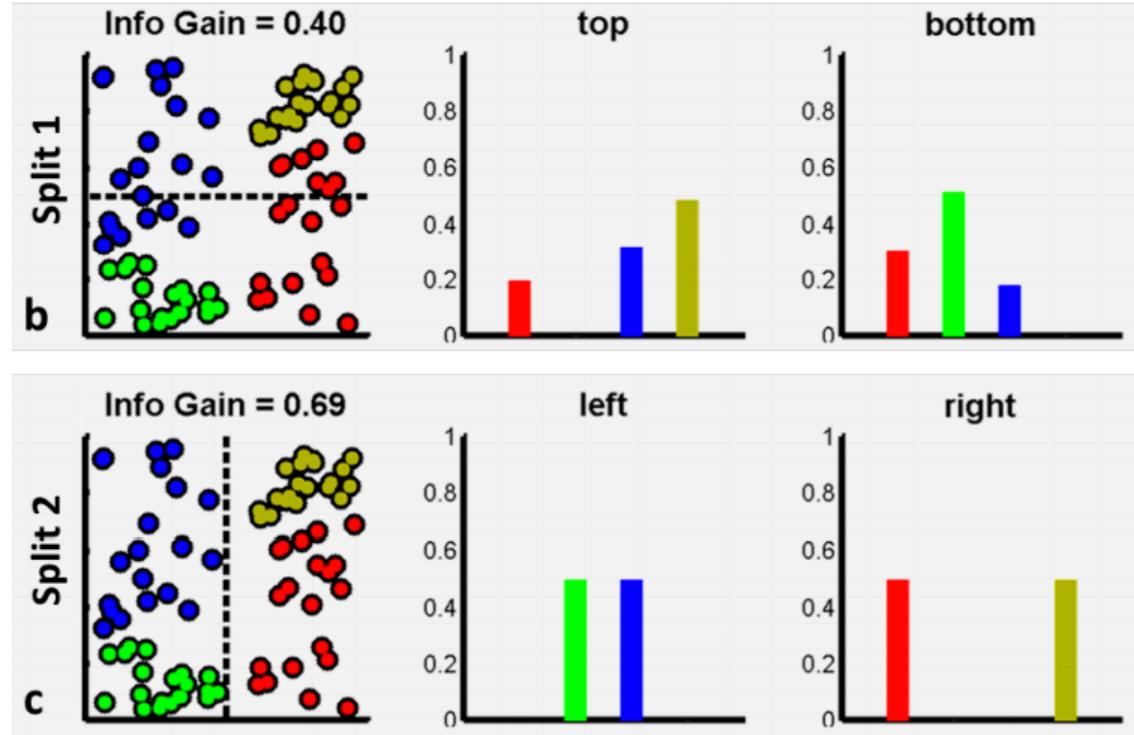
(Criminisi et al., 2011)

Single decision tree model

The decision tree model use information gain to decide splits.



(a) Dataset before a split



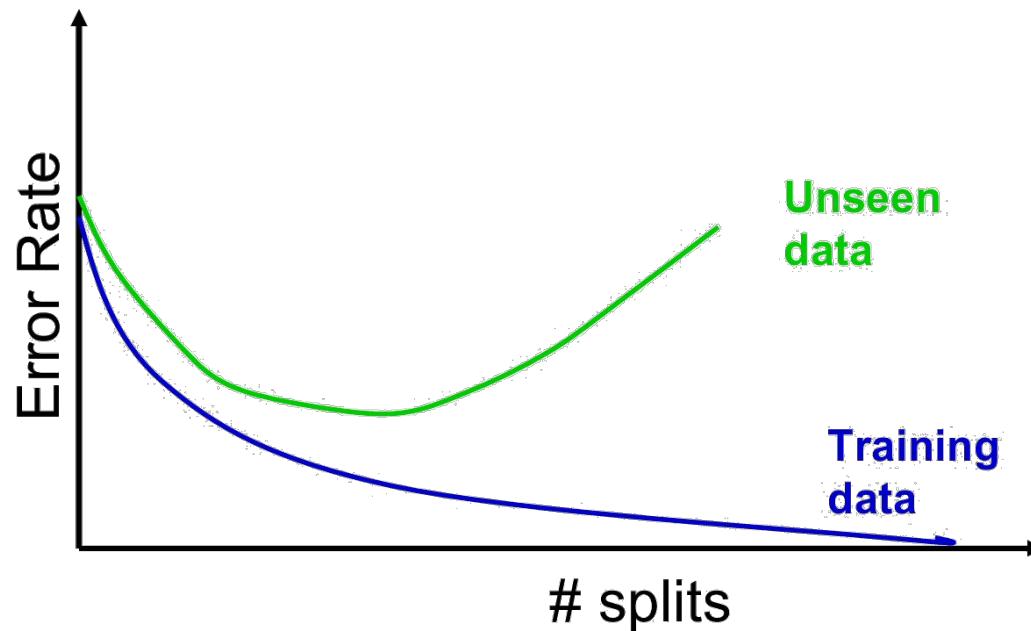
(b) After a horizontal split
(c) After a vertical split

$$I_j = H(\mathcal{S}_j) - \sum_{i \in \{L,R\}} \frac{|\mathcal{S}_j^i|}{|\mathcal{S}_j|} H(\mathcal{S}_j^i)$$

(Criminisi et al., 2011)

CART: Overfitting Problem

- ❖ Natural end of process is 100% purity in each leaf.
- ❖ It overfits the data, ending up fitting noise in the data and leading to low predictive accuracy of new data.
- ❖ Past a certain point, the error rate for the validation data starts to increase.



CART: Full-grown Tree와 Pruning

❖ Full-grown Tree

- ▶ Training data를 완벽하게 분류해내는 의사결정나무를 뜻한다.
- ▶ Full-grown Tree는 과연 새로운 데이터에도 좋은 성능을 보일 것인가?
 - 일반적으로 Training data에 과적합한(Overfitting) 모델은 예측 성능이 떨어지는 것으로 알려져 있다.
 - 따라서, 학습에 사용하지 않은 Validation data에서의 error를 고려해야 한다.

❖ Pruning(가지치기)

- ▶ Full-grown Tree에서 맨 아래부터 가지를 하나씩 쳐나간다.
- ▶ Validation data에서의 error와 나무구조의 복잡도를 고려하여 최적의 나무를 찾아낸다.

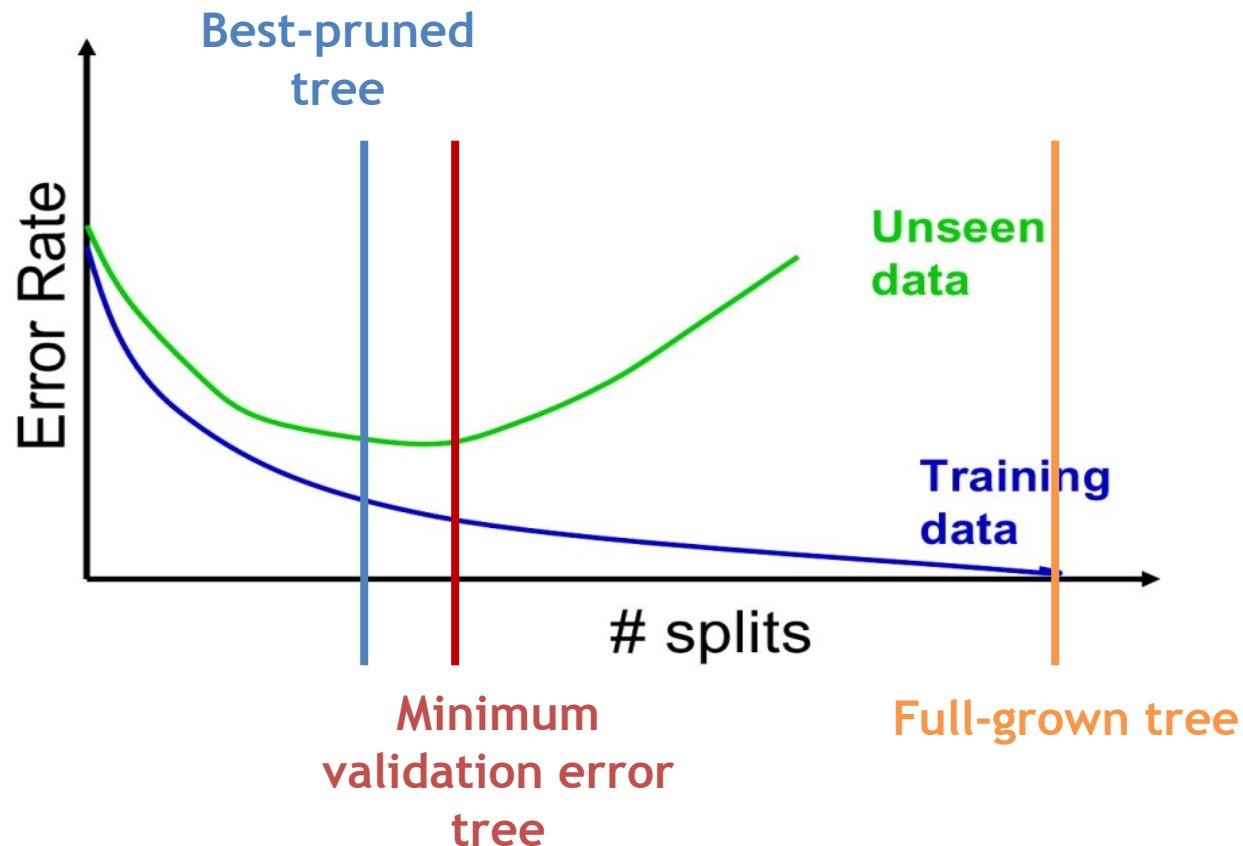
Pruning: Cost complexity

- ❖ CART를 가지치기할 때에는 검증 데이터에서의 예측 성능과 나무 구조의 복잡도를 동시에 고려함.
- ❖ 다음과 같은 Cost complexity를 이용하여, 가장 CC가 작은 지점에서 가지치기를 종료

$$CC(T) = Err(T) + \alpha \times L(T)$$

- ▶ $CC(T)$ = cost complexity of a tree
- ▶ $ERR(T)$ = proportion of misclassified records in the validation data
- ▶ Alpha = penalty factor attached to the tree size (set by the user)

Best-pruned Tree



Best-pruned Tree

의사결정 마디	학습용 집합의 오차율	평가용 집합의 오차율
41	0	2.133333
40	0.04	2.2
39	0.08	2.2
38	0.12	2.2
37	0.16	2.066667
36	0.2	2.066667
35	0.2	2.066667
34	0.24	2.066667

...

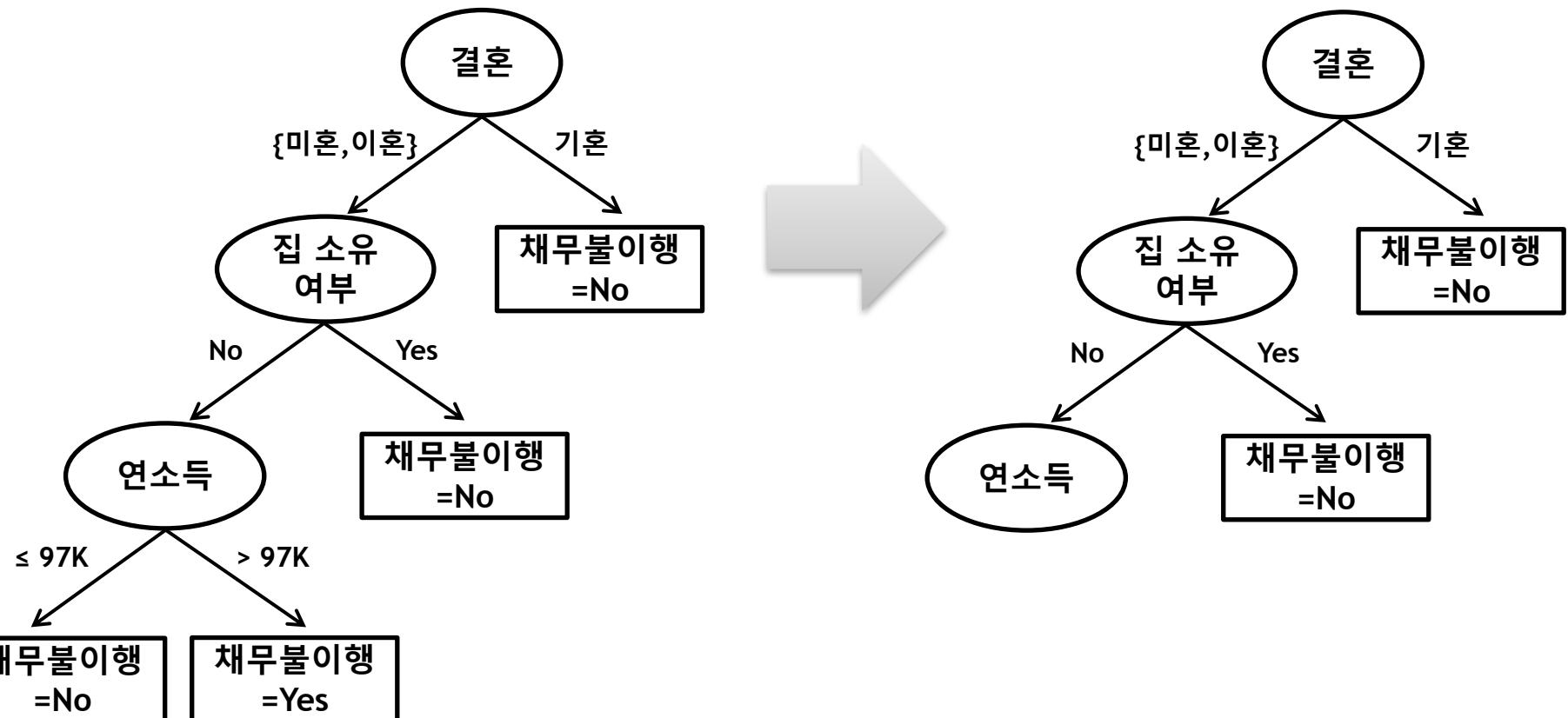
...

...

13	1.16	1.6			
12	1.2	1.6			
11	1.2	1.466667	최소 오차 나무	표준 오차	0.003103929
10	1.6	1.666667			
9	2.2	1.666667			
8	2.2	1.866667			
7	2.24	1.866667			
6	2.24	1.6	<- 최적의 가지친 나무		
5	4.44	1.8			
4	5.08	2.333333			
3	5.24	3.466667			
2	9.4	9.533333			
1	9.4	9.533333			
0	9.4	9.533333			

의사결정나무의 가지치기

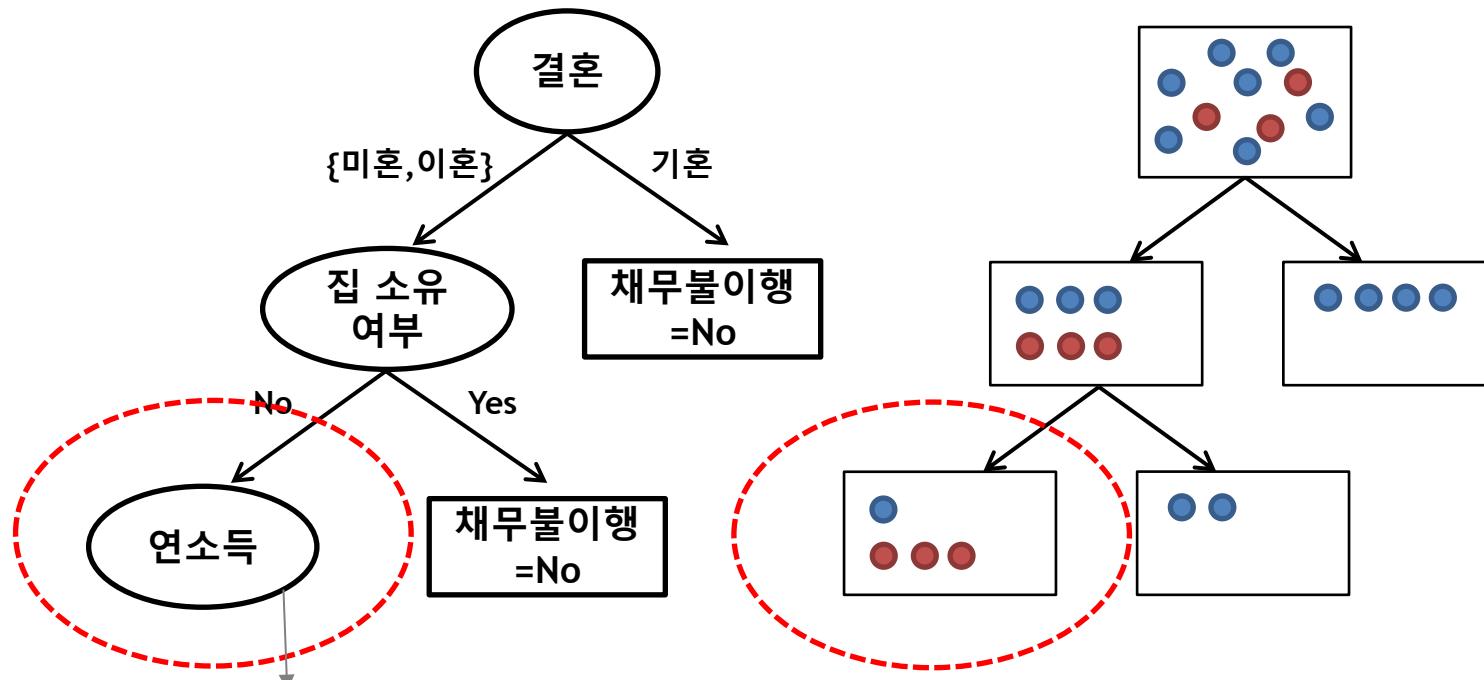
- (앞서의 예시에서) Validation data에서의 error를 기반으로 pruning을 수행했을 때 결과가 다음과 같다고 가정하자.



Terminal Node에서의 클래스 결정

❖ Cut-off value (Binary class의 경우)

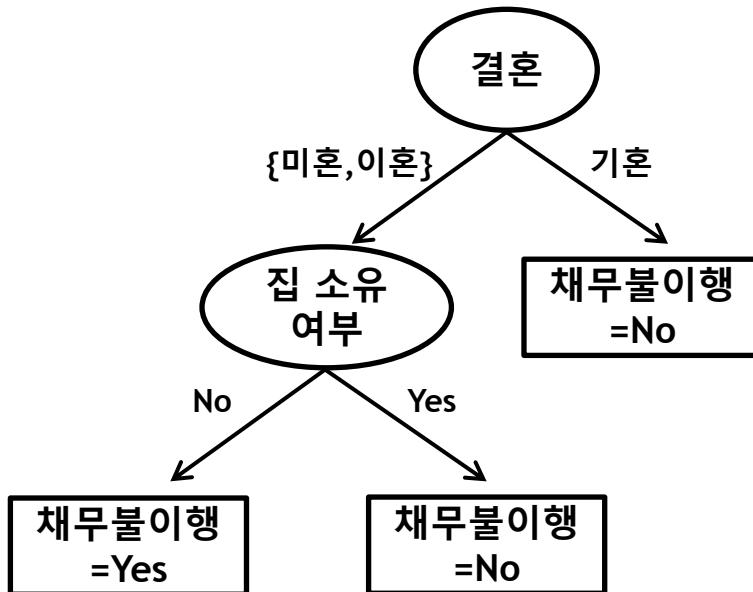
- ▶ Terminal Node에서 클래스를 결정짓는 기준
- ▶ 일반적으로 0.5로 설정하며, 실제 데이터에서의 클래스 분포 확률(사전 확률: prior prob.)을 고려하여 다른 값으로 선택하기도 함



cut-off value = 0.5일 때, “채무불이행=Yes”인 terminal node로 변경
cut-off value = 0.8일 때, “채무불이행=No” 인 terminal node로 변경

규칙 도출

- ❖ Cut-off value를 0.5로 선정했을 때, 최종 의사결정 나무의 형태와 도출되는 규칙은 다음과 같다.



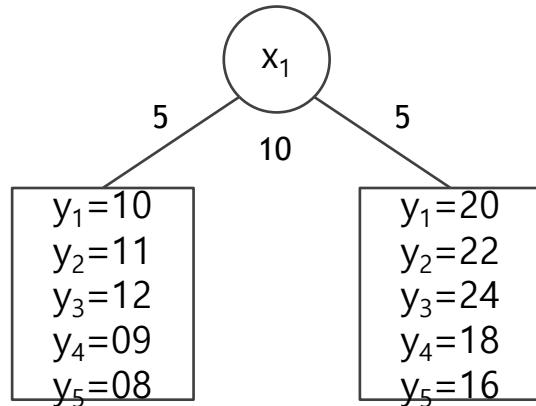
- If 결혼='기혼', then 채무불이행 = No
- If 결혼='미혼 or 이혼' and 집소유여부=No, then 채무불이행 = YES
- If 결혼='미혼 or 이혼' and 집소유여부=Yes, then 채무불이행 = No

Regression Tree

Similar process with classification tree except

❖ Prediction of the node

- ▶ The average of the outcome variables belonging to the node.



- Predicted value of the left leaf node = 10
- Predicted value of the right leaf node = 20

❖ Impurity

- ▶ Sum of the squared error (SSE: $\sum_{i=1}^n (y_i - \hat{y})^2$)
- ▶ SSE(Parent) = 300, SSE(Left) = 10, SSE(Right) = 40, Gain = 250

Regression Tree

❖ Predict the selling price of Toyota corolla...

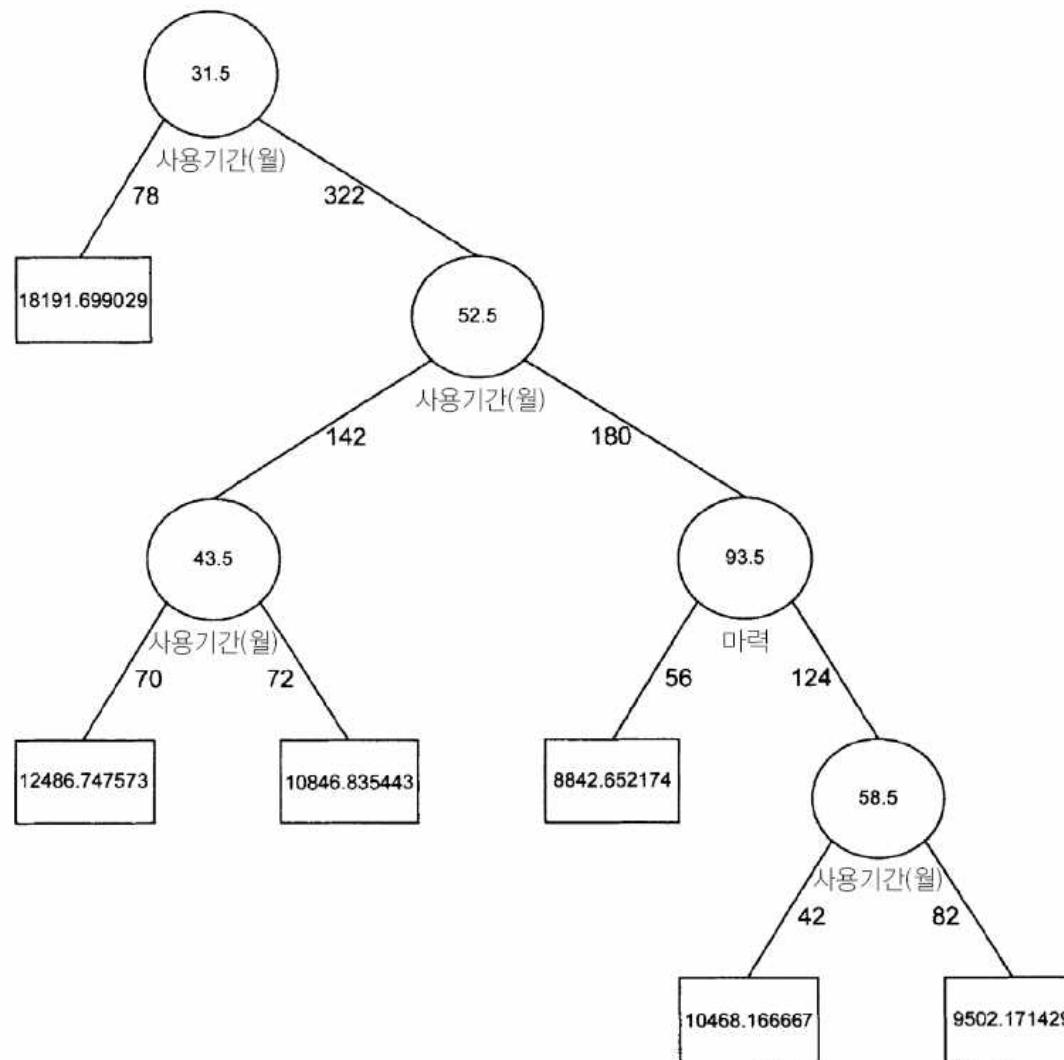


Dependent variable
(target)

Independent variables
(attributes, features)

Variable	Description
Price	Offer Price in EUROS
Age_08_04	Age in months as in August 2004
KM	Accumulated Kilometers on odometer
Fuel_Type	Fuel Type (Petrol, Diesel, CNG)
HP	Horse Power
Met_Color	Metallic Color? (Yes=1, No=0)
Automatic	Automatic ((Yes=1, No=0)
CC	Cylinder Volume in cubic centimeters
Doors	Number of doors
Quarterly_Tax	Quarterly road tax in EUROS
Weight	Weight in Kilograms

Regression Tree



CART: Summary

❖ Advantages

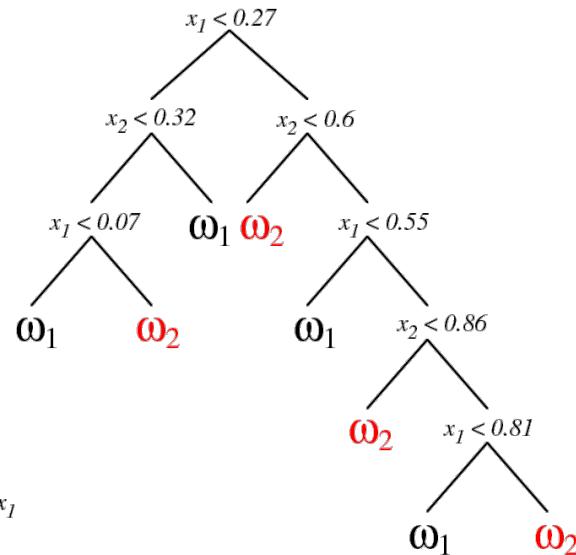
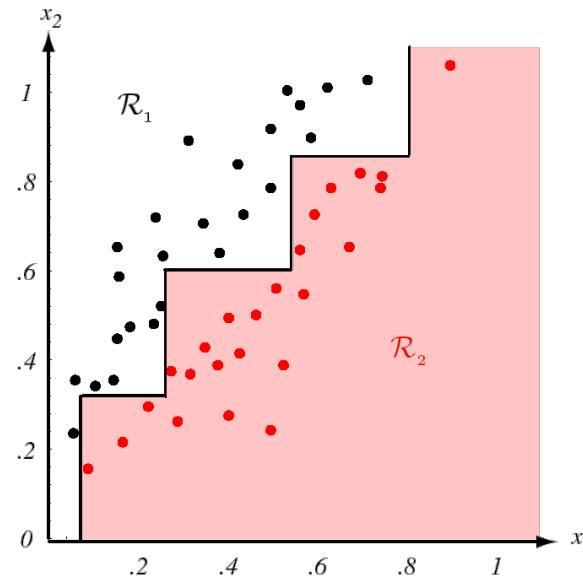
- ▶ Classification and regression tree (CART) is **easy to use and understand**.
- ▶ Produce rules that are **easy to interpret & implement**.
- ▶ Variable selection & reduction is automatic.
- ▶ Do not require the assumptions of statistical models.
- ▶ Can work without extensive handling of missing data.

❖ Disadvantages

- ▶ May not perform well where there is structure in the data that is not well captured by **horizontal or vertical split**.
- ▶ Since the process deals with “one variable at a time”, no way to capture **interactions between variables**.

Counter Example

Possible



Impossible

