5주차 강의

모델링 - 분류기법

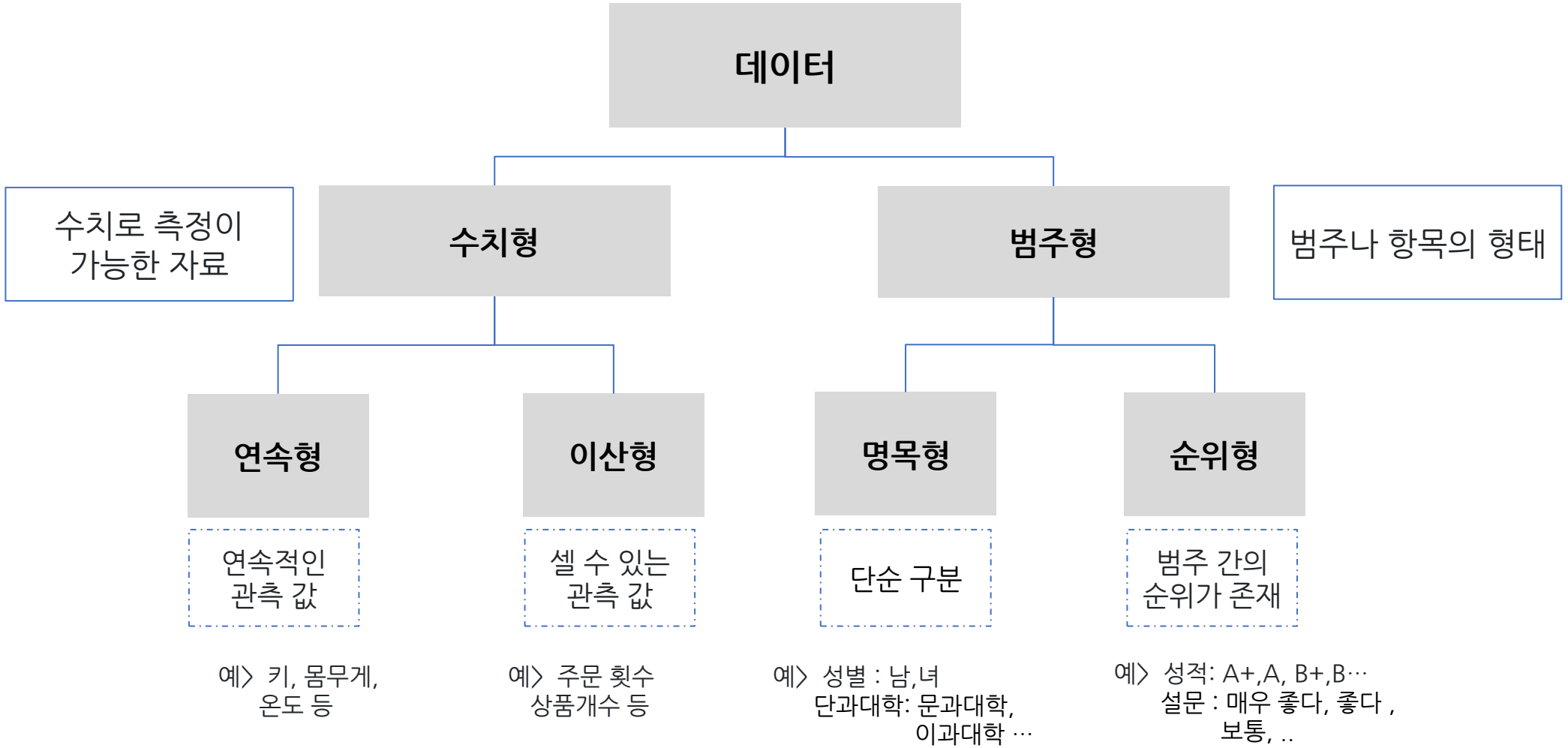
의사결정나무(Decision Tree)

2022.04.05

Institutor: JS LEE

WrapUp -데이터 탐색

데이터 값의 형태



WrapUp -데이터 탐색

1) 수치형 변수 - 기술 통계량

- 데이터셋의 주요 특성을 수량화 하기 위해 평균, 표준편차, 분포 등과 같이 요약하는 통계적 방법
- 기술적 척도들은 데이터셋에 대한 이해에 도움을 줌
- (예) 연평균 수입, 주택가격 중앙값, 신용점수 범위 등

R 과 기술 통계량

- summary()와 describe() 함수를 이용하면, 기술 통계량을 한번에 확인 가능함
- describe()함수는 psych 패키지에 내장되어 있으므로, 먼저 psych 패키지 인스톨 후, 로딩해야 함
- describe() 함수가 더 다양한 기술 통계량을 포함하고 있음
- 수치형 변수만을 선택하여 위 두 함수를 이용하는 것을 권장함

그래프 유형 :histogram, boxplot, plot

Wrap up

6) 범주형 변수 - 빈도분석

범주형은 각 변수의 범주가 어떻게 구성되어 있는지로, 데이터의 특성을 파악
그 대표적인 방법으로 도수분포표를 이용한 빈도분석이 있음.

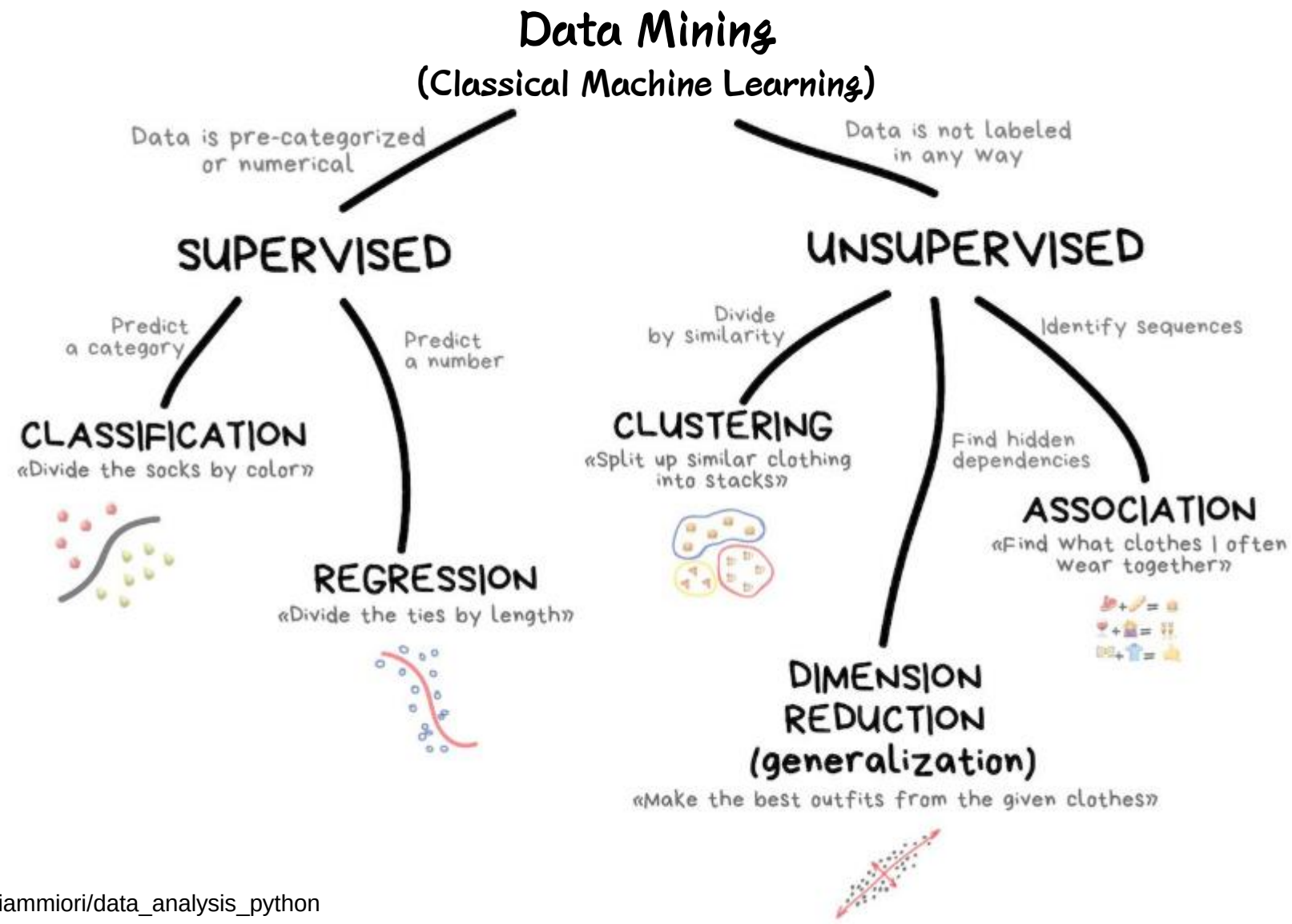
■ 도수분포표

- 계급, 도수 및 상대도수로 구성됨
- 계급(class) : 자료가 취하는 전체 범위를 몇 개의 소집단으로 나눈 것
- 도수(frequency) : 각 계급에 속하는 자료의 수
- 상대도수(relative frequency) : 도수를 전체 자료의 수, 즉 전체 도수로 나눈 비율

계급	도수	상대도수
남자	70	0.7
여자	30	0.3

- table(), freq()함수 이용
- describe()함수는 descr 패키지에 내장되어 있으므로, 먼저 descr 패키지 인스톨 후, 로딩해야 함
- describe()로 도수분포표와 막대그래프(bar chart) 생성가능
- table()함수를 이용한 결과를 barplot()함수를 이용하여 바 차트를 pie() 함수를 이용하여 파이차트(pie chart) 로 표현 가능

개요 - 분류

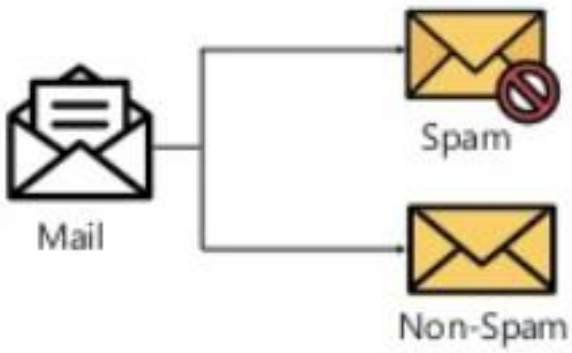


refer::https://github.com/iammiori/data_analysis_python

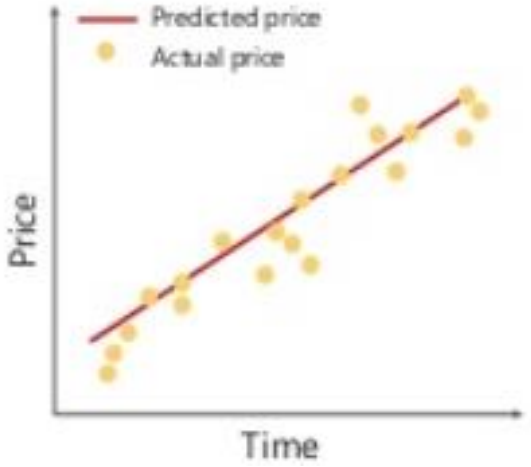
개요-분류

- Supervised (예측) : 회귀 vs. 분류
  - 입력 변수와 타겟(종속, 결과, 출력) 변수 간의 관계를 추정할 때,
    - 타겟 변수가 연속형 ➔ 회귀
    - 타겟 변수가 범주형 ➔ 분류

분류 (Classification)



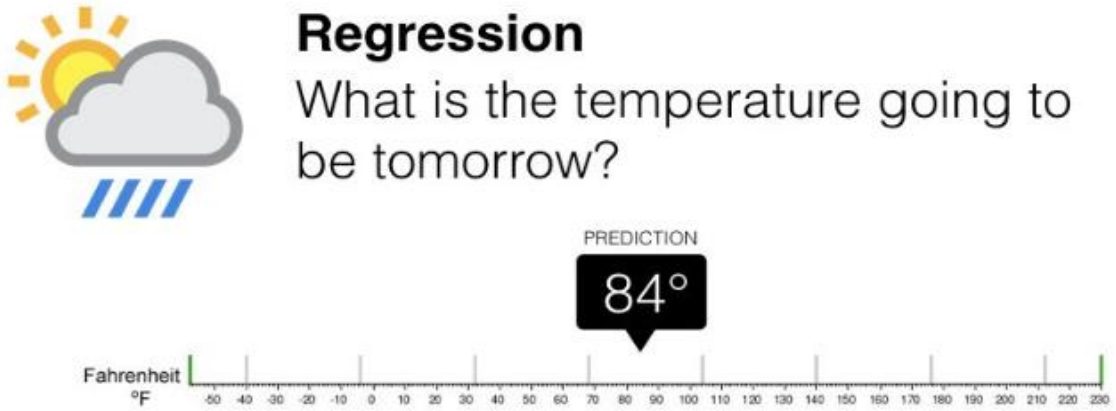
회귀 (Regression)

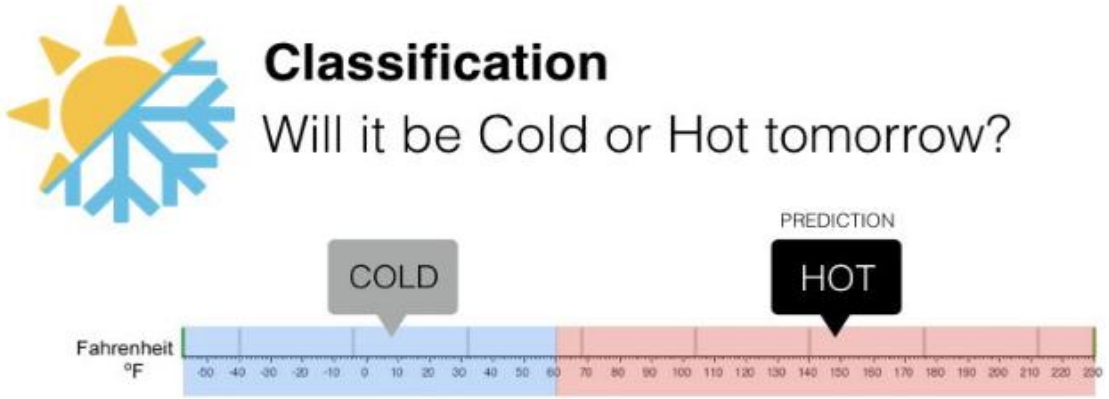


refer::https://github.com/iammiori/data_analysis_python

개요 - 분류

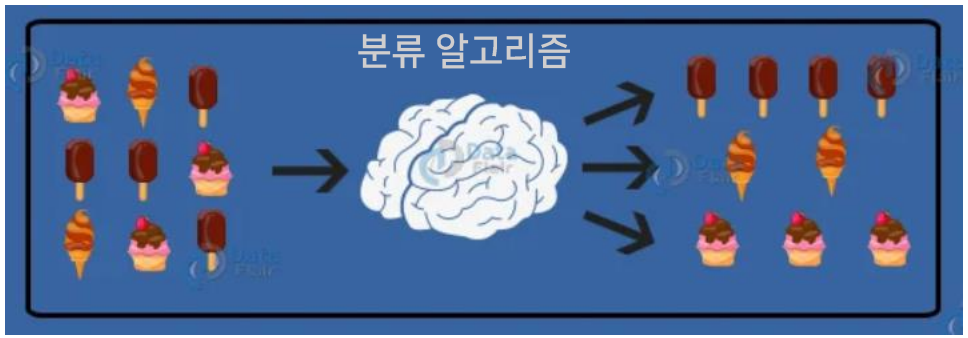
- Supervised ：회귀 vs. 분류
- 타겟(결과,종속,출력) 변수가 연속형 일지라도 범주형으로 변환시켜 분류 모델링을 적용할 수 있음





개요 - 분류

- 분류 예시



- 다음 달에 어떤 고객이 떠날 것인가?

  (유지고객 vs 이탈고객)

- 어떤 고객이 카드 값을 제 때에 입금할 것인가? 아닐 것인가?

  (정상고객 vs 연체고객)

- 쿠폰을 보냈을 때 어떤 고객이 구매를 할 것인가?

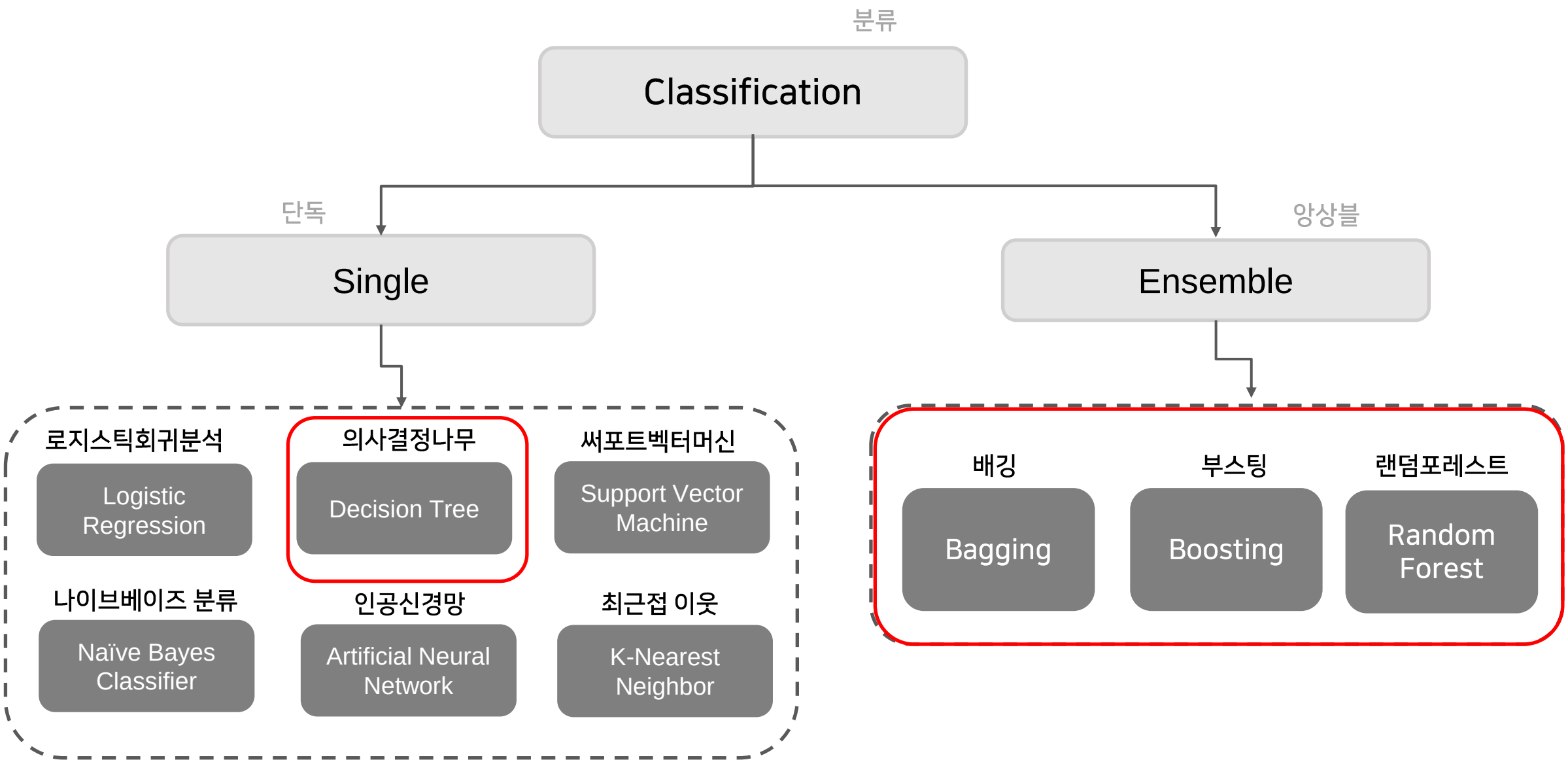
  (쿠폰 반응고객 vs 쿠폰 미반응고객)

- 병세를 바탕으로 암 환자 진단

  (정상 vs 암환자)

- 텍스트 분류
  - E-mail이 SPAM 일까? 아닐까?
  - 이 블로거는 우리의 제품을 좋아하나 좋아하지 않은가?
  - 뉴스기사를 살펴보니, 애플 주식은 상승할까? 하락할까?

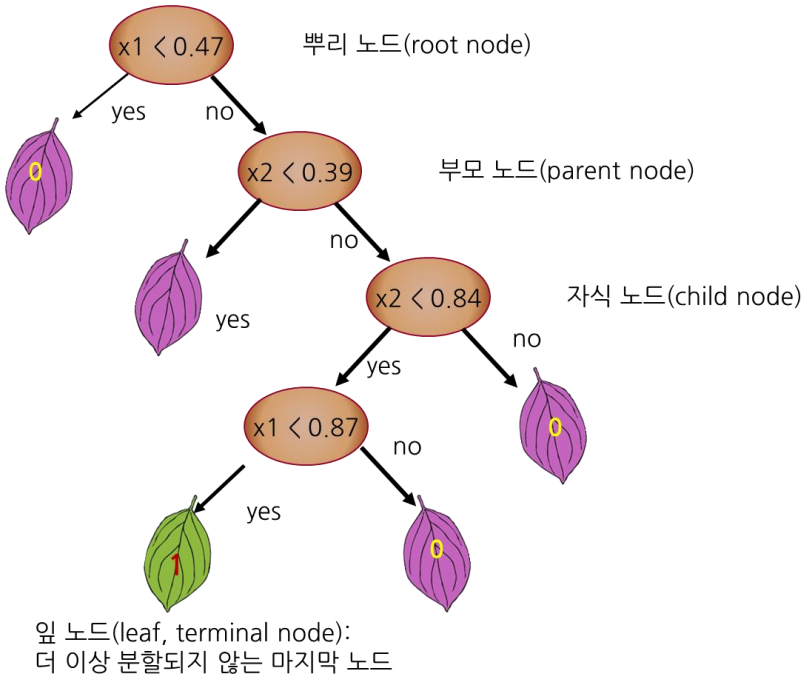
개요-분류


분류
Classification
단독
Single
앙상블
Ensemble
로지스틱회귀분석
Logistic Regression
의사결정나무
Decision Tree
써포트벡터머신
Support Vector Machine
나이브베이즈 분류
Naïve Bayes Classifier
인공신경망
Artificial Neural Network
최근접 이웃
K-Nearest Neighbor
배깅
Bagging
부스팅
Boosting
랜덤포레스트
Random Forest

의사결정나무 - 개요

- 대표적 데이터 마이닝 기법(머신러닝) 중의 하나로 Breiman 등 (1984)이 개발
- 전체 자료를 몇 개의 소집단으로 분류(classification)하거나 예측(prediction)을 하는 분석방법임
- 의사결정나무(decision tree) 또는 나무 모형(tree model)은 의사결정 규칙을 나무(tree) 구조로 나타내는 것임
- 분석의 목적과 자료구조에 따라서 분리기준(split criterion)과 정지 규칙(stopping rule)을 지정하여 의사결정나무를 구축

의사결정나무 - 개요



- ✓ 노드(node): 관측치(Observation)들의 집합체
- ✓ 가지(branch): 노드와 노드 연결
- ✓ 가지 분할: 어떤 법칙을 가지고 노드를 나누는 것

- 상위 노드에서 하위 노드로 내리는 방법(데이터를 부분집합으로 나누는 과정:가지치기)은 하위노드의 노드(집단)내에서는 동질성이 노드 간에는 이질성이 가장 커지도록 하는 분류변수와 분류기준이 선택되어 짐

- 상위 노드에서 하위 노드로 내리기(부분집합 나누는 과정, 가지치기)를 멈추는 조건은 다음과 같음
  - ✓ 노드에 있는 모든(또는 거의 모든) 관측치가 같은 클래스(범주)를 가질 때
  - ✓ 관측치(값)을 구별하는 특징이 남아 있지 않을 때
  - ✓ 미리 정의된 크기 한도까지 트리의 크기가 만들어졌을 때

- 나무 모형의 크기는 과대적합이 되지 않도록 가지치기(pruning)에 의해 적당히 조절되어야 함

의사결정나무 - 개요

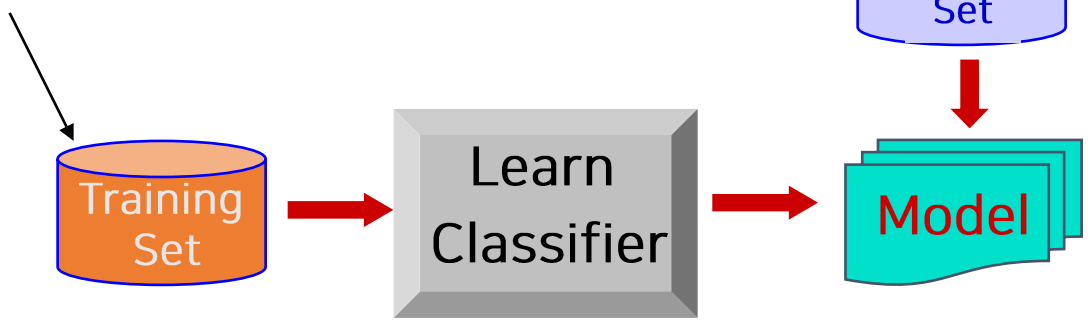
- 특징
  - 다수의 입력변수들과 타겟 변수의 관계에 대한 통찰을 얻을 수 있음
  - 분류와 수치 예측 모두 가능
  - 분석가 관점 → 사용 용이, 사용자 관점 → 이해 용이
    - ✓ 분류 규칙을 추출할 수 있음 (예) IF X>20 THEN Y=1
    - ✓ 분할의 기준이 되는 변수는 중요한 변수로 간주할 수 있음
  - 데이터 준비 과정의 노력이 상대적으로 덜 필요함
    - ✓ 변수변환 과정(정규화 과정) 불필요
    - ✓ 비선형 관계도 의사결정나무 성능에 영향을 주지 않음
    - ✓ 결측 값(null)도 하나의 값으로 보고, 분할의 기준으로 사용 될 수 있음
    - ✓ 모델 자체 내에서 특징 선택 또는 변수 가려내기를 수행

의사결정나무 - 개요

categorical categorical quantitative class

Tid	Employed	Level of Education	# years at present address	Credit Worthy
1	Yes	Graduate	5	Yes
2	Yes	High School	2	No
3	No	Undergrad	1	No
4	Yes	High School	10	Yes
…	…	…	…	…

Tid	Employed	Level of Education	# years at present address	Credit Worthy
1	Yes	Undergrad	7	?
2	No	Graduate	3	?
3	Yes	High School	2	?
…	…	…	…	…



의사결정나무 - 모델링

❖ 여러 동물들에 대한 정보를 조사하여 다음과 같은 데이터가 있다고 가정해보자

Name	Body Temperature	Gives Birth	Class Label
Porcupine	Warm-blooded	Yes	*Yes*
Cat	Warm-blooded	Yes	*Yes*
Bat	Warm-blooded	Yes	*No**
Whale	Warm-blooded	Yes	*No**
Salamander	Cold-blooded	No	*No*
Komodo dragon	Cold-blooded	No	*No*
Python	Cold-blooded	No	*No*
Salmon	Cold-blooded	No	*No*
Eagle	Warm-blooded	No	*No*
Guppy	Cold-blooded	Yes	*No*

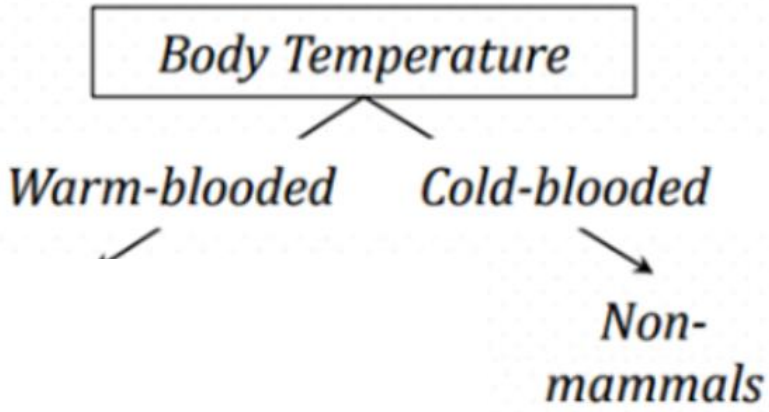
Body Temperature = Warm ➔ 온혈동물(혹은 정온동물)
Body Temperature = Cold ➔ 냉혈동물(혹은 변온동물)

Gives Birth = Yes ➔ 태생
Gives Birth = No ➔ 태생아님

참조:seoul university data mining lab thkoh, jspark

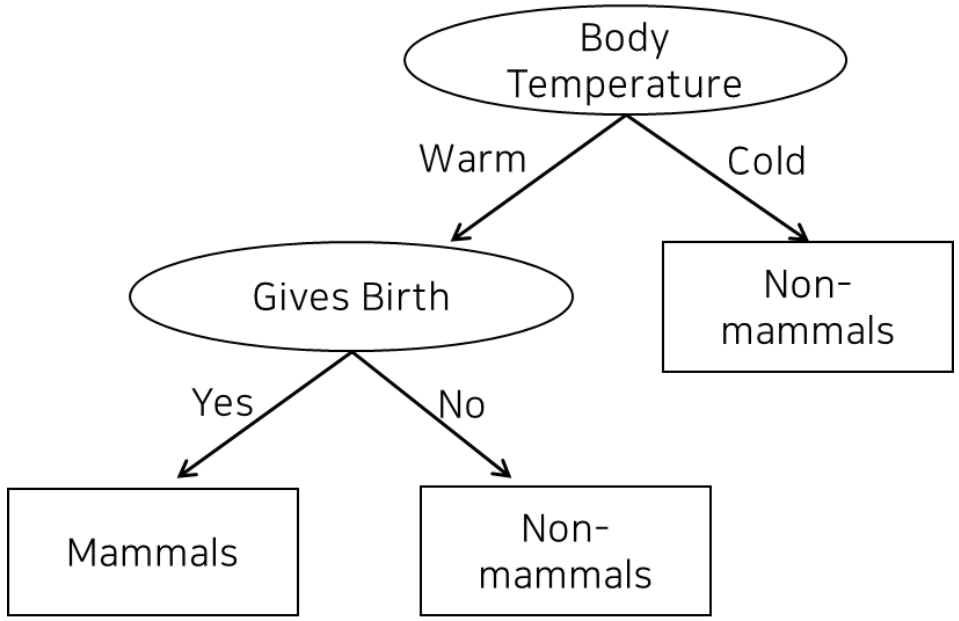
의사결정나무-모델링

❖ 앞의 데이터를 학습하여 도출된 의사결정나무 모델은 다음과 같다.



Name	Body Temperature	Gives Birth	Class Label
Porcupine	Warm-blooded	Yes	*Yes*
Cat	Warm-blooded	Yes	*Yes*
Bat	Warm-blooded	Yes	*No**
Whale	Warm-blooded	Yes	*No**
Salamander	Cold-blooded	No	*No*
Komodo dragon	Cold-blooded	No	*No*
Python	Cold-blooded	No	*No*
Salmon	Cold-blooded	No	*No*
Eagle	Warm-blooded	No	*No*
Guppy	Cold-blooded	Yes	*No*

의사결정나무-모델링



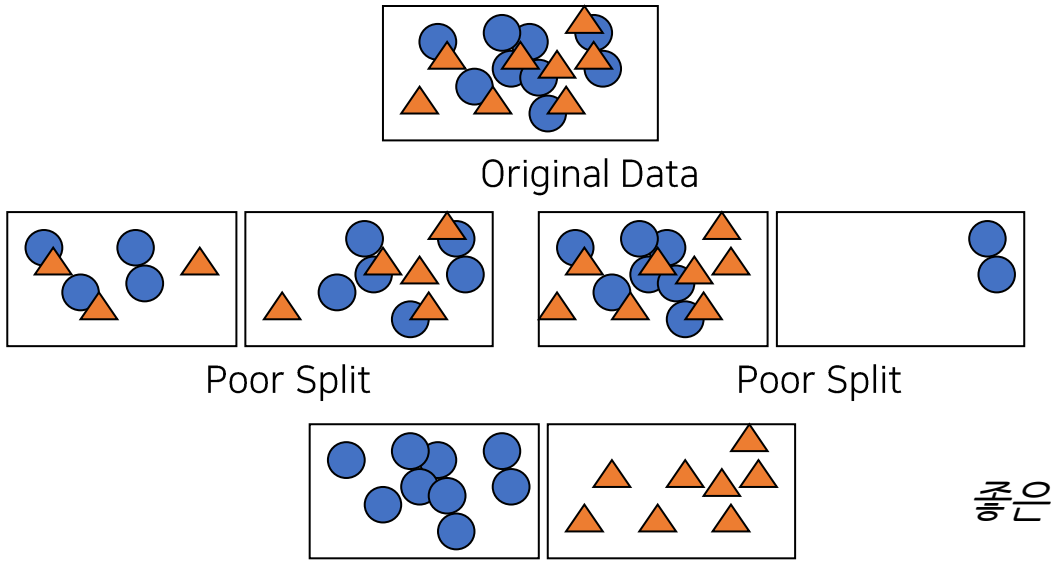
- "체온" 과 "새끼를 낳는 방법" 이 포유류를 분류하는 데에 있어 중요한 변수라는 것을 알 수 있다.

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

의사결정나무-모델링

❖ 나무의 성장

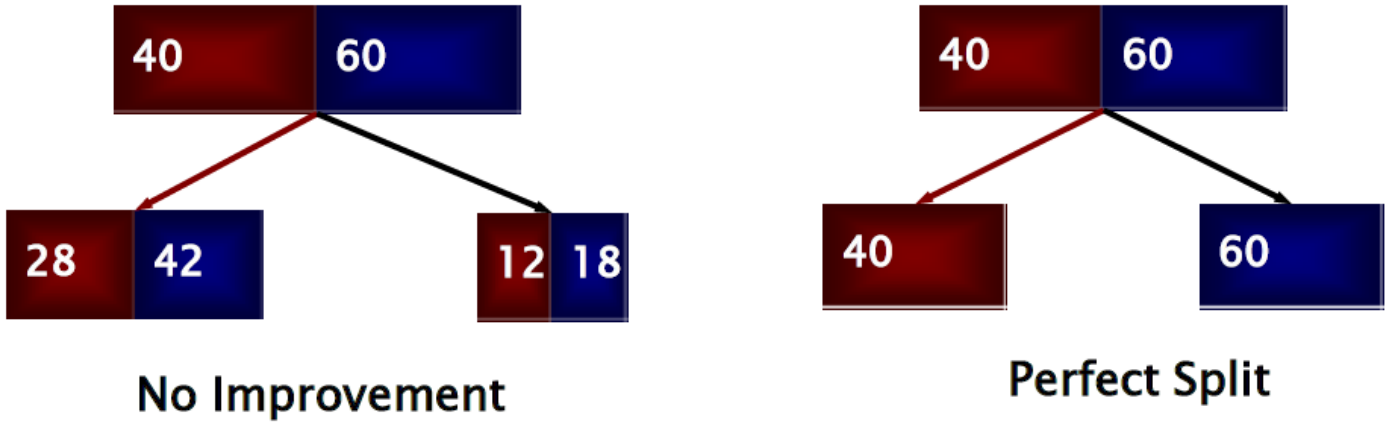
- 각 노드에서 최적의 분할 규칙을 찾아서 나무를 성장시킴
- 타겟변수 측면에서 부모 노드보다 동질성(homogeneity) 또는 순수도(purity)가 높은 자식 노드들이 되도록, 데이터를 반복적으로 더 작은 집단으로 분할
  - ✓ 수치형 변수의 경우 분할 포인트는 일반적으로 평균이 기준
  - ✓ 수치 값들을 범위로 구분하여 이산화



좋은 분할은 모든 자식노드의 순수도를 증가시킨다

모델링 – 분할

- 분할기준
  - 부모마디보다 자식마디의 순수도가 증가하도록 분류나무를 형성해 나감



의사결정나무 –모델링 : 분할

❖ **불순도를 측정하는 방법**

- 범주형 변수의 경우: **엔트로피, 지니계수, 정보이득, 카이제곱 통계량**
- 연속형 변수의 경우: 분산의 감소량, 분산분석의 F 통계량

1. 지니계수(Gini index)

- 코라도 지니(Gini): 이탈리아의 통계학자이자 경제학자
- 인구 다양성을 조사하는 생물학자들과 환경 공학자들이 자주 사용
- 같은 모집단에서 무작위로 선택된 두 항목들이 같은 클래스에 있을 확률
- 1에서 클래스의 비율의 제곱의 합을 뺌

$$G = 1 - \sum_k p_k^2$$

✓ 0 (불순도 최소, 순수)에서 0.5 (불순도 최대)의 값을 가짐 (이진분류인 경우)

- 1 - (0.1*0.1 + 0.9*0.9) = 1 - 0.82 = 0.18
  vs. 1 - (0.5*0.5 + 0.5*0.5) = 1 - 0.5 = 0.5

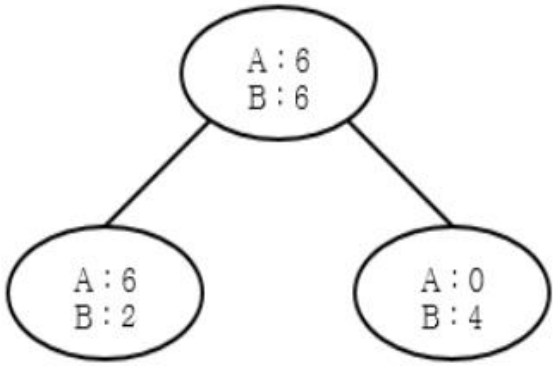
✓ 일반적인 경우 0 에서 1사이의 값을 가짐 : 숫자가 작을 수록 불순도가 적음 (즉, 순수함)

의사결정나무 –모델링 : 분할

아래의 그림에서 두 노드(집단)에 대한 지니 지수는 다음과 같이 계산되며, 지니 지수의 값이 클수록 이질적이며 순수도(purity)가 낮다고 할 수 있음

$GI = 1-(3/8)^2-(3/8)^2-(1/8)^2-(1/8)^2 = .69$

$GI = 1-(6/7)^2-(1/7)^2 = .24$



- 지니지수

분기전: $1-\left(\frac{1}{2}\right)^2-\left(\frac{1}{2}\right)^2=\frac{1}{2}$

분기후: $\left[1-\left(\frac{3}{4}\right)^2-\left(\frac{1}{4}\right)^2\right]\times\frac{2}{3}+\left[1-\left(\frac{0}{4}\right)^2-\left(\frac{0}{4}\right)^2\right]\times\frac{1}{3}=\frac{1}{8}$

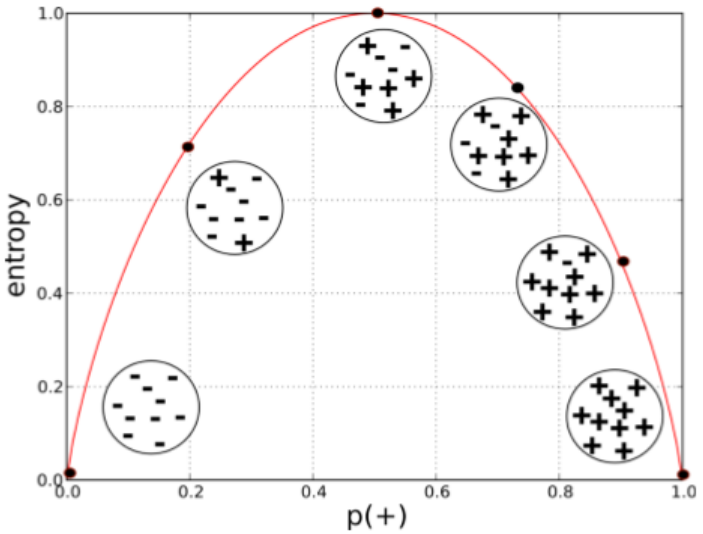
(분기 전에 비해 감소함)

의사결정나무 –모델링 : 분할

2. 엔트로피(Entropy)

- 시스템이 얼마나 정리되지 않았는지에 대한 척도
- 특정 의사결정나무 노드의 엔트로피
  - ✓ 노드에서 포함된 모든 클래스에 대하여, 특정 클래스의 레코드의 비율을 구하고 이 값과 이 값에 밑이 2인 로그를 취한 값을 곱한 값들의 합
  - ✓ 양수를 만들기 위해서 -1을 곱함

$$Entropy(H) = -p_1 \log_2 p_1 - p_2 \log_2 p_2 - \ldots = -\sum_k p_k \log_2 (p_k)$$



- ✓ E=0: 무질서 최소, 같은 항목으로만 구성 (순수)
- ✓ E=1: 무질서 최대, 각 항목이 동일하게 구성

예> 고객 10명 중, 7명이 모기지 상환을 정상적으로 하고 3명이 상환하지 않은 경우

P(정상상환) = 7/10 =0.7
P(미상환) =3/10= 0.3

$$Entropy = -0.7 \log_2 0.7 - 0.3 \log_2 0.3 = 0.88$$

의사결정나무 –모델링 : 분할

3. 정보이득(Information Gain, IG)

- Entropy(부모)– [p(자식1)×Entropy(자식1)+p(자식2)×Entropy(자식2) +…]
- 추가된 정보(속성)에 따라 엔트로피 “변화” 를 의미 함
- 정보 증가량 값이 클수록 분류에 좋은 속성임

4. 정보이득비율(Gain Ratio)

- 정보이득의 변형으로, 관측치가 많은 것을 선호하게 되는 편향성(bias)을 줄인 일반적으로 가장 좋은 옵션
- 분할하기 전에 가지들의 수를 고려함으로써 정보이득의 문제점을 해결
- 고유 정보량을 고려하여 정보이득을 수정함

의사결정나무 –모델링 : 분할

(예) 골프경기 문제: 날씨 조건에 따라 경기여부 예측

- 입력변수: 날씨 조건 Outlook, Temperature, Humidity, Wind
- 타겟 변수: 경기 여부 Play(yes/no)

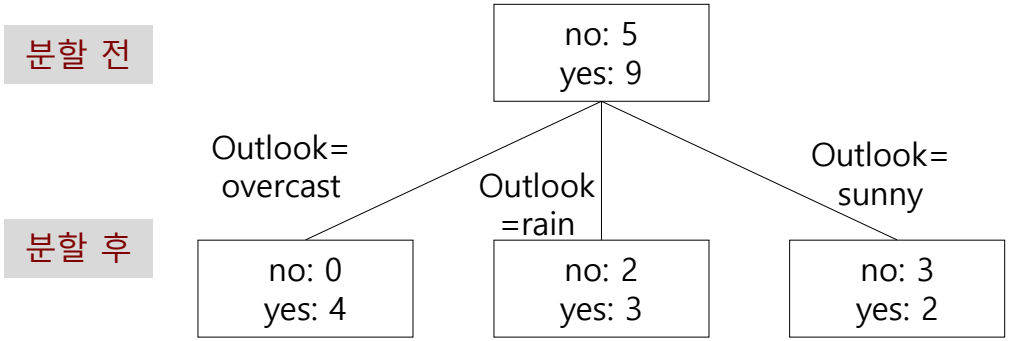
Golf (학습용 데이터셋)

Play	Outlook	Temperature	Humidity	Wind
no	sunny	85	85	false
no	sunny	80	90	true
yes	overcast	83	78	false
yes	rain	70	96	false
yes	rain	68	80	false
no	rain	65	70	true
yes	overcast	64	65	true
no	sunny	72	95	false
yes	sunny	69	70	false
yes	rain	75	80	false
yes	sunny	75	70	true
yes	overcast	72	90	true
yes	overcast	81	75	false
no	rain	71	80	true

의사결정나무 –모델링 : 분할

(예) 골프경기 문제: Outlook(날씨) 속성의 정보이득 계산

Play	Outlook
yes	overcast
yes	overcast
yes	overcast
yes	overcast
yes	rain
yes	rain
no	rain
yes	rain
no	rain
no	sunny
no	sunny
no	sunny
yes	sunny
yes	sunny



〈분할 전 엔트로피〉

$$H_{no\ split} = -(5/14) \times log_2(5/14) \quad -(9/14) \times log_2(9/14) = 0.940$$

〈분할 후 엔트로피〉

$$H_{Outlook:overcast} = -(0/4) \times log_2(0/4) \; - \; (4/4) \times log_2(4/4) = 0.0$$

$$H_{Outlook:rain} \; = -(2/5) \times log_2(2/5) \; - \; (3/5) \times log_2(3/5) = 0.971$$

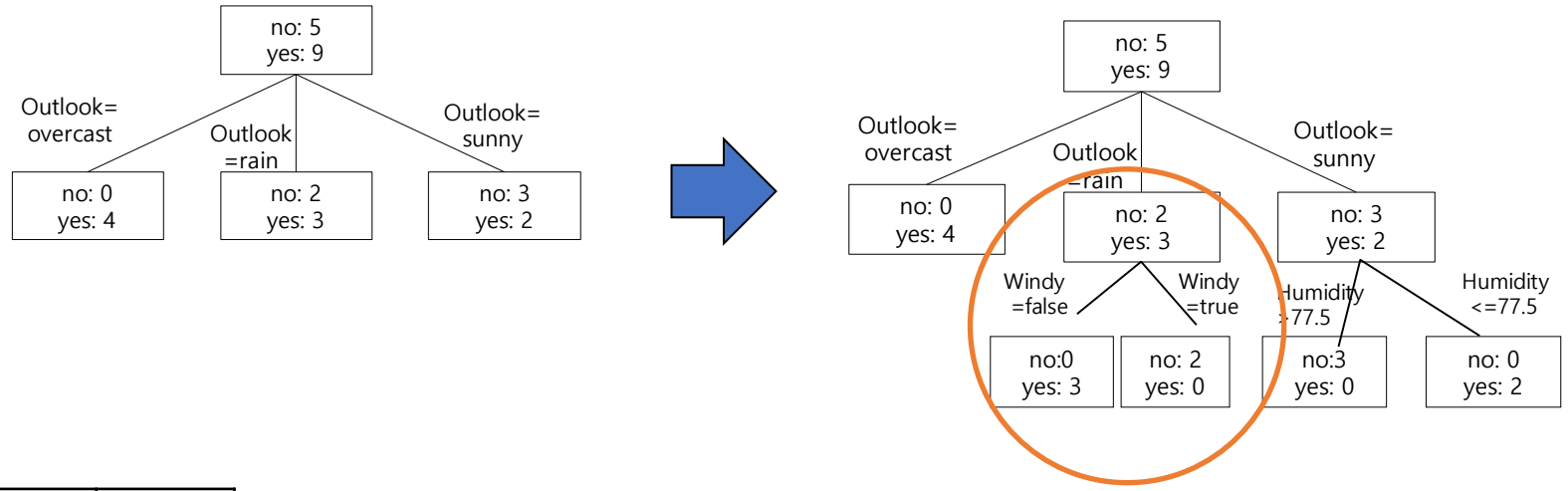
$$H_{Outlook:sunny} = -(3/5) \times log_2(3/5) \; - \; (2/5) \times log_2(2/5) = 0.971$$

속성	정보이득
Temperature	0.029
Humidity	0.102
Wind	0.048
Outlook	0.247

$$H_{Outlook} = P_{Outlook:overcast} \times H_{Outlook:overcast} + P_{Outlook:rain} \times H_{Outlook:rain}$$
$$+ P_{Outlook:sunny} \times H_{Outlook:sunny} = (4/14) \times (0) + (5/14) \times 0.971 + (5/14) \times 0.971 = 0.693$$

$$IG_{Outlook} = H_{no\ split} - H_{Outlook} = 0.940 - 0.693 = 0.247$$

의사결정나무 –모델링 : 분할



Outlook	windy	play
sunny	FALSE	no
sunny	FALSE	no
sunny	FALSE	yes
sunny	TRUE	no
sunny	TRUE	yes
rain	FALSE	yes
rain	FALSE	yes
rain	FALSE	yes
rain	TRUE	no
rain	TRUE	no
overcast	FALSE	yes
overcast	FALSE	yes
overcast	TRUE	yes
overcast	TRUE	yes

〈분할 전 Entropy〉

$H_{nosplit}$: $-(2/5) \times log_2(2/5) - (3/5) \times log_2(3/5) = 0.971$

〈분할 후 엔트로피〉

$H_{windy:true}$: $-(0/2) * \log_2(0/2) - (2/2) * \log_2(2/2) = 0$

$H_{windy:false}$: $-(3/3) * \log_2(3/3) - (0/3) * \log_2(0/3) = 0$

$H_{windy:}$: $P_{windy:true} * H_{windy:true} + P_{windy:false} * H_{windy:false}$

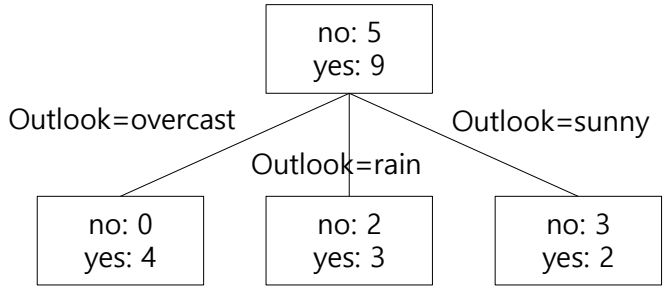
$= 2/5 * 0 + 3/5 * 0 = 0$

$IG_{windy} = H_{nosplit} - H_{windy:} = 0.971 - 0 = 0.971$ 최대 정보이득 증가량

의사결정나무 –모델링 : 분할

5. 카이제곱 통계량

- 통계학적 유의성에 대한 검정
- 1900년에 영국의 통계학자 칼 피어슨(Karl Pearson)이 개발
- 빈도에 대한 기대값과 관측값의 표준화된 차이의 제곱들의 합으로 정의
  관측된 표본들 간의 차이가 우연에 의한 것일 확률을 측정



$$\chi^2 = (\text{카이제곱통계량}) = \sum_i \sum_j \frac{(E_{ij} - O_{ij})^2}{E_{ij}}$$

기대도수 (E_{ij})			
	no	yes	total
overcast	1.429*	2.571	4
rain	1.786	3.214	5
sunny	1.786	3.214	5
total	5	9	14

* 14x(4/14)x(5/14) = 1.429

실제도수 (O_{ij})			
	no	yes	total
overcast	0	4	4
rain	2	3	5
sunny	3	2	5
total	5	9	14

$$\chi^2 = \frac{(1.429-0)^2}{1.429} + \frac{(2.571-4)^2}{2.571} + \frac{(1.786-2)^2}{1.786} + \frac{(3.214-3)^2}{3.214} + \frac{(1.786-3)^2}{1.786} + \frac{(3.214-2)^2}{3.214} = 3.547$$

**값이 클수록 순수도가 증가*

의사결정나무-모델링 : 주요 알고리즘

CART

1984년 L. Breiman, J. Friedman, R. Olshen, C. Stone에 의해서 발표
Classification and Regression Trees의 약자: 이진 나무를 생성시키고, 순수도를 증가시키는 것으로 확인되는 분할들을 계속 진행
순수도: 범주형인 경우 **지니지수**, 연속형인 경우 분산 이용

C5.0

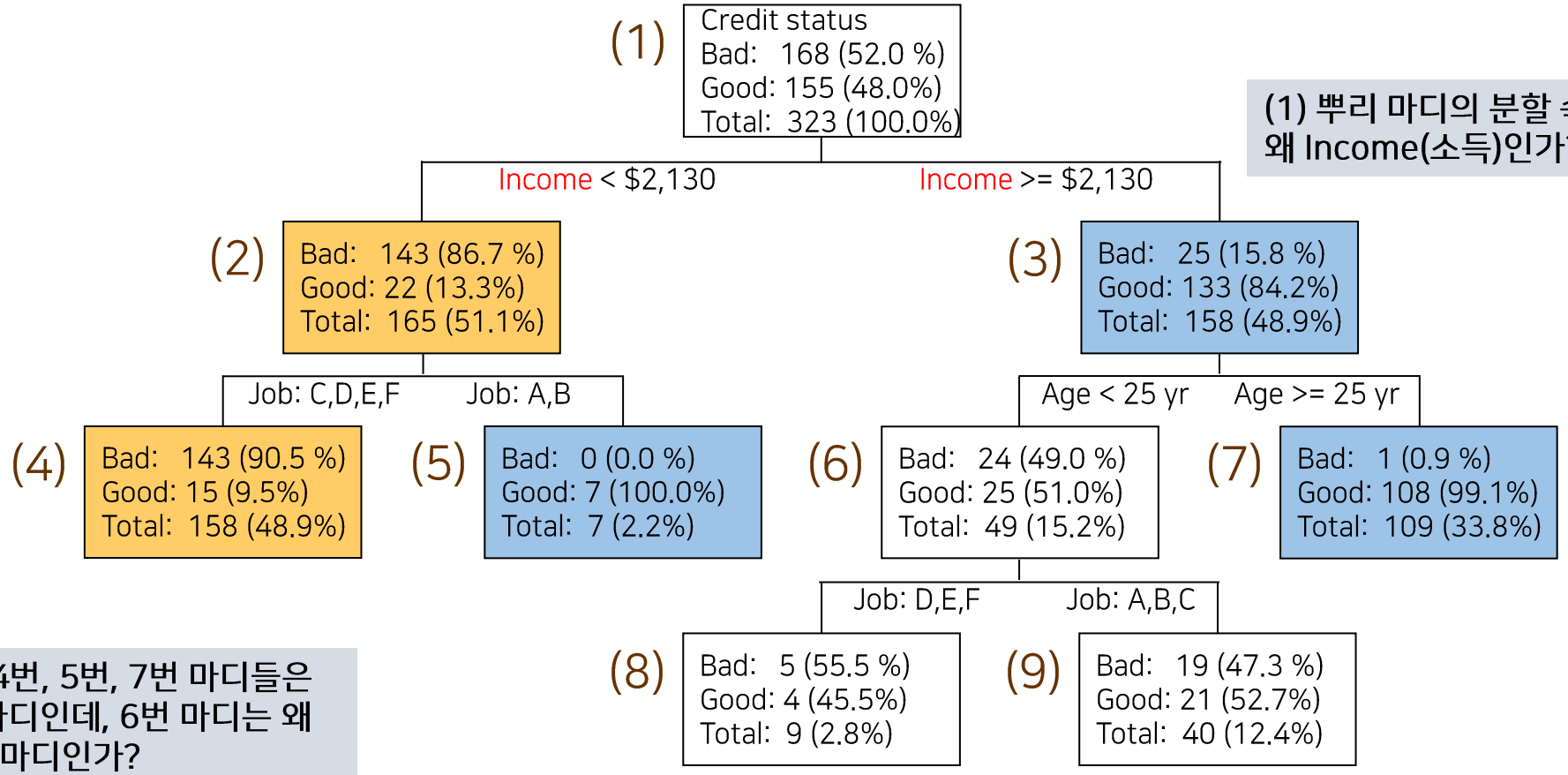
오스트레일리아 연구자 J. Ross Quinlan에 의해서 개발
초기 버전은 1986년에 개발된 ID3 (Iterative Dichotomiser 3)
CART와 다르게 범주형 변수에 대한 다중 분할 가능
순수도: **엔트로피** 지수 이용

CHAID (Chi-squared Automatic Interaction Detection)

1975년 John A. Hartigan에 의해서 발표, 변수들 간의 통계적 관계를 감지
고전적인 CHAID 알고리즘에서는 모든 입력변수가 범주형
연속형 변수들은 구간화(binning)하거나 대, 중, 소와 같은 순차적인 클래스로 대체
순수도: **카이제곱 통계량** 이용

의사결정나무 -모델링

의사결정나무 (Decision tree : DT) - 중요질문들



(1) 뿌리 마디의 분할 속성이 왜 Income(소득)인가?

(2) 4번, 5번, 7번 마디들은 끝 마디인데, 6번 마디는 왜 중간마디인가?

(3) 7번 마디에 속하는 자료는 신용상태를 어떻게 보아야 하는가?

의사결정나무 -모델링

의사결정나무 (Decision tree : DT) - 중요질문들

❖ 의사결정나무 구축을 위한 질문들

- 뿌리 마디의 분할 속성이 왜 Income(소득)인가?
- 4번, 5번, 7번 마디들은 끝 마디인데, 6번 마디는 왜 중간마디인가?
- 7번 마디에 속하는 자료는 신용상태를 어떻게 보아야 하는가?

❖ 의사결정나무의 생성요소

- 분할 규칙
- 정지 규칙: 분할을 언제 그만둘 것인지를 결정
- 가지치기 규칙: 나무의 크기가 클 때 축소시키는 방법

의사결정나무 -모델링

의사결정나무 생성 시 중요 결정사항

① 어디서 데이터를 분할(split) 할 것인가?

: 순수도가 가장 높을 때, 이질감이 가장 낮을 때

=> 선택한 알고리즘의 분할기준(지표)에 맞게

분할기준이 엔트로피일 때 이 값이 작아질수록 좋다.

⇔ 분할기준이 정보이득일 때 이 값이 커질수록 좋다.

분할기준이 지니 지수일 때 이 값이 작아질수록 좋다.

분할기준이 카이제곱 통계량일 때 이 값이 커질수록 좋다.

의사결정나무 -모델링

의사결정나무 생성 시 중요 결정사항

② 언제 분할을 멈출 것인가?

- 실제 데이터에서는 100% 동질성을 갖는 단말 노드 (또는 잎 노드)를 얻는 경우는 거의 없으므로, 언제 분할을 멈추어야 할지를 결정해야 함

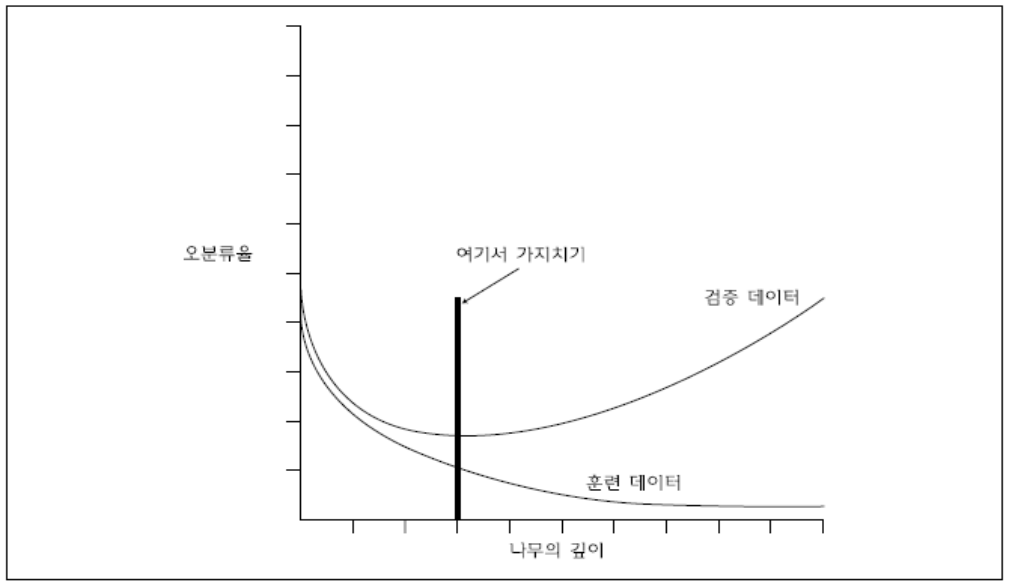
- 현재의 마디가 더 이상 분리가 일어나지 못하게 하는 규칙

  - ✓ 분할 기준의 최소 임계치 (분할 전 후의 최소 차이) 를 충족하는 속성이 하나도 없는 경우
  - ✓ 나무가 최대 깊이에 도달한 경우
    :의사결정나무가 커질수록 결과해석이 어려워질 뿐 아니라 과적합의 문제가 생김
  - ✓ 노드에 속한 관측치(사례수)가 특정 수 이하인 경우 과적합을 막기 위한 메커니즘

의사결정나무-모델링

③ 가지치기(Pruning)가 필요한 경우

- 지나치게 많은 노드를 가지는 (복잡한 모형) 의사결정 나무는 새로운 자료에 적용할 때 예측오차가 매우 클 가능성이 있음
  => 과적합
- 성장이 끝난 나무의 가지를 제거하여 적당한 크기를 갖는 나무 모형을 최종적인 예측모형으로 선택하는 것이 예측력의 향상에 도움이 됨
- 적당한 크기를 결정하는 방법은 검증용 데이터를 사용하여 예측에러를 구하고 이 예측에러가 가장 작은 모형을 선택



의사결정나무 -정리

의사결정나무는

- Supervised 유형의 기법으로
- IF- Then의 형식으로 설명이 가능하여, 사용자 관점에서 이해가 용이하며,
- 모델링 전 데이터 준비과정의 노력이 상대적으로 덜 필요함
- 분할규칙 : 불순도/순도를 측정함 (엔트로피, 지니계수, 정보이득, 카이제곱통계량 등)
- 분할기준은 불순도가 작아질 때, 또는 순도가 높아질 때 임
- 지나치게 많은 노드를 가지거나, 그로 인해 노드에 들어간 관측치가 너무 작을 때는 과적합이 발생할 수 있으므로, 의사결정나무의 성장을 제한하거나 줄여야 할 필요가 있음 (가지치기)

실 습

의사결정나무-모형 평가

분류 모델에 대한 성능평가 방법

- 혼동 행렬 (confusion Matrix)
- ROC 곡선
- 향상도 차트(lift chart)

1) 혼동 행렬(confusion Matrix)

이진 분류에서의 네 가지 예측 결과

입력된 데이터의 실제 클래스와 분류기의 예측 클래스의 조합 ☞ 4가지

		실제 클래스	
		Y	N
예측 클래스	Y	TP (true positive, 참긍정) 정분류	FP (false positive, 거짓긍정) 오분류
	N	FN (false negative, 거짓부정) 오분류	TN (true negative, 참부정) 정분류

의사결정나무-모형 평가

평가 척도(성능 척도, performance criteria)

용어	정의	계산식
민감도 sensitivity	선택되어야 할 것을 선택하는 능력 **(실제 True 중에 True를 예측한 능력)**	TP / (TP+FN)
특이도 specificity	거부되어야 할 것을 거부하는 능력 **(실제 False 중에 False 를 예측한 능력)**	TN / (TN+FP)
정밀도 precision	찾아낸 결과 중 실제로 관련이 있는 객체의 비율 (True로 예측한 것 중 실제 True비율)	TP / (TP+FP)
재현율 recall	모든 관련된 객체 중 실제로 찾아내어진 객체의 비율 (실제 true중에 true예측한 능력)	TP / (TP+FN)
정확도 accuracy	분류기 성능의 종합적 척도	(TP+TN) / (TP+FP+FN+TN)
오분류율		1 - 정확도

F1 score는 precision 과 recall의 조화평균으로, 0에서 1사이의 값을 가지며, 클수록 좋음

※ 정확도 = 정분류율

의사결정나무-모형 평가

평가 척도의 계산 예

		실제 클래스		
		1	0	계
예측 클래스	1	139	9	148
	0	81	1771	1852
	계	220	1780	2,000

- 정확도(accuracy) = (1,771 + 139) ÷ 2,000 = 0.955
- 오분류율(error) = (9 + 81) ÷ 2,000 = 0.045
- 민감도(sensitivity) = 139 ÷ 220 = 0.632
- 특이도(specificity) = 139 ÷148 = 0.939
- 정밀도(precision) = 1,771 ÷ 1,852 = 0.9563
- 재현율(recall) = 139÷ 220= 0,632

• 아무리 정확도가 좋아도, 관심이 있는 클래스의 적중률이 높은 경우는 모델로서 성능이 좋다고 할 수 없음

• 오류율뿐 아니라 특이도와 민감도 등을 잘 봐야 함

의사결정나무-모형 평가

2) ROC(Receiver operating characteristic) 곡선

사용이유

- 성능 척도는 너무 단순하여 특성을 제대로 나타내지 못하는 단점이 존재함
- 성능 척도에 대한 타협점을 찾아 요구조건에 적합한 모델 선택
- 어떤 모델에 대해 종합적인 정확도는 높지만, 특정 클래스에 대한 재현율이나 정확도는 낮을 수 있다.

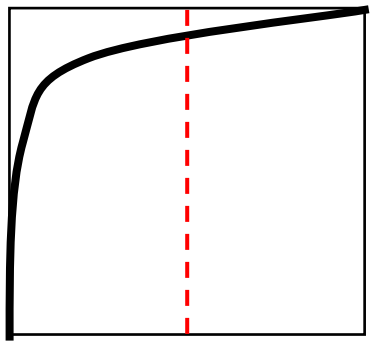
  => 사기탐지 모델에서는 TP 탐지 능력(클래스 재현율)을 높일 필요

  재현율을 큰 폭으로 높일 수 있다면 종합적인 정확도는 약간의 희생 감수

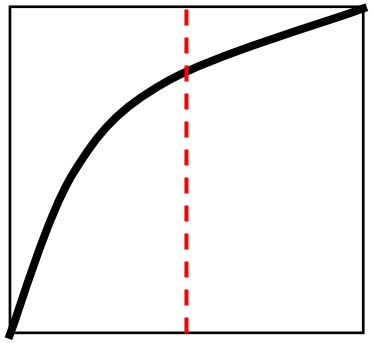
ROC 작성

- ROC(Receiver operating characteristic) 곡선의 y축은 민감도(Sensitivity) , x축의 변수는 분류기준 값에 따라서 계산된 특이도(Specificity)를 이용하여 계산된 (1 - 특이도)의 값으로 나타낸다.

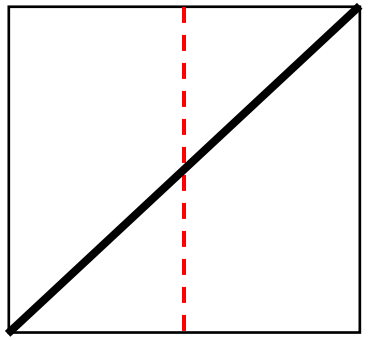
의사결정나무-모형 평가



(a)



(b)



(c)

즉 같은 오류내에서 가장 잘맞추는 것이 좋은 모델임

- 같은 (1 - 특이도) 값에서 (a)의 민감도 값이 최대 → (a)가 가장 좋은 모형
- (c)는 모형 구축의 효과가 전혀 없다. => 혼동 행렬의 대각 원소와 비대각 원소의 빈도가 동일

- 곡선 하 면적(Area Under Curve, AUC) : **ROC 곡선 아래의 면적으로 성능이 좋은 모델일 수록 면적이 1.0에 근접**

실 습