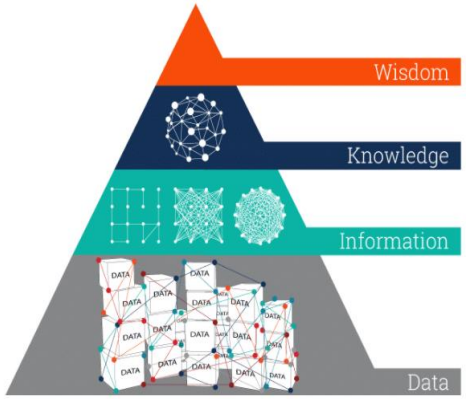
1 -7주차 강의 요약

2022.04.19

Institutor: JS LEE

DIKW 피라미드



[그림 1] DIKW 피라미드

단계	설명	예시
지혜 (Wisdom)	• 근본 원리에 대한 깊은 이해를 바탕으로 도출되는 창의적인 아이디어	• 다른 문구류도 B 마트가 쌀 것이라 판단
지식 (Knowledge)	• 정보를 바탕으로 의사결정에 활용하는 것 • 상호 연결된 정보패턴을 이용해 예측한 결과물 • 개인의 경험 결합해 고유의 지식으로 내재화	• B 문구점에서 연필을 사야 겠다
정보 (Information)	• 데이터의 가공 및 상관관계 이해를 토대로 의미를 부여한 데이터 • 누가', '무엇', '언제', '어디서' 등에 대해 관련 질문을 함으로써 데이터에서 귀중한 정보를 도출하고 더 유용하게 만들 수 있음	• B 문구점이 연필가격이 더 싸다
데이터 (Data)	• 개별 데이터 자체로는 특별한 의미부여가 안된 객관적 사실 • 타 데이터와 상관관계가 없는 가공하기 전의 순수한 수치나 기호 그 자체 • 관찰, 측정을 통해서 수집된 사실이나 값, 수치, 문자 등 가공되지 않은 원본 데이터	• A 문구점 : 연필 값 200원, B 문구점 : 연필 값 100원

참조 : https://www.ontotext.com/knowledgehub/fundamentals/dikw-pyramid/,데이터분석전문가 가이드

DIKW 피라미드

Data

자료

고객구매자료

Information

정보

고객별 구매패턴

Knowledge

지식

고급맥주와 기저귀를 함께 두면 이익이 증가

Wisdom

지혜

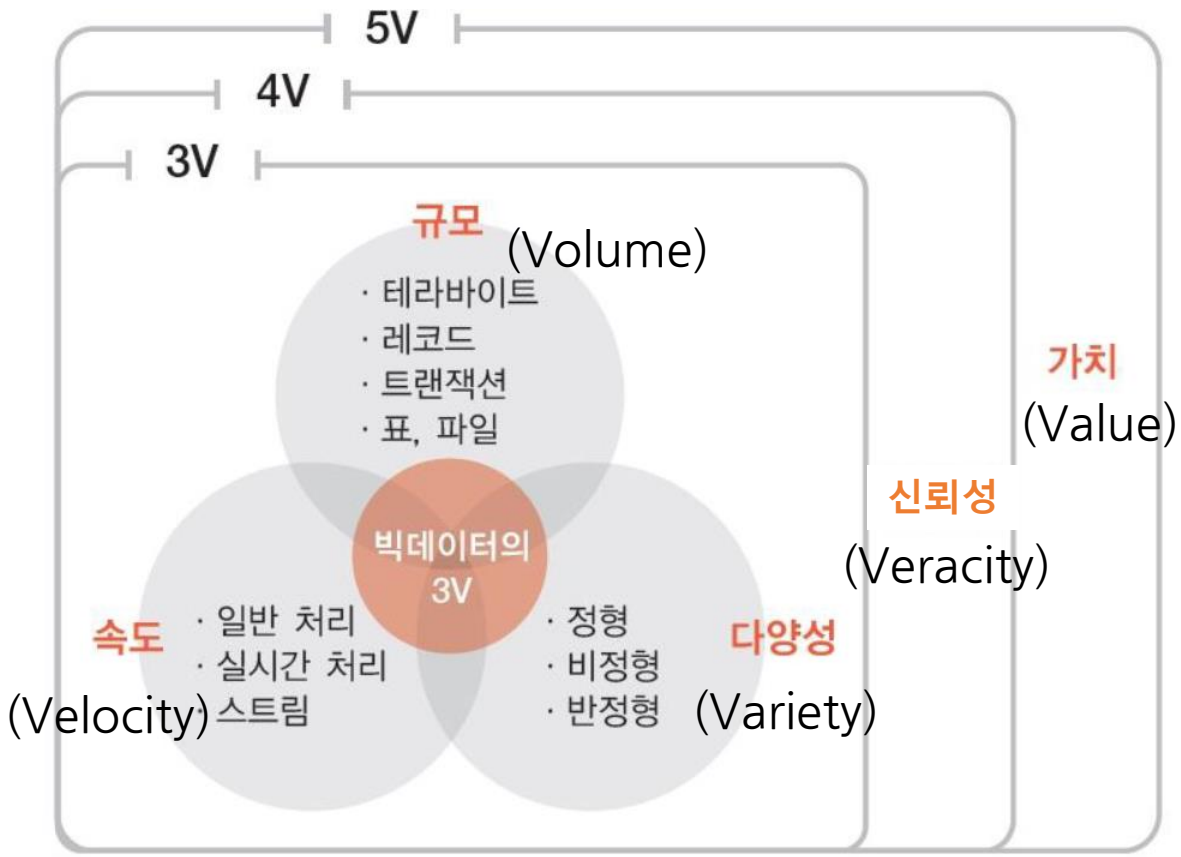
구매패턴을 살펴보아 향후 제휴관계를 맺을 회사를 알아보자

품목	함께 구매하는 %
기저귀 - 맥주	85%
빵 - 버터	60%
초콜릿 - 피넛 버터	40%

빅데이터의 개념

- **Big Data ≠ 대용량 자료**

- **Big Data 특징 5V**



초창기 빅데이터는 데이터 규모에 초점(정량적 측면의 강조)을 두고 기술적 측면 강조

3V, 4V에 따른 정의에서 일정한 패턴을 찾아 비지니스적 가치를 창출하는 정의로 발전하고 있음

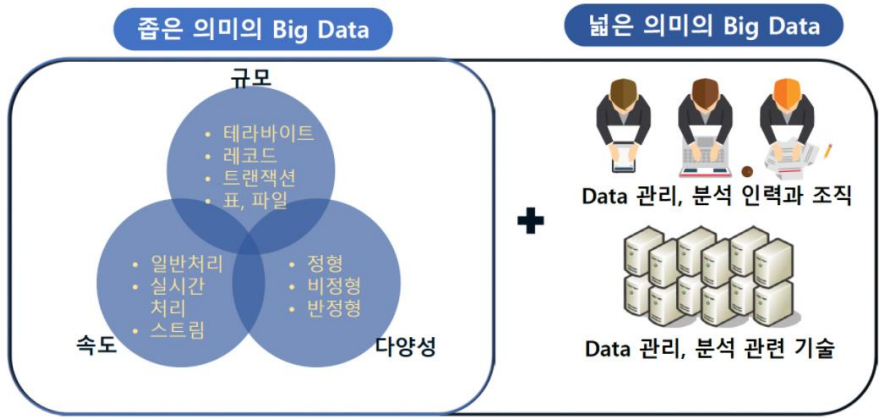
7V : Validity(정확성), Volatility (휘발성)

빅데이터의 개념

초점	주체	정의내용
데이터규모 자체특성 변화에 초점	McKinsey (2011)	▪ 일반적인 데이터베이스 SW가 수집,저장,관리, 분석할 수 있는 처 리능력을 초과하는 대규모의 데이터 집합
	가트너그룹 (2012)	▪ 더 나은 의사결정, 시사점 발견 및 프로세스 최적화를 위해 사용되는 새로운 형태의 정보처리가 필요한 대용량, 초고속 및 다양성의 특성을 가진 정보자산
	교육과학기술부 (2012)	▪ 데이터 형식이 다양하고 생성 속도가 매우 빨라, 새로운 관리 및 분석 방법이 필요한 대용량 데이터
	정보통신협회 (2015)	▪ 새로운 가치 추출을 위해 기존의 기술 또는 기법으로 처리하기 어려운 특징(규모, 신속성, 가변성, 다양성, 진정성)을 갖는 데이터 모음
분석비용 기술적 변화에 초점	IDC (2011)	▪ 다양한 형태의 방대한 데이터로부터 고속 캡처, 데이터 탐색 및 분석을 통해 경제적으로 필요한 가치를 추출하기 위해 설계된 차세대 기술 및 아키텍처
인재·조직 포괄적 변화에 초점	노무라 연구소	▪ 기본적인 데이터, 데이터 처리/저장/분석기술 이외에 의미 있는 정보도출에 필요한 인재/조직까지 포함해야 함
	메이어-쇤베르그 & 쿠키어(2013)	▪ 대용량 데이터를 활용해 작은 용량에서 얻을 수 없었던 새로운 통찰/가치 추출해 내는 일 ▪ 이를 활용해 시장, 기업, 시민, 정부 관계 등 많은 분야에 변화를 가져옴

빅데이터의 개념

빅데이터란?



https://wikidocs.net/93016

좁은 의미에서 빅데이터는 기존 데이터베이스의 데이터 수집·저장·관리·분석의 역량을 넘어서는 구조적 및 비구조적 데이터를 포함하는 대용량의 데이터 집합

넓은 의미에서 보면 좁은 의미의 빅데이터를 포함하고, 추가로 이러한 빅데이터로부터 의사 결정에 필요한 정보와 지식을 추출하고 결과를 분석하는 데 필요한 인력과 조직 및 관리·분석기술을 통칭함

(빅데이터 경영을 바꾸다. 삼성경제연구소 2012)

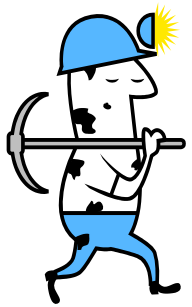
데이터마이닝이란?

데이터 마이닝은

- ✓ 대규모 데이터에 대한 귀납적 추론
- ✓ **데이터 속의 유용한(의미 있는) 패턴(규칙, 관계)을 찾고 이를 일반화하는 프로세스**
- ✓ 관측 데이터에 적합한 모델을 구축하는 과정

의미 있는 패턴 추출

- 유효하고(valid), 새롭고(novel), 잠재적으로 유용하고, 이해할 수 있는 패턴이나 관계를 파악해 가는 프로세스
- 궁극적으로 분석 결과를 이용하여 행동을 취할 수 있어야 함
- 지식 발견, 기계 학습, 예측 분석이라 불리기도 함
  knowledge discovery machine learning predictive analytics



- 일반적으로 대용량 데이터셋에 적용
- 탐색 → 전처리 → 모델링 → 평가 → 지식 추출의 과정을 거침
  exploration preprocessing modeling evaluation knowledge extraction

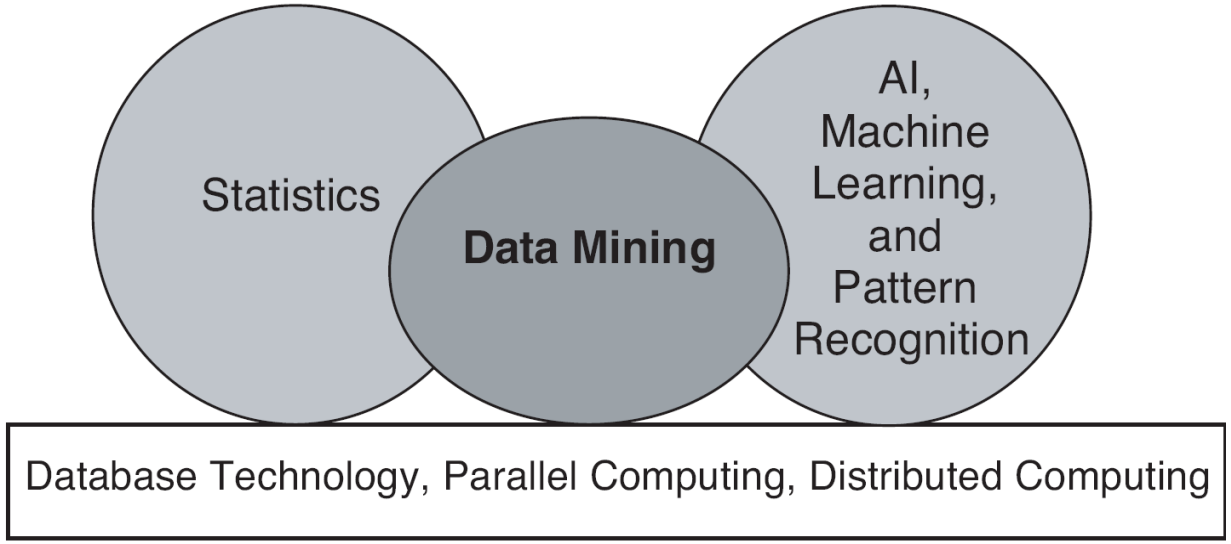
데이터마이닝 기원

데이터 마이닝의 기원

- 통계, 인공지능, 기계 학습, 데이터베이스, 패턴인식 분야에서 시작되어 발전
- 통계 지식, 주제 전문지식, 데이터베이스 기술, 그리고 기계 학습 기술 및 대용량 데이터 처리 기술(예> 병렬분산컴퓨팅) 필요

Traditional techniques may be unsuitable due to data that is

- Large-scale
- High dimensional
- Heterogeneous
- Complex
- Distributed



•주제 전문지식 :
데이터와 데이터가 생성되는 비즈니스 프로세스에 대한 지식

Introduction to Data Mining (Tan, Steinbach, Kumar, 2006)

데이터마이닝이 아닌 기법들

- 기술 통계

✓ 평균, 표준편차와 같이 데이터셋을 요약된 구조로 정량화

✓ 데이터를 이해하는데 필수적이므로 데이터 전처리, 후처리 단계에서 필수적

- 탐색적 시각화

✓ 시각적으로 데이터 표현, 대용량의 데이터 이해

✓ 데이터 전처리, 후처리 단계에서 필수적

- 차원 슬라이싱

차원(상품, 지역, 날짜 등)별로 정량적 데이터(수익, 수량 등)를 보여주나 정보 검색으로 간주됨

- 가설 검정

✓ 통계 검정, 실험 데이터가 가설을 지원할 만한 증거가 충분한지를 평가

- 쿼리 (질의)

✓ 데이터베이스에 정보를 요청 (예. 매출액이 높은 상위 5개 제품은 무엇인가)

데이터 마이닝 유형

	지도학습 (supervised learning)	자율학습 (unsupervised learning)
의미	• 학습용 데이터를 기준으로 모델을 만들고 이를 새로운 데이터에 적용하여 예측분석에 이용 • 입력변수들을 기준으로 타겟(출력, 결과)변수 예측	• 데이터 포인트들 간의 관계를 기반으로 데이터에서 패턴을 찾아내는 작업
특징	• 타겟(출력, 결과)변수 존재함	• 타겟(출력, 결과)변수 존재하지 않음
분석 기법	• 신경망, 회귀분석, 의사결정나무, 판별분석, 로지스틱회귀분석 …	• 군집분석, 연관규칙, …

데이터 마이닝 유형

❖ 일반적으로 많이 사용되는 알고리즘 (계속)

분야	설명	알고리즘	사례
분류	•데이터 포인트가 미리 정의된 클래스 중 어디에 속하는지에 대해 예측 •예측은 학습용 데이터셋을 기반으로 함	•**의사결정나무,Random Forest, Xgboost,** 신경망 베이지안 모델, 규칙 유도, k–최근접 이웃	•유권자들을 정당에 따라서 알려진 버킷으로 할당 •새 고객을 정의된(이미 알려진) 고객 그룹 중 하나의 그룹에 할당
회귀 분석	•수치형 타겟변수를 예측 •예측은 학습용 데이터셋을 기반으로 함	•선형회귀	•내년도 실업률 예측 •보험료 추정
이상 탐지	•특정 데이터 포인트가 데이터셋의 다른 데이터 포인트와 비교하여 특이값인지 예측	•거리 기반, 밀도 기반, 지역 특이값 요소(LOF)	•신용카드의 사기거래 탐지, 네트워크 침입 탐지

데이터 마이닝 유형

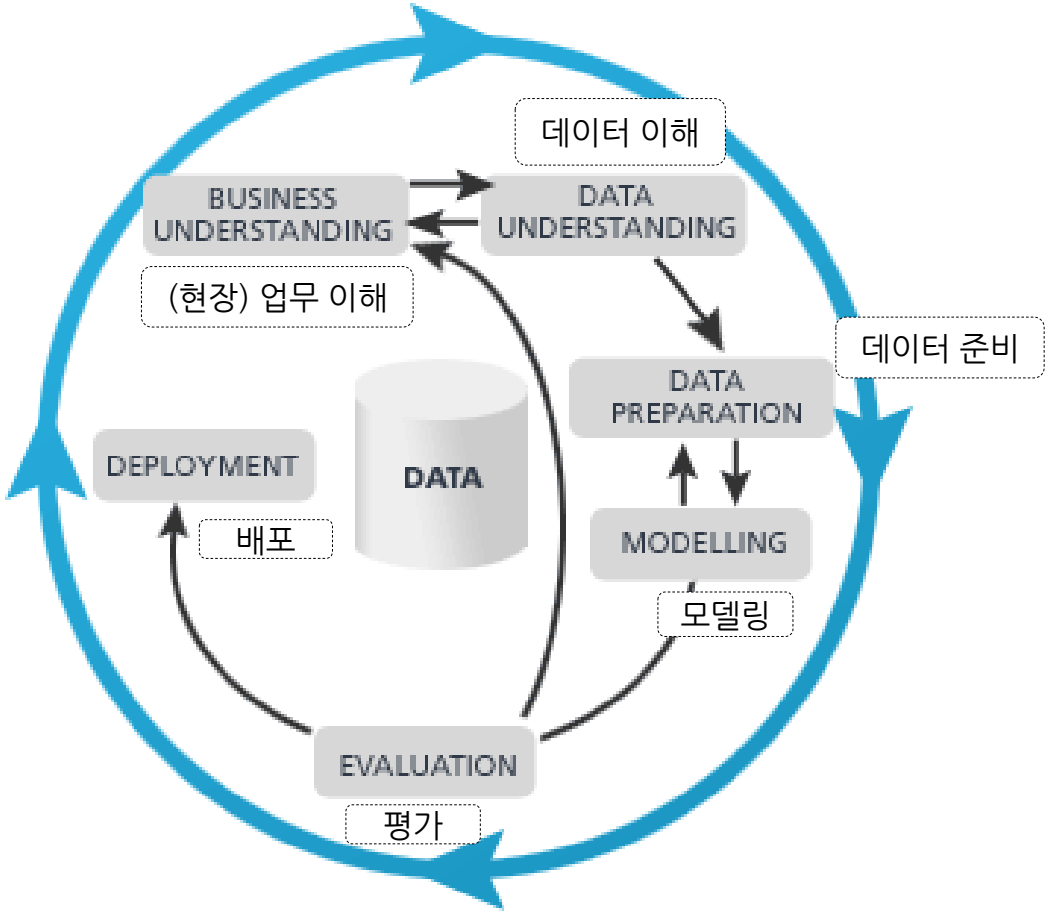
❖ 일반적으로 많이 사용되는 알고리즘

분야	설명	알고리즘	사례
시계열 분석	•과거 값에 기반하여 미래의 타겟변수 값을 예측	•지수평활, 자기회귀 누적이동평균(ARIMA), 회귀분석	•매출액 예측, 생산 예측, 추정할 필요가 있는 성장 현상
군집화	•데이터셋 내의 속성들을 기준으로 하여 데이터셋의 데이터 포인트들을 군집으로 구별	•**k-평균**, 밀도 기반 군집화 (예: DBSCAN)	•거래, 웹 및 고객 통화 데이터를 기반으로 한 고객세분화
연관성 분석	•거래 데이터를 기반으로 항목집합 내의 관계를 식별	•**빈발패턴-성장 알고리즘(FP-Growth),** 선험적(Apriori) 알고리즘	•소매업에서 구매 이력 데이터를 기반으로 한 교차판매 기회 발견

분석 프로세스

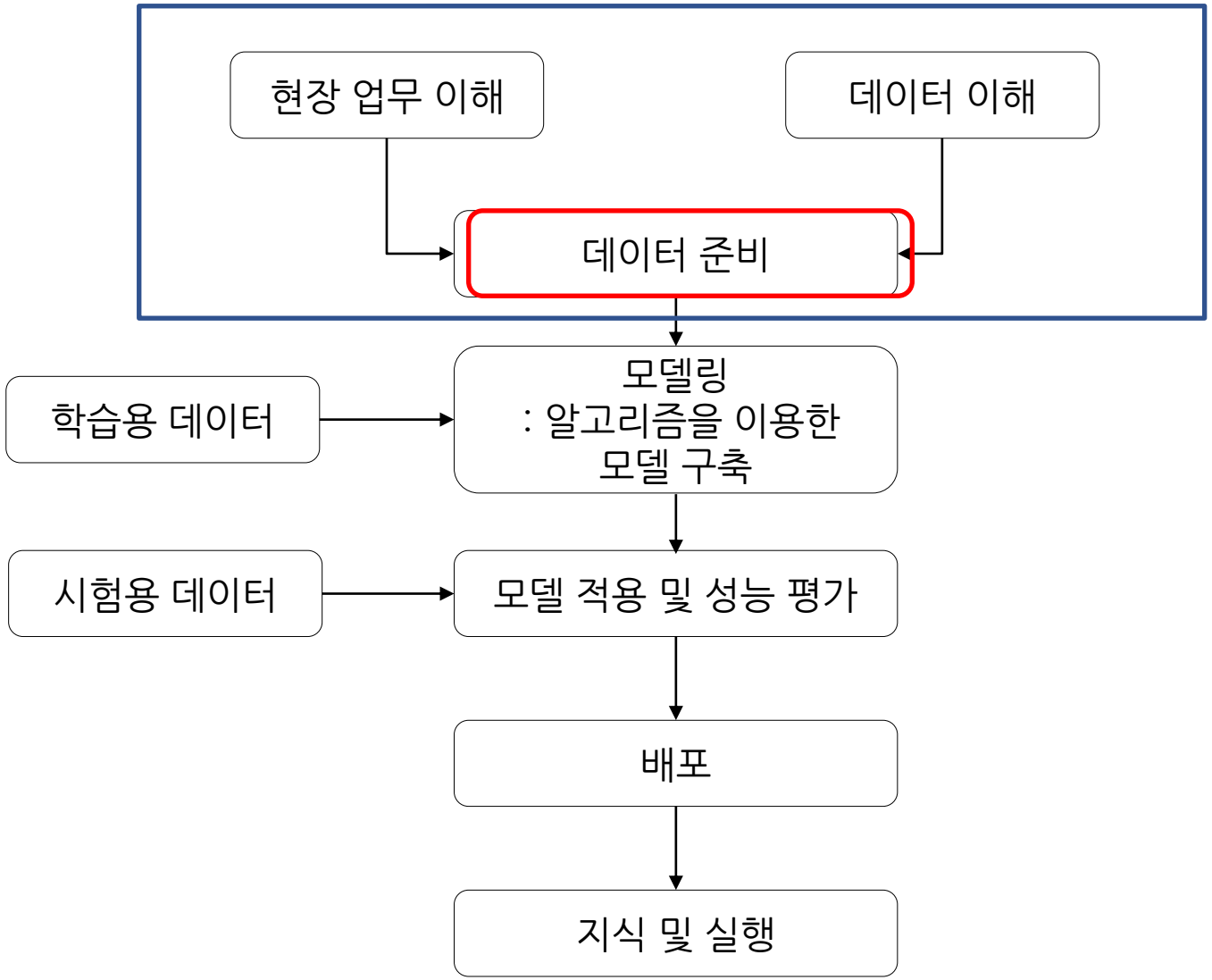
대표적인 데이터 마이닝 프레임워크

CRISP-DM(Cross Industry Standard Process for Data Mining)



다른 프레임워크
- SEMMA(Sample, Explore, Modify, Mo
- DMAIC(Define, Measure, Analyze, Improve, Control)
- KDD(Knowledge Discovery in Databases, Selection, Preprocessing, Transformation, Data Mining, Interpretation, and Evaluation framework)

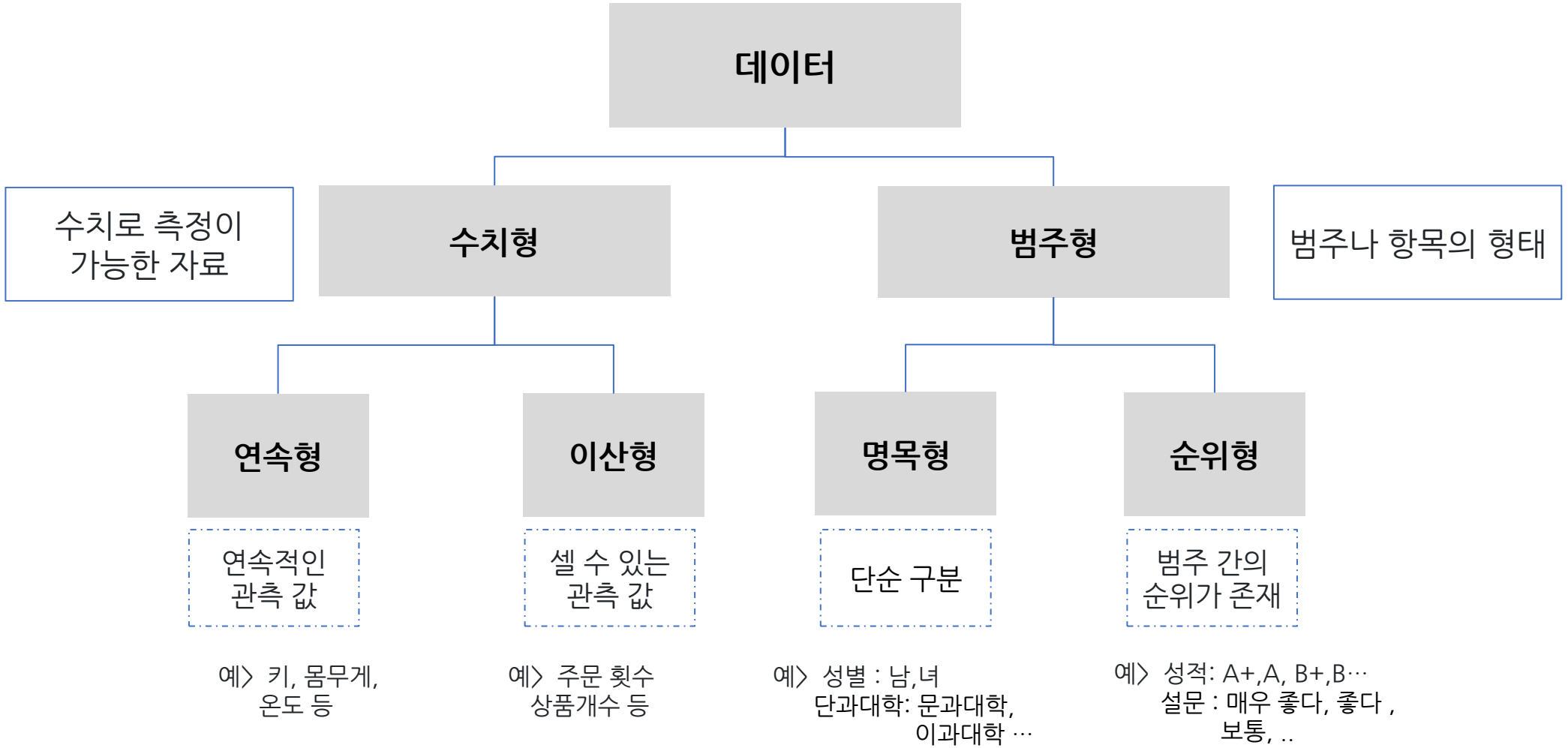
분석 프로세스



* 전체 프로세스에서 시간이 가장 많이 소요되는 부분은?

데이터 탐색

데이터 값의 형태



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

데이터 탐색

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

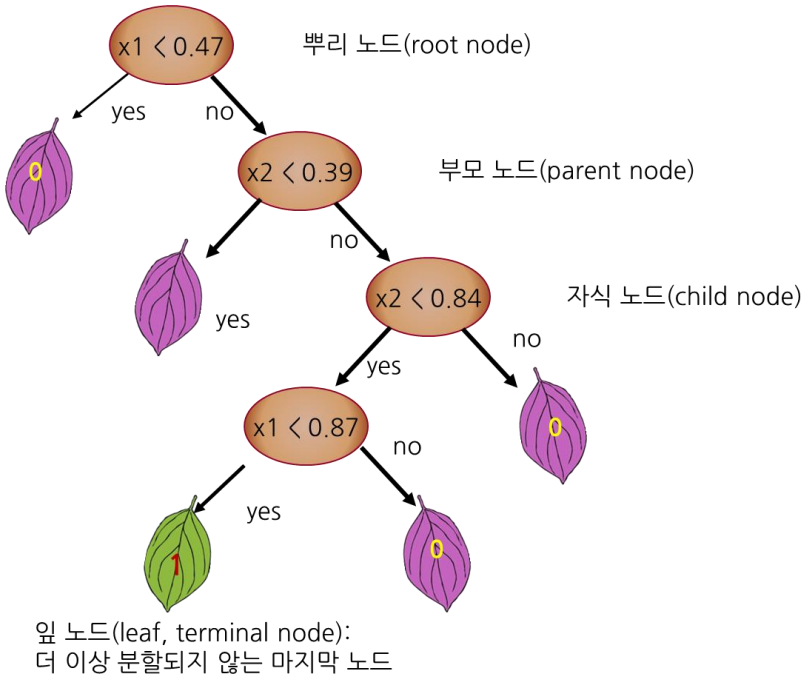
계급	도수	상대도수
남자	70	0.7
여자	30	0.3

- 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

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

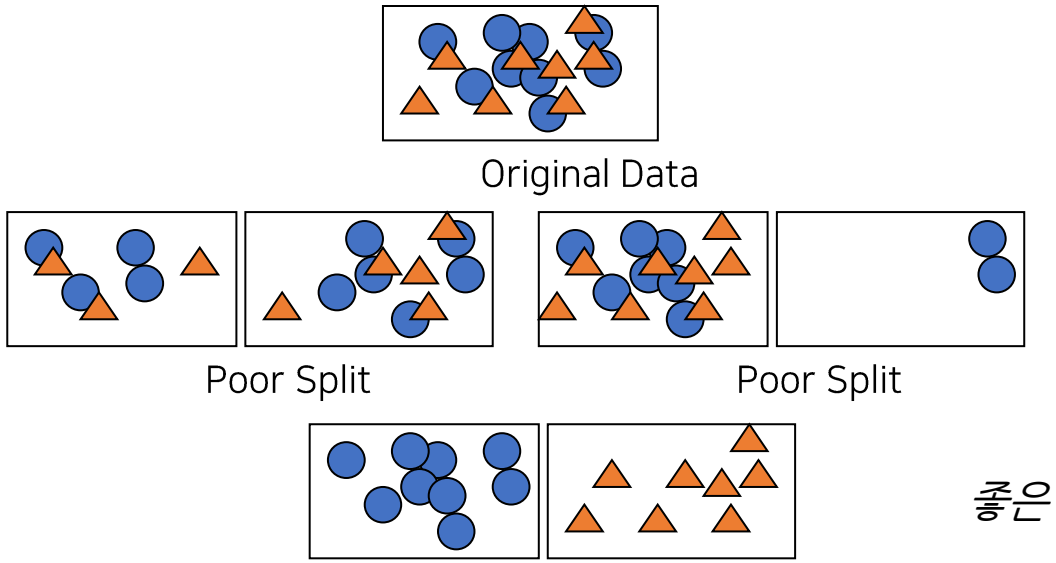
의사결정나무 - 개요

- 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-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

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

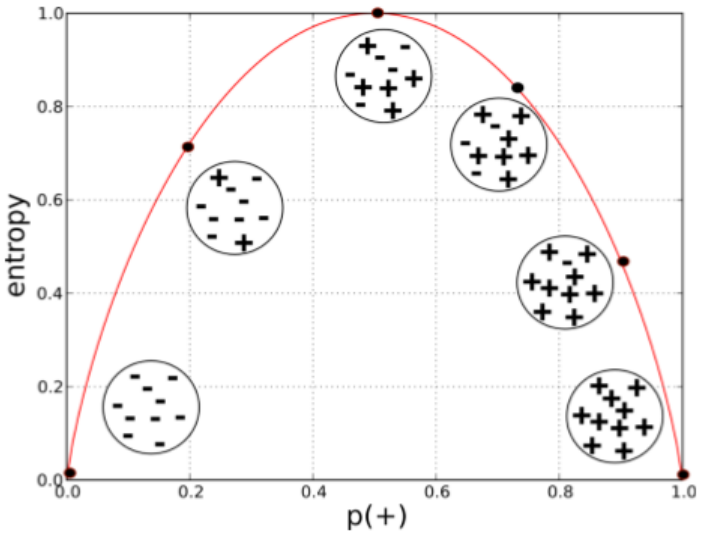
✓ 일반적인 경우 0 에서 1사이의 값을 가짐 : 숫자가 작을 수록 불순도가 적음 (즉, 순수함)

의사결정나무 –모델링 : 분할

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

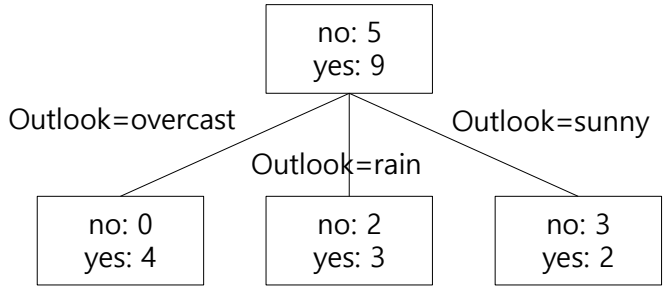
4. 정보이득비율(Gain Ratio)

- 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

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

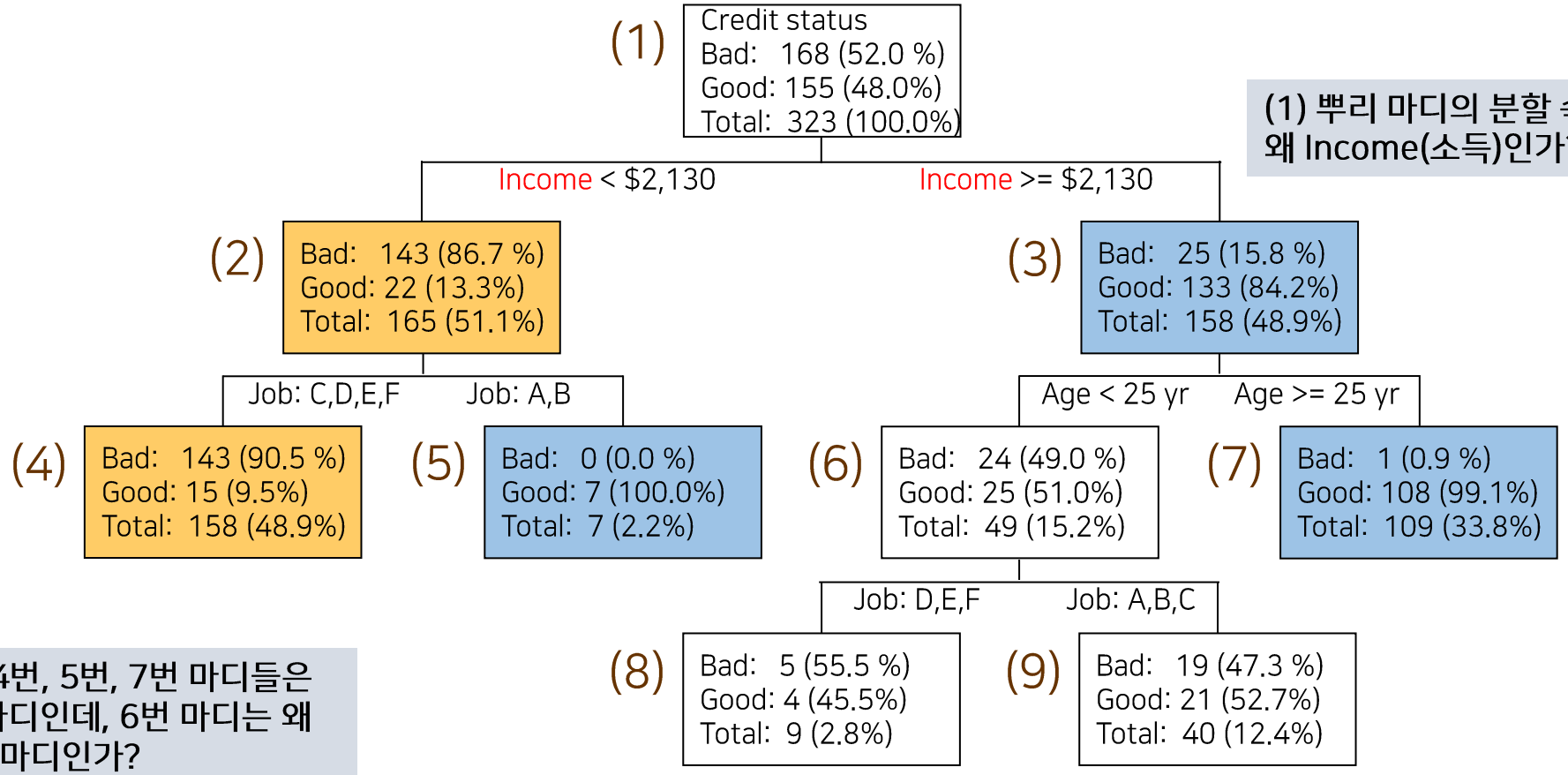
실제도수 (O_{ij})			
	no	yes	total
overcast	0	4	4
rain	2	3	5
sunny	3	2	5
total	5	9	14

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

**값이 클수록 순수도가 증가*

의사결정나무 -모델링

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

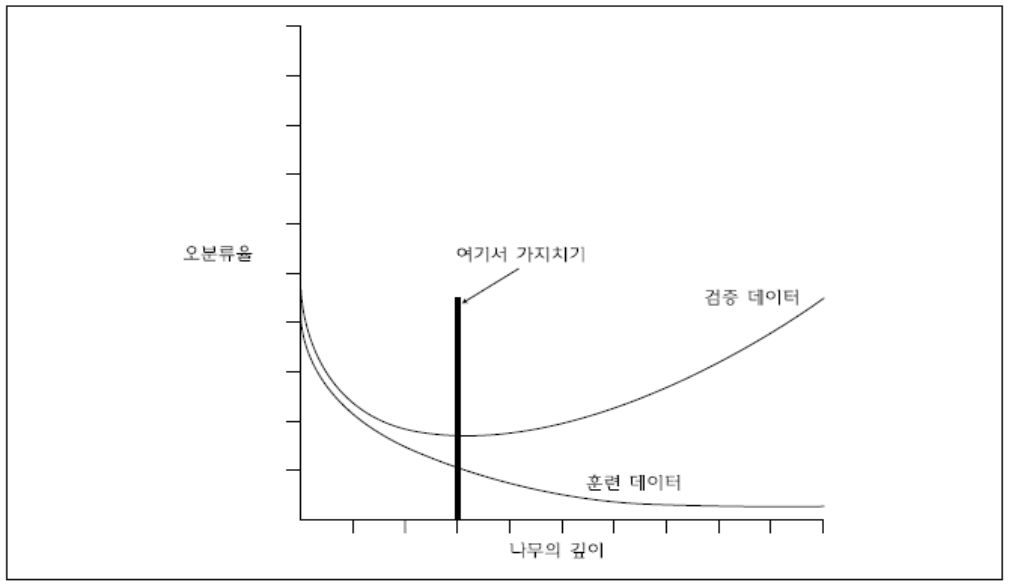
- 현재의 마디가 더 이상 분리가 일어나지 못하게 하는 규칙

  - ✓ 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-모형 평가

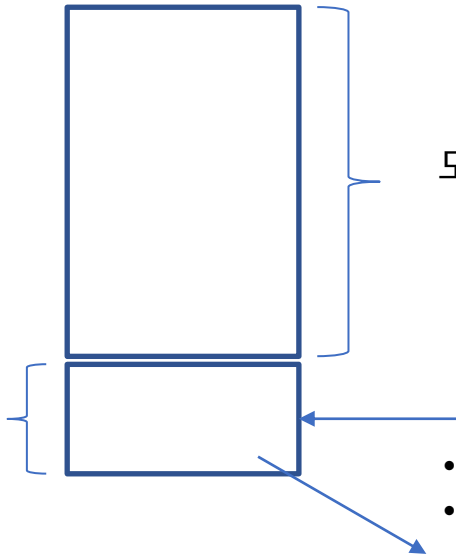
분류 모델에 대한 성능평가 방법

- 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



- 종속변수 존재함
- 그러나, 구축한 모형으로 종속변수를 예측하게 함
- 그 이후 실제 종속변수와 모델링을 통해 도출된 종속변수를 비교하여 모형을 평가함

		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) (민감도와 유사)	TP / (TP+FN)
정확도 accuracy	분류기 성능의 종합적 척도	(TP+TN) / (TP+FP+FN+TN)
오분류율		1 - 정확도
F1 Score	precision 과 recall의 조화평균으로, 0에서 1사이의 값을 가지며, 클수록 좋음	$2 * \frac{Precision * Recall}{Precision + Recall}$

※ 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

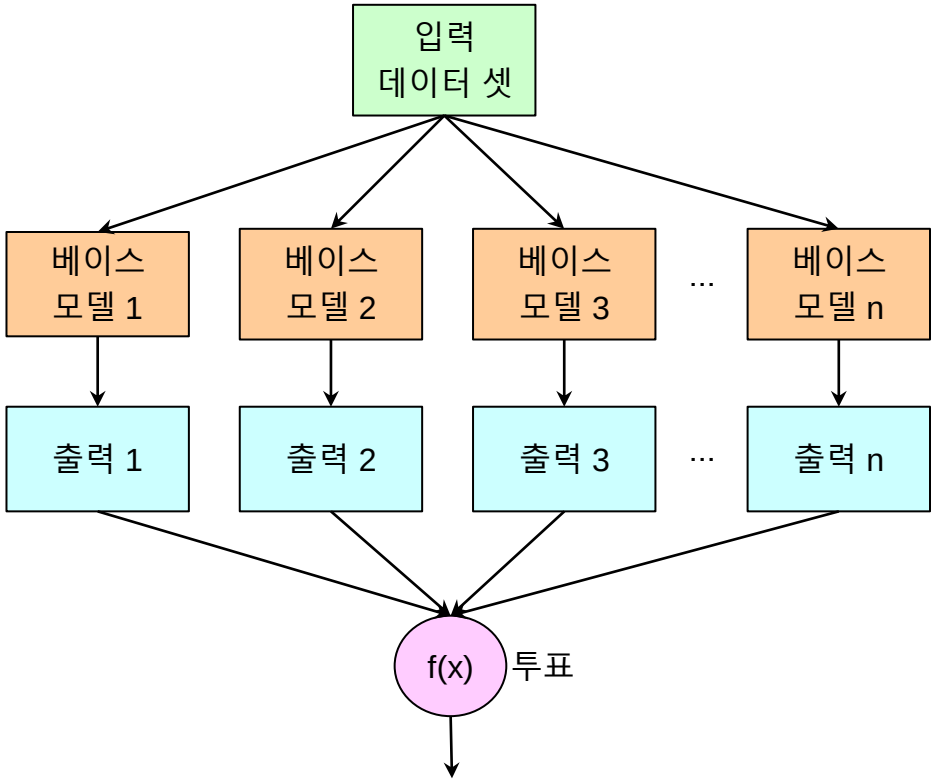
- 정확도(accuracy) = (1,771 + 139) ÷ 2,000 = 0.955
- 오분류율(error) = (9 + 81) ÷ 2,000 = 0.045
- 민감도(sensitivity) = 139 ÷ 220 = 0.632
- 특이도(specificity) = 139 ÷148 = 0.939
- 정밀도(precision) = 1,771 ÷ 1,852 = 0.9563
- 재현율(recall) = 139÷ 220= 0.632
- F1 Score = 2 * (0.9563*0.632)/(0.9563+0.632) =0.761

- 아무리 정확도가 좋아도, 관심이 있는 클래스의 적중률이 높은 경우는 모델로서 성능이 좋다고 할 수 없음
- 오류율뿐 아니라 특이도와 민감도 등을 잘 봐야 함
- 정밀도와 재현율(민감도) 를 다 고려한 F1score 살펴보는 것이 좋음

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) (민감도와 유사)	TP / (TP+FN)
정확도 accuracy	분류기 성능의 종합적 척도	(TP+TN) / (TP+FP+FN+TN)
오분류율		1 - 정확도
F1 Score	precision 과 recall의 조화평균으로, 0에서 1사이의 값을 가지며, 클수록 좋음	$2 * \frac{Precision * Recall}{Precision + Recall}$

앙상블학습법 -개요

- 앙상블(ensemble) 모형은 여러 개의 분류모형에 의한 결과를 종합하여 분류의 정확도를 높이는 방법임
- 동일 데이터 셋을 이용하여 개별적으로 결과를 예측하는 모델들의 결과들을 결합하는 방법임 (집단지성)
- 보통은 투표 또는 가중투표를 통해 앙상블 결과를 출력함
- 대표적인 방법으로는 배깅(bagging), 부스팅(boosting)이 존재하며, 랜덤포레스트(random forest)는 배깅의 개념과 속성(또는 변수)의 임의적으로 선택하는 방법을 결합한 방법임
- 앙상블 기법은 다양한 Weak Learner를 통해 Strong Learner를 만들어가는 과정
  - ✓ 약학습기(약분류기, Weak Learner) : 무작위 선정이 아닌 성공확률이 높은, 즉 오차율이 일정 이하(50% 이하)인 학습 규칙
  - ✓ 강학습기(강분류기, Strong Learner) Weak Learner로부터 만들어내는 강력한 학습 규칙



[앙상블 모델]

앙상블학습법 -개요

앙상블 모델링의 장점

- 평균을 취함으로써 편의(편향)를 최소화:
  : 치우침이 있는 여러 모형의 평균을 취하면, 어느 쪽에도 치우치지 않는 결과(평균)를 얻게 됨

- 분산 감소
  : 한 개 모형으로부터의 도출 된 결과보다 여러 모형의 결과를 결합하면 변동이 작아짐

- 과적합 방지:
  : 여러 모형으로부터 예측을 결합하면 과적합의 여지가 줄어듦

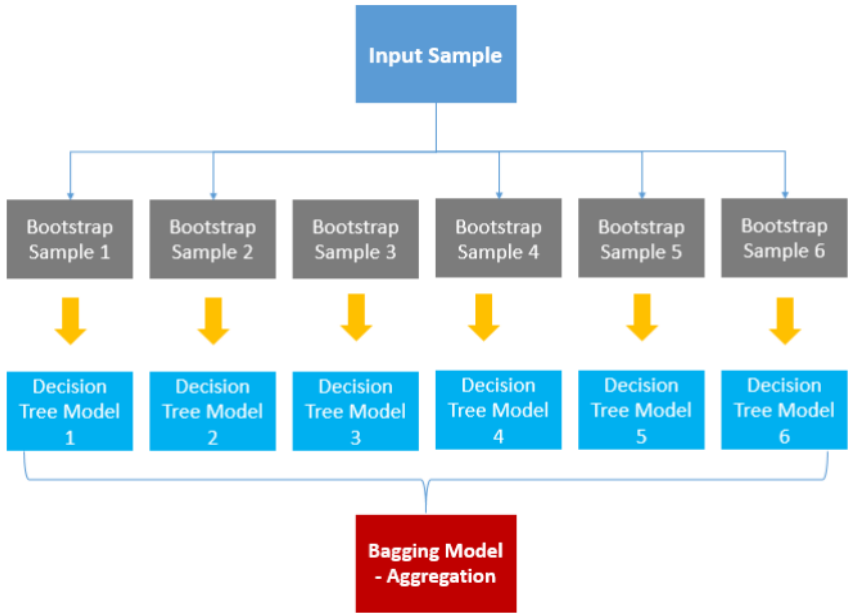
앙상블학습법 -Bagging

Bagging => bootstrap aggregating 의 준말

- Bagging은 샘플을 여러 번 뽑아 (Bootstrap) 각 모델을 학습시켜 결과물을 집계(Aggregation)하는 방법으로. Bootstrap Aggregation의 축약어 임

  * Bootstrap Sampling : 전체 데이터에서 N개의 sample을 복원추출

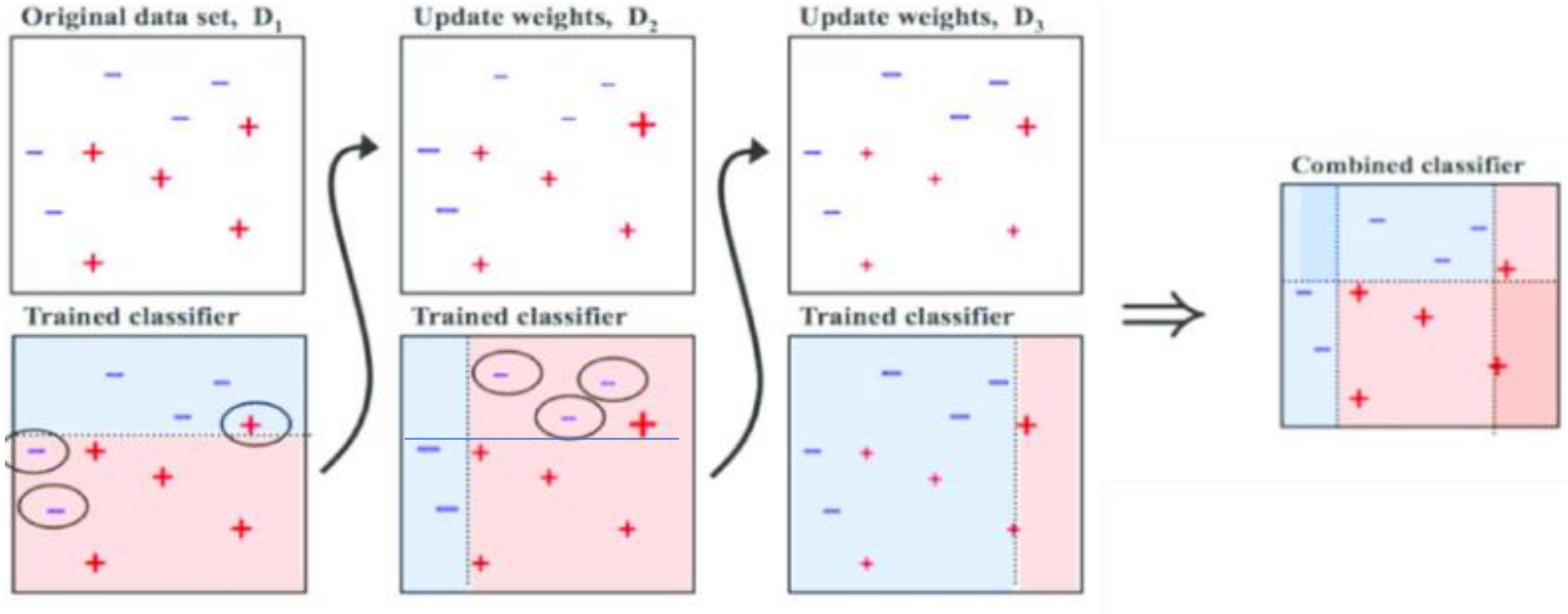
- N개의 bootstrap 샘플링된 표본에서 병렬로 학습하고. N개의 학습자의 결과를 투표(voting) 방식으로 예측 값을 결정함



앙상블학습법 – Boosting

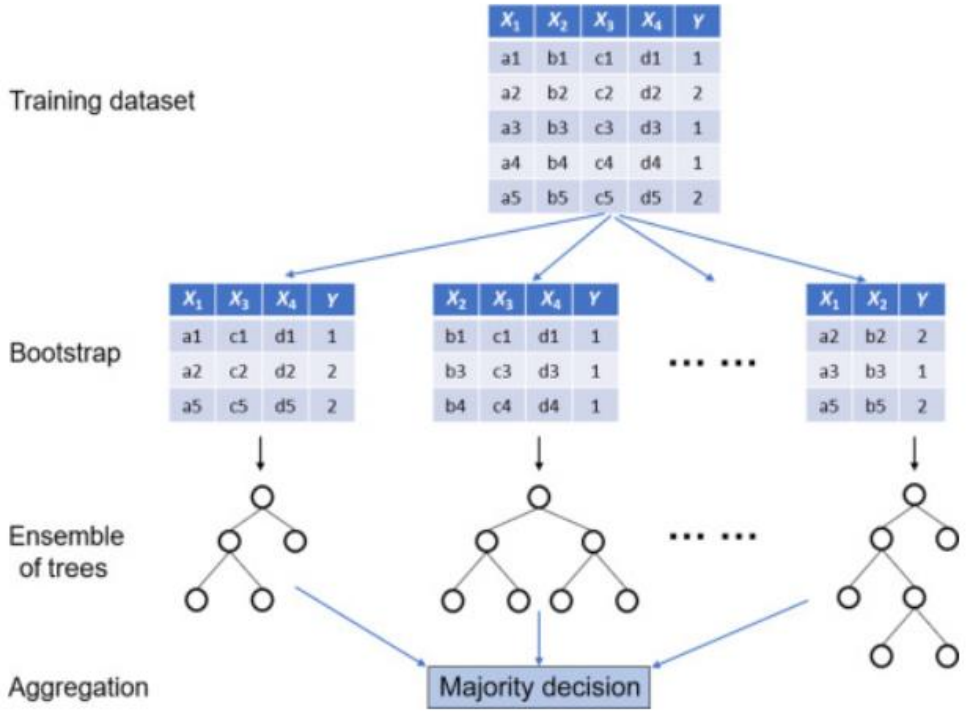
- Boosting의 기본 아이디어는 강력한 학습자를 만들기 위해 일련의 약한 학습자를 순차적으로 학습시키는 것임.
- 병렬하게 훈련되고 예측은 학습자에 대한 선호 없이 집계되는 Bagging의 경우와 달리 Boosting은 순차적으로 훈련하고, 오분류된 분류에 더 많은 가중치를 부과함
- 정분류된 데이터는 추출될 확률을 줄이고, 오분류된 데이터는 추출될 확률을 높여서 모형이 오분류된 데이터를 더 강하게 학습할 수 있도록 도와주는 방법임
- 모형결합시에도 정확도가 높은 모형에 가중치를 더 주는 방식으로 결합함
- Bagging은 학습자 간의 독립성을 활용하여 분산을 줄이기 위해 병렬로 학습하는 반면, Boosting은 학습자 간의 의존성을 이용하여 편향 및 분산을 줄이기 위해 순차적으로 학습함
- 아다부스팅(AdaBoosting: adaptive boosting)은 가장 많이 사용되는 부스팅 알고리즘임

앙상블학습법 -Boosting



앙상블학습법 -RandomForest

- 랜덤포리스트(random forest)는 배깅에 랜덤 과정을 추가한 방법임
- 원 자료로부터 붓스트랩 샘플을 추출하고, 각 붓스트랩 샘플에 대해 트리를 형성해 나가는 과정은 배깅과 유사하나, 예측변수들을 임의로 추출하고, 추출된 변수 내에서 최적의 분할을 만들어 나가는 방법을 사 용함
- 새로운 자료에 대한 예측은 분류(classification)의 경우는 다수결(majority votes)로, 회귀 (regression)의 경우에는 평균을 취하는 방법을 사용함 (다른 앙상블모형과 유사)



refer: Machine Learning for Subsurface Characterization

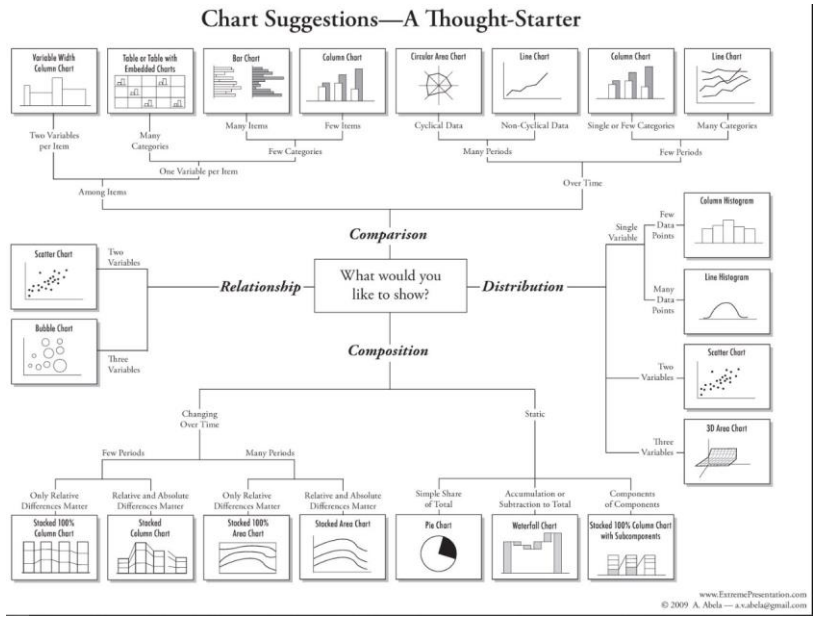
시각화 개요

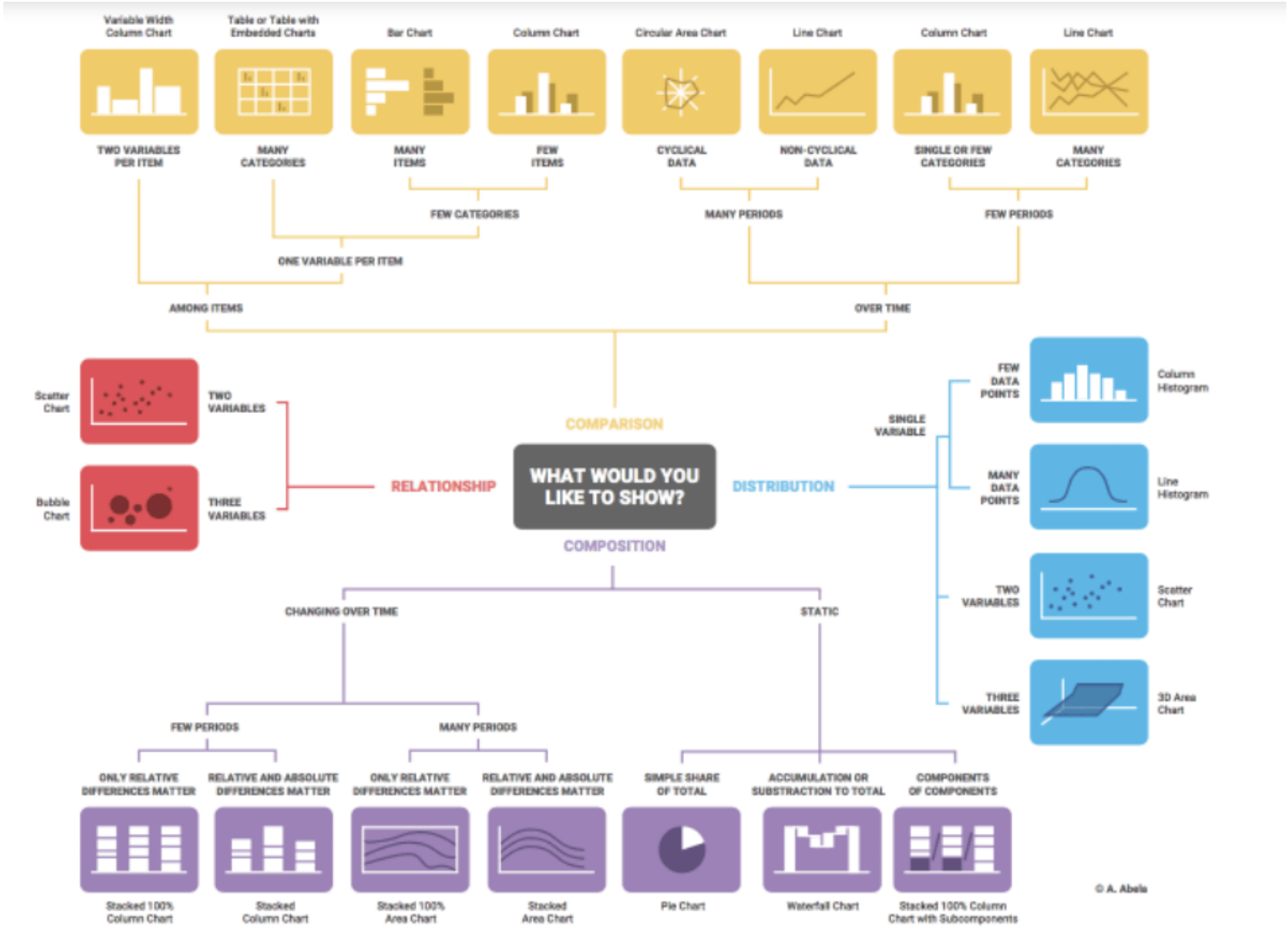
시각화(그래프) 유형

- 시각화의 유형을 선택에는 어려움이 존재함
- 시각화이론에서 많이 사용되는 시각화 선택방법은 앤드류 아벨라(Andrew Abela)의 차트 선택방법 임
  - 비교(Comparison)
  - 구성(Composition)
  - 분포(Distribution)
  - 관계(Relationship)
- 시각화 표출 유형의 결정 시 고려사항
  - ✓ 얼마나 많은 변수들이 하나의 그래프에서 표출되기를 원하는가?
  - ✓ 각 변수에 대하여 얼마나 많은 데이터 점들이 표현되어질 것인가?
  - ✓ 시점 또는 항목간 또는 집단간 값들을 비교하고자 하는가?

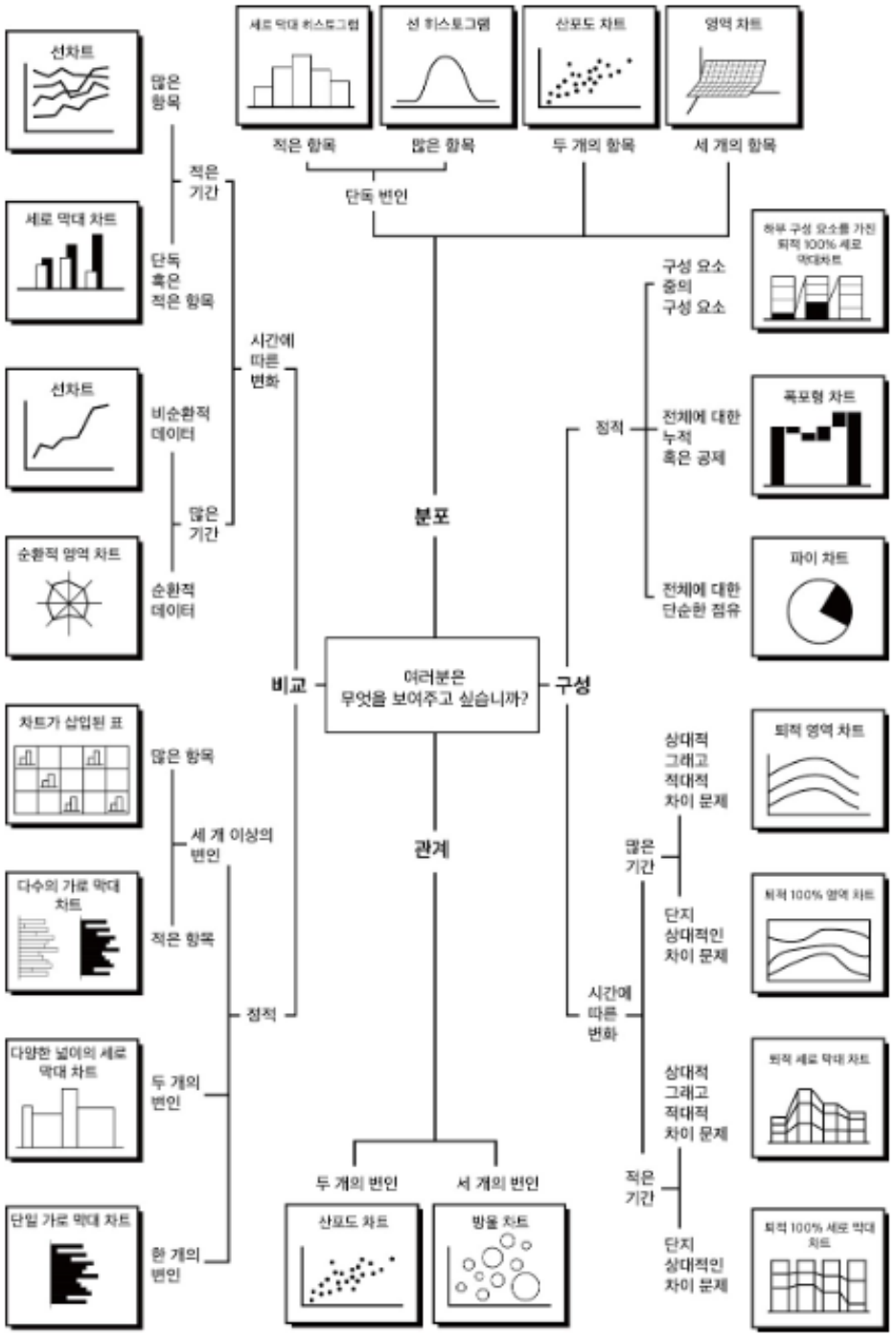
시각화 개요

앤드류 아벨라(Andrew Abela)의 차트 선택방법





refer: https://biuwer.com/en/blog/how-to-choose-the-right-chart-for-your-data/
Original source: Andrew Abela(2009)[left]



refer:https://www.i-boss.co.kr/ab-74668-1747