[2020 혁신성장 청년인재 집중양성 사업]

프로젝트 기반 데이터 과학자 양성과정

빅데이터 분석

- 8주차 -

#데이터 마이닝 #의사결정나무

Part 1

데이터 마이닝

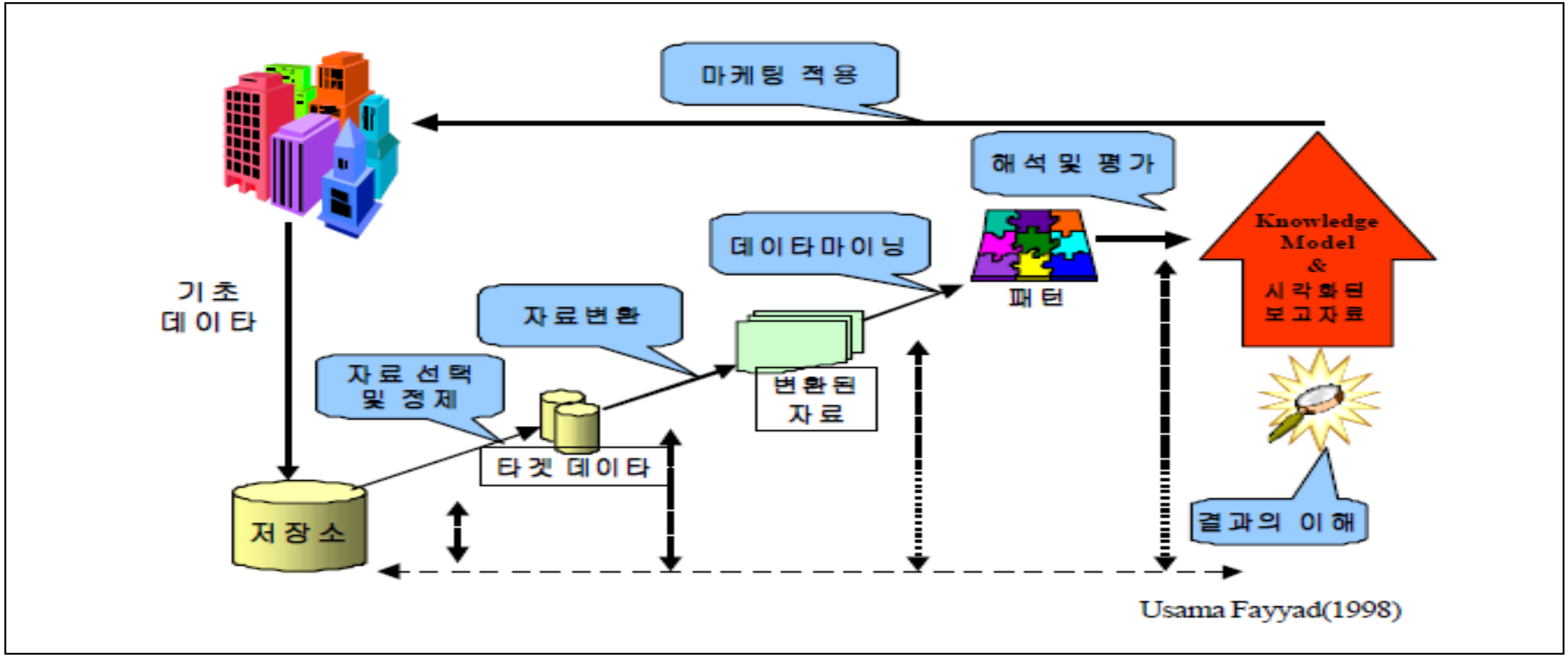
데이터 마이닝

데이터 마이닝의 이해

- 대용량의 데이터에서 유용한 정보와 관계를 탐색하고 모형화하여 지식을 발견하는 과정
- 대용량의 관측된 자료를 다룬다
- 이론보다는 실무위주의 컴퓨터 중심적인 방법이다
- 경험적 방법에 근거하고 있다
- 주요관심은 예측모형의 일반화에 있다
- 기업의 다양한 의사결정에 활용된다
- 통계학, 전산학, 인공지능, 공학과 같은 분야에서 주로 개발된다

데이터 마이닝

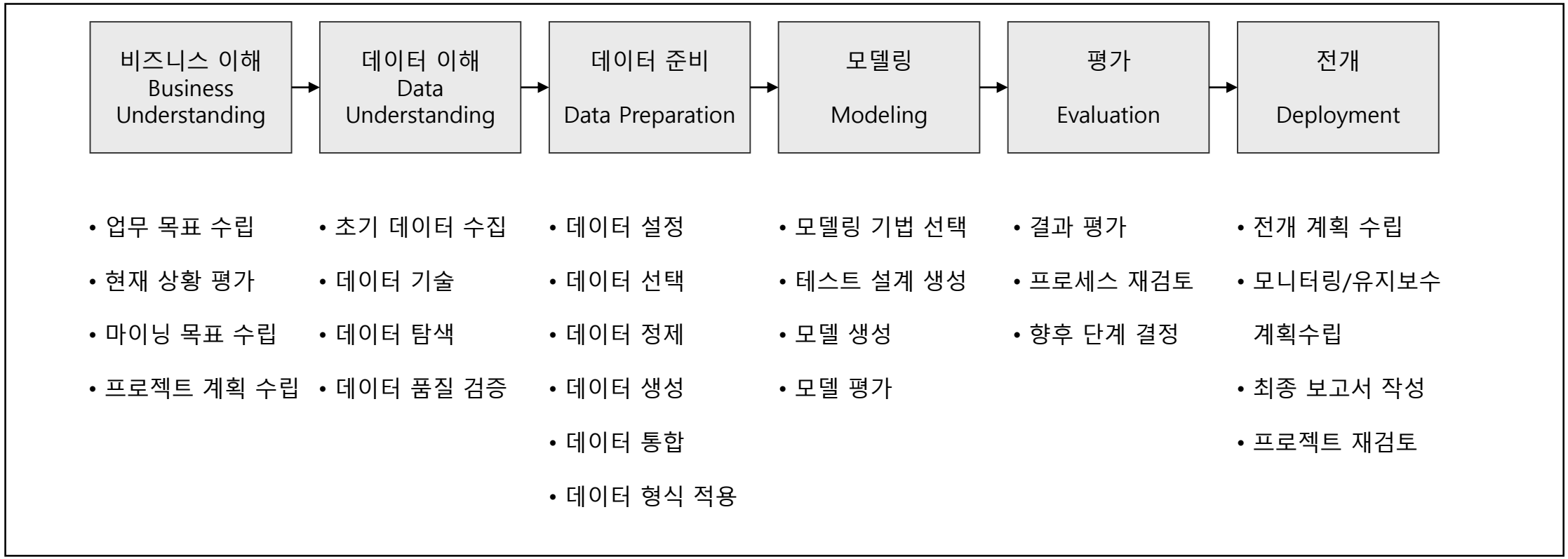
데이터 마이닝 프로세스(1/4)



데이터 마이닝

데이터 마이닝 프로세스(2/4)

✓ CRISP-DM(Cross-Industry Standard Process for Data Mining, 데이터 마이닝 표준 실행과정)은 SPSS, NCR, Daimler-Chrylsler 등 여러 업계의 선도회사가 3년여 동안 데이터 마이닝 작업의 표준화를 연구하여 발표한 방법론이다.

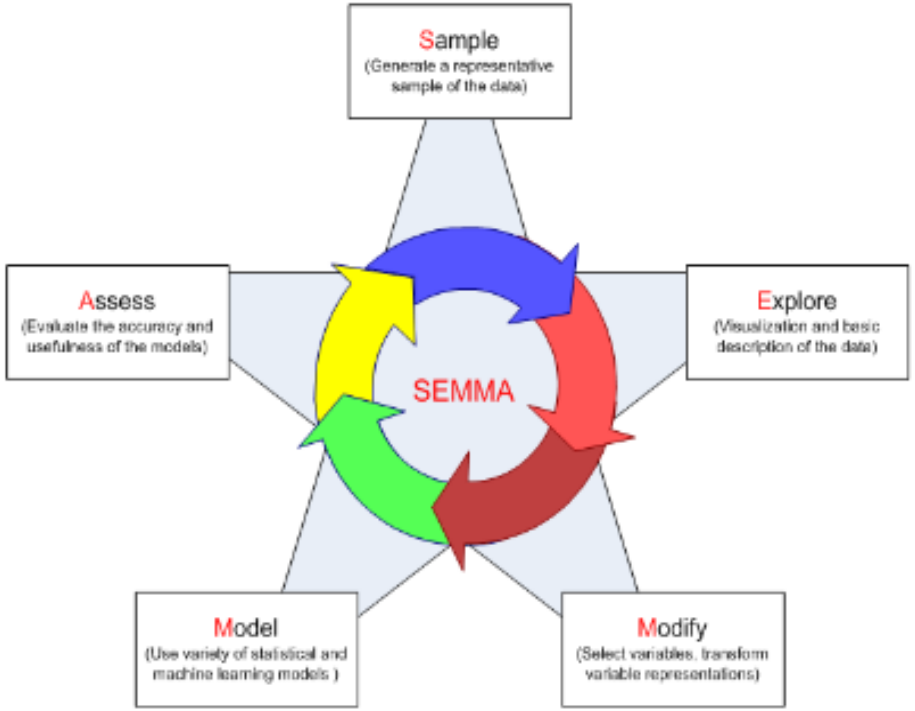


데이터 마이닝

데이터 마이닝 프로세스(3/4)

✓ SAS의 지도예측 문제 해결을 위한 5단계 분석 전략은 다음과 같습니다

1	표본 Sample	▪ 예측모형을 구축하기 위해 적절한 양의 표본을 원 자료로부터 추출하는 단계
2	탐색 Explore	▪ 모형개발과정에 적합한 데이터 특성을 조사하고 정보를 획득하는 단계 (기초통계, 도수분포표, 평균, 분산, 비율 등)
3	수정 Modify	▪ 모형의 예측력을 증대시키기 위해 데이터 형태 변경(변환, 수치화, 그룹화 등)하는 단계
4	모형화 Model	▪ 분석 목적에 따라 적절한 기법(로지스틱 판별분석, 신경망, 의사결정나무)을 사용하여 예측모형을 만드는 단계
5	평가 Assess	▪ 모형화의 결과에 대한 신뢰성, 유용성 등을 평가하는 단계

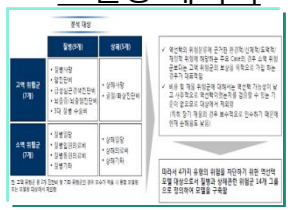


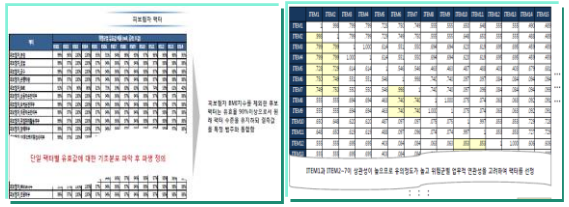
데이터 마이닝

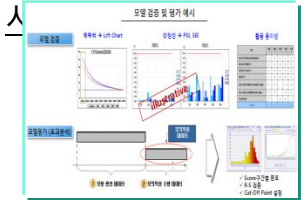
데이터 마이닝 프로세스(4/4)

- 현황파악을 통해 모델링 방향을 정의하고, 실제 업무와 데이터를 분석하여 적절한 후보 모델들을 생성하고 평가 수행을 통하여 최종 Score 모델을 생성 합니다.

Phase	현황파악	현황분석	업무수행	평가수행
Score Modeling	방향정의	현황분석/설계	모델생성	모델 평가
수행활동	• 모델링 방향 설정 • 모델링 범위 정의	• 현황분석 - 업무현황분석 - 데이터 현황분석 • 요구사항 분석 • 모델 상세설계 • Factor 설계 • 모델 운영 설계	• 데이터 탐색 • 데이터 변환/정제 - 데이터 보간 - Binning • 유의변수 선택 - 상관관계분석 • 모델 생성	• 모델 평가 - 예측력 평가 - 안전성 평가 - 활용용이성 평가 • 모델선정 • 모델 Score 생성
주요산출물	모델링 계획서	데이터 탐색 및 유의변수 선택 결과서		모델 생성 결과 보고서



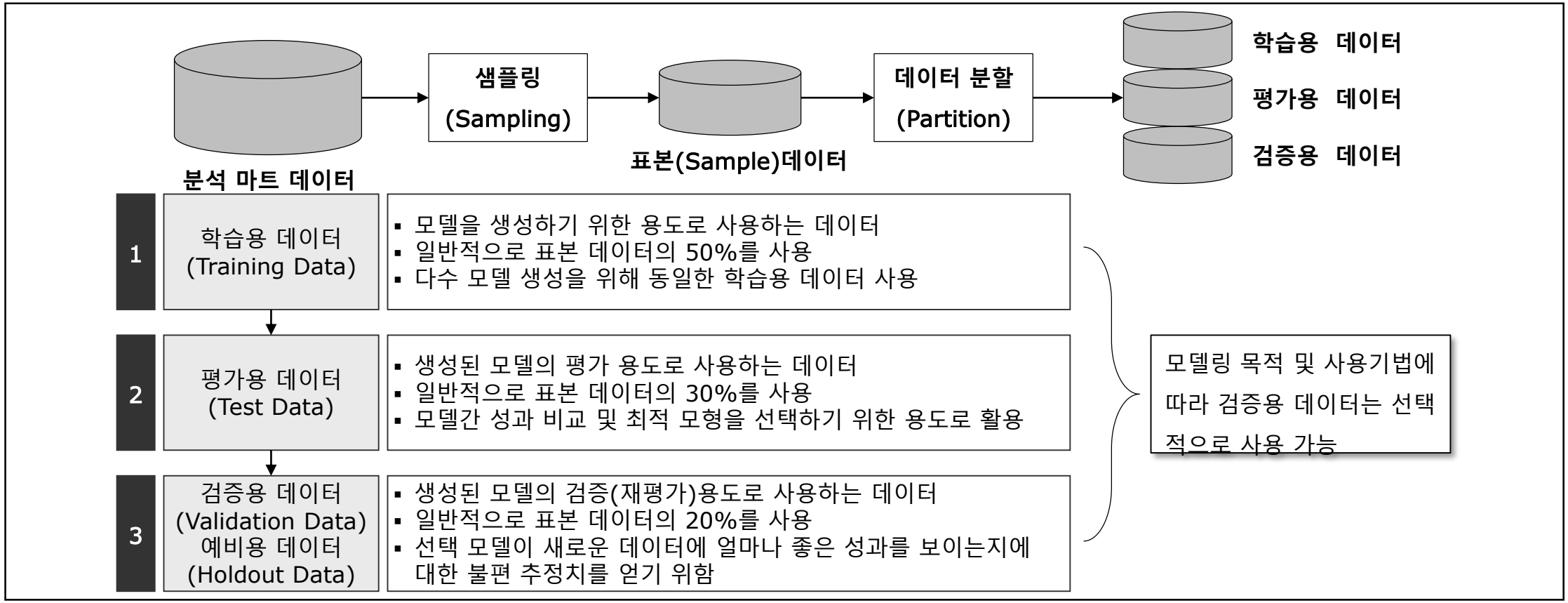




데이터 마이닝

데이터 마이닝의 분할

- 데이터 마이닝을 이용하여 모델을 개발할 시에는 모델의 개발, 테스트, 검증을 위해 데이터를 분할하여 사용합니다.



데이터 마이닝

모델링 기법(1/2)

- 데이터 유형에 따른 기법 분류

변수 구분	연속형 반응변수	범주형 반응변수	반응변수 없는 경우
연속형 예측변수	• 선형 회귀분석 • 신경망모형 • K-최근접이웃기법	• 로지스틱 회귀분석 • 판별분석 • K-최근접이웃기법	• 주성분 분석 • 군집분석
범주형 예측변수	• 선형 회귀분석 • 신경망모형 • 회귀나무	• 신경망모형 • 분류나무 • 로지스틱 회귀분석 • 단순 베이즈 분류모형	• 연관성규칙

※ 반응변수 : 보통 Y로 표기, 타겟 변수, 종속변수 등 vs. 예측변수 : 보통 X로 표기, 설명변수, 독립변수 등

데이터 마이닝

모델링 기법(2/2)

- 데이터 유형에 따른 기법 분류

목적	작업 유형	설명	도출 규칙 예	사용기법
예측 (Predictive Modeling)	분류 규칙 Classification	가장 많이 사용되는 작업으로 과거의 데이터로부터 고객특성을 찾아내어 분류모형을 만들어 이를 토대로 새로운 레코드의 결과값을 예측하는 것으로 목표마케팅 및 고객 신용평가 모형에 활용됨	고객신용 평가결과가 '불량'인 고객은 '25~30세 가량의 미혼남으로 월평균 수입이 200 만원 이하인 고객'으로 신규고객 유치 시 이러한 규칙을 활용	회귀분석, 판별분석, 시계열, 신경망, 의사결정나무
설명 (Descriptive Modeling)	연관 규칙 Association	데이터 안에 존재하는 항목간의 종속관계를 찾아내는 작업으로, 제품이나 서비스의 교차판매(Cross Selling), 매장진열(Display), 첨부우편(Attached Mailings), 사기적발(Fraud Detection) 등의 다양한 분야에 활용됨	'넥타이를 구입한 고객은 셔츠도 구입한다', '정장과 벨트를 구입하면 코트도 함께 구입한다.'등	동시발생 매트릭스
	연속 규칙 Sequence	연관 규칙에 시간관련 정보가 포함된 형태로, 고객의 구매이력(History)속성이 반드시 필요하며, 목표 마케팅(Target Marketing)이나 일대일 마케팅(One to One Marketing)에 활용됨	'새 냉장고를 구입한 고객 중 한달 이내에 새 오븐을 구입하는 확률이 75% 이다' 등	동시발생 매트릭스
	데이터 군집화 Clustering	고객 레코드들을 유사한 특성을 지닌 몇 개의 소그룹으로 분할하는 작업으로 작업의 특성이 분류규칙(Classification) 과 유사하나 분석대상 데이터에 결과값이 없으며, 판촉활동이나 이벤트 대상을 선정하는 데 활용됨	고객을 평균 예탁금, 월평균약정, 회전율, 월평균 거래건수 등으로 고객세분화 실시 후, 이들 군의 특징을 파악하고 이탈과 복귀를 반복하는 고객군 선별, 특성파악 (다른 데이터 마이닝의 선행작업으로 많이 활용됨)	K-Means Clustering

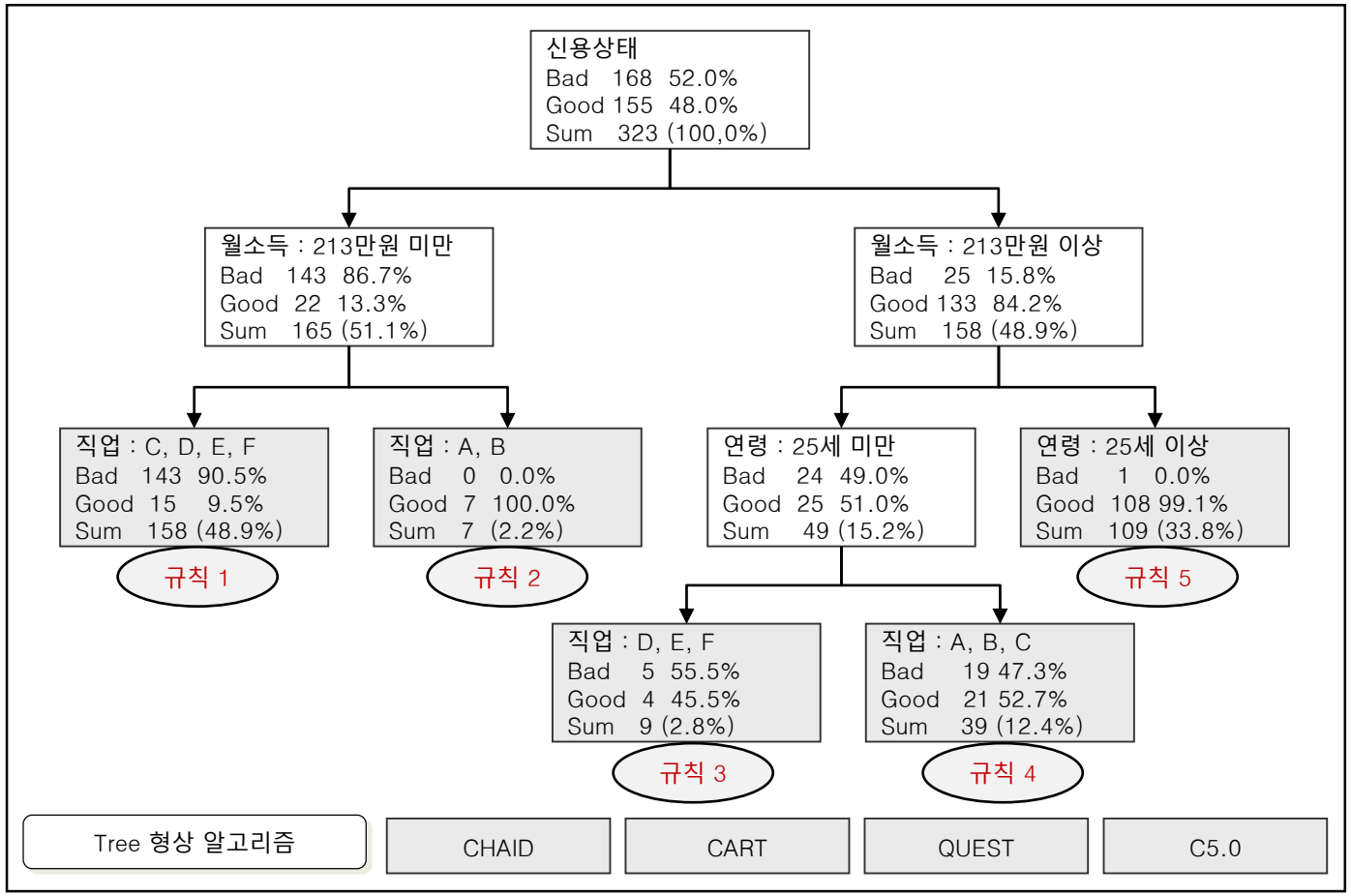
Part 2

의사결정나무

의사결정나무

의사결정나무란?

- 발견된 변수의 규칙 혹은 조건문을 토대로 나무 구조로 도표화하여 분류와 예측을 수행하는 방법



의사결정나무

의사결정나무의 개념

- 목적과 자료구조에 따라 적절한 분리 기준과 정지 규칙을 지정하여 의사결정나무를 얻음
- 분리 기준 : 어떤 입력 변수를 이용하여 어떻게 분리하는 것이 목표 변수의 분포를 가장 잘 구별해 주는지에 대한 기준
- 목표 변수의 분포를 구별하는 정도 : 순수도 or 불순도
- 순수도 : 목표 변수의 특정 범주에 개체들이 포함되어 있는 정도
- 부모 마디의 순수도에 비해서 자식마디들의 순수도가 증가하도록 자식 마디를 형성함
- 가지치기 : 분류 오류를 크게 할 위험이 높거나 부적절한 가지를 제거

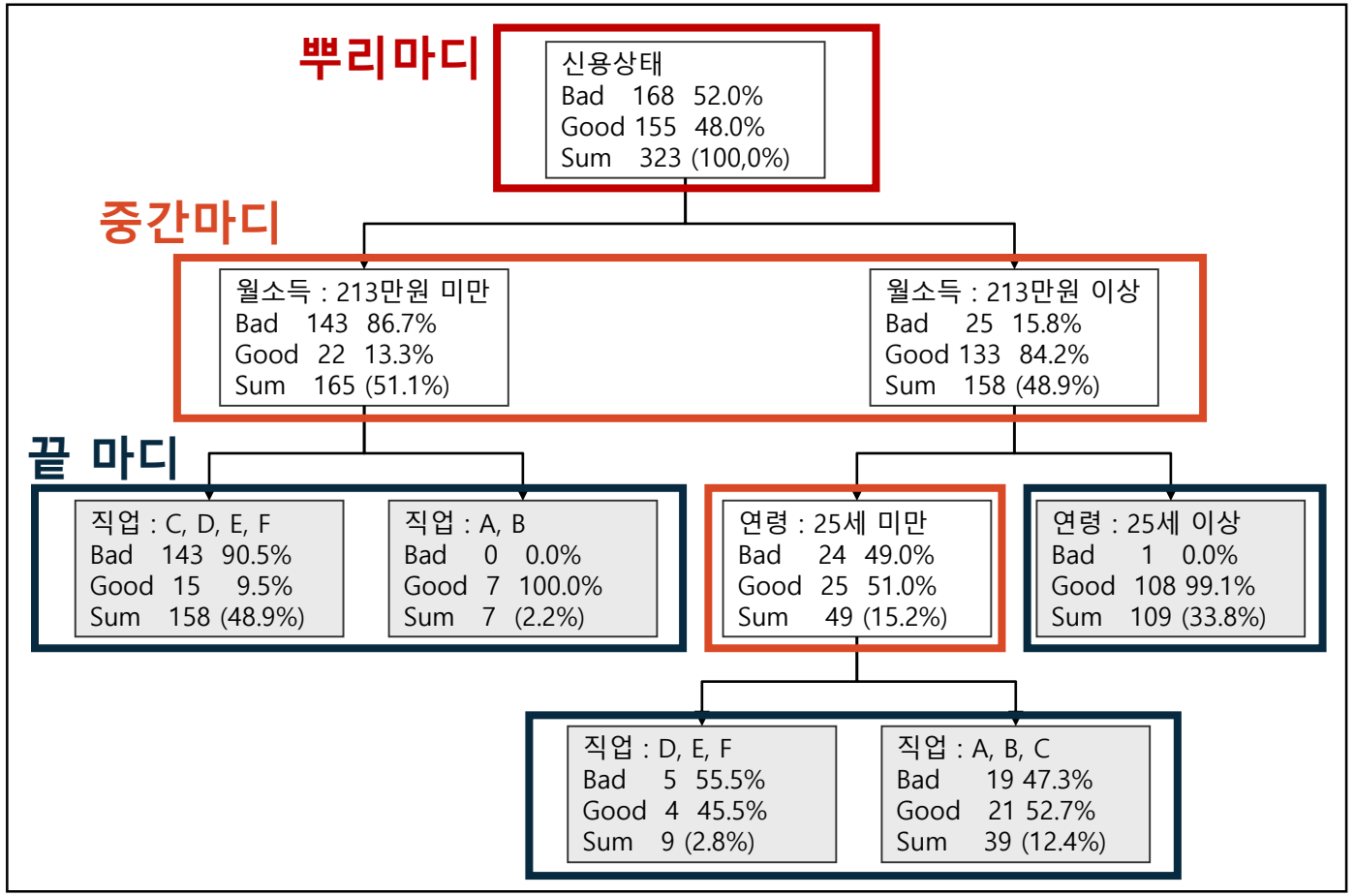
의사결정나무

의사결정나무의 장단점

- 해석이 용이하며, 어떤 입력변수가 중요한지 파악이 쉽다
- 두 개 이상의 변수가 결합하여 목표변수에 어떠한 영향을 주는지 알기 쉽다
- 계산 속도가 빠르고 대형자료 처리에 용이
- 이상치에 민감하지 않다
- 이산형 변수에 대하여 수준이 많은 경우 결과가 정확하지 않을 수 있다
- 연속형 변수를 비연속적인 값으로 취급하여 예측 오류가 크다

의사결정나무

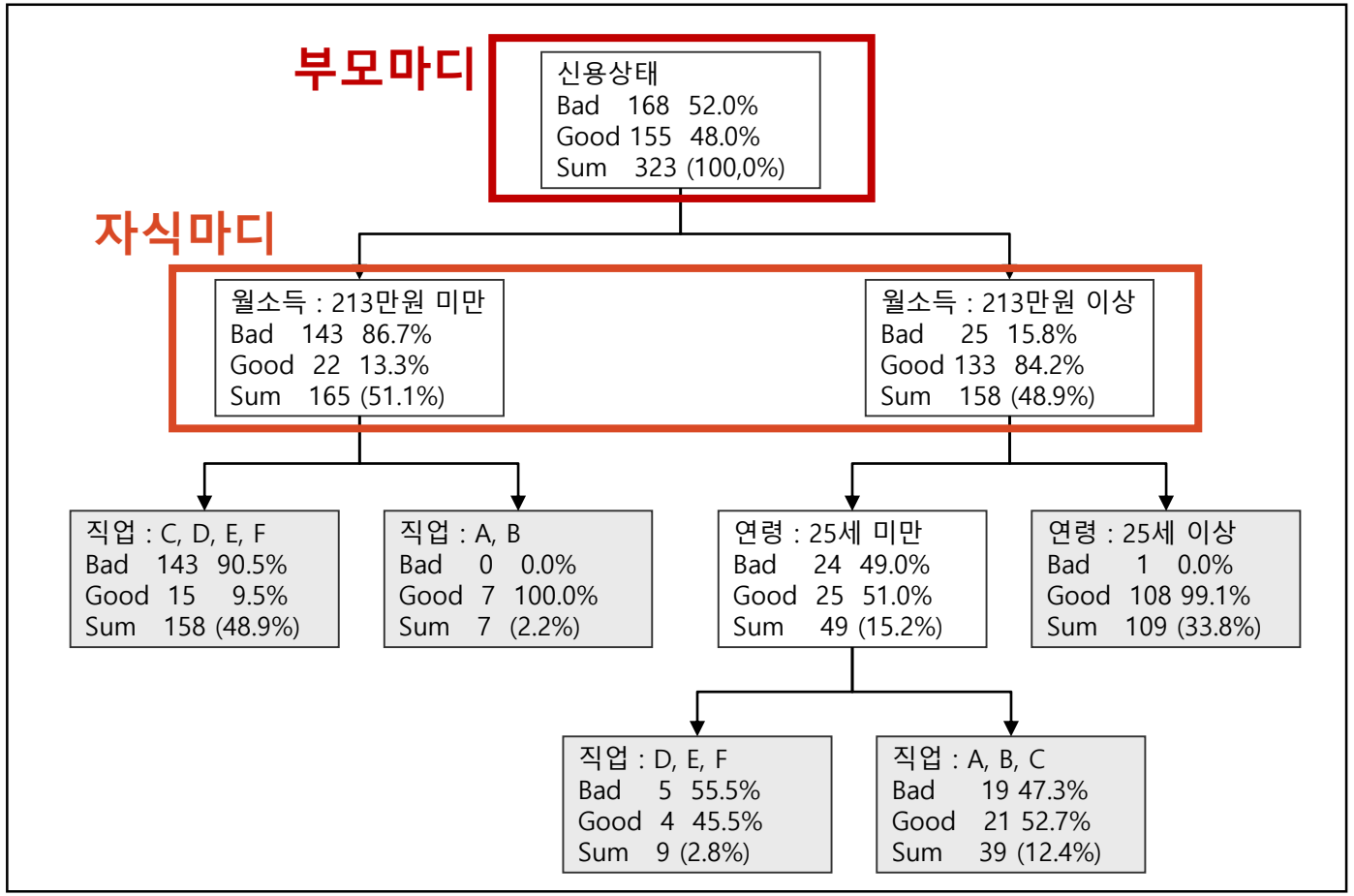
의사결정나무의 구성요소(1/5)



- 뿌리마디(root node) : 나무구조가 시작되는 마디
- 중간마디(inernal node) : 중간에 있는 끝 마디가 아닌 마디
- 끝 마디(terminal node, leaf) : 각 나무줄기의 끝에 위치하는 마디

의사결정나무

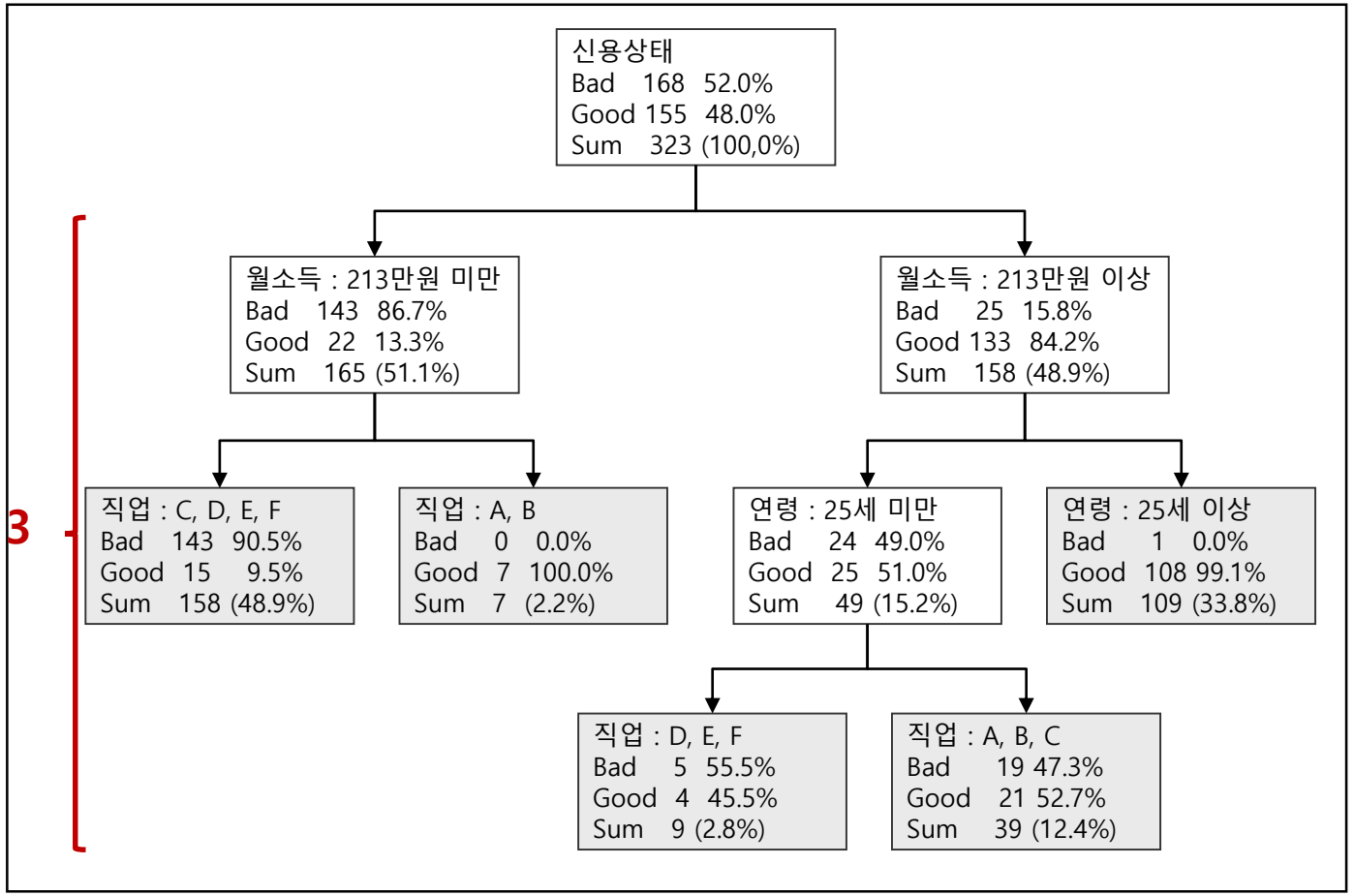
의사결정나무의 구성요소(2/5)



- 부모마디(parent node) : 자식마디의 상위마디
- 자식마디(child node) : 하나의 마디로부터 분리되어 나간 마디

의사결정나무

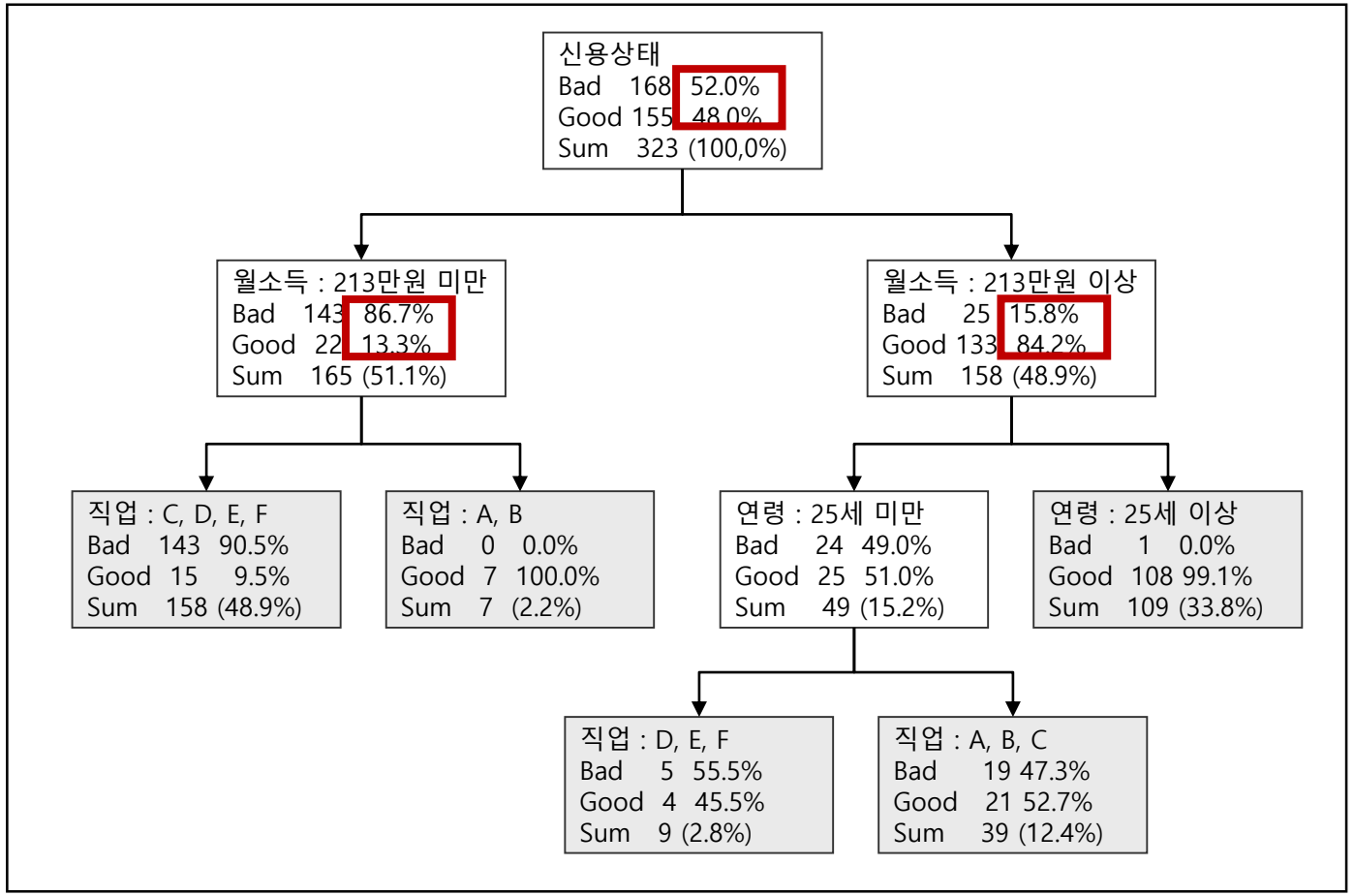
의사결정나무의 구성요소(3/5)



- 가지(branch) : 하나의 마디로부터 끝마디까지 연결된 마디들
- 깊이(depth) : 가지를 이루고 있는 마디의 개수

의사결정나무

의사결정나무의 구성요소(4/5)

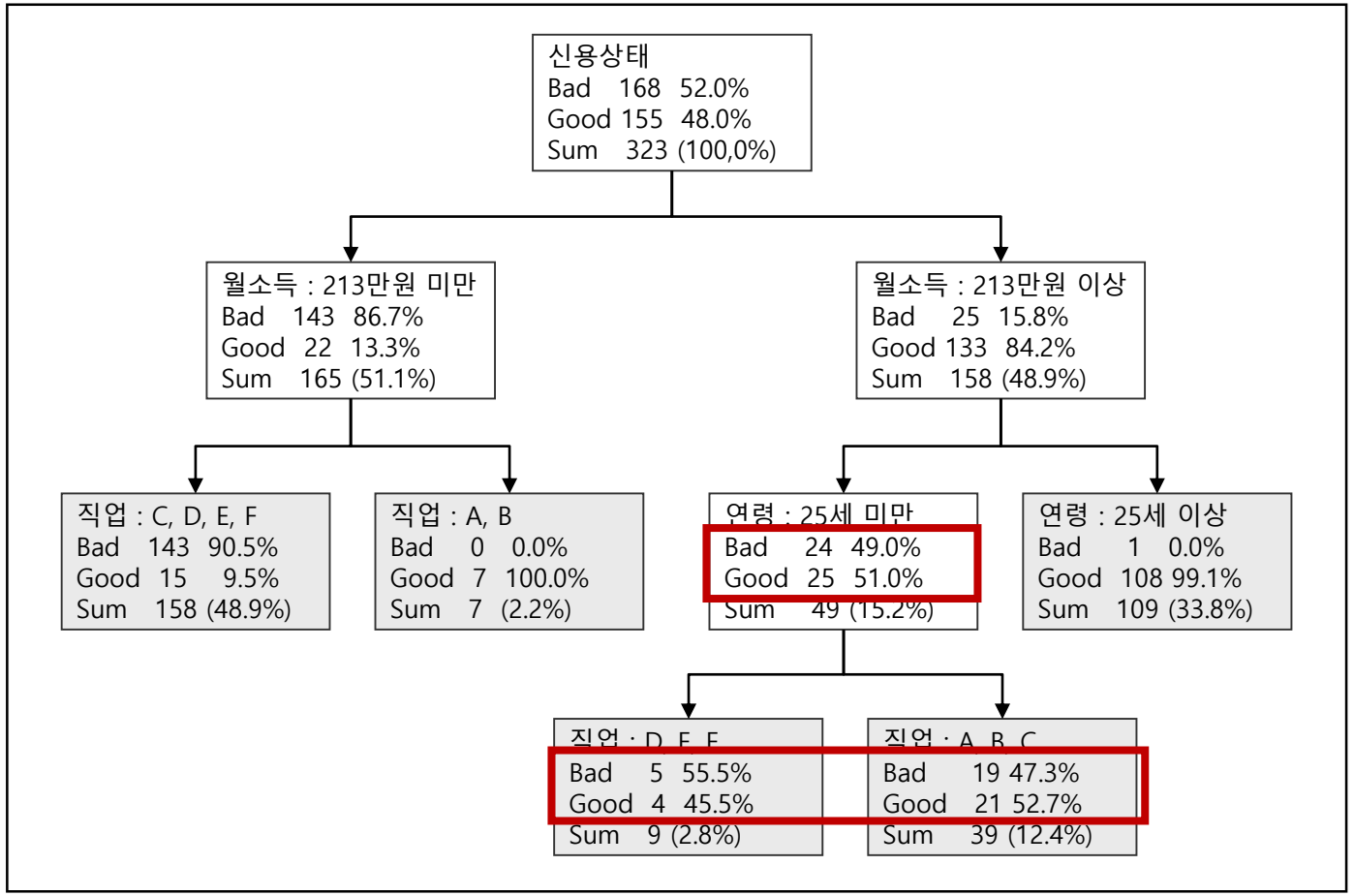


- 분리기준 : 어떤 입력변수를 이용하여 어떻게 분리하는 것이 타겟 변수를 잘 구별해 주는지 비교

- 부모 마디의 순수도에 비해서 자식 마디의 순수도가 증가하도록 자식 마디를 형성

의사결정나무

의사결정나무의 구성요소(5/5)



- 정지규칙(stoping rule) : 더 이상 분리가 일어나지 않고, 현재의 마디가 끝 마디가 되도록 하는 규칙

- 가지치기(pruning) : 적절하지 않은 마디를 제거하여 적당한 크기의 부나무(subtree) 구조를 가지도록 하는 규칙

의사결정나무

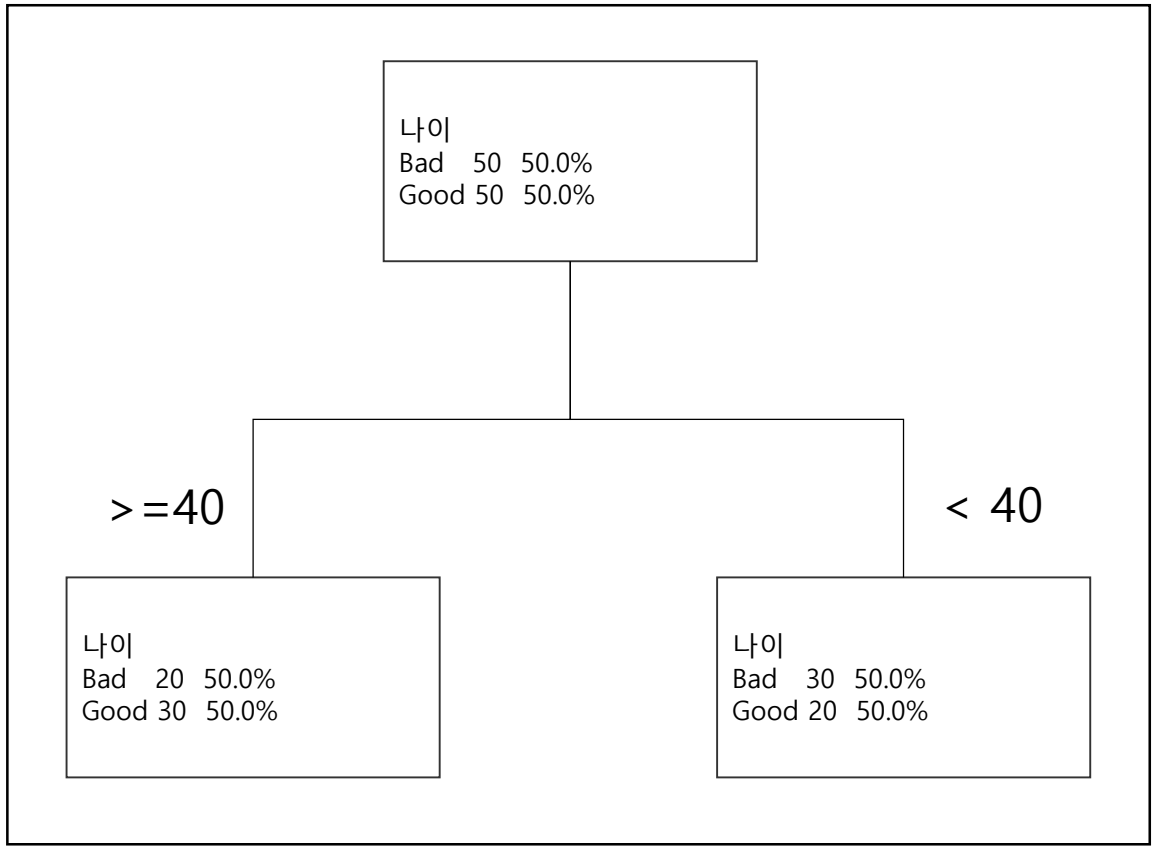
CART(Classification and Regression Tree)

- 가장 널리 사용되는 의사결정나무 알고리즘
- 분류와 회귀에 모두 적용
- 이산형(범주형) 타겟 변수의 경우, 타겟 변수의 각 범주에 속하는 빈도에 기초하여 분리가 일어남
- 지니 지수(Gini index)
- 연속형(구간형) 타겟 변수의 경우, 타겟 변수의 평균과 표준편차에 기초하여 분리가 일어남
- 분산의 감소량(Variance reduction)
- 분리 방법은 이지 분리(binary split)

의사결정나무

Gini 지수

- 불순도 계산 예시 $\phi(g)=1-\sum_{i=1}^{j}\hat{p}_i(g)^2$



구분	계산(예시)
뿌리 노드의 불순도	$1-[(50/100)^2+(50/100)^2]=0.5$
끝 노드의 불순도	$1-[(20/50)^2+(30/50)^2]=0.48$ $1-[(30/50)^2+(20/50)^2]=0.48$
불순도 감소 폭	$0.5-[(50/100)*0.48]*2=0.02$

의사결정나무

Gini 지수

- 불순도 계산 예시

	CART	C 5.0	CHAID	QUEST
목표 변수	범주형 연속형	범주형	범주형 연속형	범주형
예측 변수	범주형 연속형	범주형 연속형	범주형	범주형 연속형
분리 기준	지니지수 분산의 감소량	엔트로피 지수	카이제곱 통계량 F-검정	카이제곱 통계량 F-검정
분리 개수	이지분리	다지분리	다지분리	이지분리

Part 3

실습

의사결정나무

실습

- Kyphosis 척추수술 후 관측되는 척추기형(Kyphosis="present") 17명, 대조군 (Kyphosis="absent") 64명에 대한 자료

	Kyphosis	Age	Number	Start
1	absent	71	3	5
2	absent	158	3	14
3	present	128	4	5
4	absent	2	5	1
5	absent	1	4	15
6	absent	1	2	16
7	absent	61	2	17
8	absent	37	3	16
9	absent	113	2	16
10	present	59	6	12

- Age(나이)
- Start(척추 시작번호)
- Number(척추 수)

의사결정나무

실습

- 209개의 컴퓨터 cpu의 종합적인 성능평가 perf 및 하드웨어 특성 변수 syct, mmin, mmax, cach, chmin, chmax, perf, estperf에 대한 자료

	name	syct	mmin	mmax	cach	chmin	chmax	perf	estperf
1	ADVISOR 32/60	125	256	6000	256	16	128	198	199
2	AMDAHL 470V/7	29	8000	32000	32	8	32	269	253
3	AMDAHL 470/7A	29	8000	32000	32	8	32	220	253
4	AMDAHL 470V/7B	29	8000	32000	32	8	32	172	253
5	AMDAHL 470V/7C	29	8000	16000	32	8	16	132	132
6	AMDAHL 470V/8	26	8000	32000	64	8	32	318	290
7	AMDAHL 580-5840	23	16000	32000	64	16	32	367	381
8	AMDAHL 580-5850	23	16000	32000	64	16	32	489	381
9	AMDAHL 580-5860	23	16000	64000	64	16	32	636	749
10	AMDAHL 580 5880	23	32000	64000	128	32	64	1144	1238

Part 4

로지스틱 회귀

로지스틱 회귀

로지스틱 회귀모형의 개념

- 목표변수가 이항형일 때, 선형 회귀모형의 단점을 극복하기 위해 확률에 대한 로짓변환을 고려하여 모형화
- Odds : P(A)를 특정사건 A가 일어날 확률이라고 하면 A가 일어나지 않을 확률은 정의상 1 – P(A)가 됨
- Odds가 1보다 작다는 것은 입력변수가 감소방향으로 영향을 미침을 뜻하고, 반대로 1보다 크다는 것은 증가방향으로 영향을 미침을 의미한다
- Odds는 0에서 무한대의 수로 표현
- Logit 무한대에서 무한대로 표현

로지스틱 회귀

최대 우도 추정법

- 우도(Likelihood) : 주어진 정보를 바탕으로 모집단의 모수에 관해 어떤 추정량이 적합한가?

- 모수로부터 특정현상이 관찰되는 것을 확률의 문제라고 한다면, 우도는 주어진 현상을 가지고 이 현상이 추출될 가능성을 추적하는 법

- 이를 가장 높게 하는 모수를 거꾸로 추적하는 방법이 최대우도추정법(maximum likelihood method)이다

- 다중회귀분석에서는 최소제곱법, 로지스틱 회귀분석에서는 최대우도추정법을 씀

$$L(a,b) = \prod_{i=1}^{n} F(a+bx_i)^{y_i}(1-F(a+bx_i))^{1-y_i},$$

여기서 $F(x) = \exp(x)/(1+\exp(x))$

로지스틱 회귀

Odds Ratio

- 오즈(odds) : 상대적 확률의 표현 방법 중 하나. A를 했을때 B라는 결과가 나올 확률을 π라고 할때 B의 오즈(odds)는 $\pi/(1-\pi)$가 된다.

- 오즈비(odds ratio) : A의 유무가 B의 유무에 얼마나 강한 영향을 끼치는지를 수치화하기 위한 방법. 오즈비가 1에서 멀어질 수록 A와 B 사이의 연관성이 크다.

- Bassassinator 의 오즈 = 50/50 = 1

- No bait 의 오즈 = 2/98 = 0.02

- Bassassinator vs No bait의 오즈비 = 1/0.02 = 50

- Bassassinator를 사용한 경우 No bait를 사용한 경우보다 물고기를 잡을 확률이 50백 높다

	Of time caught	Of time not caught	Total of cast
Bassassinator	50	50	100
No bait	2	98	100

로지스틱 회귀

Logit

- 로짓(Logit) 변환 : 두 개의 범주를 갖는 반응변수를 범주형의 설명변수로 설명. 두 범주를 취할 확률의 비가 설명변수의 수준에 따라 어떻게 달라지는지를 선형모형으로 표현한 모형

$$L(A) = \log\left(\frac{P(A)}{1-P(A)}\right)$$

- 범위는 무한대

Part 5

실습