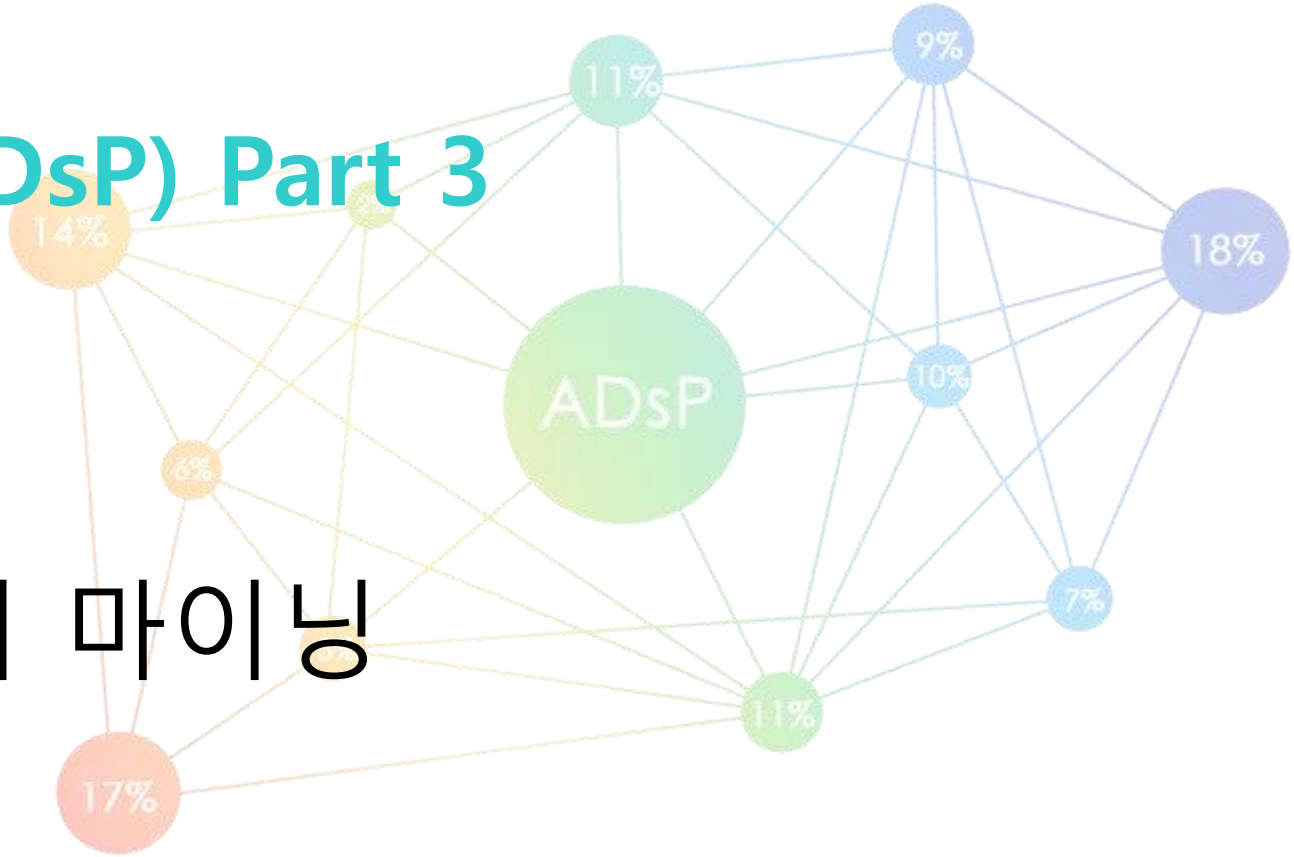
데이터분석준전문가(ADsP) Part 3

데이터분석

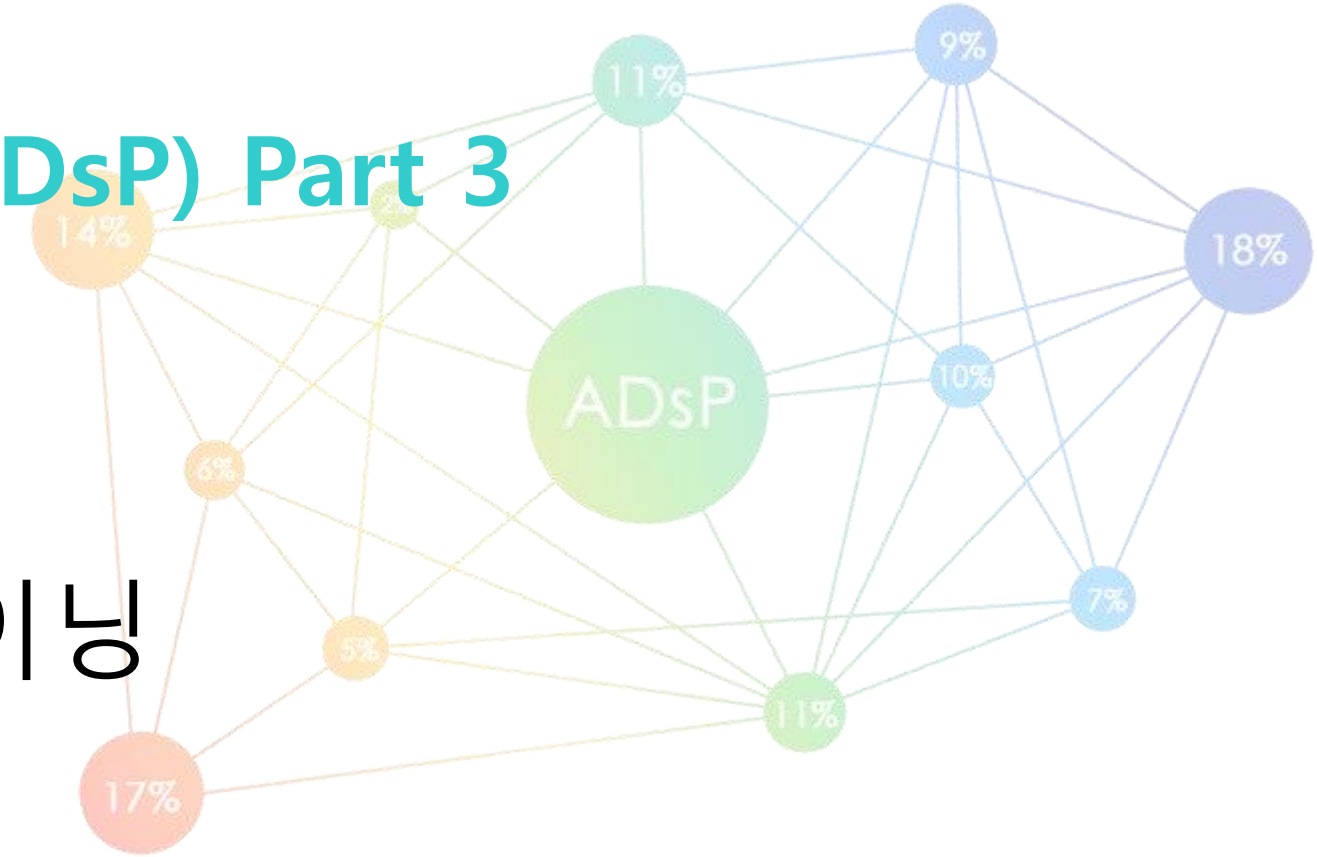


03 제3장 정형 데이터 마이닝

제1절 데이터 마이닝
제2절 분류분석
제3절 군집분석
제4절 연관분석

데이터분석준전문가(ADsP) Part 3

데이터분석



03 제1절 데이터 마이닝

1. 데이터 마이닝 개요
2. 모형평가

1.데이터 마이닝 개요

(1) 데이터 마이닝이란 정의

대용량 데이터베이스로부터 숨어 있는 예측 가능한 정보들을 추출, 의사결정활용

Key Word → 자동화(Automated), 숨겨진(Hidden), 예측가능(Predictive)

통계분석 Vs. 데이터 마이닝

전통적 통계분석

대상집단이 있으며, 모집단의 분포 혹은 모형 등 여러 가지 가정을 전제로 하게 되며 이 전제 조건하에서 분석을 실시 표본(Sample)의 관찰을 통해 모수(Population) 전체를 추론(Inference)하는 과정

데이터마이닝

표본조사/실험에서 필연적으로 수반되는 분포라든가 모형에 대한 전제조건이 필요하지 않음. 모집단의 전체자료를 이용하여 필요한 정보/지식을 추출하는 과정. 대용량 자료여야 한다는 전제조건 있음

1.데이터마이닝 개요

- 데이터마이닝 VS 데이터웨어하우스

  데이터마이닝 기법을 사용하기 위해 데이터웨어하우스가 필요할 뿐

- 데이터마이닝 VS SQL vs OLAP 차이

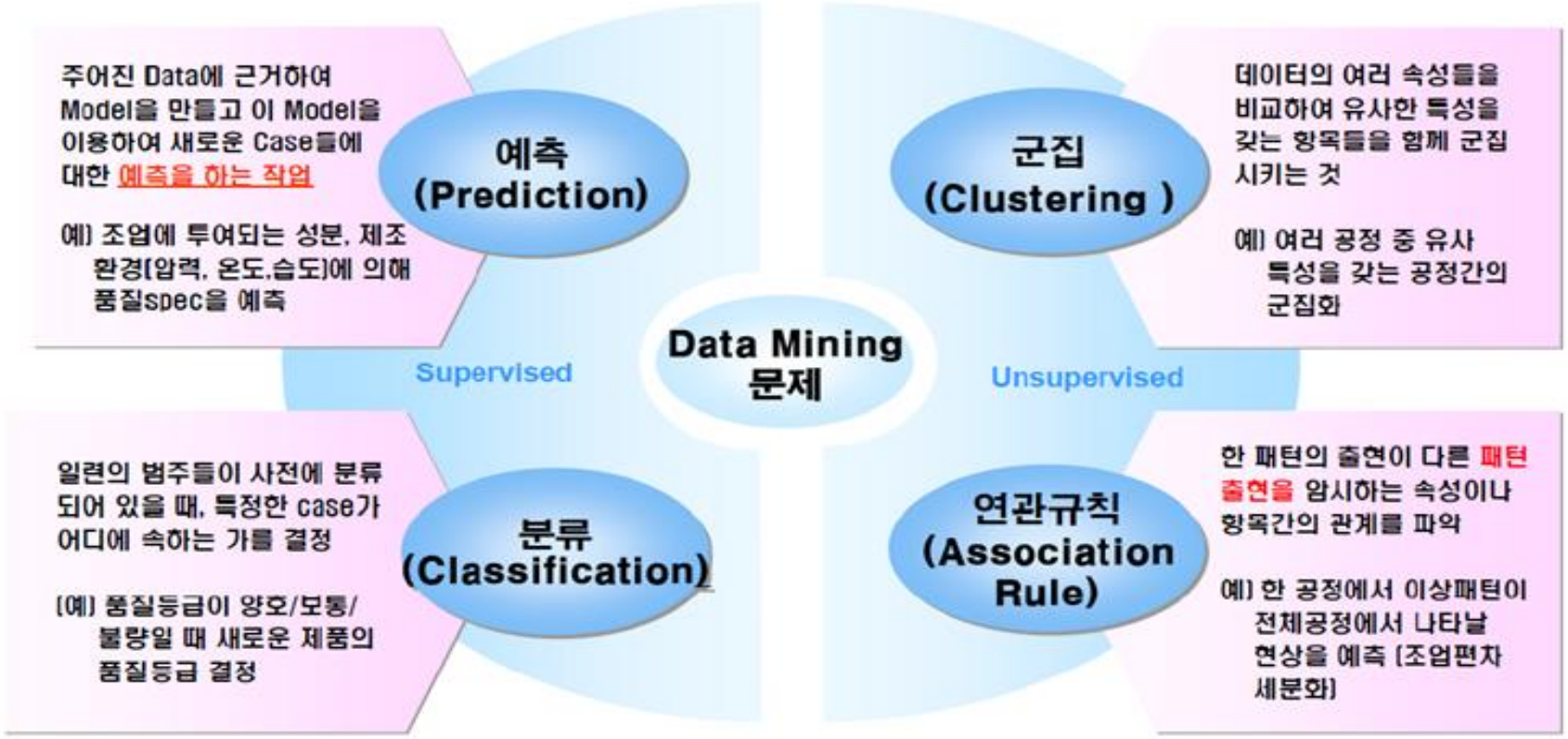
  SQL: 2019년 1월에 50만원 이상 구매 고객

  OLAP: 2019년 1월에 50만원 이상 구매,여자, 미혼, 년 소득이5천만원 이상

  Mining: 미혼남, 서울거주, 년소득 3천만원, 취미가 여행인 고객의
  신용불량여부 예측

1.통계분석 개요

(2) 데이터 마이닝 기법



1.통계분석 개요

Supervised learning(지도학습) **Top down**

- 인공신경망 (Artificial Neural Network)
- 의사결정나무(Decision Tree)
- 판별분석 (Discrimination Analysis)
- 일반화선형모형 (GLM, Generalized Linear Model)
  - -선형회귀분석 (Regression Analysis)
  - -로지스틱 회귀분석 (Logistic Regression)
- 사례기반추론 (Case-Based Reasoning)

→훈련용 데이터(training data)에 알고리즘을 적용하여 함수를 추론하고,
이제 그 추론된 함수를 통해 예측확률, 예측
따라서 지도학습은 명확한 입력변수와 목표변수가 존재한다.
이러한 지도학습에는 분류(Classification)과 예측(Regression)이 있다.

1.통계분석 개요

Unsupervised Prediction (비지도학습,자율학습)-**Bottom up**

- OLAP (On-Line Analytic Processing)
- 연관성규칙발견 (Association Rule Discovery, Market Basket)
- 군집분석 (k-Means Clustering)
- 인자분석(Factor Analysis), 주성분분석(Principal Component)
- k-Nearest Neighbor
- SOM (Self Organizing Map, Kohonen Network)

→ 목표변수(종속변수) 정해져 있지 않음

데이터가 어떻게 구성되었는지를 알아내는 그룹핑 알고리즘

1.통계분석 개요

(3) 데이터 마이닝 추진 5단계

① 목적 정의
데이터 마이닝을 통해 무엇을 왜 하는지 대한 명확한 목적 설정
이해관계자와 전문가 함께 목적에 따라 데이터 마이닝 모델과 필요한 데이터 정의
데이터 마이닝 기법 결정

② 데이터 준비:데이터 수집 단계, 데이터를 정제를 통해 품질 확보

③ 데이터 가공:모델링 목적에 따라 목적변수 정의, 데이터 마이닝 소프트웨어에 적합한 형식으로 가공

④ 수집된 데이터에 데이터 마이닝 기법 적용

⑤ 검증: 데이터 마이닝 추출한 정보를 검증하는 단계
test 자료와 모델링 차이 구분
→검증 완료 후 관련부서와 상시 협의하여 데이터 마이닝 결과 업무 적용

데이터분석 출제문제

1. 데이터마이닝 단계 중 목적 변수를 정의하고 필요한 데이터를 데이터 마이닝 소프트

   웨어에 적용할 수 있게 데이터를 준비하는 단계는?

   ① 데이터 가공

   ② 데이터 준비

   ③ 검증

   ④ 데이터 마이닝 기법의 적용

2.모형평가

(1) 모형평가

다양한 분석 모형 중에서 데이터마이닝의 목적 및 데이터의
특성에 따라 가장 적합한 모형을 선택하기 위해서는
모형 평가 기준 필요

모형평가의 기준

① 일반화 가능성 : 모집단 내의 다른 데이터 적용-> 안정적 결과

② 효율성 : 모형이 효율적 구축->적은 입력변수 필요할수록

③ 예측과 분류의 정확성: 실제 문제에 적용 -> 정확성

2.모형평가

(2) 데이터 분할

모형의 평가를 위해서 전체자료(Raw data)에서 모형 구축을 위한 훈련용 데이터(train data)와 모형의 성과검증을 위한 검증용 데이터(test data) 추출

→ 주어진 데이터에서만 높은 성과를 보이는 모형의 과적합화(overfitting) 방지

Train Data

분석 모델을 만들기 위한 학습용 데이터이다.

Validation Data

여러 분석 모델 중 어떤 모델이 적합한지 선택하기 위한 검증용 데이터이다.

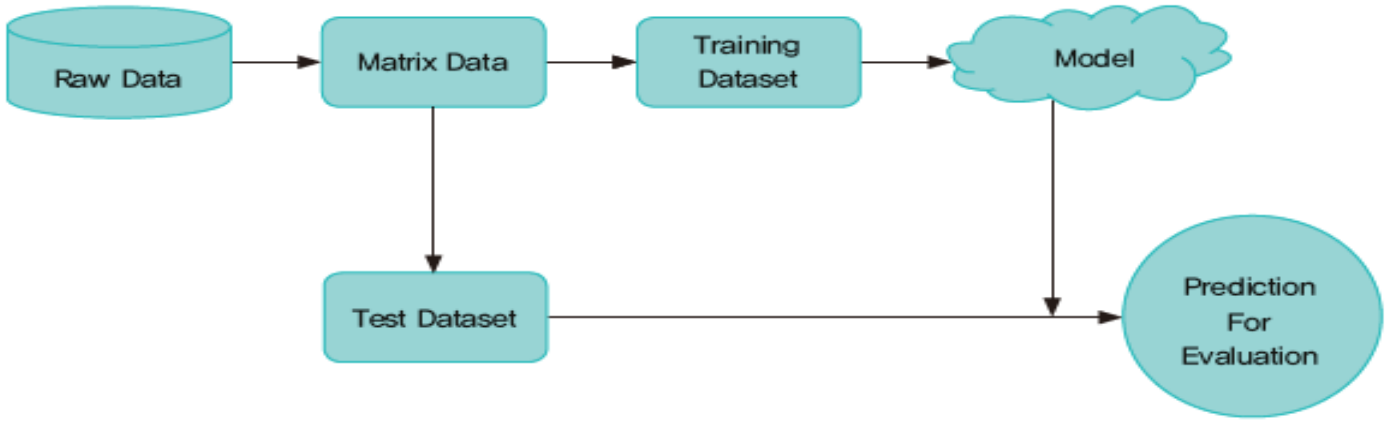
Test Data

최종적으로 선택된 분석 모델이 얼마나 잘 작동하는지 확인하기 위한 결과용 데이터이다

2.모형평가

1) 홀드아웃(hold-out)

Raw data를 두 분류 분리하고 교차 검정 실시
일반적으로 전체 데이터 중 70%의 데이터는 훈련용
30% 검증용 자료 사용. 검증용 자료는 모형의 성과측정



홀드아웃 방법

2.모형평가

2) 교차검증(Cross Validation)

교차검증은 주어진 데이터를 가지고 반복적으로 성과를 측정하여
그 결과를 평균한 것으로 분류분석 모형을 평가하는 방법이다

	A	B	C	D	E
Cross Validation Iteration 1	Test	Train	Train	Train	Train
Cross Validation Iteration 2	Train	Test	Train	Train	Train
Cross Validation Iteration 3	Train	Train	Test	Train	Train
Cross Validation Iteration 4	Train	Train	Train	Test	Train
Cross Validation Iteration 5	Train	Train	Train	Train	Test

→ 5번 반복 측정하고 각각의 반복측정 결과를 평균 낸 값을
최종 평가로 사용. 대표적 기법 k-fold 교차검증
일반적으로 10-fold 교차검증이 사용(편중과 편차가 적은 지표 생성)

2.모형평가

3) 붓스트랩(Boostrap)

붓스트랩은 평가를 반복한다는 측면에서 교차검증과 유사하나
훈련용 자료를 반복 재선정한다는 점에서 차이가 있다.
붓스트랩은 관측치를 한 번 이상 훈련용 자료로 사용
하는 복원추출법에 기반한다.
붓스트랩 기법은 비교적 작은 데이터 세트에서 적합

2.모형평가

(3) 분류모형 평가지표

1) 오분류표(Confusion Matrix)

분류모형이 특정 데이터 집합에 대해 수행한
정분류와 오분류의 요약정보

		예측치	
		True	False
실제값	True	TP	FN
	False	FP	TN

① TP:실제값과 예측값이 모두 True인 빈도
② TN:실제값과 예측치 모두 False인 빈도
③ FP: 실제값은 False이나 True로 예측한 빈도
④ FN: 실제값은 True이나 False로 예측한 빈도

2.모형평가

1.1) 오분류표를 활용한 평가지표

오분류표를 활용한 평가 지표 (★★★)

매트릭	계산식	의미
Precision	TP/(TP+FP)	Y로 예측된 것 중 실제로도 Y인 경우의 비율
Accuracy	TP+TN/(TP+FP+FN+TN)	전체 예측에서 옳은 예측의 비율
Recall(Sensitivity)	TP/(TP+FN)	실제로 Y인 것들 중 예측이 Y로 된 경우의 비율
Specificity	TN/(FP+TN)	실제로 N인 것들 중 예측이 N으로 된 경우의 비율
FP Rate	FP/(FP+TN)	Y가 아닌데 Y로 예측된 비율 =(1-Specificity)
F1	2*[Precision*Recall/ (Precision+Recall)]	Precision과 Recall의 조화평균. 시스템의 성능을 하나의 수치로 표현하기 위해 사용하는 점수로, 0~1 사이의 값을 가진다.
Kappa	Accuracy-P(e)/(1-P(e))	코헨의 카파는 두 평가자의 평가가 얼마나 일치하는지 평가하는 값으로 0~1 사이의 값을 가진다.

Error rate(에러 분류율)=FP+FN/TP+FP+FN+TN=1-accuracy

- 전체 관측치 중 실제값과 예측치가 다른 정도

2. 모형평가

코엔의 kappa의 P(e) 산정방법

Cohen's 카파 계수 공식

유방암 감별		판독의 1		전체
		양성	악성	
판독의 2	양성	50	10	60
	악성	10	30	40
전체		60	40	100

- 판독의1과 2 모두가 양성 판정을 내릴 확률 : 0.6 * 0.6 = 0.36
- 판독의1과 2 모두가 악성 판정을 내릴 확률 : 0.4 * 0.4 = 0.16
- P_c = 0.36 + 0.16 = 0.52 (두 평가자간 확률적으로 우연히 일치할 확률)

2.모형평가

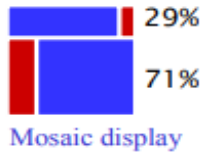
1.2_ 오분류 평가지표 예시

The Two-Class Problem

Actual Class \ Predicted Class	0	1	
0	True Neg	False Pos	Total Negative
1	False Neg	True Pos	Total Positive
	Total Negative	Total Positive	

…Two-Class Problem

Actual \ Predicted	0	1	
0	40	4	44
1	20	86	106
	60	90	150



Mosaic display

오류율 (Error rate)
= (false negative + false positive)/(grand total)=(20+4)/150=16%

정확도 (Accuracy)
= (true negative + true positive)/(grand total)=(40+86)/150=84%

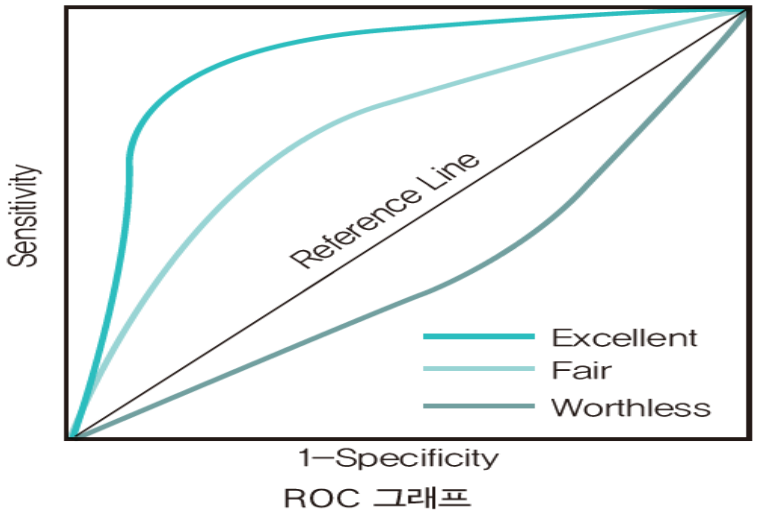
민감도 (Sensitivity)
= (true positive)/(total actual positive)=86/106=81%

특이도 (Specificity)
= (true negative)/(total actual negative)=40/44=91%

2.모형평가

2) ROC 그래프

ROC 그래프의 x축에는 FP Ratio(1-특이도),
y축에는 민감도를 나타내어 이 두 평면값의 관계로 모형을 평가
평가는 ROC 그래프의 밑부분 면적이 넓을수록 좋은 모형으로
평가



ROC 그래프

양성율(True Positive Rate; TPR)
"양성이라고 제대로 분류된 개수/전체양상 개수"
ex) 암환자를 진찰해서 암이라고 진단

위양성율(False Positive Rate; FPR) =1-특이도
"양성으로 잘못 분류된 개수/전체양상 개수"
ex) 암환자가 아닌데 암이라고 진단

TPR과 FPR은 Trade-off 관계
->완벽한 분류모형은 FPR이 '0' TPR이 '1'

2.모형평가

3) 이익도표(lift chart)와 향상도 곡선(lift curve)

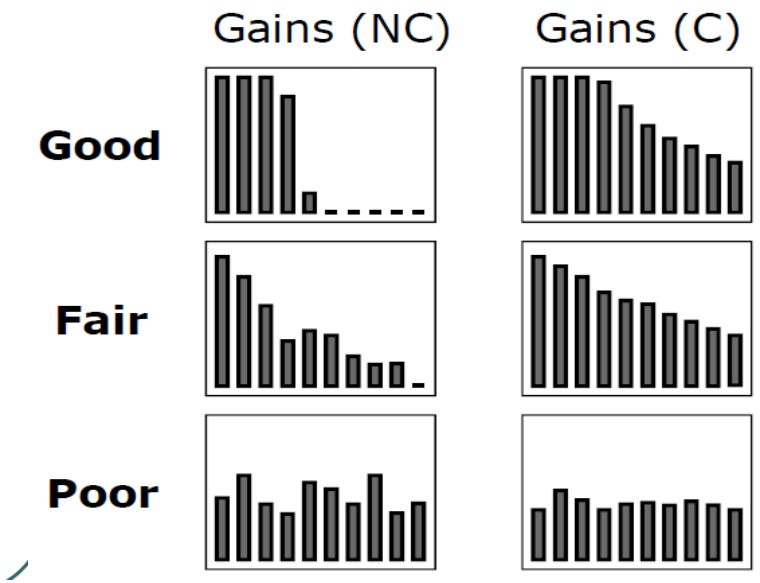
이익도표 (Lift chart=gain chart)
목표범주에 속하는 개체들이 각 등급 얼마나 분포값으로 계산된 이익값을 누적으로 연결 도표

향상도 테이블(Lift Table)

≪ Example ≫ n = 2000, 1 = 381

Baseline = 381/2000 = 19 %

Decile	Y=1	%Captured	%Response	Lift
1	174	174/381=45.6	174/200=87.0	87.0/19=4.57
2	110	110/381=28.8	110/200=55.0	55.0/19=2.89
3	38	38/381= 9.9	38/200=19.0	19.0/19=1.00
4	14	14/381= 3.6	14/200= 7.0	7.0/19=0.36
5	11	11/381= 2.8	11/200= 5.5	5.5/19=0.28
6	10	10/381= 2.6	10/200= 5.0	5.0/19=0.28
7	7	7/381= 1.8	7/200= 3.5	3.5/19=0.18
8	10	10/381= 2.6	10/200= 5.0	5.0/19=0.26
9	3	3/381= 0.7	3/200= 1.5	1.5/19=0.07
10	4	4/381= 1.0	4/200= 2.0	2.0/19=0.10



2.모형평가

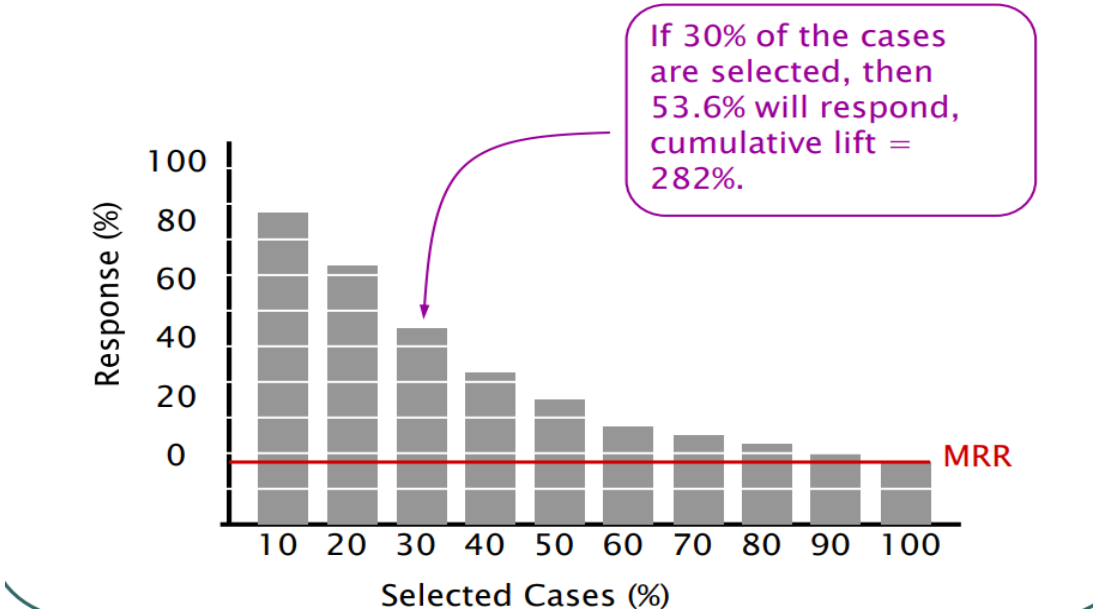
3) 이익도표(lift chart)와 향상도 곡선(lift curve)

향상도 곡선(lift curve)? 랜덤모델과 비교하여 해당모델의 성과가 얼마나 향상되었는지를 등급별로 파악

Lift:향상도

반응률 / 베이스라인 향상도

좋은 모델이라면 Lift가 빠른 속도로 감소해야 한다.



데이터분석 기출문제

1. 데이터 분할에 관한 설명 중 적절하지 않은 것은?

① 모델을 만들 때는 보통 데이터를 training set와 test set로 나누어 사용하며 학습에 사용한 training 데이터와 test 데이터가 비슷하다면 앞에서 만든 모델의 정확도는 높게 나올 것이다.

② 모델이 너무 간단하여 정확도가 낮은 모델을 과소적합(Under fitting)되었다고 말한다.

③ 과대적합이나 과소적합의 문제를 최소화하고 모델의 정확도를 높이는 가장 좋은 방법은 더 많고 다양한 데이터를 확보하고, 확보한 데이터로부터 더 다양한 특징 (feature)들을 찾아서 학습에 사용하는 것이다.

④ test set 결과가 일반적으로 training set 결과보다 좋다

데이터분석 기출문제

2. 과적합(overfitting) 발생 여부를 확인하기 위해서는 주어진 데이터에서 일정 부분을 모델을 만드는 훈련 데이터로 사용하고, 나머지 데이터를 사용해 모델을 평가한다. 이렇게 데이터를 훈련, 테스트 데이터로 분리하여 검증하는 방법을 무엇이라 하는가?

① 홀드아웃(Hold-Out)

② 신경망 모형

③ 향상도 곡선

④ 오분류표

데이터분석 기출문제

단답1) 보기의 표를 보고 재현율(recall)를 구하여라

Confusion Matrix		Predicted class	
		positive	Negative
Actual class	Positive	TP	FN
	Negative	FP	TN

데이터분석 기출문제

3. 오분류표 중 정확도와 재현율의 조화평균을 나타내며 정확도와 재현율에 같은 가중치를 부여하여 평균한 지표를 무엇이라 하는가?

① F1 ② Precision ③ Recall ④ Specificity

4. 두 평가자의 평가가 얼마나 일치하는지 평가하는 값으로 0~1 사이의 값을 가진다. P(e)는 두 평가자의 평가가 우연히 일치할 확률을 뜻하는 모델 평가 메트릭을 무엇이라 하는가?

데이터분석 기출문제

5.FP Ratio(1-특이도), 민감도를 나타내어 이 두 평면 값의 관계로 하는 모형 평가를 무엇이라 하는가?

데이터분석 기출문제

단답2) 보기의 표를 보고 정확도(Accuracy)를 구하여라

Confusion Matrix		Predicted class	
		1	0
Acual class	1	a	b
	0	c	d

데이터분석 기출문제

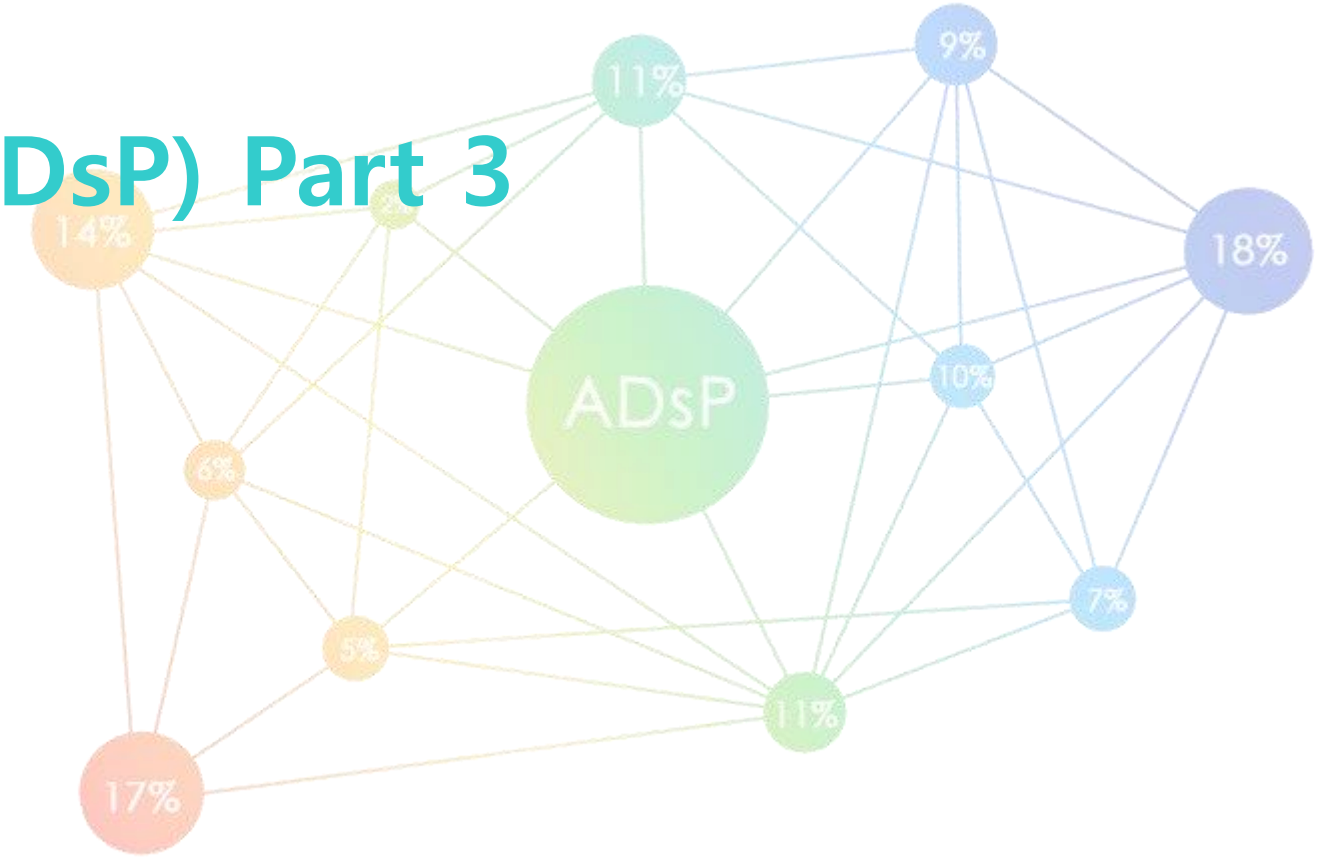
6. 오분류표를 활용한 평가지표 F1지표중 민감도(sensitivity)과 같은 지표는?

① recall ② specificity ③ precision ④ kappa

7. 랜덤모델과 비교하여 해당 모델의 성과가 얼마나 향상 되었는지를 각 등급별로 파악하는 그래프를 무엇이라 하는가?

데이터분석준전문가(ADsP) Part 3

데이터분석



03 제2절 분류분석

1. 로지스틱 회귀모형
2. 신경망모형
3. 의사결정나무
4. 앙상블 모형

1.로지스틱 회귀모형

(1) 로지스틱 회귀분석

종속변수가 “성공 또는 실패”,“흡연 또는 비흡연”,“ 생존 또는 사망”
binary data(이항변수)되어 있을 때 종속변수와 독립변수간의
관계식을 이용하여 두 집단 또는 그 이상의 집단을 분류하고자 할 때
사용되는 분석기법

단순 또는 다중선형회귀 Vs. 로지스틱

(공통점)독립변수의 선형결합으로 종속변수를 설명하는 관점은

선형회귀와 로지스틱 유사

(단순,다중회귀) 대부분 연속형인 독립변수와 연속형인 종속변수만 고려

(로지스틱) 종속변수가 어떤 집단에 속할 것인지를 분류,예측

1.로지스틱 회귀모형

	일반선형 회귀분석	로지스틱 회귀분석
종속변수	연속형 변수	이산형 변수
모형 탐색 방법	최소자승법	최대우도법, 가중최소자승법
모형 검정	F-test, t-test	χ^2 test

- 로지스틱 회귀분석의 활용 예시

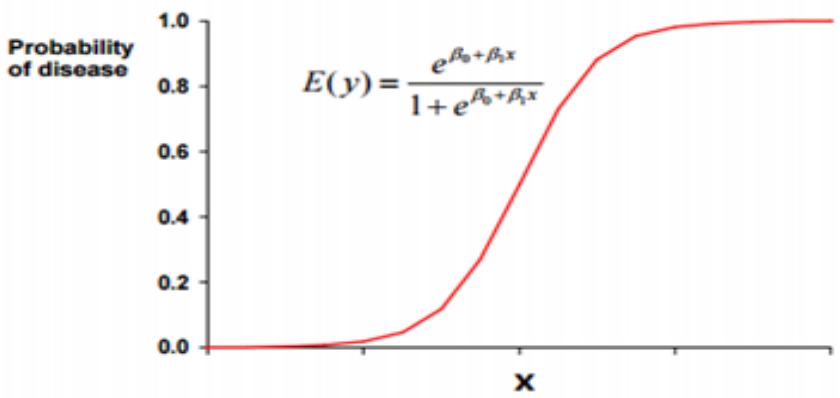
비즈니스 이해	변수	결과
TV 홈쇼핑 반품에 영향을 미치는 요인	독립변수:소득,학력, 성별,거주지,구매금액, 종속변수: 반품유무	20-30대의 젊은여성, 고학력, 고소득일수록 반품률이 높음

1.통계분석 개요

(2) 로지스틱 회귀모형

- 종속변수가 두 가지 범주를 나타내는 이항변수일 경우 기댓값은 확률 의미하므로 0~1사이의 값을 가지는 곡선형태의 모형

$$E(Y_i) = p_i = \frac{\exp(\beta_0 + \beta_1 x_i)}{1 + \exp(\beta_0 + \beta_1 x_i)}$$



로지스틱 모형은 $\beta_1 > 0$, 독립변수 x의 값이 작아질수록 예측값은 '0'에 가까워지고, x값이 증가함에 따라 예측값은 S자 형태의 모양으로 증가하면서 점점 '1'로 접근하는 모형
->이 곡선을 시그모이드(sigmoid)함수라고 한다.
이 함수는 경계값인 0 근처에서 기울기가 급속하게 커져서 두 개의 범주에 대한 구분을 쉽게

2.확률 및 확률분포

(3)로짓변환(로그오즈)

로지스틱 함수를 로짓변환(logit transformation)또는
로그오즈하면 일반적인 회귀모형 형태

$$\text{로짓변환: } ln\left(\frac{p_i}{1-p_i}\right) = \beta_0 + \beta_1 x_i$$

p_i =특정집단에 속할 확률

2.확률 및 확률분포

(4)로지스틱 회귀분석 과정

1단계 각 집단에 속하는 확률의 추정치를 예측.

2단계 추정확률 분류 기준값(cut-off) 적용 특정범주로 분류

예) $P(Y=1) \geq 0.5$ 집단 1로 분류

$P(Y=1) < 0.5$ 집단 0으로 분류

2.확률 및 확률분포

(5)회귀계수의 해석

오즈(odds)

성공=1, 실패=0 이항자료에서 성공률을 p라 할 때

오즈= $\frac{\text{성공률}}{\text{실패률}} = \frac{p}{1-p}$

① 오즈는 음이 아닌 실수값

② 성공이 일어날 가능성이 높은 경우 1.0 보다 크다

③ 실패가 일어날 가능성이 높은 경우 1.0 보다 작다

2.확률 및 확률분포

오즈 이해하기

ex) 성공률이 0.75 실패율이 0.25 오즈는 0.75/0.25=3.0
-> 성공할 확률이 실패할 확률의 3배란 의미

ex)반대로 오즈가 1/3 되면 실패할 확률이 성공할 확률의 3배의 의미

ex)월드컵에서 독일이 우승할 오즈가 0.18, 브라질이 우승할 오즈 0.25 오즈비는 0.25/0.18=1.39 브라질이 월드컵에서 우승할 가능성은 독일의 1.39배라고 해석

2.확률 및 확률분포

오즈 이해하기

로짓변환: $ln\left(\frac{p_i}{1-p_i}\right) = \beta_0 + \beta_1 x_i$

여기에 양변에 지수 취하면 오즈는 다음과 같다.

$\frac{p}{1-p}$=exp($\beta_0 + \beta_1 x_i$)=exp(β_0)exp($\beta_1 x_i$)

x를 1단위 증가시키게 되면 오즈의 예측값은 exp(β_1)만큼 증가

2.확률 및 확률분포

로지스틱 회귀계수 해석하기

해석 예시1)
종속변수 setosa=1,versicolor=2 범주형
Coefficients:

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-27.831	5.434	-5.122	3.02e-07 ***
sepal.length	5.140	1.007	5.107	3.28e-07 ***

->(해석) sepal.length가 1단위 증가할 때 Versicolor일
오즈가 exp(5.140)=170배 증가
-> 로직계수 부호도 확인

2.확률 및 확률분포

로지스틱 회귀계수 해석하기

해석 예시2)
종속변수 vs(0:flat engine,1:straight engine),
독립변수(mpg,am=변속기(0:automatic,1:manual)
Coefficients:

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-12.7051	4.6252	-2.747	0.00602 **
mpg	0.6809	0.2524	2.698	0.00697 **
am	-3.0073	1.5995	-1.880	0.06009 .

->(해석) am이 주어질 때 mpg값이 한 단위 증가함에 따라
vs가 1일 오즈가 exp(0.6899)=1.98배 증가한다=(98%) 증가
mpg가 주어질 때, 오즈에 대한 am의 효과는
exp(-3.0073)=0.05배 변속기가 수동인 경우
자동에 비해 vs=1 오즈가 0.05배=95% 감소한다.

2.확률 및 확률분포

로지스틱 회귀계수 해석하기

해석예시3)
종속변수 , 이직생각 있음=0,이직생각없음=1
독립변수(신체적건강, 심리적 건강,조직몰입도)

	B	EXP(B)
신체적 건강	.122	1.129
심리적 건강	-0.94	.910
조직 몰입도	.453	1.573

-> 신체적 건강 1단위 증가 이직생각없음이1.129배 증가
조직 몰입도 1단위 증가 이직생각없음이 1.573배 증가

2.확률 및 확률분포

(6)로지스틱 회귀모형 유의성 검정

Coefficients:

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-27.831	5.434	-5.122	3.02e-07 ***
sepal.length	5.140	1.007	5.107	3.28e-07 ***

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

Null deviance: 138.629 on 99 degrees of freedom

Residual deviance: 64.211 on 98 degrees of freedom

카이제곱분포 x^2자유도 98, 0.05의 값이 잔차 이탈도(64.211)보다

크므로 귀무가설 채택 결국 적합값이 관측된 자료를 적합하고 있다.

2.확률 및 확률분포

(7)glm()

glm(모형,family=분포모양,data=자료명,link=형태)

최소제곱법을 이용한 기존의 선형회귀분석과는 달리

최대 우도법추정법을 이용한 회귀계수의 추정을

위해서 R에서는 glm 함수 이용

종속변수가 이항변수이므로 family=binomial

최대우도추정법->관측값들이 가정된 모집단에서 하나의 표본으로

추출될 가능성이 가장 크게 되도록 하는 회귀계수 추정 방법

2.확률 및 확률분포

(8) 꼭 출제되는 로지스틱 기출

	Coefficient	Std. Error	Z-statistic	P-value
Intercept	-3.5041	0.0707	-49.55	< 0.0001
student[Yes]	0.4049	0.1150	3.52	0.0004

Positive

A student is risker than non students if ***no information*** about credit card balance is available

	Coefficient	Std. Error	Z-statistic	P-value
Intercept	-10.8690	0.4923	-22.08	< 0.0001
balance	0.0057	0.0002	24.74	< 0.0001
income	0.0030	0.0082	0.37	0.7115
student[Yes]	-0.6468	0.2362	-2.74	0.0062

Negative

Negative coefficient for student in multiple logistic regression indicates that *for a fixed value of balance and income*, a student is less likely to default than a non-student…

데이터분석 기출문제

18. 학생-연체율 데이터의 로지스틱 회귀분석 결과 화면을 보고 틀린 설명을 고르시오?

```
Call:
glm(formula = default ~ income + balance + studentD, family = "binomial",
    data = Default)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.469   -0.142   -0.056   -0.020    3.738

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -1.09e+01   4.92e-01  -22.08   <2e-16 ***
income       3.03e-06   8.20e-06    0.37   0.7115
balance      5.74e-03   2.32e-04   24.74   <2e-16 ***
studentD    -6.47e-01   2.36e-01   -2.74   0.0062 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 2920.6  on 9999  degrees of freedom
Residual deviance: 1571.5  on 9996  degrees of freedom
AIC: 1580

Number of Fisher Scoring iterations: 8
```

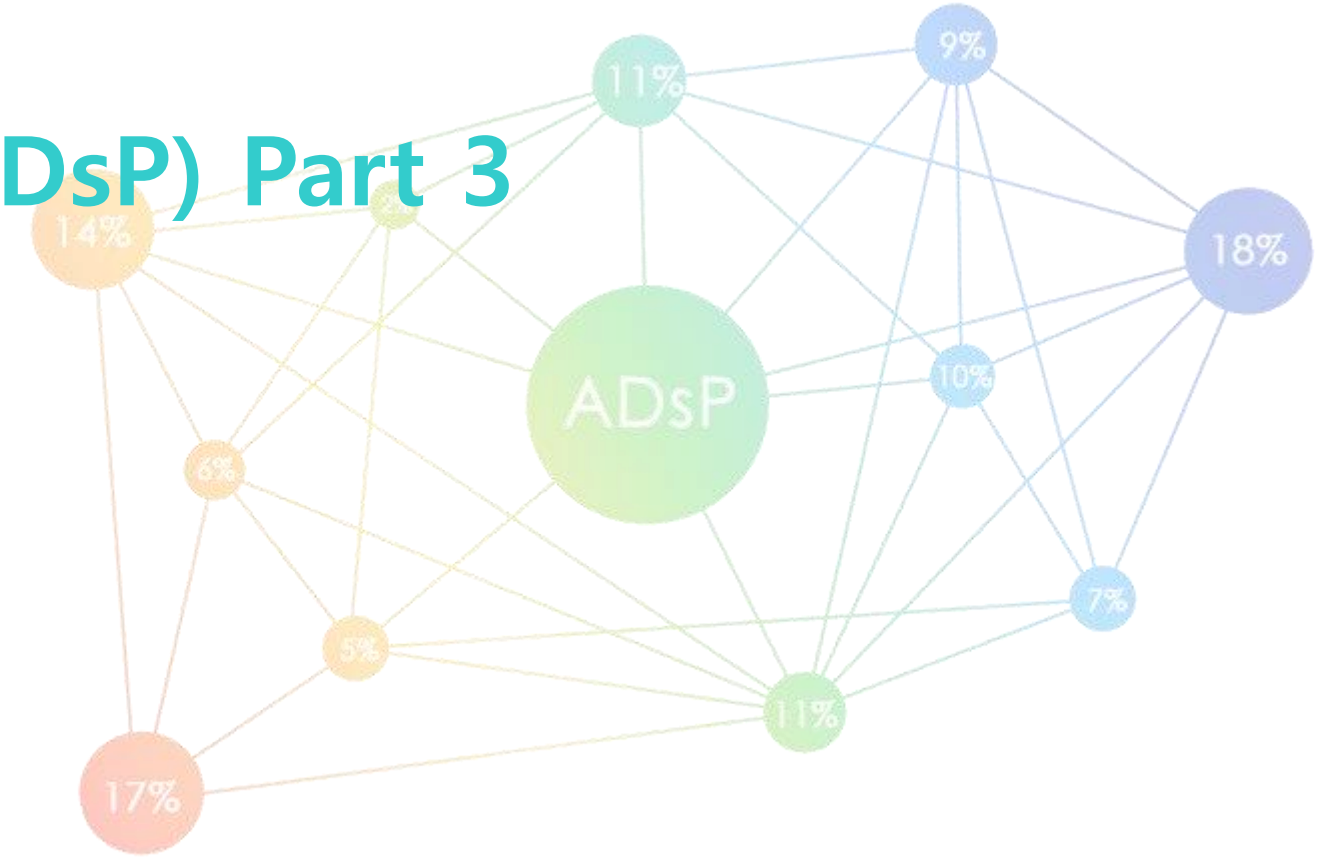
① balance는 default(연체 여부)에 통계적으로 유의미한 영향을 주는 변수이다.
② income은 default에 통계적으로 유의미한 영향을 주는 변수이다.
③ 학생인 사람이 학생 아닌 사람에 비해 연체가 아닐 가능성이 높다.
④ 로지스틱 회귀는 대표적인 지도학습으로 분류한다.

데이터분석 기출문제

8)로지스틱 회귀모형에서 exp(x1)의 의미는 나머지 변수가 주어질 때 x1이 한 단위 증가할 때마다 성공(Y=1)의 ( )가 몇 배 증가하는지를 나타낸다. ( )에 들어가는 내용은?

데이터분석준전문가(ADsP) Part 3

데이터분석



03 제2절 분류분석

1. 로지스틱 회귀모형
2. 신경망모형
3. 의사결정나무
4. 앙상블 모형

1.인공신경망

1.1 인공신경망(Artificial Neural Network) 정의 및 개요

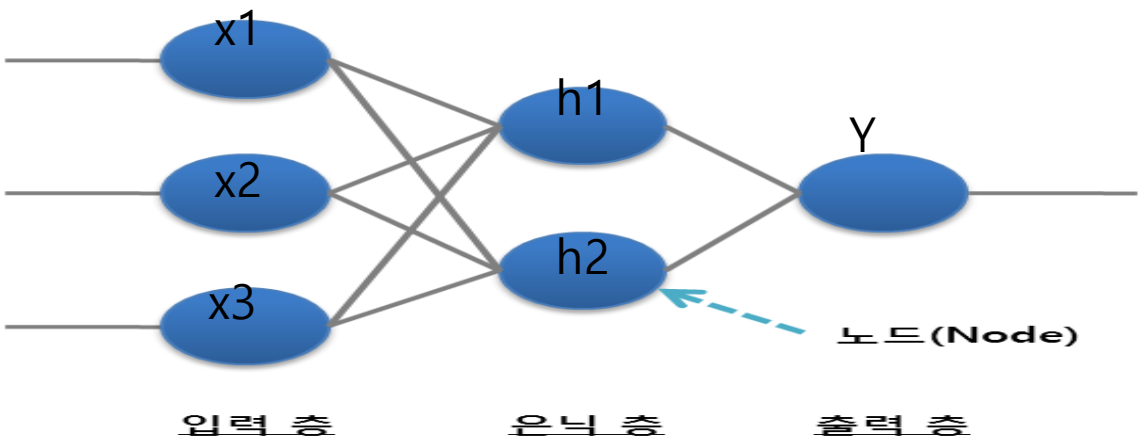
- 뇌의 뉴런들이 상호작용을 모형화한 프로세스 알고리즘
- 프로세스 입력변수는 출력변수를 얻기 위한 상호 연결된 가중치로 구성
- 계량적 분석 이외에도 문자 인식, 신호처리 등 다양한 분야에 사용됨
- 비선형적이고 잡음(noise)이 많은 영역에서도 적합한 모형을 구축할 수 있음

1.인공신경망

1.2 인경신경망 구성요소

- Processing unit : Node
  - 입력 신호를 측정
  - 충 입력신호를 가중(weight) 합산 (hidden node,output node)
  - 출력신호를 계산->변환->출력

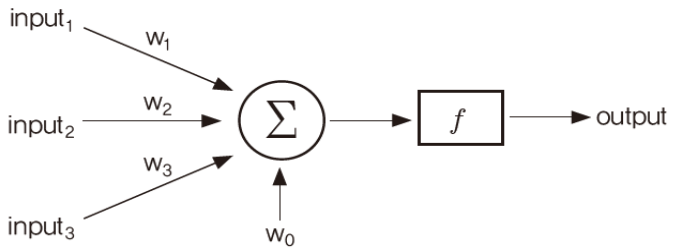
Input node Hidden node Output node



1.인공신경망

1.3 구성요소

- 연결강도(Weight)
  - 입력신호의 강도를 표현
- 총 입력값= 입력값의 가중합(합성함수)
- 활성함수(activation function)
  - 입력값에 함수를 적용하여 출력값으로 변환



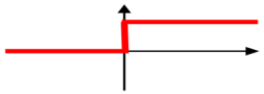
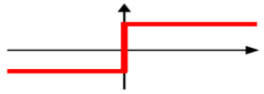
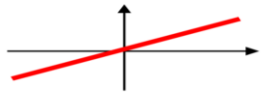
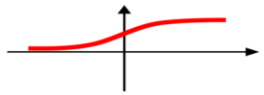
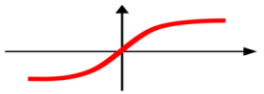
즉, input 1, 2, 3에 대해서 output은 다음과 같이 계산된다.

$$\text{Output} = f(w_0 + w_1 input_1 + w_2 input_2 + w_3 input_3)$$

1.인공신경망

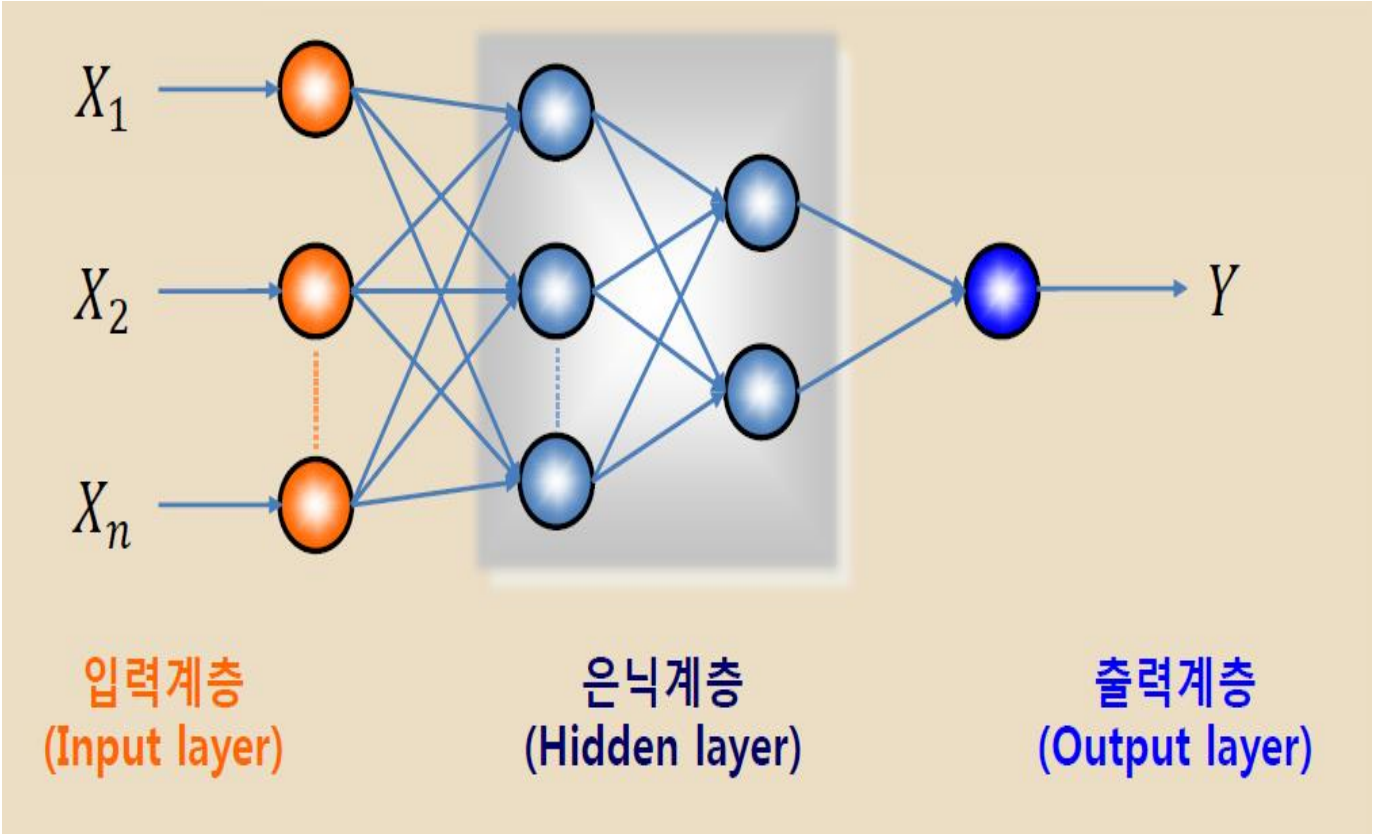
1.3 구성요소

- 활성함수

Activation function	Equation	Example	1D Graph
Unit step (Heaviside)	$\phi(z) = \begin{cases} 0, & z < 0, \\ 0.5, & z = 0, \\ 1, & z > 0, \end{cases}$	Perceptron variant	
Sign (Signum)	$\phi(z) = \begin{cases} -1, & z < 0, \\ 0, & z = 0, \\ 1, & z > 0, \end{cases}$	Perceptron variant	
Linear	$\phi(z) = z$	Adaline, linear regression	
Piece-wise linear	$\phi(z) = \begin{cases} 1, & z \geq \frac{1}{2}, \\ z + \frac{1}{2}, & -\frac{1}{2} < z < \frac{1}{2}, \\ 0, & z \leq -\frac{1}{2}, \end{cases}$	Support vector machine	
Logistic (sigmoid)	$\phi(z) = \frac{1}{1 + e^{-z}}$	Logistic regression, Multi-layer NN	
Hyperbolic tangent	$\phi(z) = \frac{e^{z} - e^{-z}}{e^{z} + e^{-z}}$	Multi-layer NN	

1.인공신경망

1.4 구성요소- Ann layer



1.인공신경망

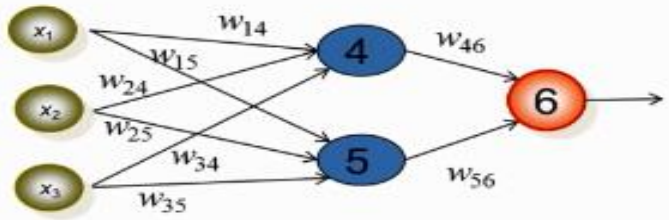
2.역전파 알고니즘

- 역전파 알고리즘(Back-propagation)을 위해서는 입력 데이터와 원하는 출력데이터(O)를 알고 있음
- 입력이 신경망의 가중치와 곱하고 더해지는 과정을 반복하여 출력값(y)이 나오고, 이는 원하는 출력값(O)와 다를 수 있음
- 이 때 오차(e=y-o)가 발생
- 오차에 비례하여 가중치 갱신 (오차제곱합,엔트로피)
- 가중치 갱신방향: 출력층 ➔ 은닉층->입력변수

1.인공신경망

1. Network Topology

※ 출력 = 1
경사도 = 1
learning rate η(eta) = 0.9
momentum = 0



2. 초기 입력값, 가중치 및 bias

x_1	x_2	x_3	w_{14}	w_{15}	w_{24}	w_{25}	w_{34}	w_{35}	w_{46}	w_{56}	θ_4	θ_5	θ_6
1	0	1	0.2	-0.3	0.4	0.1	-0.5	0.2	-0.3	-0.2	-0.4	0.2	0.1

3. 초기 입력값, 가중치 및 bias

Neuron j	*Net Input I_j*	*Output O_j*
4	0.2+0-0.5-0.4 = - 0.7	$1/(1+e^{0.7})$ = 0.332
5	-0.3+0+0.2+0.2 = 0.1	$1/(1+e^{-0.1})$ = 0.525
6	(-0.3)(0.332)-(0.2)(0.525)+0.1=-0.105	$1/(1+e^{0.105})$ = 0.474

학습의 종료조건
- 이전 시기의 모든 W_{ij} 보다 작을 때
- 미리 설정된 학습횟수 초과한 경우

1.인공신경망

3.Ann 유형

- Self Organizing Maps
  - 비지도학습

1.인공신경망

각 층의 노드 수 설정 고려 사항

1. 출력층 노드의 수는 출력 범주의 수로 결정
2. 입력의 수는 입력의 차원 수로 결정
3. 은닉층 노드의 수는

-은닉노드가 너무 적으면 복잡한 의사결정 경계를 만들 수 없다.

-은닉도드가 너무 많으면 일반화가 어렵다

1. 인공신경망

인공신경망 장점과 단점

장점

1. 변수의 수가 많거나, 입력과 출력 변수간에 복잡한 비선형 관계 존재할 때 유용
2. 잡음에 대해서도 민감하게 반응하지 않는다

단점

1. 은닉층의 수와 은닉노드 수의 결정이 어렵다
2. 초기값에 따라 전역해가 아닌 지역해로 수렴될 수 있음
3. 모형이 복잡하면 훈련과정에 시간이 많이 소요

1.통계분석 개요

배경지식 신경망은 가중치를 임의의 값으로 초기화한 후 반복적으로 가중치를 조절하면서 SSE 또는 엔트로피 기준을 최적화하는데, 이렇게 반복적으로 답을 찾아가는 이유는 단번에 최적의 가중치를 찾는 것이 어렵기 때문이다. 즉 정규화하지 않으면 오차가 최소인 전역해를 찾지 못하고 지역해에 빠질 위험이 있다.

지역해(local minimun)와 전역해(global minmum)

데이터분석 기출문제

1. 신경망 모형에 관한 설명 중 적절하지 않은 것은?

① 다층신경망은 단층신경망에 비해 훈련(training)이 어렵다.

② 은닉층 노드의 수가 너무 적으면 네트워크가 복잡한 의사결정 경계를 만들 수 없다.

③ 은닉층 노드의 수가 너무 많으면 일반화가 어렵다.

④ 은닉층의 수와 은닉노드 수의 결정은 자동으로 설정된다.

데이터분석 기출문제

2. 인공신경망의 특징으로 부적절한 것은?

① 분석가의 주관과 경험에 따른다.

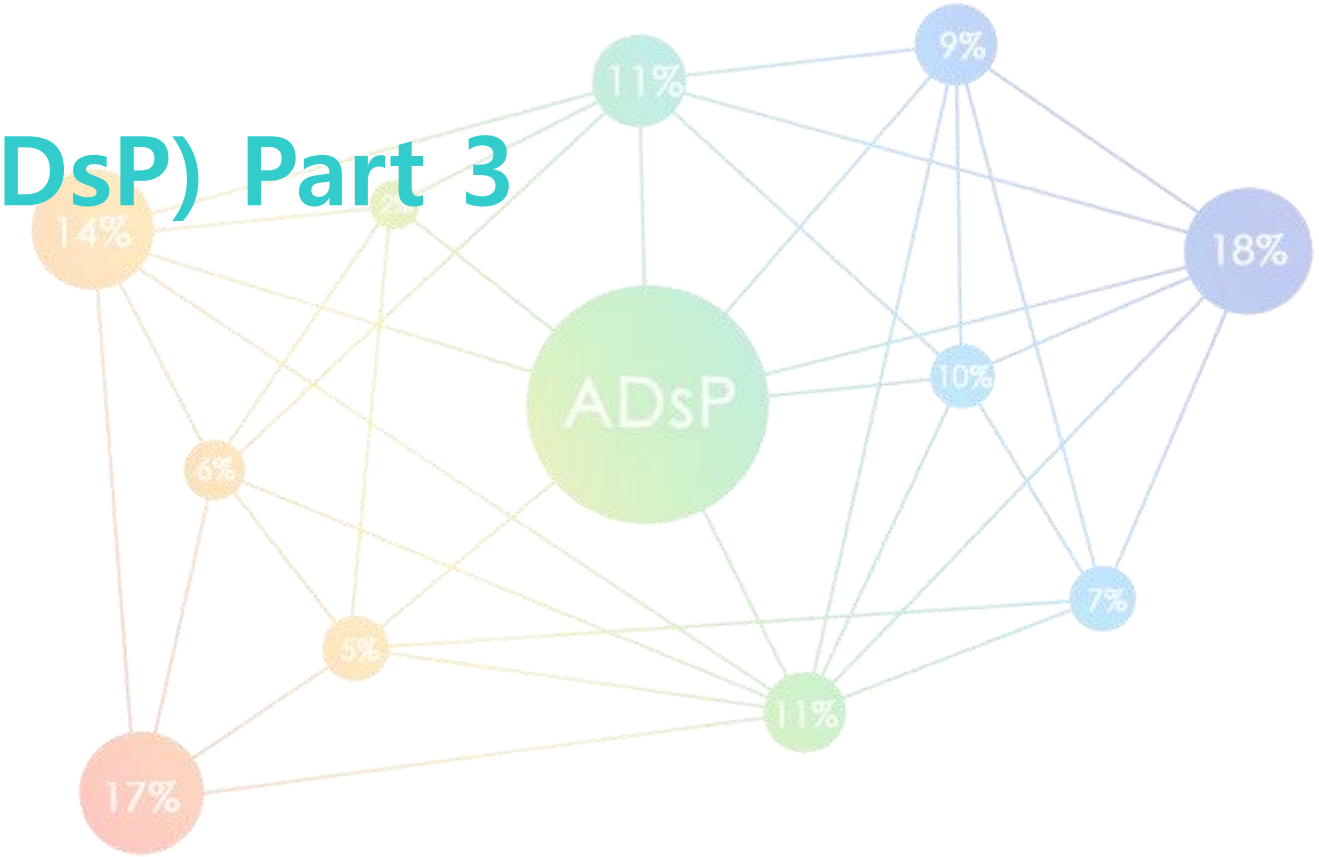
② 입력변수의 속성에 따라 활성화 함수의 선택이 달라진다.

③ 역전파 알고리즘 동일 입력층에 대해 원하는 값이 출력되도록 개개의 weight를 조정하는 방법으로 사용된다.

④ 이상치 잡음에 민감하지 않다

데이터분석준전문가(ADsP) Part 3

데이터분석



03 제2절 분류분석

1. 로지스틱 회귀모형
2. 신경망모형
3. 의사결정나무
4. 앙상블 모형

1.의사결정나무 개요

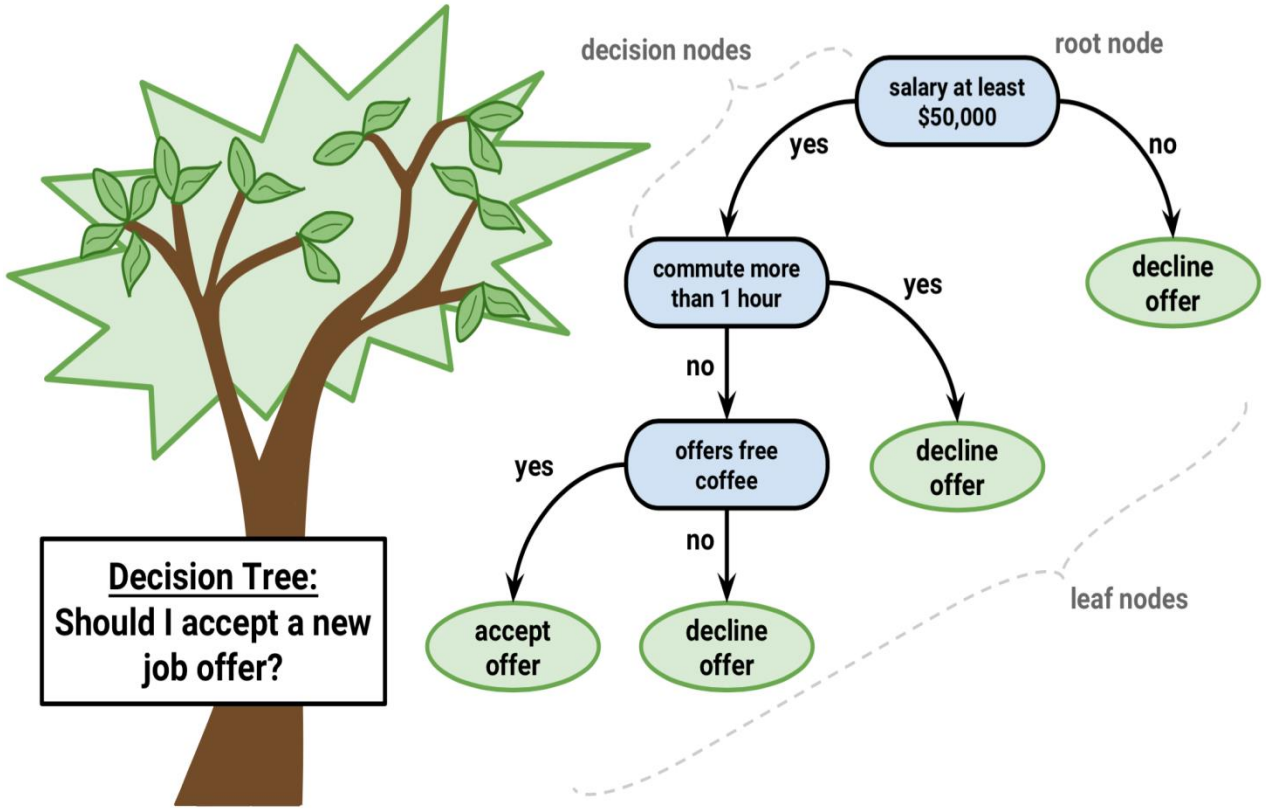
1.1. 정의

- 의사결정규칙(Decision Rule)을 나무구조(Tree)로 세분화하여 분류(Classification)와 예측(Prediction) 수행하는 분석방법
- 상위 노드로부터 하위 노드로 Tree 구조를 형성하는 단계마다 **분류변수와 분류기준값** 선택이 중요
- 나무 모형의 크기는 과대적합 되지 않도록 합리적 조절이 필요

1.의사결정나무 개요

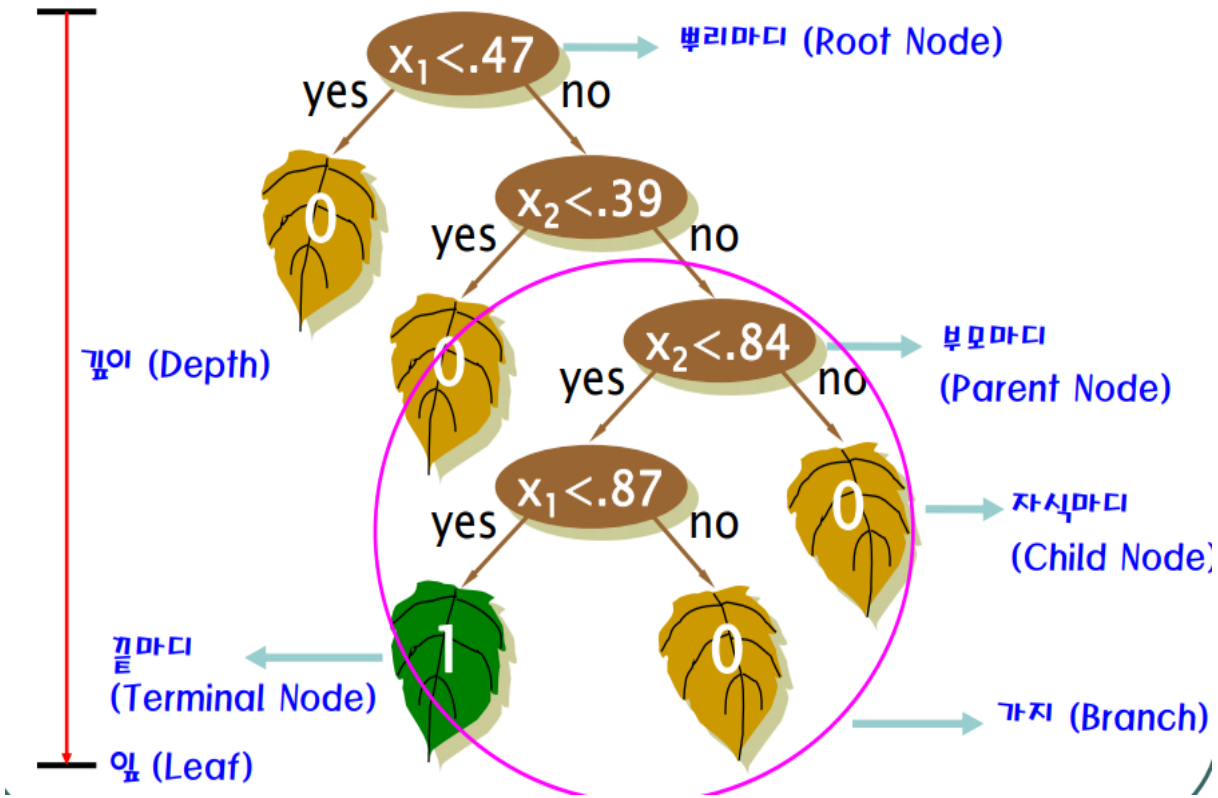
1.2. 구성

- 노드(node), 가지(branch), 깊이(Depth)



1.의사결정나무 개요

1.3. 구성



2.의사결정나무 원리

2.1. 의사결정나무 모형 구축

- Split(가지 분할) → 나무의 가지를 생성
- Stopping rule(정지규칙) → 더 이상 분리가 일어나지 않고 현재의 마디가 끝마디

  (기준) 최대나무의 깊이, 자식마디의 최소 관측치 수, 카이제곱 통계량 지니지수, 엔트로피 지수
- Pruning(가지치기)→생성된 가지를 잘라내어 단순화 끝마디가 너무 많으면(Overfitting)

  (기준) 분류된 관측치의 비율 또는 MSE

2.의사결정나무 원리

2.2. 의사결정나무 형성

의사결정나무 생성

분석의 목적과 자료구조에 따라서 적절한 분리기준(split criterion)과 정지규칙(stopping rule)을 지정하여 의사결정나무를 얻음

가지치기

부적절한 나뭇가지는 제거

타당성 평가

이익도표(gain chart)나 위험도표(risk chart) 또는 검증용 데이터(test data)에 의한 교차 타당성 등을 이용하여 의사결정나무 평가

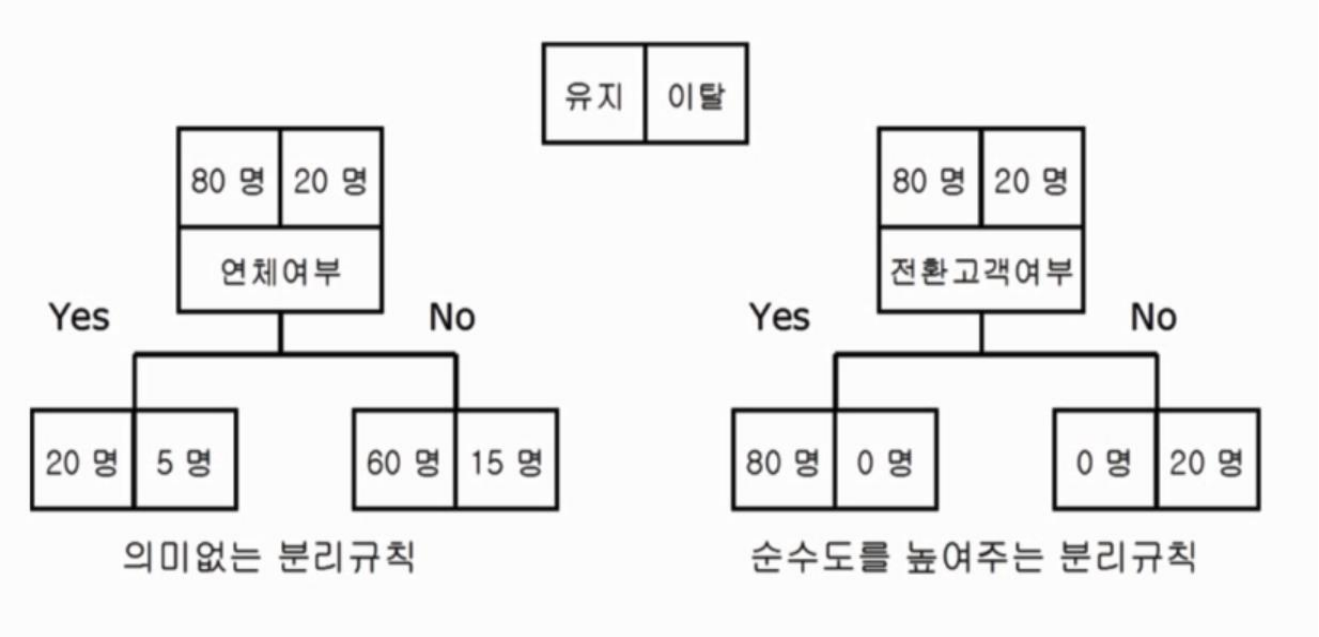
분류 및 예측

의사결정나무를 해석하고 예측모형 설정

2.의사결정나무 원리

1.1.의사결정나무 분리기준

- 분리기준 : 순수도 증가량 = 불순도 감소량=정보획득



2.의사결정나무 원리

1.1.의사결정나무 분리기준

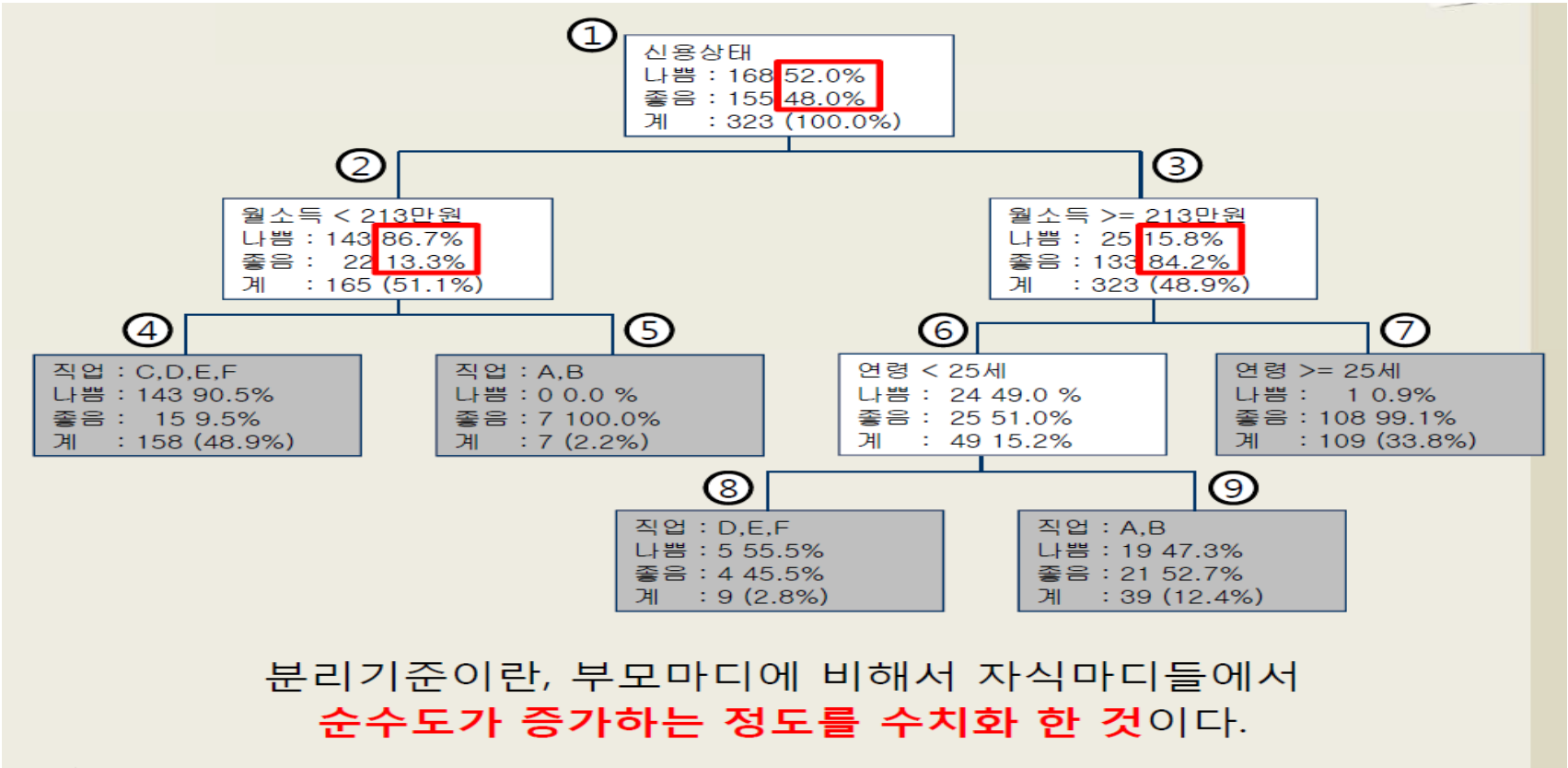
분리기준(split criterion): 어떤 입력변수를 이용하여 어떻게 분리하는 것이 목표변수의 분포를 가장 잘 구별해 주는지 그 기준

- 목표변수의 분포를 구별하는 정도 : **순수도** or **불순도**

* 순수도 : 목표변수의 특정 범주에 개체들이 포함되어 있는 정도

- 부모마디의 순수도에 비해서 자식마디들의 순수도가 증가하도록 자식마디를 형성함

2.의사결정나무 원리



분리기준이란, 부모마디에 비해서 자식마디들에서
순수도가 증가하는 정도를 수치화 한 것이다.

3.의사결정나무 분리기준

1.2. 분류기준

- 이산형 목표변수(분류나무 분류기준)-목표변수이 각 범주에 속하는 빈도에 기초하여 분리
  - 오차율 분할(잘못 분류된 관찰값의 수/전체 관찰값의 수0
  - 카이제곱 통계량 p값: P값이 가장 작은 예측변수와 그 때의 최적분리에 의해서 자식마디를 형성
  - 지니 지수: 지니 지수를 감소시켜주는 예측변수와 그 때의 최적분리에 의해서 자식마디를 선택
  - 엔트로피 지수: 엔트로피 지수가 가장 작은 예측 변수와 이 때의 최적분리에 의해 자식마디를 형성
- 연속형 목표변수(회귀나무 분류기준)- 목표변수의 평균과 표준편차에 기초하여 분리
  - 잔차제곱합(RSS) 개선되는 방향으로 분할(불필요한 가지도 학습, bias 낮고, variance 높음)
  - 분산분석에서 F 통계량: P값이 가장 작은 예측변수와 그 때의 최적분리에 의해서 자식마디를 형성
  - 분산의 감소량: 분산의 감소량을 최대화 하는 기준의 최적분리에 의해서 자식마디를 형성

3.의사결정나무 분류기준

불순도 측도-> 부모마디와 자식마디사이의 불소도감소량을 최대화하는 기준 선택

카이제곱 통계량의 p-값 (p-value of Chi-square statistics):

$$\sum_{i=1}^{k} \frac{(O_i - E_i)^2}{E_i}$$

주사위를 60번 던진 결과

4 3 3 1 2 3 4 6 5 6
2 4 1 3 3 5 3 4 3 4
3 3 4 5 4 5 6 4 5 1
6 4 4 2 3 3 2 4 4 5
6 3 6 2 4 6 4 6 3 2
5 4 6 3 3 3 5 3 1 4

주사위 눈	관측된 도수	기대도수
1	4	10
2	6	10
3	17	10
4	16	10
5	8	10
6	9	10
합	60	60

실제 관측된 도수와 기대도수의 차이가 크다.

CHAID: 예측변수는 반드시 범주형, 분리방법은 다지분리(multi split)

3.의사결정나무 분리

불순도 측도

지니지수 (Gini index) : **CART(범주형,연속형,이지분리), 분산의 감소량(분리기준)**

지니 지수를 감소시켜주는 예측변수와 그 때의 최적분리에 의해서 자식마디를 선택

자료세트 T가 k개의 범주로 분할 되고 범주 비율이 p_1 , •••, p_k라고 한다면, 다음과 같이 표기됨

$$Gini(T) = 1 - \sum_{j=1}^{k} p_j^2$$

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

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

3.의사결정나무 분리기준

불순도 측도

엔트로피지수 (Entropy index) : **C5.0,C4.5**

열역학에서 쓰는 개념으로 무질서도에 대한 측도

엔트로피 지수가 가장 작은 예측 변수와 이 때의 최적분리에 의해 자식마디를 형

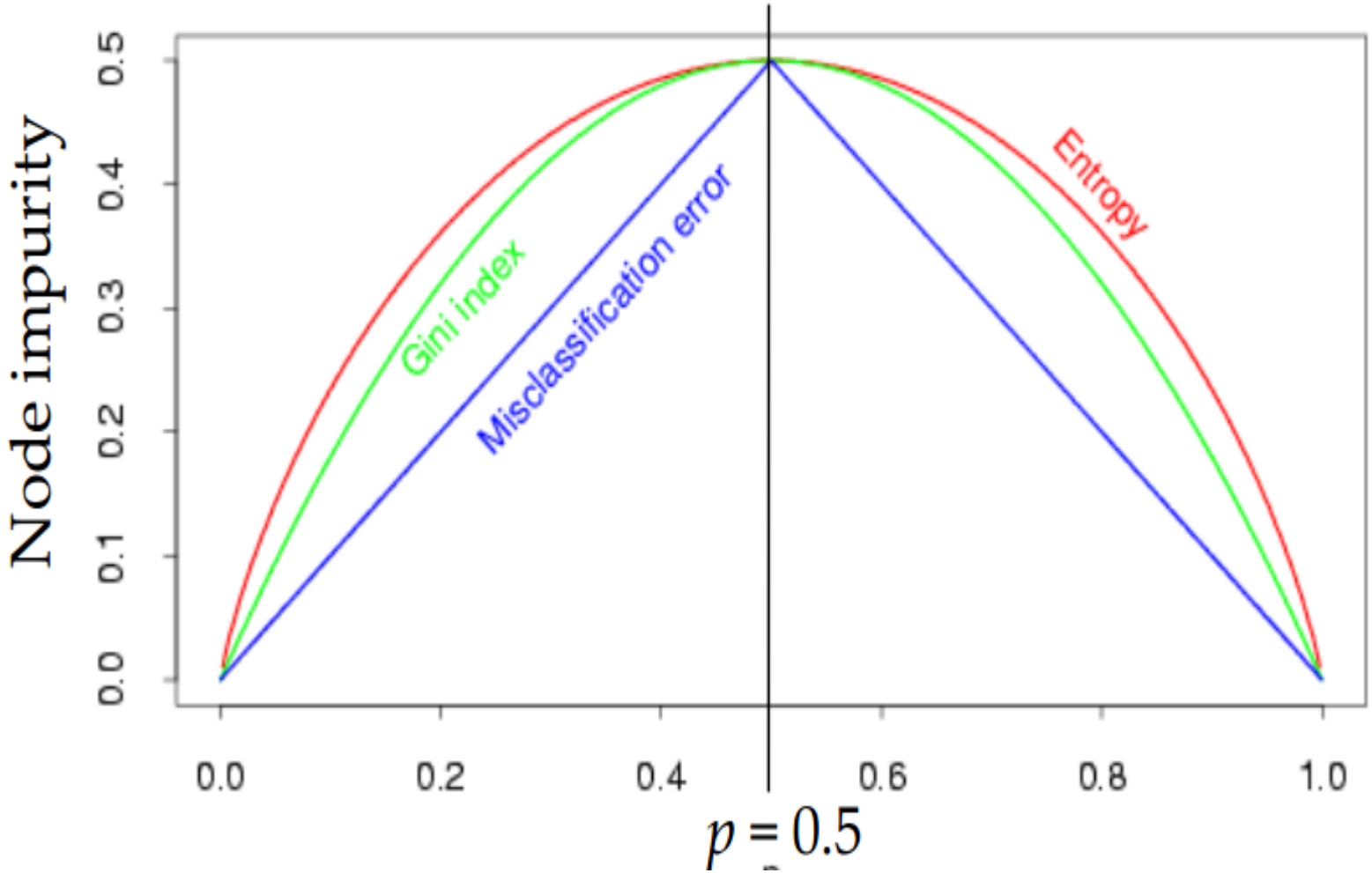
자료세트 T가 k개의 범주로 분할 되고 범주 비율이 p_1 , •••, p_k라고 한다면, 다음과 같이 표기됨

$$Entropy(T) = -\sum_{i=1}^{k} p_i \log_2 p_i$$

예) 4개의 범주가 (0.25, 0.25, 0.25, 0.25) 비율로 구성(T_0)

$$Entropy(T_0) = -(0.25 \log_2 0.25) * 4 = 1.39$$

3.의사결정나무 분리기준



4.의사결정나무 알고리즘

알고리즘 요약

	CART	C5.0	CHAID
목표변수	범주형,연속형	범주형	범주형,연속형
예측변수	범주형,연속형	범주형,연속형	범주형
분리기준	지니 지수 분산의 감소량	엔트로피 지수	카이제곱통계량 F-검정
분리갯수	이지분리	다지분리	다지분리

5.의사결정나무 장.단점

의사결정나무분석의 장점

해석의 용이성

나무구조에 의해서 모형이 표현되기 때문에 모형을 사용자가 쉽게 이해할 수 있다.
나무구조로부터 어떤 입력변수가 목표변수를 설명하기 위해서 더 중요한지를 쉽게 파악할 수 있다.

상호작용 효과의 해석

두 개 이상의 변수가 결합하여 목표변수에 어떻게 영향을 주는지를 쉽게 알 수 있다.
의사결정나무는 유용한 입력변수나 상호작용(interaction)의 효과 또는 비선형성 (nonlinearity)을 자동적으로 찾아내는 알고리즘이라고 할 수 있다.

비모수적 모형

의사결정나무는 선형성(linearity)이나 정규성(normality) 또는 등분산성(equal variance) 등의 가정을 필요로 하지 않는 비모수적인(nonparametric) 방법이다.
의사결정나무에서는 순서형 또는 연속형 변수는 단지 순위(rank)만 분석에 영향을 주기 때문에 이상치(outlier)에 민감하지 않다는 장점을 가지고 있다.

5.의사결정나무 장.단점

의사결정나무분석의 단점

비연속성

의사결정나무에서는 연속형 변수를 비연속적인 값으로 취급하기 때문에 분리의 경계점 근 방에서는 예측오류가 클 가능성이 있다.

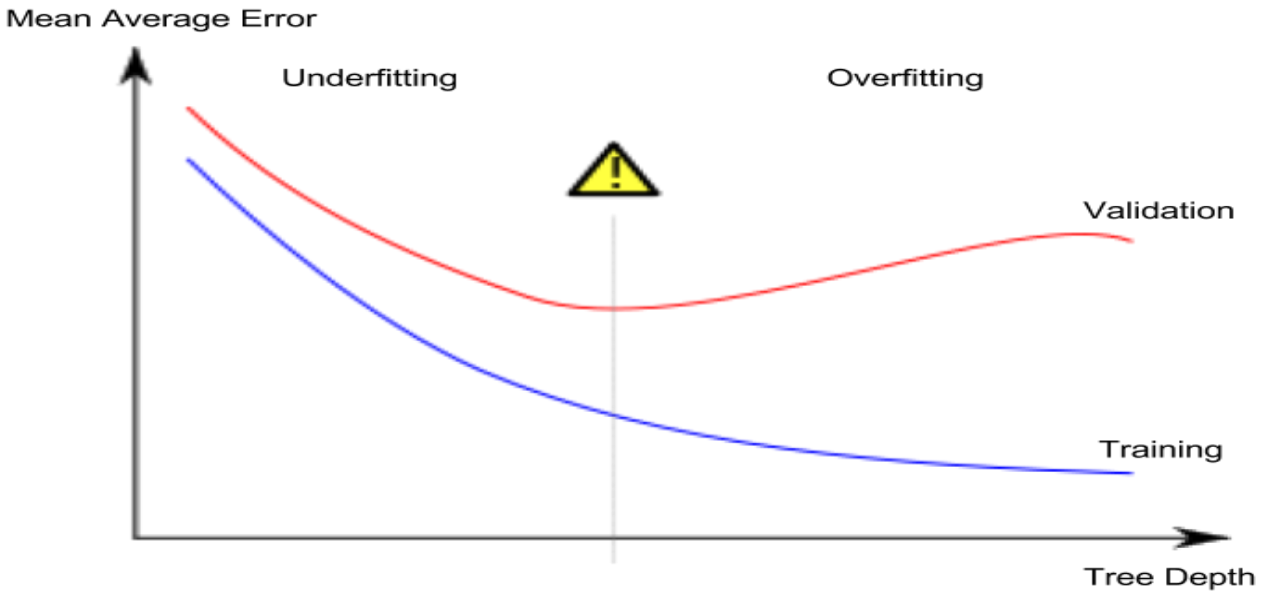
선형성 또는 주효과의 결여

회귀모형에서는 회귀계수나 오즈비(odds ratio)를 이용하여 결과에 대한 유용한 해석을 얻을 수 있다. 즉, 선형모형(linear model)에서 주효과(main effect)는 다른 예측변 수와 관련시키지 않고서도 각 변수의 영향력을 해석할 수 있다는 장점을 가지고 있는데 의사결정나무에서는 선형(linear) 또는 주효과(main effect) 모형에서와 같은 결과를 얻을 수 없다는 한계점이 있다.

비안정성

분석용 자료(training data)에만 의존하는 의사결정나무는 새로운 자료의 예측에서는 불안정(unstable)할 가능성이 높다. 따라서 검증용 자료(test data)에 의한 교차타당성(cross validation) 평가나 가지치기에 의해서 안정성 있는 의사결정나무를 얻는 것이 바람직하다.

5.의사결정나무 장.단점



학습용 데이터에 기초한 완전 성장한(full-grown) 나무는 데이터를 과적합

-> 정지규칙과 가기치기

5. 지도학습 장.단점 비교

지도학습의 장단점 비교

	장점	단점
로지스틱회귀 모형	생성모형에 대한 해석이 쉽다	모형이 가정이 있음 (선형성,분산성,정규성)
의사결정나무 모형	생성된 모형이 단순하고 해석이 가장 쉽다 설명력이 우수	결과의 불안정성
인공신경망	복잡한 상황에 유연하게 대처(유연성)	생성된 모형에 대한 해석이 어렵다 설명력이 약함

6.의사결정나무 기출문제

1. 다음 중 의사결정나무(Decision Tree)에 대한 설명 중 틀린 것은?

① 정지규칙이란 더 이상 분리가 일어나지 않고 현재의 마디가 최종마디가 되도록 하는 여러 가지 규칙으로 카이제곱통계량, 지니 지수, 엔트로피 지수 등이 있다.

② 가지치기란 최종마디가 너무 많으면 모형이 과대 적합된 상태로 현실 문제에 적용할 수 있는 적절한 규칙이 나오지 않게 된다.

③ 의사결정나무를 위한 알고리즘은 CHAID, CART, ID3, C4.5가 있으며 상향식 접근 방법을 이용한다.

④ 의사결정나무는 목표변수가 이산형인 경우의 분류나무(classification tree)와 목표변수가 연속형인 경우의 회귀나무(regression tree)로 구분된다

6. 의사결정나무 기출문제

2. 의사결정나무 결과이다. 해석으로 부적절한 것은?
① 끝 노드로 갈수로 불순도가 상승한다.
② 구조가 단순하여 해석이 용이하다.
③ 수치형 또는 범주형 변수를 모두 사용할 수 있다.
④ 선형성, 정규성, 등분산성 등의 수학적 가정이 불필요한 비모수적 모형이다.

3. Bias-Variance trade off 관계에서 유연한 경우 옳은 것은?
① Bias 높고 Variance 높다
② Bias 높고 Variance 낮다
③ Bias 낮고 Variance 높다
④ Bias 낮고 Variance 낮다

R 실습

R사용하는 사용하는 의사결정나무 패키지

tree 패키지 tree()->binary recursive partitioning,엔트로피 지수

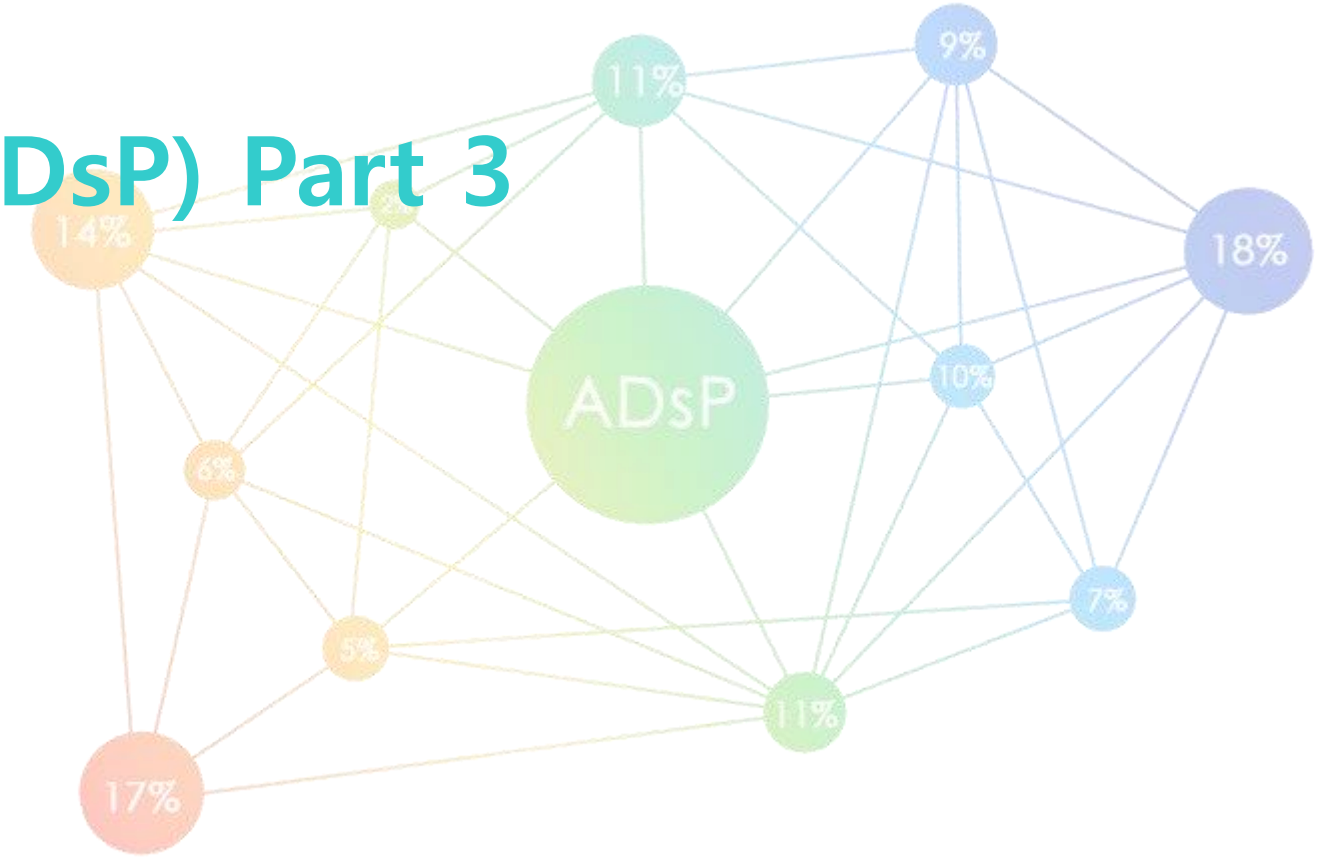
rpart 패키지 rpart()->CART, 지니지수

party 패키지 ctree()->p-test를 거친 significance를 기준

step1) training/test -> step2) decision tree model->step3) 가지치기-step4) 평가

데이터분석준전문가(ADsP) Part 3

데이터분석



03 제2절 분류분석

1. 로지스틱 회귀모형
2. 신경망모형
3. 의사결정나무
4. 앙상블 모형

1.앙상블 모형 개요

1.1. 정의

❖ 앙상블(ensemble) 모형은 여러 개의 분류 모형에

의한 결과를 종합하여 분류의 정확도를 높이는 방법

❖ 적절한 표본추출법으로 데이터에서 여러 개의 훈련용 데이터 집합을

만들어 각각의 데이터 집합에서 하나의 분류기를 만들어 앙상블 하는 방법이다.

❖ 데이터가 충분히 큰 경우, 각 데이터가

하나의 붓스트랩 표본에서 제외될 확률은 36.78%이다 -> test data로 활용

1.앙상블 모형 개요

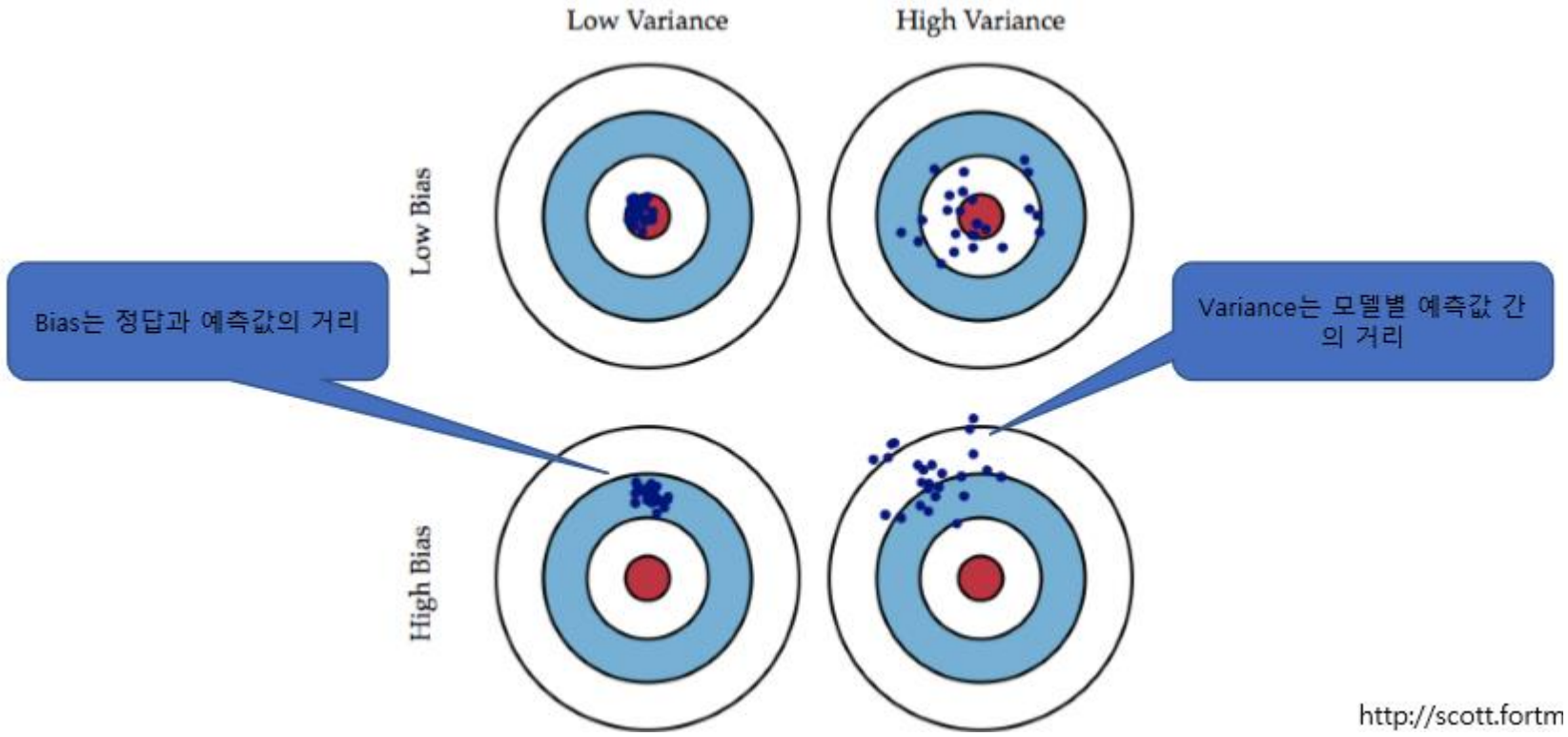
1.2. 왜 앙상블 모형

- 평균을 취함으로써 편의를 제거해준다:
- 분산을 감소시킨다: 한 개 모형으로부터의

  단일 의견보다 여러 모형의 의견을 결합하면 변동이 작아진다.
- 과적합의 가능성을 감소-> 일반화가 잘된다.

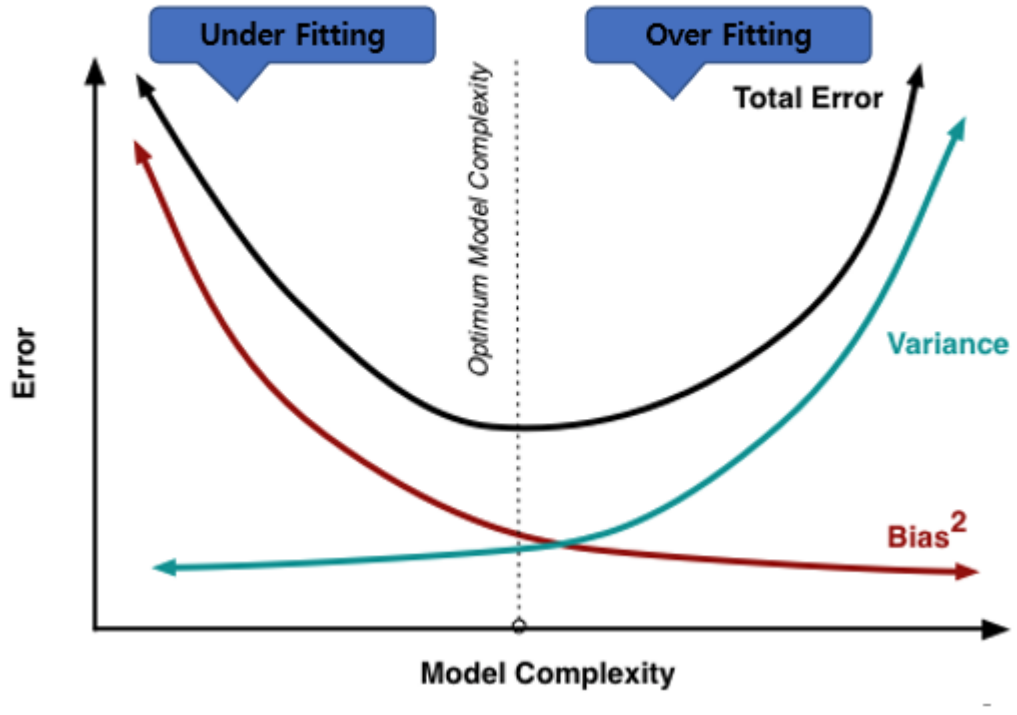
1.앙상블 모형 개요

학습한 모델의 예측 오류는 Bias, Variance 구성



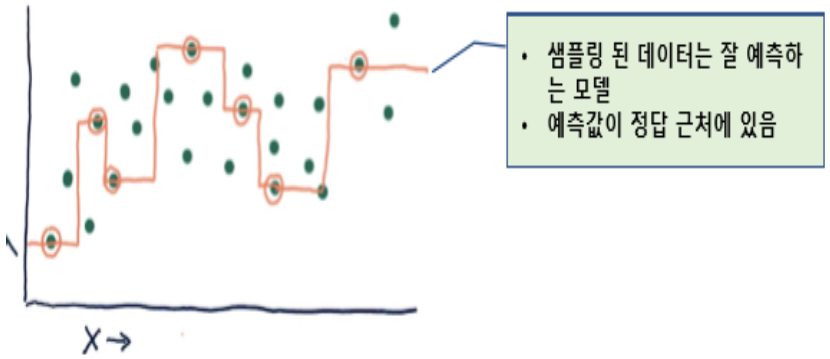
1.앙상블 모형 개요

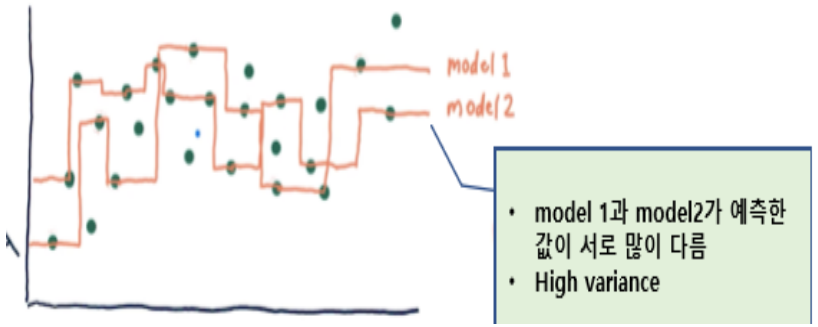
Bias, Variance의 관계는 Trade-off



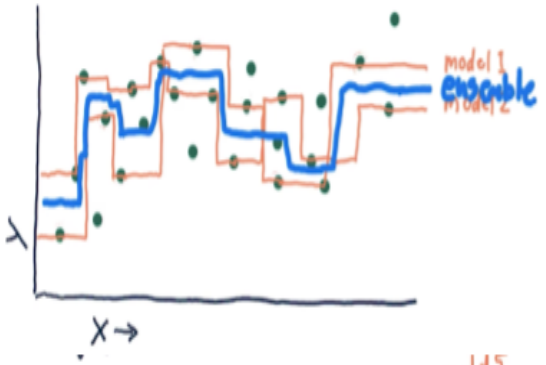
1.앙상블 모형 개요

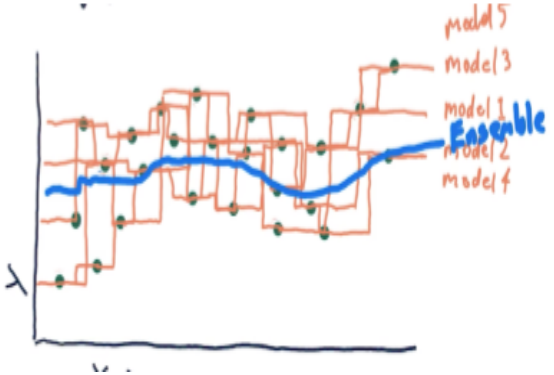
- 하나의 모델만 본다면...
  - Low Bias : model1과 2가 정답과 가까운 거리에 위치
- 여러 개의 모델을 함께 보면...
  - High variance : 각 모델별로 예측한 값의 차이가 크다





- 모든 모델이 예측한 값의 평균을 사용하자
- 아래의 예시를 보면,
- 모델이 많을 수록 평균 값을 가진 모델이 실제 데이터와 유사





2.배깅(bagging)

2.1 배깅(bagging)

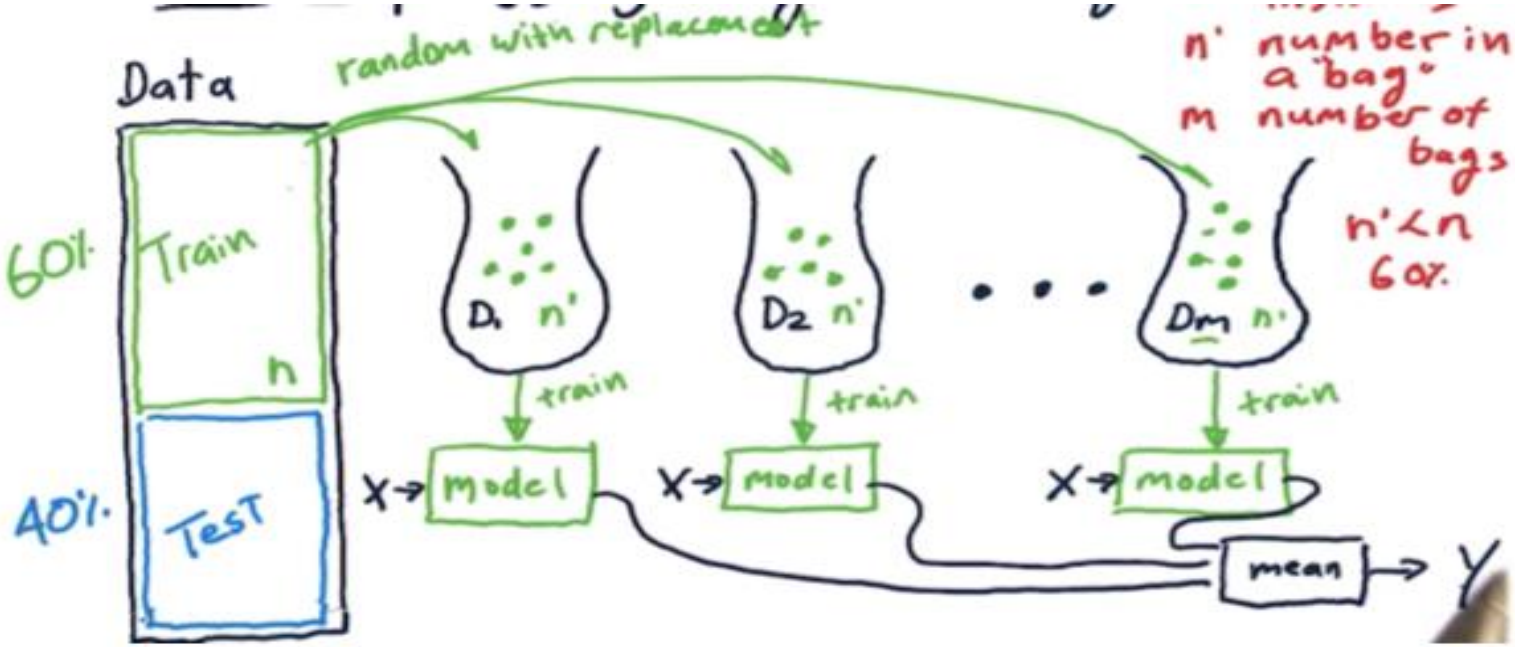
배깅(bagging)은 bootstrap aggregating의 준말로 원 데이터 집합으로부터
크기가 같은 표본을 여러 번 단순임의 복원 추출하여
각 표본(이를 붓스트랩 표본이라 함)에 대해 분류기
(classifiers)를 생성한 후 그 결과를 앙상블 하는 방법

반복추출 방법을 사용하기 때문에 같은 데이터가 한 표본에
여러 번 추출될 수도 있고, 어떤 데이터는 추출되지 않을 수도 있다.

데이터가 충분히 큰 경우, 각 데이터가 하나의 붓스트랩
표본에서 제외될 확률은 36.78%이다

2.배깅(bagging)

2.1 배깅(bagging)



2.부스팅(boosting)

2.2 부스팅(boosting)

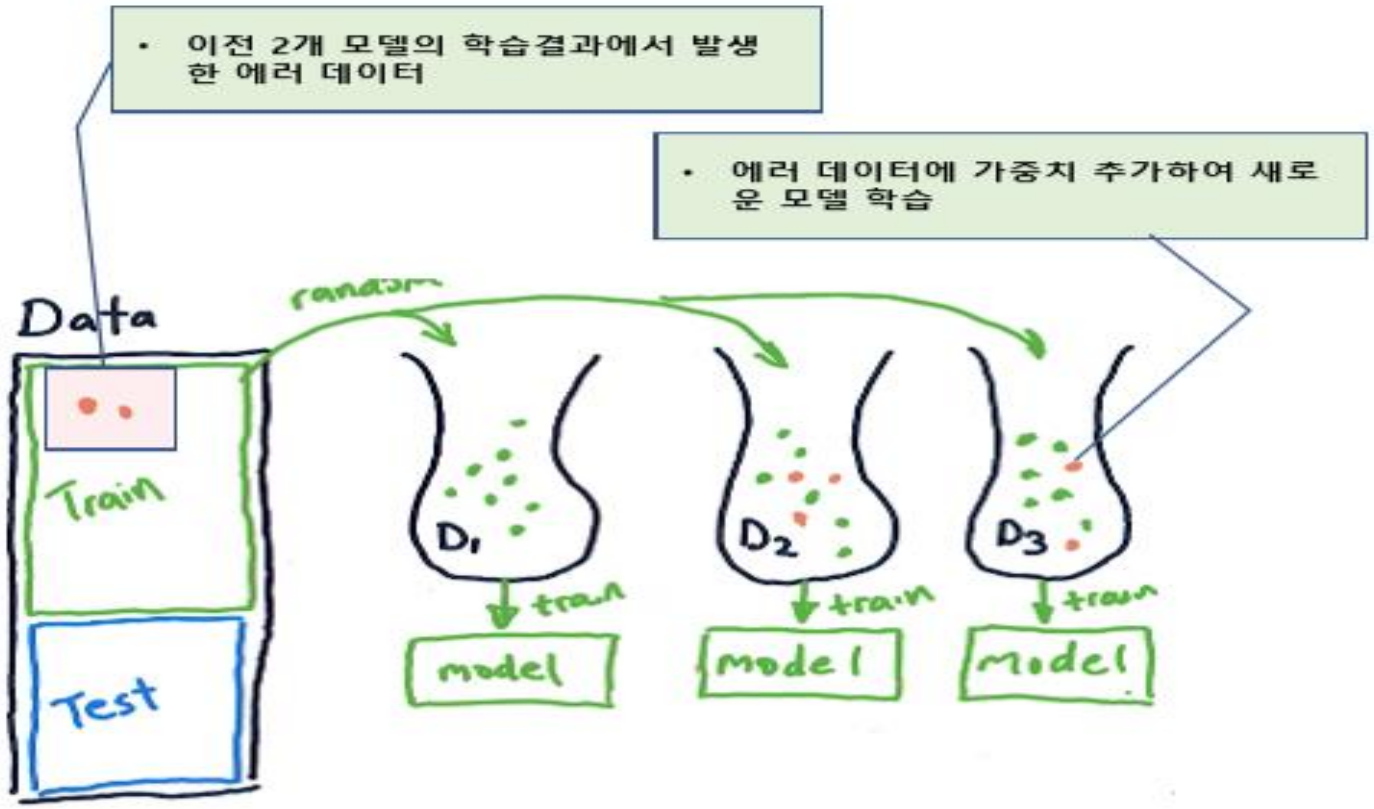
부스팅(boosting)은 배깅의 과정과 유사하나 붓스트랩 표본을 구성하는 재표본(re-sampling)과정에서 각 자료에 동일한 확률을 부여하는 것이 아니라, 분류가 잘못된 데이터에 더 큰 가중을 주어 표본을 추출

부스팅에서는 붓스트랩 표본을 추출하여 분류기를 만든 후, 그 분류결과를 이용하여 각 데이터가 추출될 확률을 조정한 후, 다음 붓스트랩 표본을 추출하는 과정을 반복

아다부스팅(adaBoosting: adaptive boosting)은 가장 많이 사용되는 부스팅 알고리즘이다.

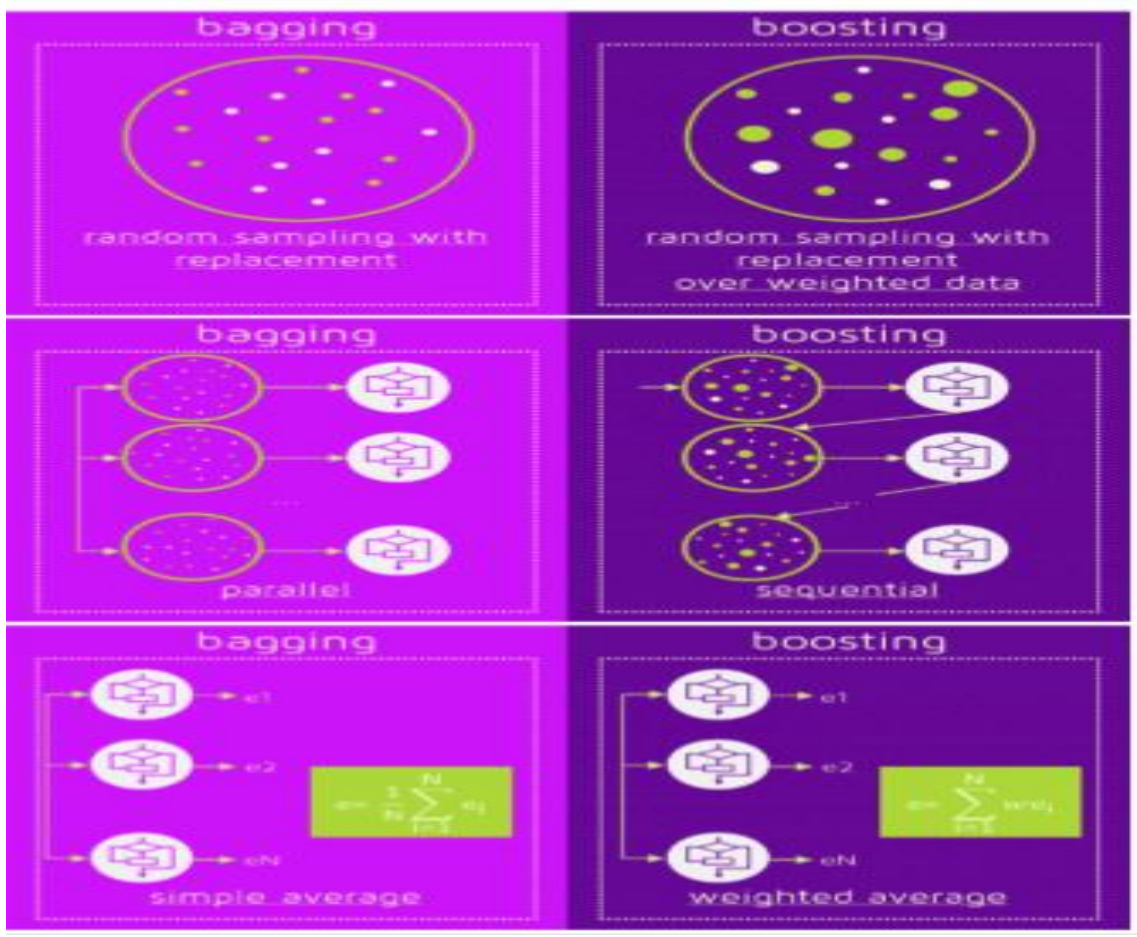
2.부스팅(boosting)

2.2 부스팅(boosting)



2.랜덤포리스트(random foreast)

2.2.1 배깅과 부스팅



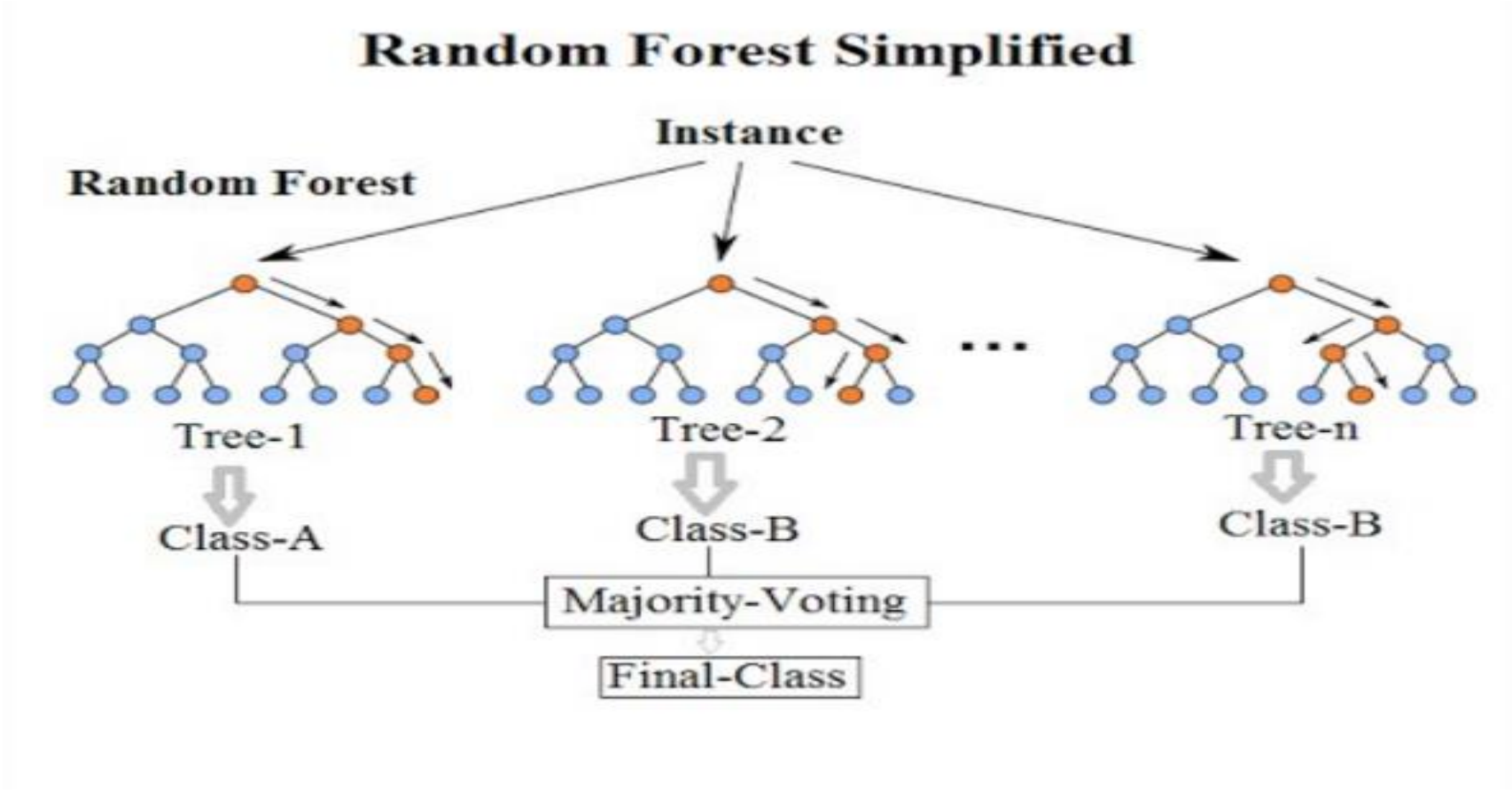
3.랜덤포리스트(random foreast)

2.3 랜덤포리스트(random foreast)

랜덤포리스트(random forest)는 배깅에 랜덤 과정을 추가한 방법이다.
원 자료로부터 붓스트랩 샘플을 추출하고, 각 붓스트랩 샘플에 대해 트리를 형성해
나가는 과정은 배깅과 유사하나, 각 노드마다
모든 예측변수 안에서 최적의 분할(split)을 선택하는 방법 대신
예측변수들을 임의로 추출하고, 추출된 변수 내에서
최적의 분할을 만들어 나가는 방법을 사용
새로운 자료에 대한 예측은 분류(classification)의 경우는 다수결(majority votes)로,
회귀(regression)의 경우에는 평균을 취하는 방법을 사용하며,
이는 다른 앙상블 모형과 동일하다.

3.랜덤포리스트(random foreast)

2.3 랜덤포리스트(random foreast)



4. 기출문제

1. 붓스트랩 방식을 이용하였을 때 일반적인 훈련 데이터의 양은?

① 63, 2%

② 10, 2%

③ 23, 8%

④ 36, 8%

2. 다음 중 앙상블 모형이 아닌 것은?

① 시그모이드(sigmoid)
② 배깅(bagging)
③ 랜덤 포리스트(random forest)
④ 부스팅

4. 기출문제

단답1) 재표본 과정에서 각 자료에 동일한 확률을 부여하지 않고, 분류가 잘못된 데이터에 가중을 주어 표본을 추출하는 분석 기법은?

3. 원 데이터로 집합으로부터 크기가 같은 표본을 중복을 허용하여 복원추출하여 각 표본에 대해 분류기(classifiers)를 생성하는 기법은?

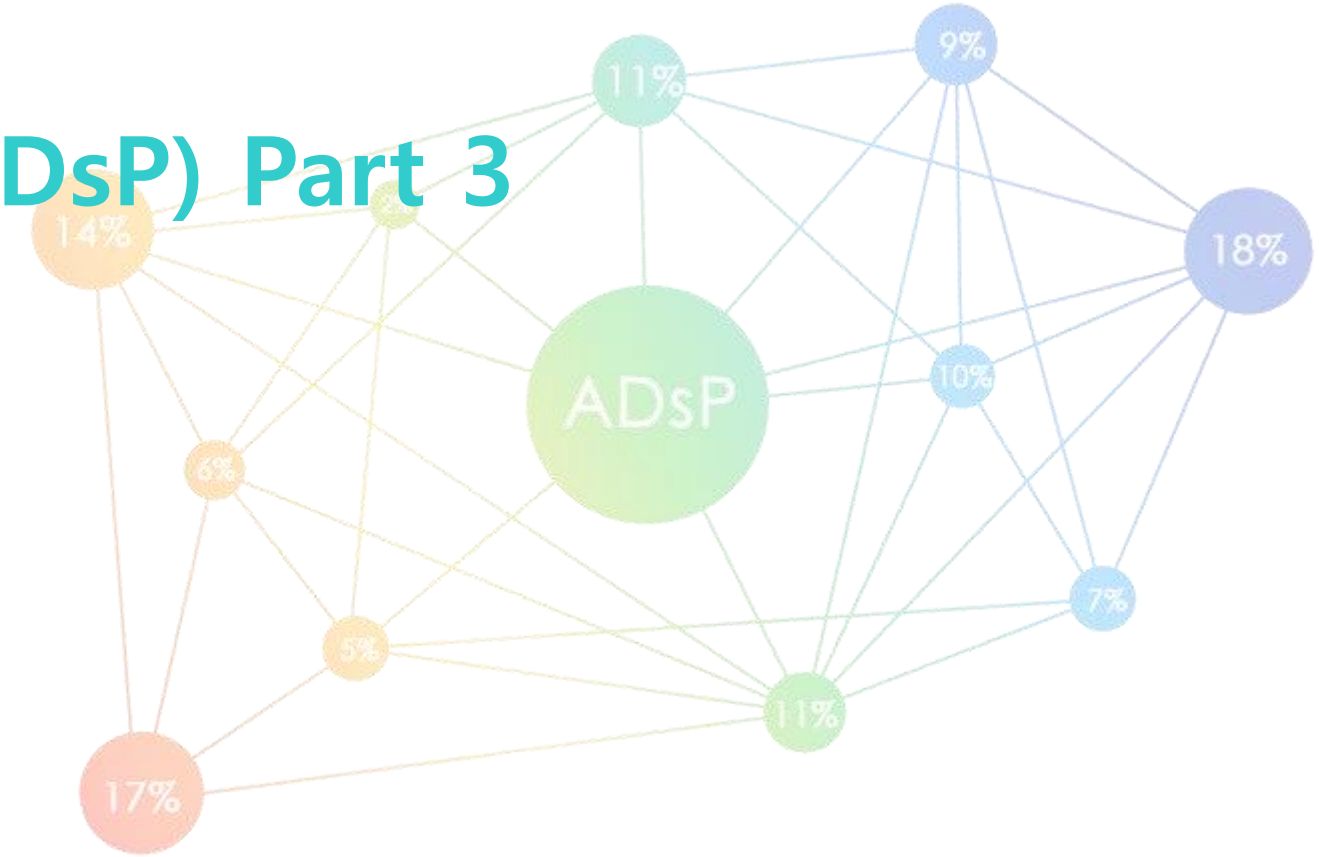
① 배깅
② 부스팅
③ 랜덤포레스트
④ 퍼셉트론

4. 기출문제

단답2) 의사결정나무의 형성 과정 중 최종마디가 너무 많으면 모형이 과대적합 상태로 현실문제에 적응할 수 없는 규칙이 나오게 된다.
이러한 과대적합(overfitting) 문제를 해결하기 위해 필요한 것은 무엇인가?

데이터분석준전문가(ADsP) Part 3

데이터분석



03 제3절 군집분석

1. 계층적군집분석
2. k-평균군집
3. 혼합분포군집
4. SOM(Self-Organizing Maps)

1.군집분석 개요

1.1. 정의

- 군집분석은 소속 집단을 모르는 데이터들을 서로

  동질적인 집단으로 분류하는 방법

- Data 구조에 대한 이해,형성된 군집의 특성과 군집들 간의 관계 파악
- 군집 형성의 최대기준은 그룹내의 데이터의 유사도는 최대로, 그룹간의 유사도는 최소
- 군집에 대한 정의가 쉽지 않고, 몇 개의 군집으로 나눌 것인지 명확하지 않다

1.군집분석 개요

1.2. 군집분석 vs 주성분 분석

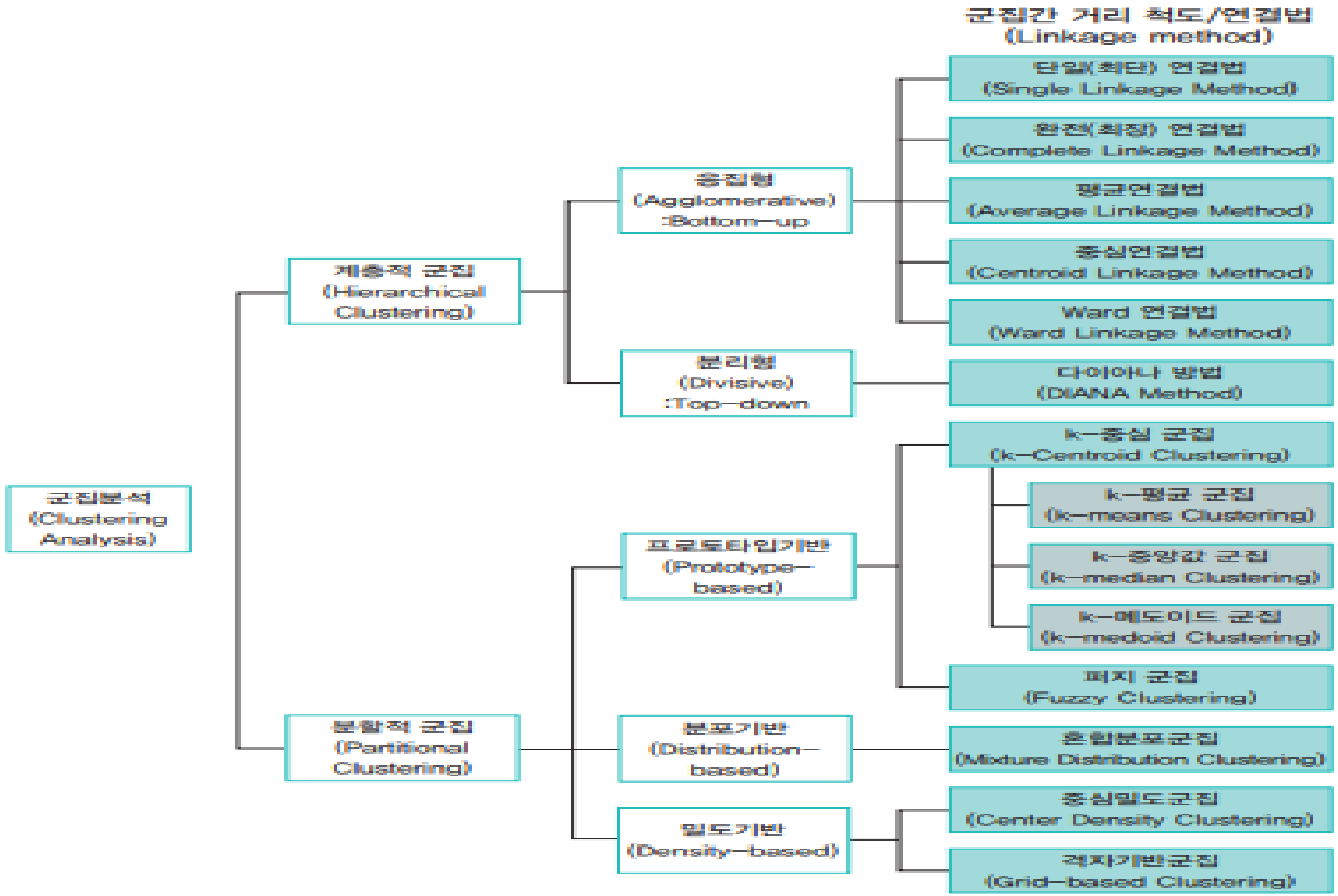
군집분석	주성분 분석
분석대상(case) 상호관련성에 의해 서로 동질적인 집단으로 그룹핑 분석case를 동질적인 집단, 집단간이질적인 집단	변수들 사이의 상호관계 분석해서 유사한 특징으로 축약하는 것 각 요인내에 있는 변수들은 상관관계가 높고 다른 요인들과는 상관관계가 낮음

1.군집분석 개요

1.2. 군집분석 vs 판별 분석

군집분석	판별분석
사전에 집단이 모르는 자료를 유사한 것들끼리 분류하여 군집 (소속집단을 모름) 비지도학습(목표변수없음)	관측된 모형을 만들고 새로운 자료가 들어왔을 때 분류 (소속집단 알고있음) 지도학습(목표변수 있음)

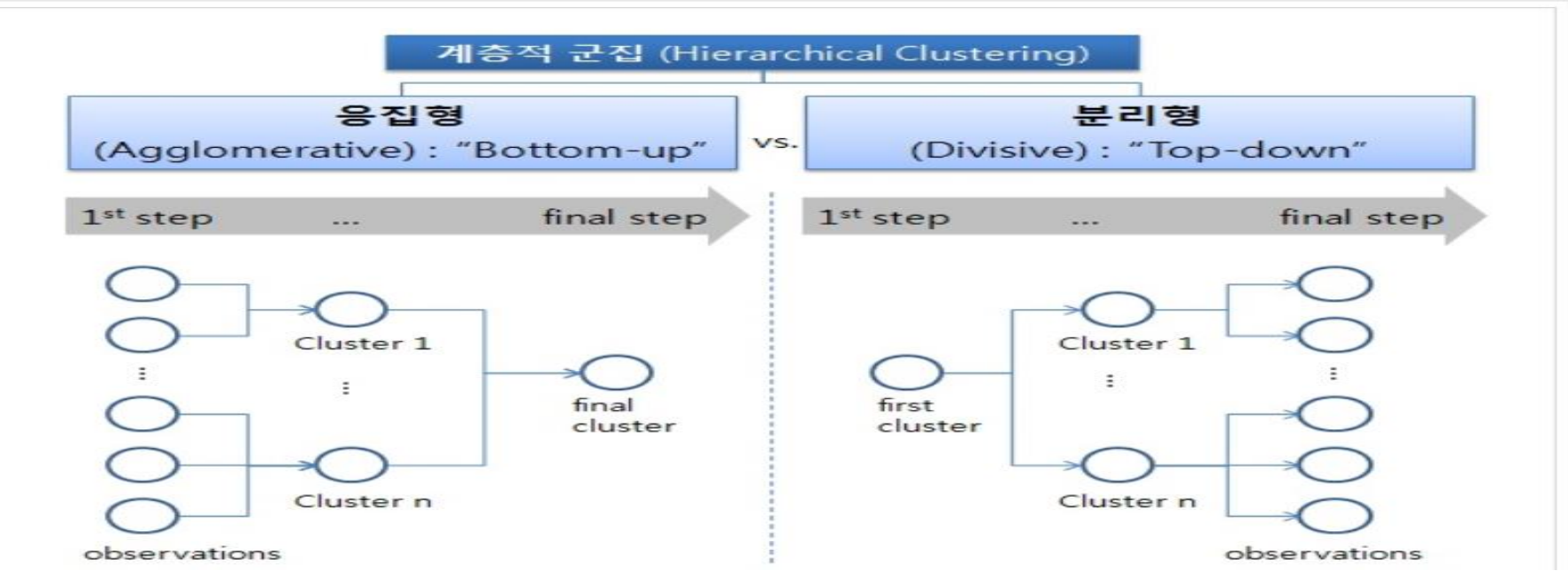
1. 군집분석 개요


군집분석
(Clustering Analysis)
계층적 군집
(Hierarchical Clustering)
분할적 군집
(Partitional Clustering)
응집형
(Agglomerative)
:Bottom-up
분리형
(Divisive)
:Top-down
프로토타입기반
(Prototype-based)
분포기반
(Distribution-based)
밀도기반
(Density-based)
군집간 거리 척도/연결법
(Linkage method)
단일(최단) 연결법
(Single Linkage Method)
완전(최장) 연결법
(Complete Linkage Method)
평균연결법
(Average Linkage Method)
중심연결법
(Centroid Linkage Method)
Ward 연결법
(Ward Linkage Method)
다이아나 방법
(DIANA Method)
k-중심 군집
(k-Centroid Clustering)
k-평균 군집
(k-means Clustering)
k-중앙값 군집
(k-median Clustering)
k-메도이드 군집
(k-medoid Clustering)
퍼지 군집
(Fuzzy Clustering)
혼합분포군집
(Mixture Distribution Clustering)
중심밀도군집
(Center Density Clustering)
격자기반군집
(Grid-based Clustering)

2. 계층적 군집분석

2.1. 계층적 군집분석 (hierarchical clustering)

가장 유사한 개체를 묶어 나가는 과정을 반복하여 원하는 갯수의 군집을 형성하는 방법이다. 덴드로그램(dendrogram) 사용



2. 계층적 군집분석

2.1. 계층적 군집분석 (hierarchical clustering)

- 병합법

- Bottom-up
- N개의 군집들을 가지고 시작해서 최종적으로 하나의 군집이 남을 때까지 순차적으로 유사한 군집들을 병합
  (계층적 군집분석에서는 주로 병합 방법이 쓰임)

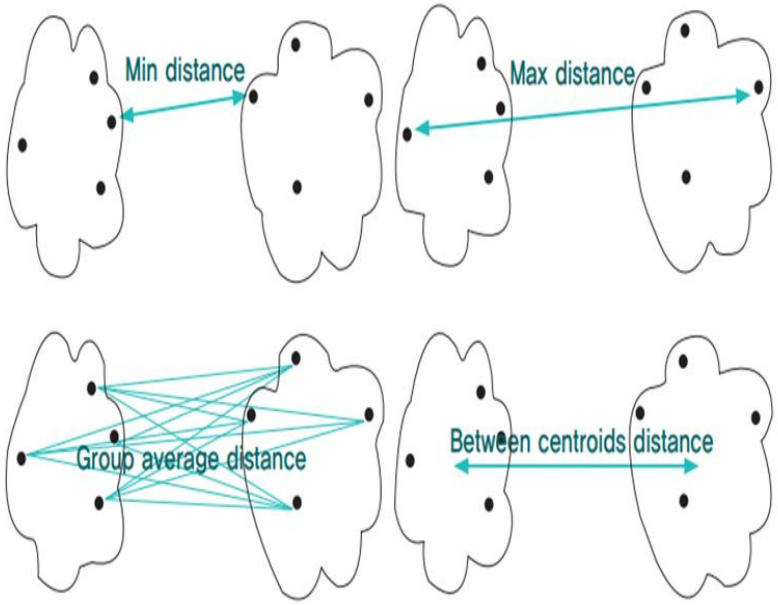
- 분할법

- Top-down
- 모든 레코드들을 포함하고 있는 하나의 군집에서 출발하여 n개의 군집으로 분할

2. 계층적 군집분석

2.2. 계층적 군집분석 (hierarchical clustering)

군집 방법	두 군집 사이의 거리
단일연결법(single linkage)	한 군집의 점과 다른 군집의 점 사이의 가장 짧은 거리(shortest distance). 사슬 모양으로 생길 수 있으며, 고립된 군집을 찾는 데 중점을 둔 방법이다.
완전연결법(complete linkage)	두 군집 사이의 거리를 각 군집에서 하나씩 관측값을 뽑았을 때 나타날 수 있는 거리의 최댓값을 측정한다. 같은 군집에 속하는 관측치는 알려진 최대 거리보다 짧으며, 군집들의 내부 응집성에 중점을 둔 방법이다.
평균연결법(average linkage)	모든 항목에 대한 거리 평균을 구하면서 군집화를 하기 때문에 계산량이 불필요하게 많아질 수 있다.
중심연결법(centroid)	두 군집의 중심 간의 거리를 측정한다. 두 군집이 결합할 때 새로운 군집의 평균은 가중평균을 통해 구해진다.
와드연결법(Ward linkage)	군집 내의 오차제곱합에 기초하여 군집을 수행한다.



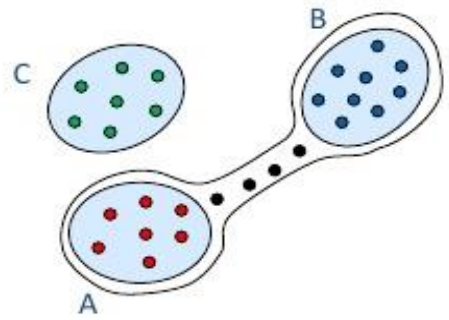
다양한 연결법들의 원리

2. 계층적 군집분석

2.2. 계층적 군집분석 (hierarchical clustering)

Single-Linkage Clustering

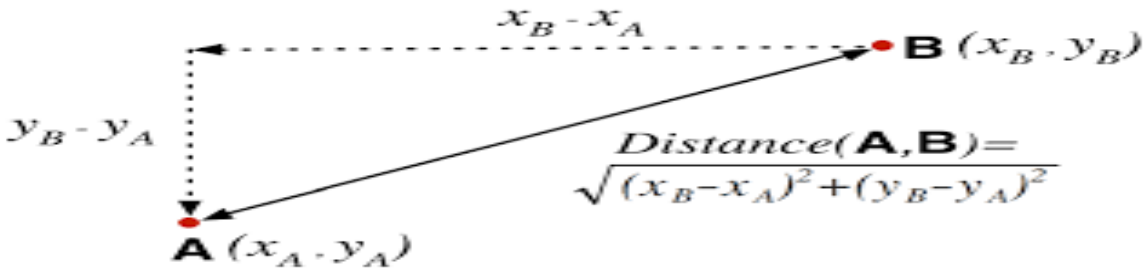
- **Advantage:** Can detect very long and even curved clusters.
  Can be used to detect outliers.
- **Drawback:** Chaining phenomen
  Clusters that are very distant to each other
  may be forced together
  due to single elements being close to each other.



2. 계층적 군집분석

2.3 비유사성 측도 : 거리

- 관측치들이 서로 얼마나 유사한지 또는 유사하지 않은지 측정하기 위한 측도로서 거리' 사용
- 가장 잘 알려진 거리척도로는 유클리드 거리(Euclidean distance)

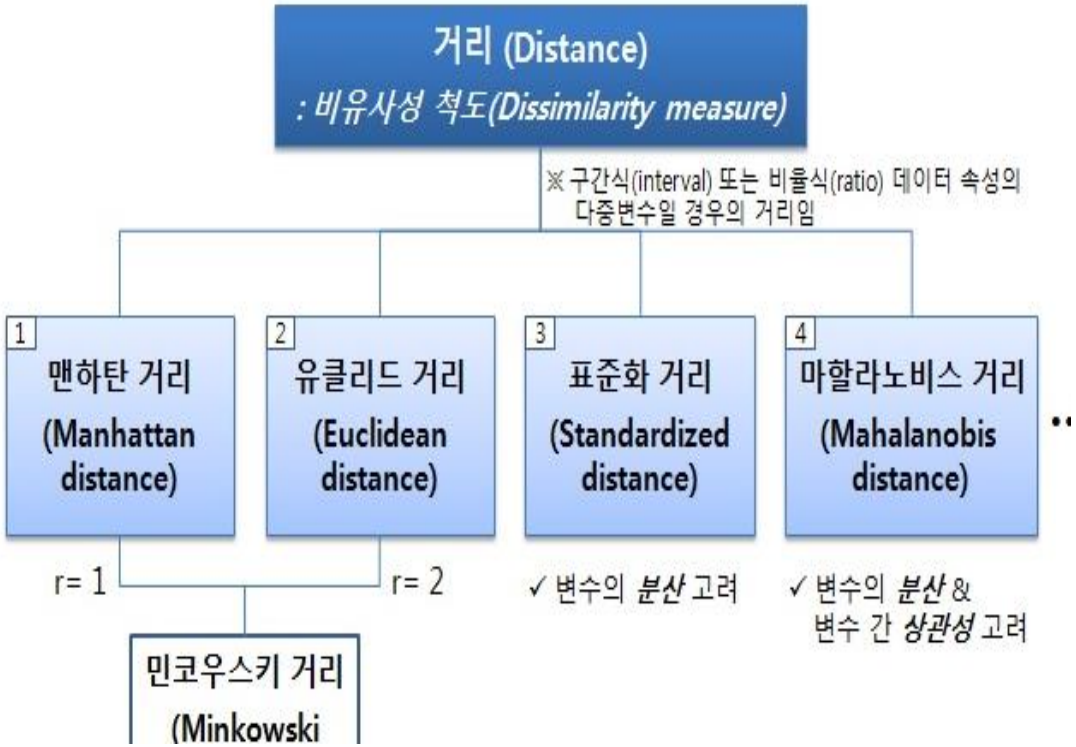


두 관찰단위 xi,xj 사이의 거리를 dij라 하면 dij는 일반적으로 다음 조건을 만족함

- 양의 성질(Nonnegative): dij >=0
- 대칭성(Symmetry): dij=dji
- 삼각 부등성(Triangle inequality): dij<=dik+dkj

2. 계층적 군집분석

2.3 비유사성 측도 : 거리



민코우스키 거리 (Minkowski Distance)

$$d_{Minkowski}(x,y)=\left(\sum_{j=1}^{m}|x_j-y_j|^r\right)^{1/r}$$

✓ 1-norm distance

$r=1$ ➡ $d(x,y)=\sum_{j=1}^{m}|x_j-y_j|$ ✓맨하탄 거리 (Manhattan Distance)

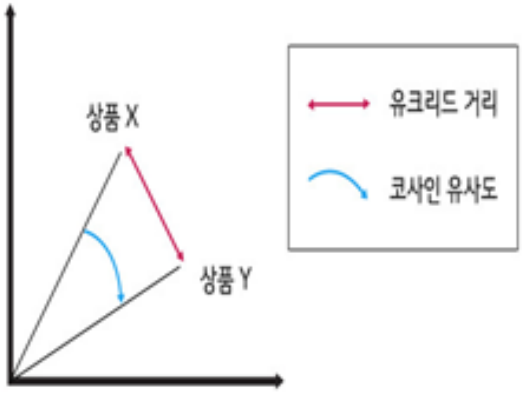
✓ 2-norm distance

$r=2$ ➡ $d(x,y)=\left(\sum_{j=1}^{m}|x_j-y_j|^2\right)^{1/2}$ ✓유클리드 거리 (Euclidean Distance)

(In the Euclidean space $\mathbb{R}^n$, the distance between two points)

2. 계층적 군집분석

- 캔버라 거리 : 가중치 있는 맨하탄 거리입니다. 원점 주변에 흩어져 있는 데이터에 주로 사용됩니다.
- 코사인 유사도 : 두 벡터이 내적을 각 벡터의 크기로 나눈 값을 1에서 뺀 것입니다.



- 명목형 데이터에 대한 유사성 척도 : 자카드계수

User based 데이터 셋 (1 : 구매)

	Item1	Item2	Item3	Item4
User1	0	1	0	1
User2	0	1	1	1
User3	1	0	1	0

$$J(A,B) = \frac{|A \cap B|}{|A \cup B|}$$

사용자 1과 사용자 2의 Jaccard 유사도를 계산하는 경우
(분모) |A∪B| 즉 두 사람이 산 상품의 합집합의 개수는 3

(분자) |A∩B| 두 사람이 산 상품의 교집합의 개수는 2
Jaccard 유사도 값은 ⅔= 0.67

2. 계층적 군집분석

최단연결법 예제

5개의관측값에대한 거리 행렬에 대하여 최단연결법으로 군집을 얻고

덴드로 그램 표현하기

	I	2	3	4	5
I	0				
2	7	0			
3	I	6	0		
4	9	3	8	0	
5	8	5	7	4	0

거리행렬에서 d(I,3)=I이 최소 관측값 I과 3을 묶어 군집 (I,3)

거리행렬을 갱신한다.

d((I,3),2)=min{d(I,2),d(3,2)}=min{7,6}=6

d((I,3),4)=min{d(I,4),d(3,4)}=min{9,8}=8

d((I,3),5)=min{d(I,5),d(3,5)}=min{8,7}=7

	(I,3)	2	4	5
(I,3)	0			
2	6	0		
4	8	3	0	
5	7	5	4	0

거리 행렬 갱신

d((I,3),(2,4))=min(d(I,3),2),d((I,3)),4)}=min{6,8}=6

d((2,4),5)=min(d(2,5)),d((4,5)}=min{5,4}=4

	(I,3)	(2,4)	5
(I,3)	0		
(2,4)	6	0	
5	7	4	0

2. 계층적 군집분석

d((2,4),5)=4 (2,4)와 5를 병합

	(1,3)	(2,4,5)
(1,3)	0	
(2,4,5)	6	0

d((1,3),(2,4,5))=min(d(1,3),(2,4),d((1,3)),5)}=min{6,7}=6

덴드로그램

1. 무슨 군집과 무슨 군집이 서로 묶였는지
2. 어떤 순서로 차례대로 묶였는지
3. 군집 간 거리는 얼마나 되는지

2. 계층적 군집분석

2.4. 계층적 군집의 장단점

• 장점
- 군집의 수를 명시할 필요 없음
- 덴드로그램을 통해 군집화 결과를 표현하며 설명 및 해석이 가능

• 단점
- 데이터 집합이 매우 클 경우 계산 속도가 느림
- 이상치 값에 민감

계층적 군집 방법은 매 단계에서 지역적 최적화 수행하므로
그 결과가 전역적인 최적해라고 볼 수 없음.
병합적 방법에서 한번 군집이 형성되면 군집에 속한 개체는 다른 군집으로 이동할 수 없다.

2. 기출문제

01. hclust에 관한 설명 중 적절하지 않은 것은?

① 최단연결법 또는 단일연결법(Single linkage method)은 두 군집 사이의 거리를 각 군집에서 하나씩 관측값을 뽑았을 때 나타날 수 있는 거리의 최솟값으로 측정한다.

② 최단연결법은 최단거리를 사용할 때 사슬 모양으로 생길 수 있으며 고립된 군집을 찾는 데 중점을 둔 방법이다.

③ 중심연결법(Centroid linkage)은 두 군집의 중심 간의 거리를 측정한다. 두 군집이 결합될 때 새로운 군집의 평균은 가중평균을 통해 구해진다.

④ 최단연결법은 평균연결법보다 계산량이 많아질 수 있다.

2. 기출문제

02. 계층적 군집은 두 개체 간의 거리에 기반하므로 거리측정에 대한 정의가 필요하다.
다음 중 dist( ) 함수에서 지원하지 않은 거리는?

① 유클리드

② 맨하튼

③ minkowski

④ cosine

2. 기출문제

03. 계층적 군집에서 군집내의 오차제곱합에 기초하여 군집을 수행하는 군집 방법은 무엇인가?

① 단일연결법

② 완전연결법

③ 평균연결법

④ 와드연결법

3. 비계층적 군집분석

3.1. *k*-means 군집(Non-Hierarchical Clustering)

- 주어진 군집 수 k에 대해서
  군집 내 오차제곱합의 합을 최소화 하는 것을 목적

- 군집 내 오차제곱합의 얼마나 군집화가 잘 되었는지를
  알려주는 척도

- 계층적군집보다 많은 양의 자료를 다룰 수 있다

  평균 등 거리 계산에 기반하므로 모든 변수가 연속형

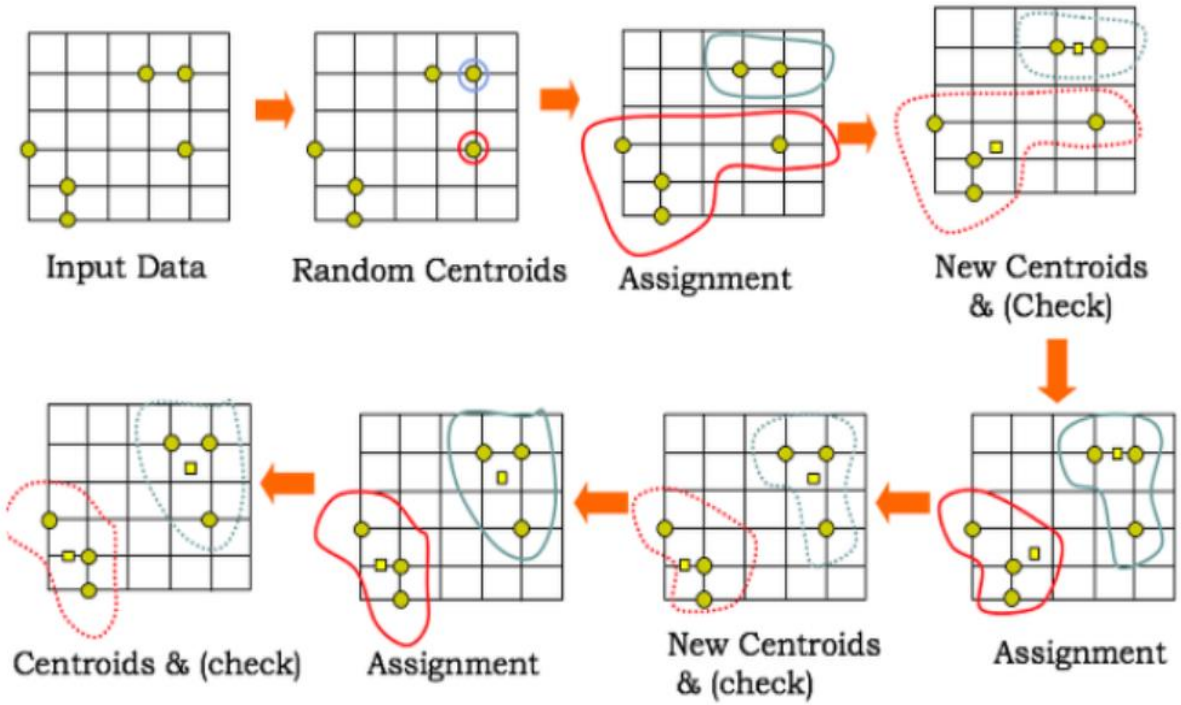
  군집의 중심 계산하는 과정에 잡음이나 이상값에 영향을 많이 받음

  -> K 중앙값(medoids)

- u형태의 군집이 존재할 경우 성능이 떨어진다

3. 비계층적 군집분석

3.2. *k*-means 알고리즘



3. 비계층적 군집분석

3.3. *k*-means 단계

1. 군집의 개수 k 정한다 (계층적 군집과 다른점)
2. 임의의 k개 점을 택하여 군집의 중심점으로 정한다
3. 각 관찰치가 가장 가까운 중심점을 계산하여 특정한

   중심점에 가까운 점들은 그 군에 속하는 것으로 간주한다

4. 군집내의 점들의 평균을 계산하여 새로운 중심점을 계산한다.
5. 개체의 할당에 변화가 없을 때까지 3.4의 과정을 반복한다.

3. 기출문제

01. k평균 군집에 대한 설명 중 적절하지 않은 것은?

① 초기값 선택이 최종 군집 선택에 영향을 미친다.

② 초기 군집수를 결정하기 어렵다.

③ 한 개체가 속해있던 군집에서 다른 군집으로 이동해 재배치가 가능하지 않다.

④ 각 군집내의 자료들의 평균을 계산하여 군집의 중심을 갱신한다

3. 기출문제

02. k-평균 군집에서 단점을 해결하기 위한 방안은?

① 이상값 자료에 민감한 k-평균 군집의 단점을 보완하기 위해 군집을 형성하는 매 단계 마다 평균 대신 중앙값을 사용하는 k-중앙값 군집을 사용한다.

② K-평균은 군집의 수를 미리 정할 필요가 없다.

③ 블록한 형태가 아닌 군집이 존재할 경우 군집 성능이 높아진다.

④ k 평균군집은 중심점으로부터의 오차제곱합을 최대가 되도록 할당해야 한다.

3. 기출문제

03. k-평균 군집의 설명 중 옳은 것은?

① K-평균 군집 결과는 덴드로그램(dendrogram)의 형태로 표현된다.

② K-평균 군집은 한번 군집이 형성되면 군집에 속한 개체는 다른 군집으로 이동할 수 없다.

③ K-평균 군집은 초기값을 지정하지 않는다.

④ 알고리즘이 단순하며, 빠르게 수행되며 계층적 군집보다 많은 양의 자료를 다를 수 있다.

4. 혼합분포 군집

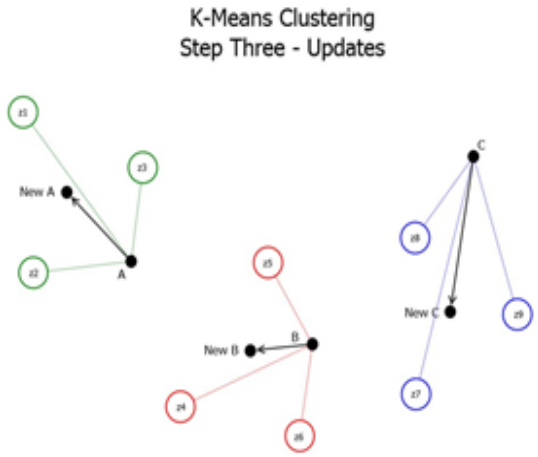
4.1. *혼합분포군집(Mixture distribution clustering)*

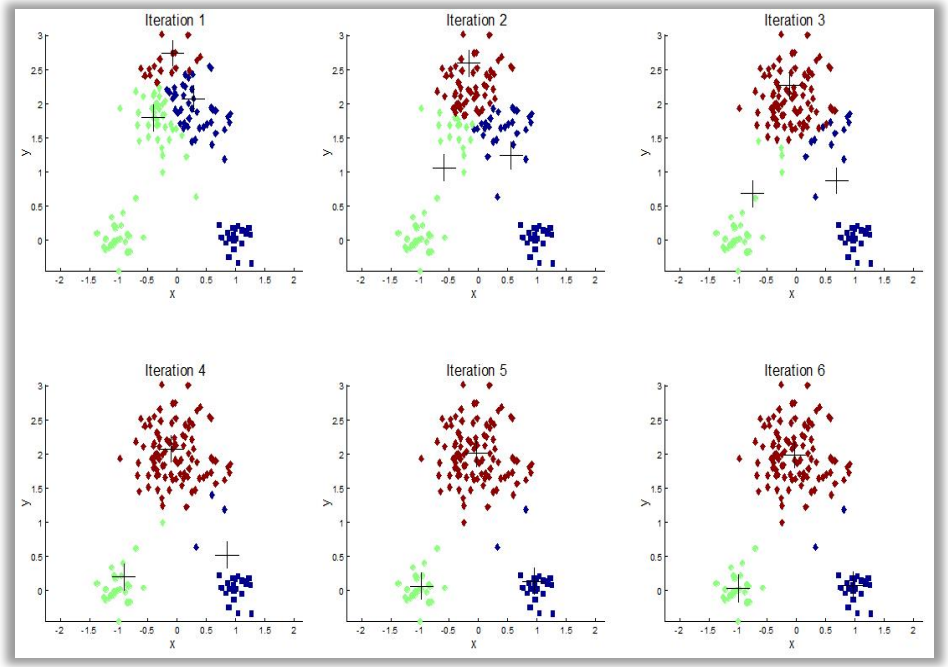
- 혼합분포군집은 모형기반의 군집 방법으로 데이터가
  k개의 모수적 모형의 가중합으로 표현되는 모집단 모형으로 나왔다는
  가정하에서 모수와 가중치를 자료로 추정하는 방법 사용

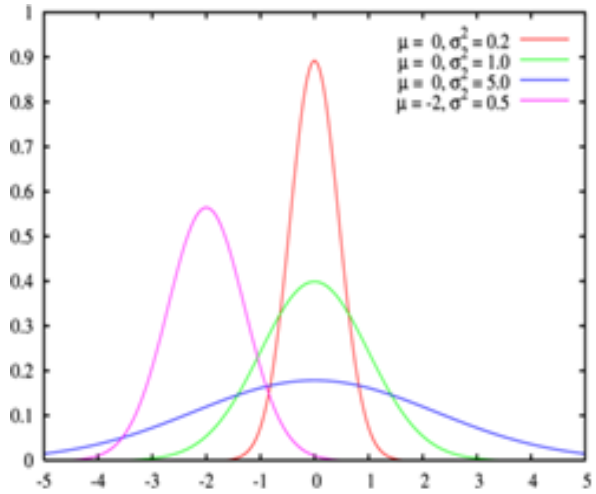
- k개의 각 모형은 군집의미, 각 데이터는 추정된 k 모형 중 어느
  모형으로부터 나왔을 확률이 높은지에 따라 군집의분류가 결정
  혼합모형에서는 모수와 가중치의 추정에는
  EM 알고리즘 이용

4. 혼합분포 군집

4.1. *혼합분포군집(Mixture distribution clustering)*

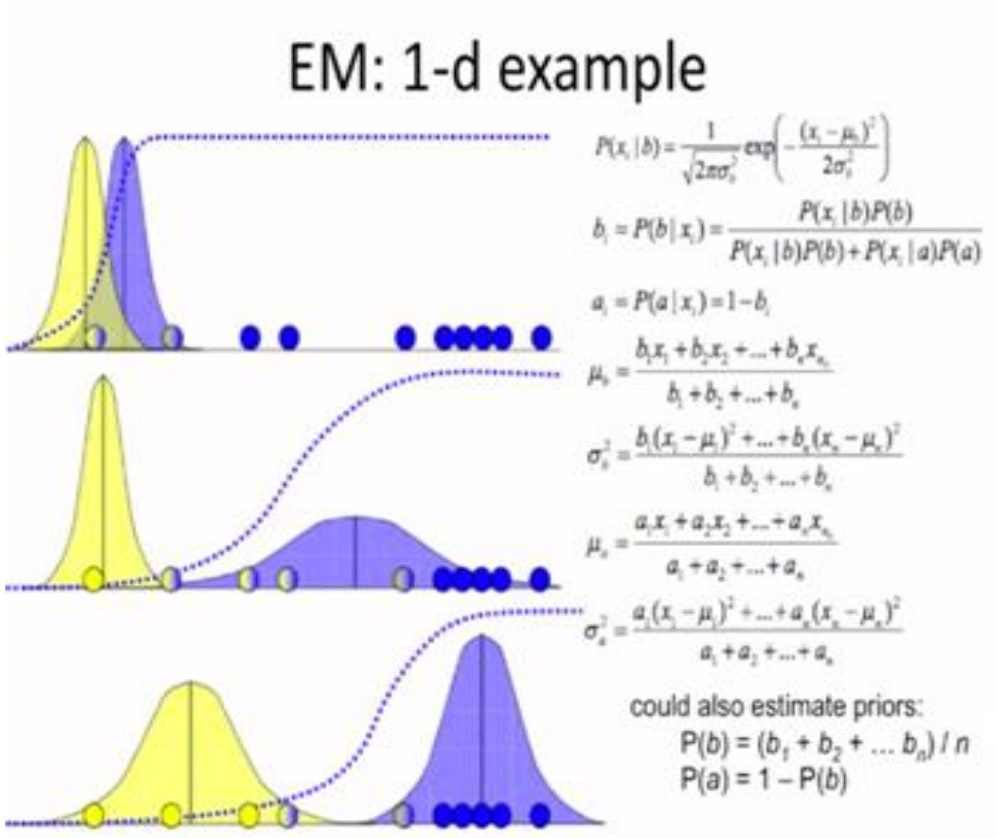






4. 혼합분포 군집

4.1. *혼합분포군집(Mixture distribution clustering)*



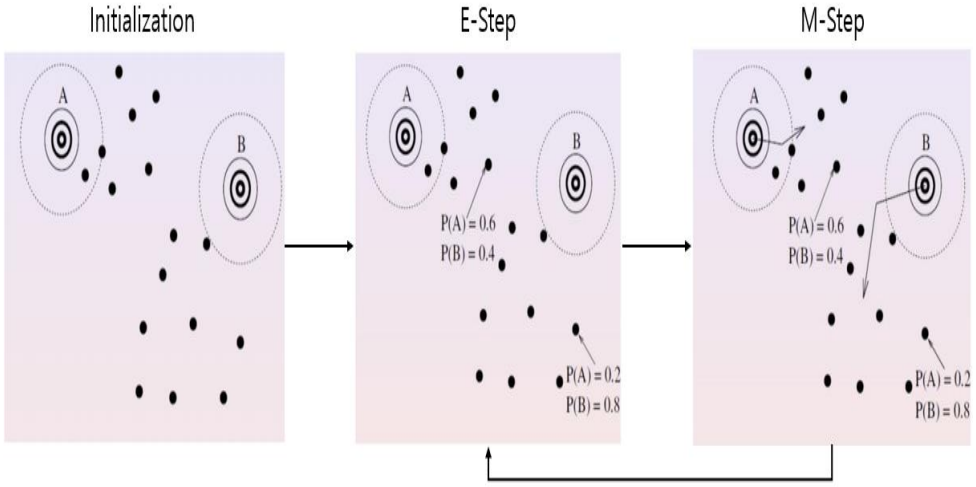
4. 혼합분포 군집

4.2. *혼합분포군집(Mixture distribution clustering) 특징*

- K 평균군집의 절차와 유사, 확률분포를 도입하여 군집을 수행하는 모형-기반의 군집 방법
- 군집을 몇 개의 모수로 표현이 가능
- 서로 다른 크기의 모양의 군집을 찾을 수 있다.
- 이상값 자료에 민감하다

4. 혼합분포군집

4.1. EM 알고리즘



- 혼합분포 모형인 경우 변수의 수가 많고 분포함수가 많아지면 최대가능도 추정법으로 모수를 추정하기가 쉽지 않다.
- EM 알고리즘은 혼합분포모형의 모수를 추정한 후 각 데이터 각각의 확률분포에 속할 확률을 계산하다 (E-step)
- 이 확률을 이용하여 최대가능도 추정법으로 모수의 추정 다시한다.(모수의 값이 거의 변하지 않을 때 까지 반복)(M-step)

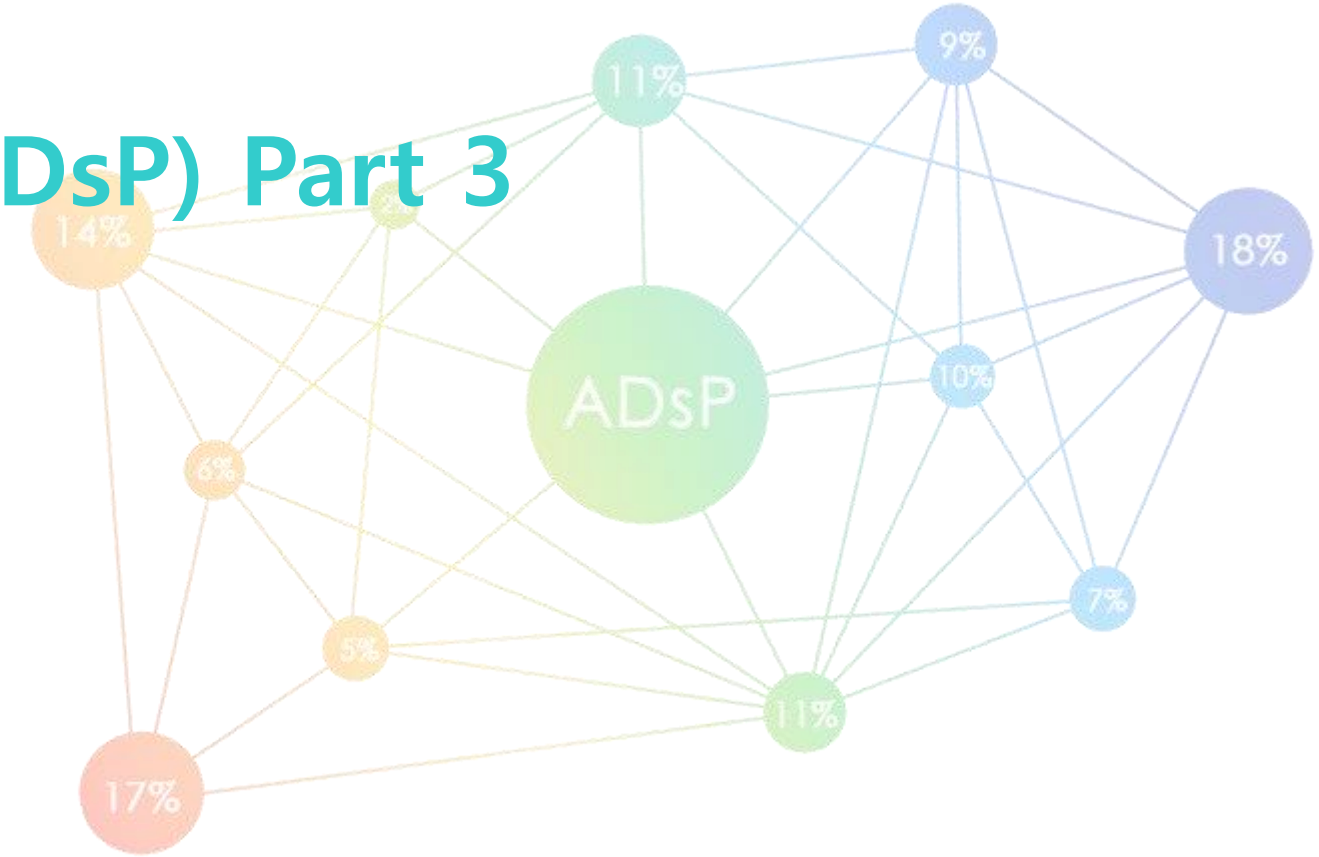
4. *혼합분포 군집*

4.2. *혼합분포 군집 모형의 특징*

- 혼합분포군집 모형은 K-평균 군집모형에 통계적 확률분포를 도입한 일반화된 모형이다.
- 서로 다른 크기나 모양의 군집을 찾을 수 있다. 그리고 군집을 몇 개의 모수로서 표현할 수 있다.
- EM 알고리즘은 데이터의 크기가 커지면 수렴하는데 시간이 걸릴 수 있다. 한 군집의 크기가 너무 작으면 추정에 문제가 있을 수 있다.
- 혼합분포 군집 모형은 잡음점이나 극단점에 민감한 결과를 가져 올 수 있다.

데이터분석준전문가(ADsP) Part 3

데이터분석



03 제3절 군집분석

1. 계층적군집분석
2. k-평균군집
3. 혼합분포군집
4. SOM(Self-Organizing Maps)

5. *Self-Organizing Maps (SOMs)*

5.1. *이론적 배경*

- **SOM 또는 SOFM(Self-Organizing Feature Map)**은 인공신경망**(ANN)**의 한 종류로서 기본개념 **1980**년대 핀란드 교수인 코호넨**(Kohonen)** 제안한 코호넨 네트워크 근간

- SOM VS ANN(인공신경망) 차이

  - ANN은 연속적인 레이어 구성, SOM은 뉴런(노드) 2차원 그리드 구성

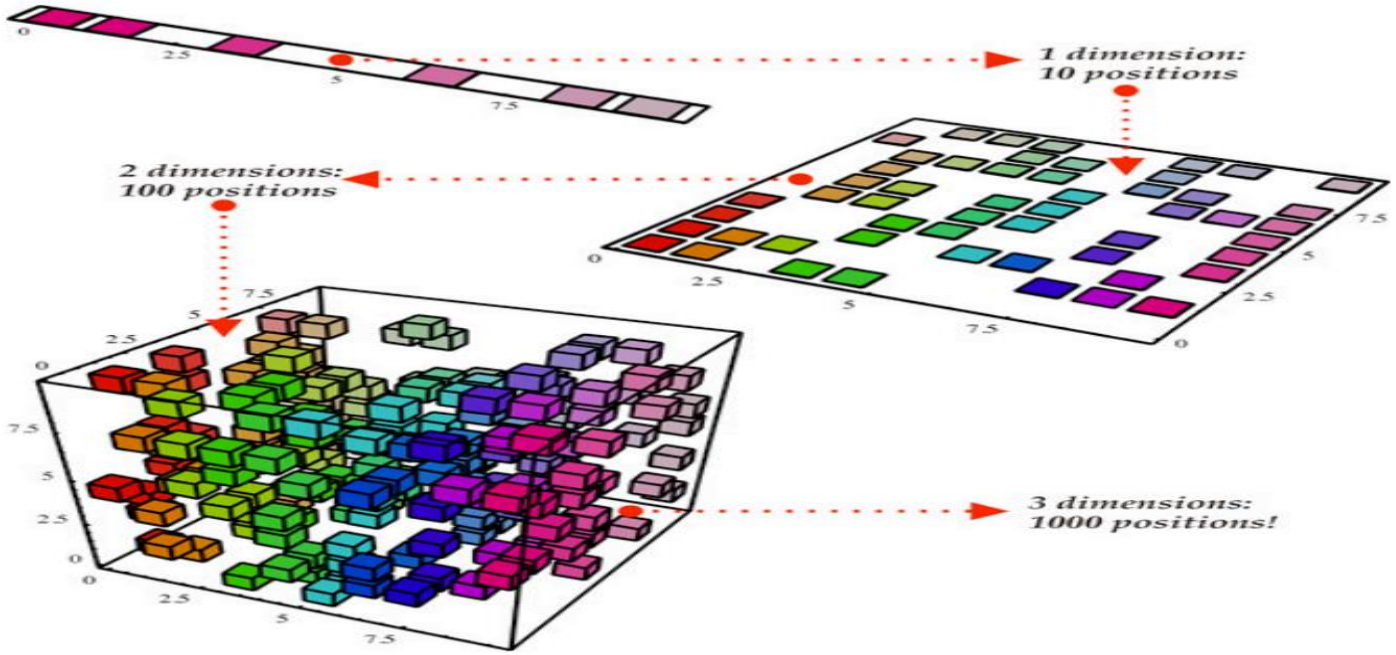
  - ANN은 에러를 수정하는 방향으로 학습, SOM은 경쟁학습(Competitive Learning)

  - SOM은 비지도학습(Unsupervised Learning)

5. *Self-Organizing Maps (SOMs)*

5.1. *이론적 배경*

- 주어진 입력 패턴에 대하여 정확한 해답을 미리 주지 않고 자기 스스로 학습할 수 있는 능력. 자기조직화 지도라고 함
- 다차원의(**multi-dimensional**) 데이터를 저차원(**low-dimensional**)으로 표현



5. *Self-Organizing Maps (SOMs)*

5.1. *이론적 배경*

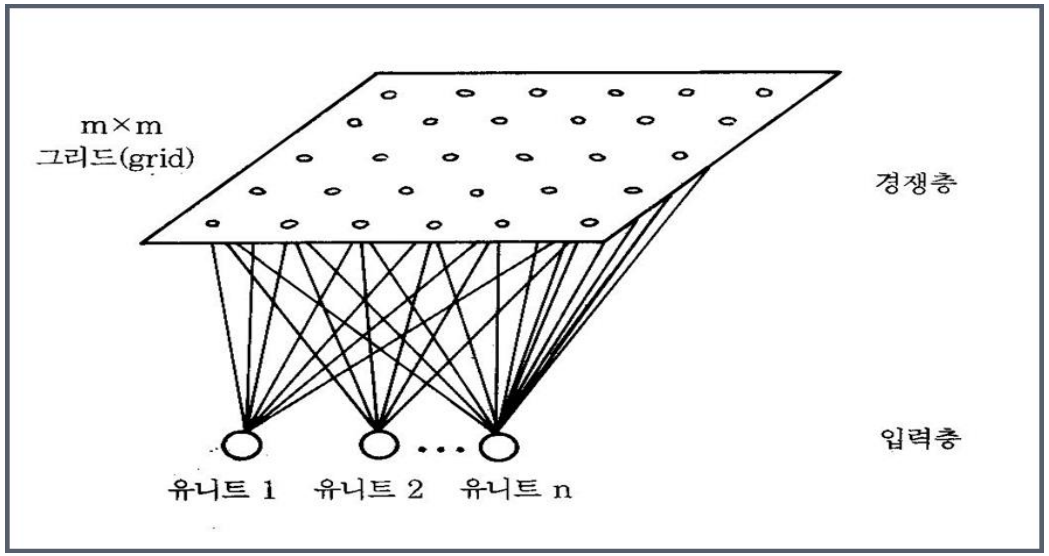
- **SOM 모형은 오류 역전파(Back-propagation) 모형과는 달리 한 번의 전방 전달 (feedforward flow)로 연산속도가 빠르다.**
- SOM은 단지 수치형 데이터 변수에서만 사용이 가능
- 범주형 자료를 더미변수로 변환하여 사용

5. *Self-Organizing Maps (SOMs)*

- 경쟁학습
  - 알고리즘 : 입력벡터들을 신경회로망에 계속적으로 제시하면서 자율적으로 연결가중치를 변경시키는 방법
  - 단순하면서 하드웨어 구현 시 구조 간단
    - 통계학의 k-means 군집화 알고리즘을 기초: 주어진 데이터를 k개의 클래스로 어느 오차수준 이하로 구분될 때까지 반복 → 패턴분류
  - 단순 구조
    - 한 개의 입력층과 한 개의 출력층
    - 입력층과 출력층이 완전 연결 (fully-connected)
    - 출력뉴런들은 승자 뉴런이 되기 위해 경쟁하고 오직 승자만이 학습함

5. *Self-Organizing Maps (SOMs)*

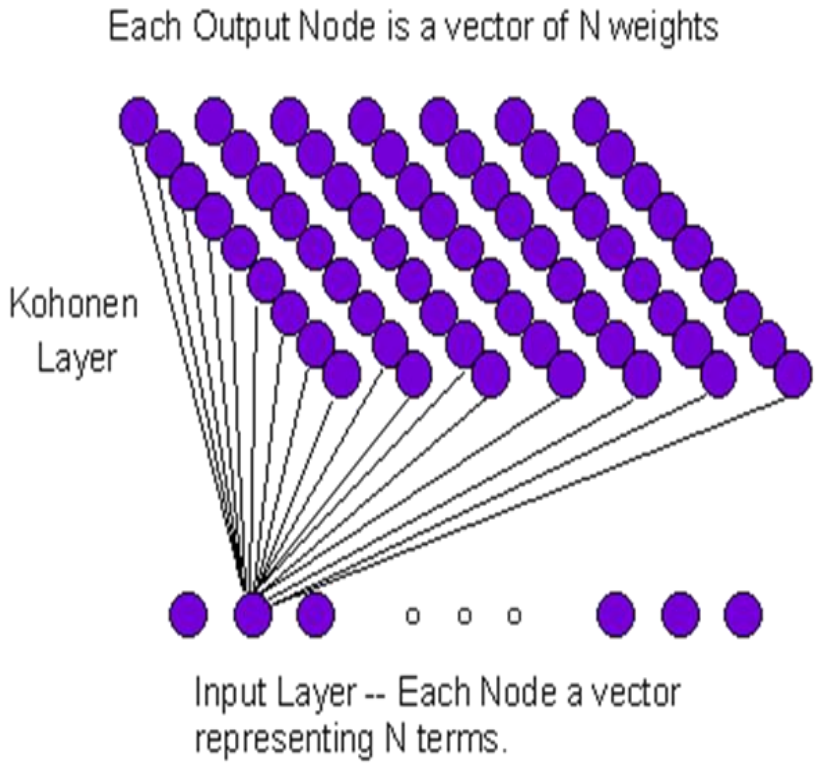
5.2. *구성*



- **Input Layer(정규화) / Competitive Layer(연결강도 초기화)로 구성**
- 모든 연결은 첫번째 층에서 두번째 층 방향으로 연결
- **The 2nd layer is fully connected.**

4. *Self-Organizing Maps (SOMs)*

5.2. *구성*



각각의 백터 입력 뉴런

- input vector: $x=[x_1, x_2, x_3, \ldots, x_m]^T$

각각의 백터 m 차원 공간에서의 한 점

- weight vector: $w_i = [w_{i1}, w_{i2}, w_{i3}, \ldots, w_{im}]^T, \quad i = 1,2,\ldots,l$

- 연결강도 벡터와 입력 벡터의 거리가 가장 가까운 뉴런만이 출력을 낼 수 있다. **승자 독점(winner take all)**

- best matching node

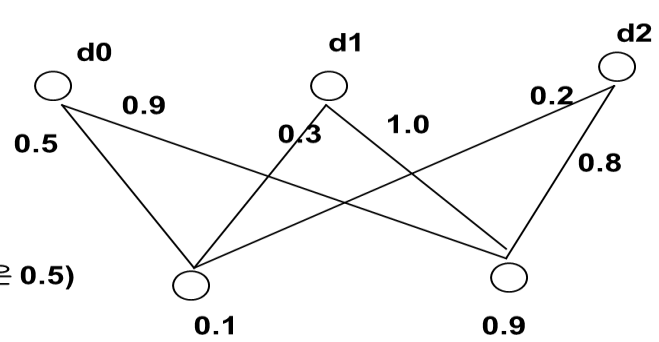
5.Self-Organizing Maps (SOMs)

5.2. 구성

SOM 예

❑ 예제

입력 : X=(0.1, 0.9)



(단, 학습률은 0.5)

– 거리 구함
- $d_0 = (0.1-0.5)^2 + (0.9-0.9)^2 = 0.16$,
- $d_1 = 0.05$,
- $d_2 = 0.02$ (승자뉴런)

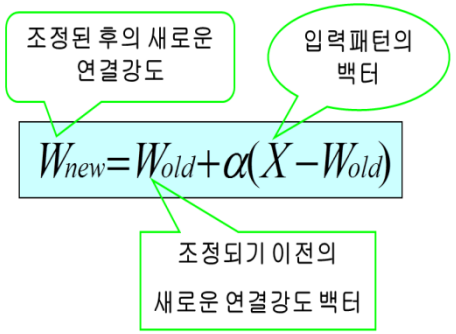
– d_2의 연결 가중치 조정

$W_{02}(t+1) = 0.2+0.5(0.1-0.2) = 0.15$

$W_{12}(t+1) = 0.8+0.5(0.9-0.8) = 0.85$

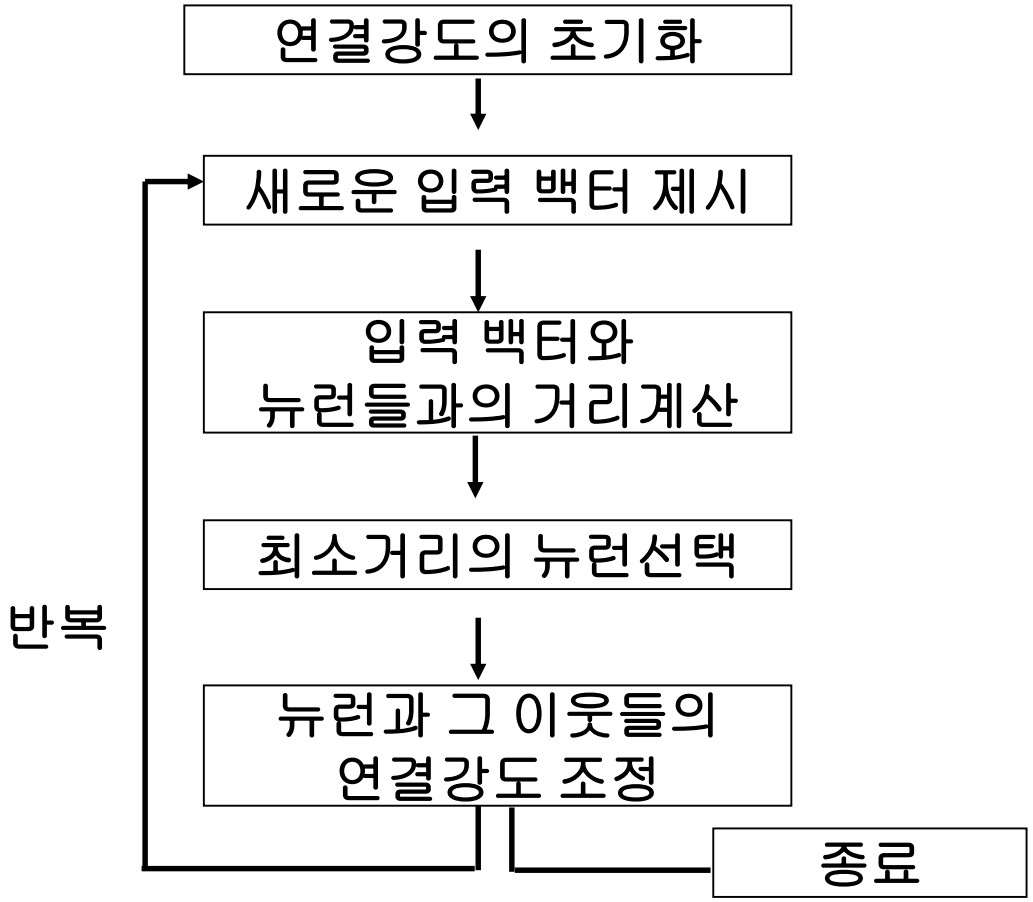
→ 제시된 입력과 가장 유사한 출력뉴런의 가중치벡터가 입력을 향하여 이동

뉴런의 연결강도 조정



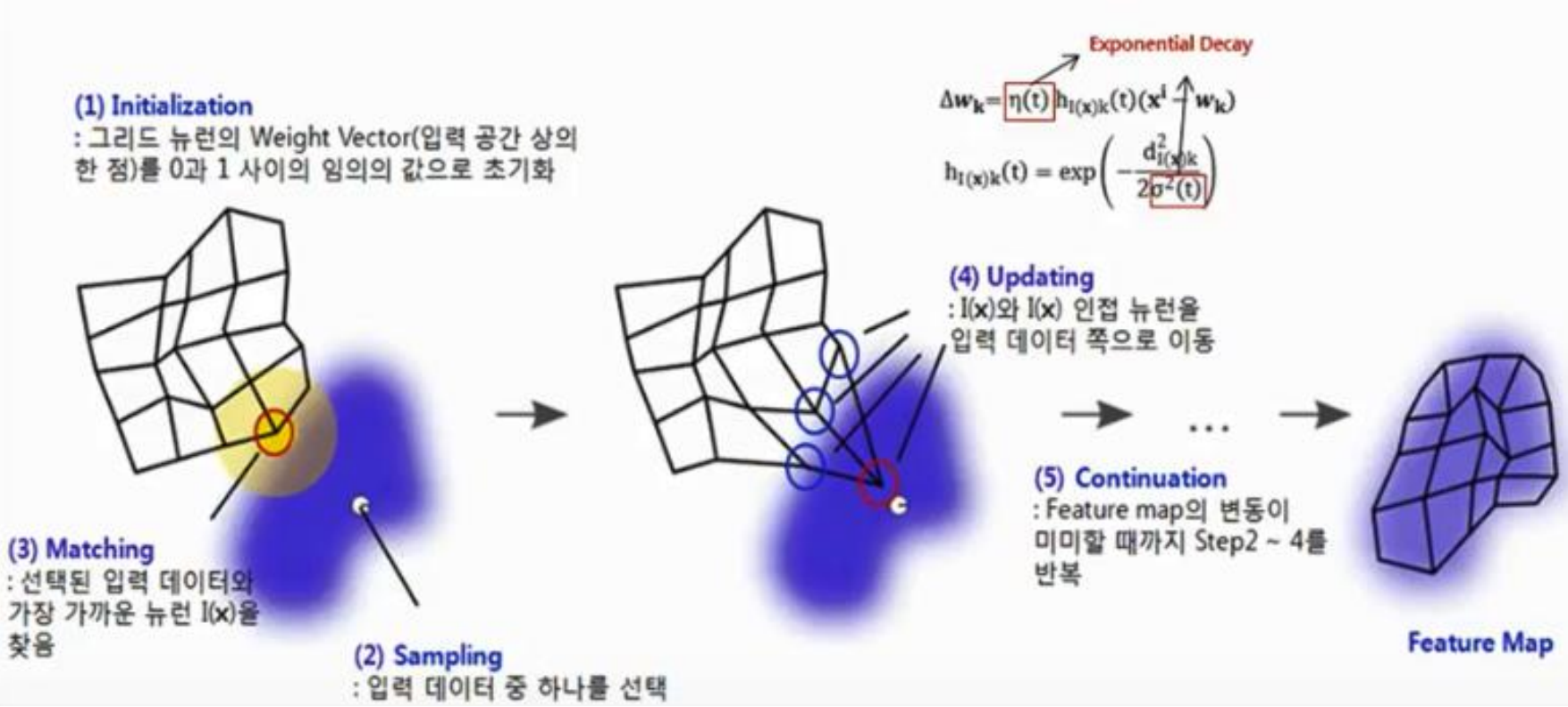
5. *Self-Organizing Maps (SOMs)*

Self-organizing Feature Maps Algorithm



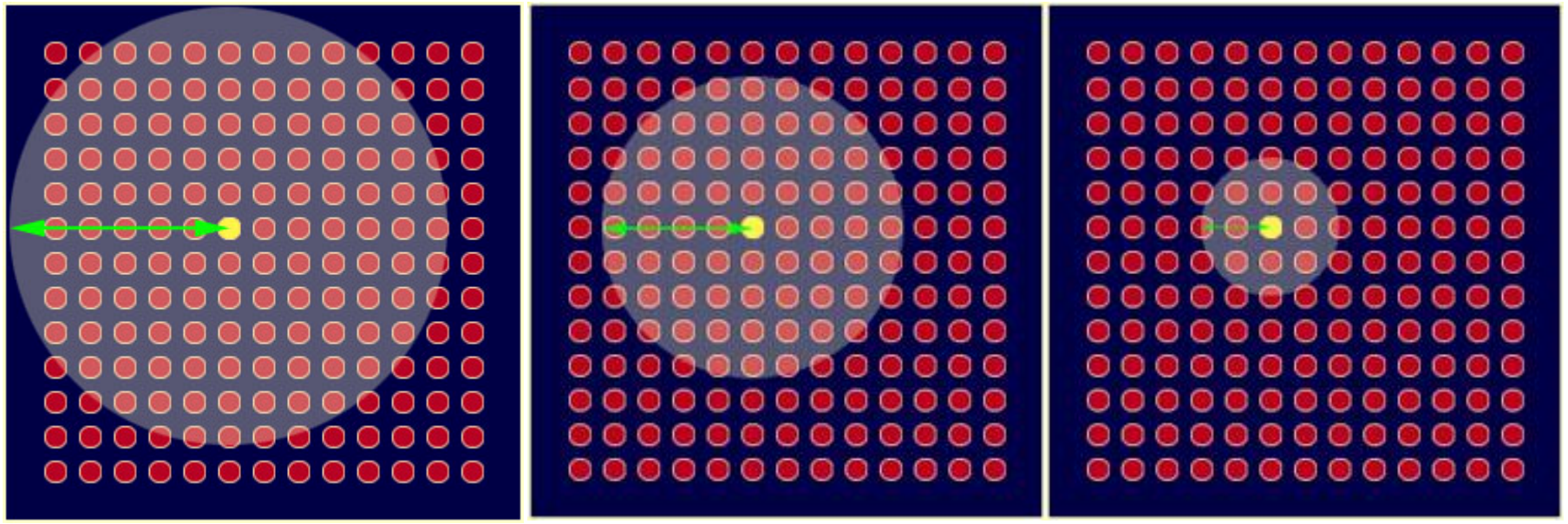
5. *Self-Organizing Maps (SOMs)*

5.3. 알고리즘



5. *Self-Organizing Maps (SOMs)*

5.3. *알고리즘(BMU)*



: Best matching unit (winning node)
: Raduis (처음에는 크게 시작함)

- 이웃반경 (Nc)의 크기가 시간이 지남에 따라서 점차 줄어든다
- 결국 BMU와 이웃한 노드들은 비슷하게 adjust 된다

5. *Self-Organizing Maps (SOMs)*

Som 장점

- 구조상 수행이 상당히 빠른 모델
  - ✓ 여러 단계의 피드백이 아닌 단 하나의 전방 패스(feedforward flow)를 사용
  - ✓ 잠재적으로 실시간 학습 처리를 할 수 있는 모델

- 연속적인 학습이 가능
  - ✓ 입력데이터의 통계적 분포가 시간에 따라 변하면 코호넨 네트워크는 자동적으로 이런 변화에 적응

5. *Self-Organizing Maps (SOMs)*

Q1 SOM에 관한 설명 중 틀린것은?

① 첫 번째로 입력벡터와 뉴런들간의 연결강도를 임의의 값으로 초기화한다.

② 입력벡터와 연결강도 벡터는 정규화된 값을 사용한다.

③ 입력벡터와 뉴런들 간의 거리를 측정할 때 유클리디언 거리를 이용한다.

④ 거리에 상관없이 모든 연결 강도를 재조정한다.

5. *Self-Organizing Maps (SOMs)*

Q2 **SOM에서 사용하는 원리가 아닌것은?**

① 경쟁학습

② 지도학습

③ 자율학습

④ 비지도학습

5. *Self-Organizing Maps (SOMs)*

Q3 **SOM에 대한 설명이 잘못된 것은?**

① SOM은 역전파 알고리즘 사용한다

② 차원축소와 군집화를 동시에 수행하는 기법이다

③ 데이터의 특징을 파악하여 유사 데이터를 군집화한다

④ 대표적인 비지도학습이다.

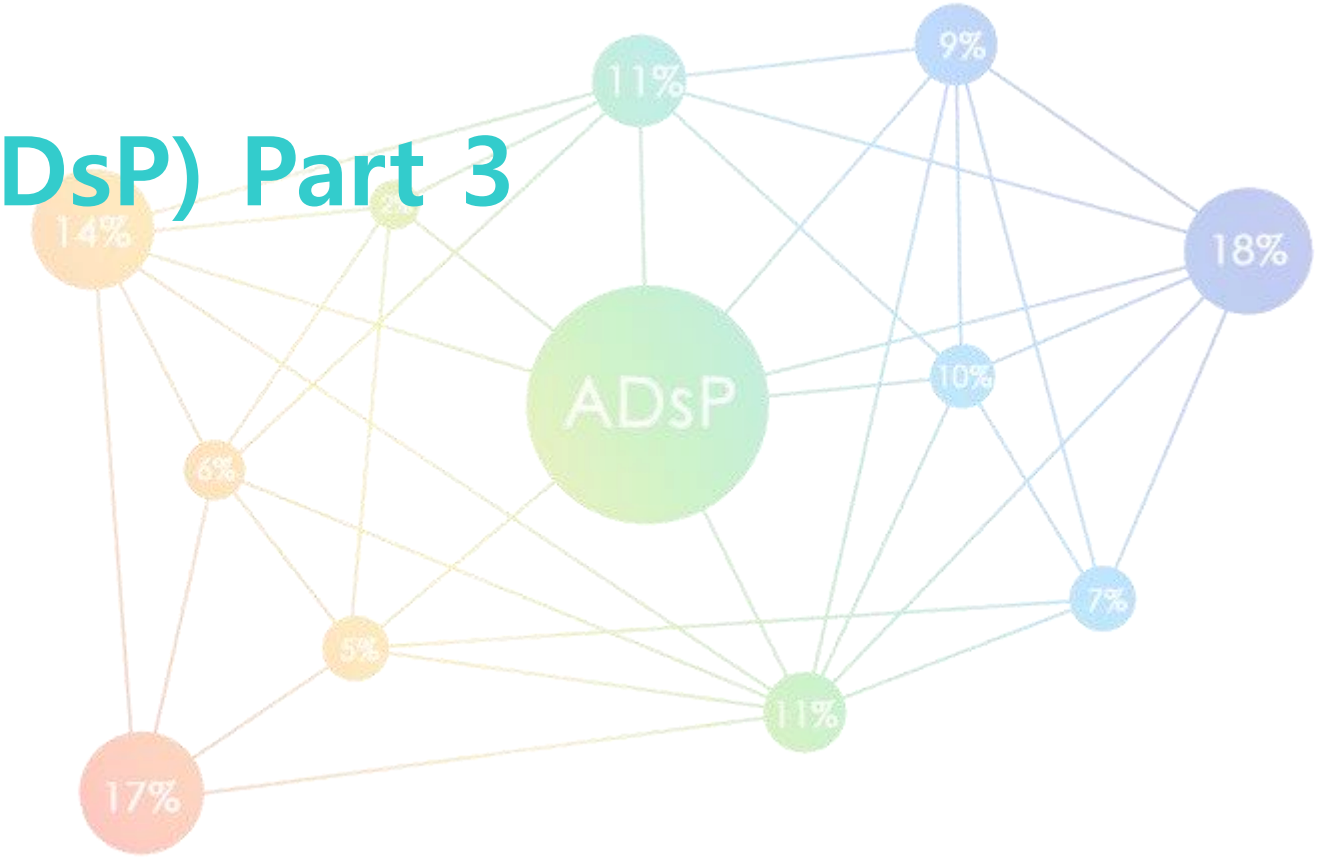
5. *Self-Organizing Maps (SOMs)*

Q4 SOM process에서 입력벡터와 경쟁층 노드 간의 유크리드안 거리를 계산하여 그 중에서 제일 가까운 뉴런을 무엇이라 하는가?

데이터분석준전문가(ADsP) Part 3

데이터분석

03 제4절 연관분석



1. 연관분석

1.1 연관분석 개요

- 연관분석은 기업의 활동 중에서도 마켓팅 분야에 가장 많이 사용되고 있고,이때 사용되는 데이터이 거래정보를 연관성 규칙을 이용하여 우리는 보통 장바구니 분석(market basket analysis)라고 함
- 활용

매장 내 상품 진열

- 연관성이 있는 물품들을 가까운 곳에 진열 예) 맥주를 진열한 곳에 안주류 같이

묶음 판매

- 연관성이 있는 물품들을 묶어서 판매

쿠폰발행

- 하나의 물건 사고 연관성이 있는 물푸을 사면 할인 혜택이 있는 쿠폰 발행

교차판매 (cross selling)

- 특정 물품을 구매한 고객에게 연관성이 있는 물품을 추천하여 추가 구매하도록 유도

1. 연관분석

1.2 연관성 규칙의 발견 기준

지지도(Support), 신뢰도(Confidence), 향상도(Lift)

$$\text{지지도(support)} = \frac{\text{선행 항목집합}(A)\text{과 후행 항목집합}(B)\text{을 동시에 포함하는 거래수}}{\text{전체거래의 수}}$$

$$= P(\text{선행 항목집합 } AND \text{ 후행 항목집합})$$

P(A),P(B)를 각각 물품 A, 물품 B가 구매될 확률로 정의

support(A->B) 의미는 IF A 구매 THEN B도 구매

support(B->A)= support(A->B)

1. 연관분석

1.2 연관성 규칙의 발견 기준

지지도(Support), 신뢰도(Confidence), 향상도(Lift)

$$\text{신뢰도(confidence)} = \frac{\text{선행 항목집합}(A)\text{과 후행 항목집합}(B)\text{을 동시에 포함하는 거래수}}{\text{선행 항목집합}(A)\text{을 포함하는거래수}}$$

$$= \frac{P(\text{선행 항목집합 } AND \text{ 후행 항목집합})}{P(\text{선행 항목집합 })}$$

$$= P(\text{후행 항목집합} \mid \text{선행 항목집합})$$

신뢰도(A->B): 물품 A를 구매했다는 조건하에 물품 B를 구매확률 =P(B|A) 조건부 확률

신뢰도가 50% 이라는 의미는 "A를 구매한 거래 가운데 50% B도 구매"

Confidence(A->B) ≠ Confidence(B->A)

1. 연관분석

1.2 연관성 규칙의 발견 기준

지지도(Support), 신뢰도(Confidence), 향상도(Lift)

$$\text{향상도} = \frac{P(B|A)}{P(B)}$$

향상도는(A->B) 전체에서 B가 거래된 비율과 /
A가 구매되었다는가정하에 B가 구매된
비율사이의 비율

$$\frac{P(A \cap B)}{P(A)P(B)} = \frac{A\text{와 } B\text{를 포함하는 거래수}}{A\text{를 포함하는 거래수} * B\text{를 포함하는 거래수}}$$

lift(A->B)=lift(B->A)

1. 연관분석

1.2 연관성 규칙의 발견 기준

지지도(Support), 신뢰도(Confidence), 향상도(Lift)

- 향상도의 의미
  - Lift >1 : 품목간 상호 양의 상관관계에 있음
    (예) 맥주와 감자칩
  - Lift =1 : 품목간 상호 독립적인 관계에 있음
    (예) 과자와 후추
  - 0 < Lift <1 : 품목간 상호 음의 상관관계에 있음
    (예) 지사제와 변비약

1. 연관분석

1.3 알고리즘(Apriori)

1. 최소지지도를설정한다.(최소지지도 0.3, 최소 신뢰도 0.7)

거래 정보	물품
I	a,b,d,e,m
2	b,c,g,m,s
3	a,d,m
4	b,c,m
5	a,b,m
6	b,e,g,t
7	d,e,s,z

1. 연관분석

1.3 알고리즘(Apriori)

2. 개별품목 중에서 최소지지도를 넘는 품목을 찾는다.

물품	빈도	지지도	
a	3	3/7	>0.3
b	5	5/7	>0.3
c	2	2/7	<0.3
d	3	3/7	>0.3
e	4	4/7	>0.3
g	2	2/7	<0.3
m	4	4/7	>0.3
s	2	2/7	<0.3
t	I	I/7	<0.3
z	I	I/7	<0.3

지지도(a) 분모 전체거래수 중에 a

-> 최소지지도 만족하는 것은 a,b,d,e,m

1. 연관분석

1.3 알고리즘(Apriori)

3. 2단계에서 찾은 품목 집합을 결합하여 최소지지도를 넘는 2가지 품목 집합을 찾는다

물품	지지도
a,b	2/7
a,d	2/7
a,e	1/7
a,m	3/7
b,d	1/7
b,e	3/7
b,m	3/7
d,e	2/7
d,m	2/7

1. 연관분석

1.3 알고리즘(Apriori)

4. 위의 두 절차에서 찾은 품목 집합을 결합하여 조합을 찾는다.

물품	지지도	
a,b	2/7	
a,d	2/7	
a,e	1/7	
a,m	3/7	>0.3
b,d	1/7	
b,e	3/7	>0.3
b,m	3/7	>0.3
d,e	2/7	
d,m	2/7	
e,m	1/7	

-> (a,m),(b,e),(b,m)

1. 연관분석

1.3 알고리즘(Apriori)

5. 반복적으로 수행해 최소지지도 (최소 신뢰도가)가 넘는 빈발품목 집합을 찾는다.

물품	신뢰도	지지도
a->m	3/3>0.7	3/7>0.3
m->a	3/4>0.7	
b->e	3/5>0.7	3/7>0.3
e->b	3/4>0.7	
b->m	3/5>0.7	3/7>0.3
m->b	3/4>0.7	

-> rules
a-m
b-e
b-m

a<->m

b<->e

1. 연관분석

n개 물품 페어도 가능한지?

3개 물품 연관 rule

(a,m,b),(m,b,e)

물품	지지도
a,b,m	2/7 <0.3
m,b,e	1/7 <0.3

-> 3개 이상은 충족하지 못함

1. 연관분석

1.4 연관분석의 장.단점

연관분석 장점

1. 탐색적인 기법:: 조건반응(if~then)으로 표현되는 연관분석 결과를 이해
2. 비목적성 기법
3. 사용이 편리한 분석 데이터의 형태

연관분석 단점

1. 품목수가 증가하면 계산은 기하급수적 증가
2. 너무 세부화된 연관규칙이나 거래량이 적은 픔목에는 의미가 없거나
   규칙발견시 제외되기 쉽다..

2. 기출문제

01. 다음 중 연관분석(Association analysis) 설명으로 적절하지 않은 것은?

① 품목 수가 증가하면 분석에 필요한 계산은 기하급수적으로 늘어난다.

② 너무 세부화된 품목을 가지고 연관규칙을 찾으려고 하면 의미 없는 분석 결과가 생성 될 수 있다.

③ 향상도가 1이면 두 품목 간에 연관성이 없는 서로 독립적인 관계이고, 1보다 작으면 서로 음의 관계로 품목 간에 연관성이 없다.

④ 시차 연관분석은 인과관계 분석이 가능하다.

02. 연관규칙의 향상도 설명이 옳은 것은?

① 향상도가 1보다 크면 이 규칙은 결과를 예측하는 데 있어서 우수하다는 것을 의미
② 항샹도가 1이면 두 품목은 연관성이 높다는 의미
③ 향상도가 1보다 작으면 두 품목은 서로 양의 상관관계를 의미
④ 지지도는 a와 b가 동시에 포함된 거래 수 /a를 포함한 거래 수를 의미한다.

2. 기출문제

03. 연관분석의 특징으로 옳지 않은 것은?
① 조건반응 (if then)으로 표현되는 연관분석의 결과를 이해하기 쉽다.
② 비목적성 분석 기법이다.
③ 세분화된 분석 품목 없이 연관 규칙을 찾을 수 있다.
④ 분석 계산이 간편하다.

04. 연관 규칙 지표에 대한 설명 중 옳은 것은?
① 향상도가 1보다 크면 이 규칙은 결과를 예측하는데 우수하다는 것을 의미한다.
② 향상도가 1이면 두 품목간에 상호 연관성이 높다는 것으로 해석한다.
③ 향상도가 1보다 작으면 품목간에 독립적인 관계로 본다.
④ 지지도(a->b)는 조건부 확률로 "품목a를 구매한사람이 품목 b도 구매한다고 해석한다.

2. 기출문제

05. 아래의 거래 데이터에서 추출된 연관규칙 중 사과->딸기에 대한 향상도?

거래 번호	판매 상품
1	배, 사과, 딸기
2	배, 사과, 포도
3	배, 자몽
4	사과 ,딸기
5	배, 사과, 딸기, 포도
6	자몽