EduAtoZ ADsP 수정/보완

2023년 5월 18일

2-02. 분석 주제 유형 4가지 - 문제 1

1. 분석 주제 유형 중 분석의 대상(WHAT)을 모르고, 분석 방법(HOW)은 아는 경우 분석 방향은?

문제 교정

1. 최적화(Optimization)
2. 통찰(Insight)
3. 솔루션(Solution)
4. 발견(Discovery)

2. 분석 대상을 모르나, 기존 분석 방식을 활용할 경우와 대상을 새로 선정하는 것은?

1. 통찰, 발견
2. 최적화, 통찰
3. 솔루션, 발견
4. 최적화, 솔루션

- Insight : 분석 대상이 불분명하고, 분석 방법을 알고 있는 경우 인사이트 도출
- Discovery : 분석 대상, 방법을 모른다면 발견을 통해 분석 대상 자체를 새롭게 도출함

2-07. 분석 방법론의 모델 3가지

출 : 17, 18, 19, 23, 25, 31, 33

폭포수 모델

- 단계를 순차적으로 진행하는 방법
- 이전 단계가 완료되어야 다음 단계로 순차 진행하는 하향식 진행
- 문제점이 발견되면 전단계로 돌아가는 피드백 수행

나선형 모델

- 반복을 통해 점증적으로 개발
- 반복에 대한 관리 체계가 효과적으로 갖춰지지 못한 경우 복잡도가 상승하여 프로젝트 진행이 어려울 수 있음

프로토타이핑 모델

- 사용자 요구사항이나 데이터를 정확히 규정하기 어렵고 데이터 소스도 명확히 파악하기 어려운 상황에서 사용
- 일단 분석을 시도해보고 그 결과를 확인해가면서 반복적으로 개선해 나가는 방법
- 신속하게 해결책 모형제시, 상향식 접근방법에 활용

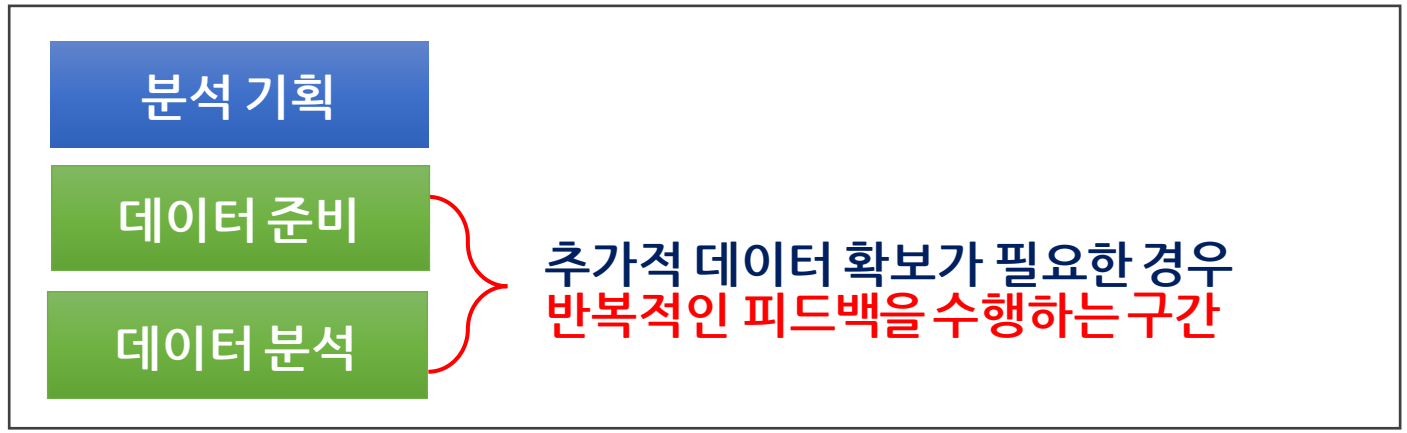
폭포-하향! 프로토-상향!!

프로토타입 모델 → 프로토타이핑 모델

2-10-3. 데이터 분석 단계- 문제 1, 2

1. 빅데이터 분석 방법론 중 추가적인 데이터 확보가 필요한 경우 '반복적인 피드백'을 수행하는 구간은?

1. 분석 기획 ~ 데이터 준비
2. 데이터 준비 ~ 데이터 분석
3. 데이터 분석 ~ 데이터 구현
4. 시스템 구현 ~ 평가 및 전개



2. **빅데이터 분석 프로세스에서** 분석용 데이터를 이용한 가설 설정을 통해 통계 모델을 만들거나 기계학습을 이용한 데이터의 분류, 예측, 군집 등의 기능을 수행하는 과정을 무엇이라 하는가?

문제 좀 더 구체화 -> 빅데이터 분석 프로세스에서 추가

2-10-3 답지	1	2
	2	모델링

2-21. 이행계획 수립

1. 로드맵 수립

- 결정된 과제의 우선순위를 토대로 분석 과제별 적용 범위 및 방식을 고려하여 최종적인 실행 우선순위를 결정 후 단계적 구현 로드맵 수립

2. 세부 이행계획 수립

- 반복적인 정련 과정을 통해 프로젝트의 완성도를 높이는 방식을 주로 사용
- 모든 단계 반복보다 데이터 수집 및 확보와 분석 데이터를 준비하는 단계를 순차적 진행하고 모델링 단계는 반복적으로 수행하는 혼합형을 많이 적용함
- 데이터 분석체계의 특징을 고려한 세부적인 일정계획을 수립해야 함

오타 교정, 내용 구체화

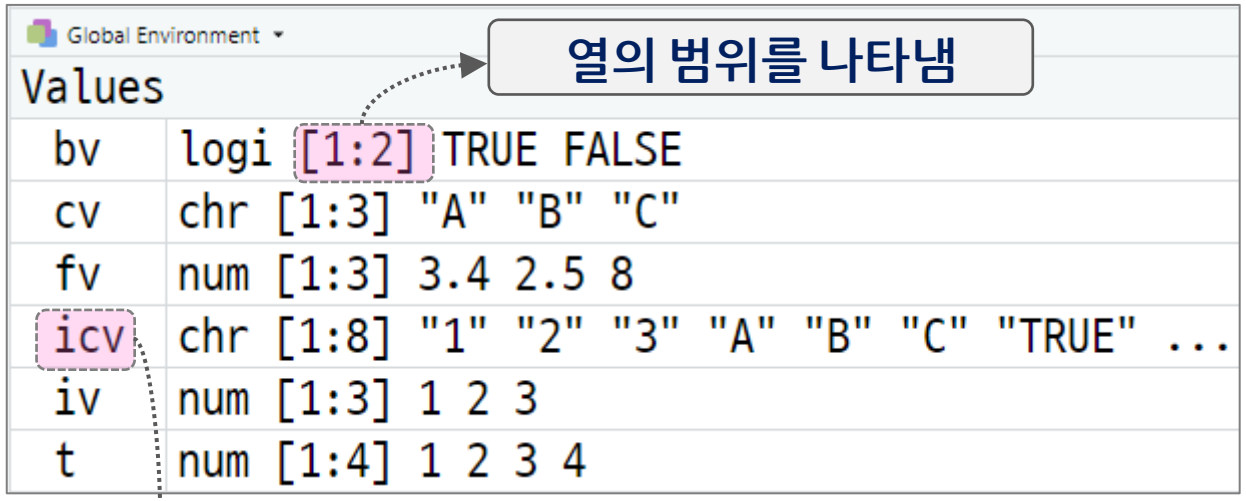
vector - c 함수로 생성

- 하나 이상의 스칼라(=길이가 1인 벡터) 원소들을 갖는 단순한 형태의 집합
- 숫자, 문자, 논리형 데이터를 원소(Element)로 사용할 수 있음
- 동일한 자료형을 갖는 값들의 집합으로 하나의 열(Column)로 구성됨

- 벡터 생성 함수 : c(value1, value2, …), seq(from, to, by), rep(x, times, each)

File : test_002.R

```
1 rm(list=ls())
2 iv <- c(1, 2, 3)
3 cv <- c('A', 'B', 'C')
4 bv <- c(TRUE, FALSE)
5 fv <- c(3.4, 2.5, 8)
6 t <- c(1, 2, 3, 4)
7 icv <- c(iv, cv, bv)
```



문자열이 포함된 서로 다른 타입을 연결할 경우 문자열 취급되어 연결됨

내용 구체화

출 : 19

독립사건

- A의 발생이 B가 발생할 확률을 바꾸지 않는 사건
- 두 사건 A, B가 독립이면 $P(B|A)=P(B)$, $P(A|B) = P(A)$, $P(A\cap B) = P(A)\cdot P(B)$ 성립
- $P(A\cup B) = P(A) + P(B) - P(A\cap B) = P(A) + P(B) - P(A)\cdot P(B)$
- 예) 주사위 던져서 나오는 눈의 값과 동전을 던져 나오는 앞/뒤 사건

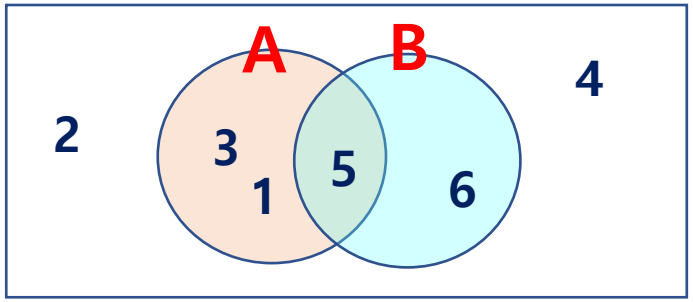
합집합 추가

배반사건

- 교집합이 공집합인 사건, 한쪽이 일어나면 다른 쪽이 일어나지 않을 때의 두 사건
- $P(A\cap B) = 0$, $P(A\cup B) = P(A) + P(B)$
- 예) 동전 하나를 던져 앞면 나오는 사건, 뒷면 나오는 사건

종속사건

- 두 사건 A와 B에서 한 사건의 결과가 다른 사건에 영향을 주는 사건
- 예) 음주와 사고 사건, $P(A\cap B) = P(A|B)\cdot P(B)$



- $P(A) = 3/6 = 1/2$
- $P(B) = 2/6 = 1/3$
- $P(A)\cdot P(B) = 1/2 * 1/3 = 1/6$

독립사건

3-11. 데이터 분석 - 문제 11

B

11. 두 개 변수, 1000개 Sample로 구성된 데이터에서 결측값을 제거하려고 한다. 결측치 비율이 변수 각각 5%이며, 두 변수가 독립일 때, 삭제되는 데이터 비율은?

1. 9.75%
2. 20%
3. 2.5%
4. 25%

3-48. 사건의 종류

독립사건의 교집합, 합집합, 조건부확률

- $P(A \cap B) = P(A) \cdot P(B)$
- $P(A \cup B) = P(A) + P(B) - P(A \cap B) = P(A) + P(B) - P(A) \cdot P(B)$
- $P(B|A)=P(B)$, $P(A|B) = P(A)$

합집합 : $(0.05+0.05) - (0.05 * 0.05) = 0.1 - 0.0025 = 0.0975 = 9.75\%$

해설 수정

3-65. 회귀모형 해석(평가방법)

출 : 18*2, 19, 22, 23, 25

표본 회귀선의 유의성 검정

- 두 변수 사이에 선형관계가 성립하는지 검정하는 것으로
  회귀식의 기울기 계수 $\beta_1 = 0$ 일 때 귀무가설, $\beta_1 \neq 0$일 때 대립가설로 설정한다

회귀모형 해석

모형이 통계적으로 유의미한가?	F 통계량, 유의확률(p-value)로 확인
회귀계수들이 유의미한가?	회귀계수의 t값, 유의확률(p-value)로 확인
모형이 얼마나 설명력을 갖는가?	결정계수(R^2) 확인
모형이 데이터를 잘 적합하고 있는가?	잔차 통계량 확인, 회귀진단 진행(선형성~ 정규성)

정규성으로 통일!

F 통계량, p-value

F 통계량 = 회귀제곱평균(MSR) / 잔차제곱평균(MSE)
F 통계량에 대한 p-value < 0.05

t 값, p-value

t 값 = Estimate(회귀계수) / Std.Error(표준오차)
t 값에 대한 p-value < 0.05

결정계수(R^2)

70~90%

분류분석의 종류

1. 로지스틱 회귀
2. 의사결정나무
3. 앙상블
4. 신경망 모형

kNN, 베이즈분류 모형, SVM(서포트벡터기계), 유전 알고리즘

유전자 알고리즘 → 유전 알고리즘

나이브베이즈 분류 모형은 학습에서 다루지 않지만, 단답으로 출제된 있습니다.

베이즈 추론을 기반으로 한 방법론의 정확도는 일반적으로 머신러닝의 대표적인 방법인 랜덤 포레스트나 트리분류 방법보다도 높다고 평가받고 있다. 베이지안 추론을 활용한 대표적 분류 방법 알고리즘.

3-82. 의사 결정 나무(Decision Tree) 모형

출 : 17, 22, 33, 34

특징

- 목적은 새로운 데이터를 분류하거나 값을 예측하는 것이다
- 분리 변수 P차원 공간에 대한 현재 분할은 이전 분할에 영향을 받는다
- 부모마디보다 자식마디의 순수도가 증가하도록 분류나무를 형성해 나간다 (불순도 감소)

종류

- 목표변수(= 종속변수)가 이산형인 경우의 분류나무(classification tree)
- 목표변수가 연속형인 경우의 회귀나무(regression tree)

장점

- 구조가 단순하여 해석이 용이함
- 비모수적 모형으로 선형성, 정규성, 등분산성 등의 수학적 가정이 불필요함
- 범주형(이산형)과 수치형(연속형) 변수를 모두 사용할 수 있음
- 상관성이 높은 변수가 있을 경우 영향을 받지 않음
- 비정상적인 잡음 데이터에 대해 민감하게 분류하지 않음

내용 추가

독립변수	종속변수
▪ 설명변수 ▪ 예측변수 ▪ Feature	▪ 목표변수 ▪ 반응변수 ▪ Label

단점

- 분류 기준 값의 경계선 부근의 자료 값에 대해서는 오차가 큼(비연속성)
- 로지스틱 회귀와 같이 각 예측변수의 효과를 파악하기 어려움
- 새로운 자료에 대한 예측이 불안정할 수 있음(Overfitting에 취약함)

19. 의사결정나무의 특징으로 알맞지 않은 것은?

1. 상관성이 높은 변수가 있어도 영향을 받지 않는다.
2. 비정상적인 잡음 데이터에 대해서는 민감하게 분류한다.
3. 목적 변수가 이산형(범주형)인 경우와 연속형인 경우 모두 사용할 수 있다.
4. 설명력이 좋으며, 과대적합에 취약한 특징이 있다.

3-82. 의사 결정 나무(Decision Tree) 모형
- 상관성이 높은 변수가 있을 경우 영향을 받지 않음
- 비정상적인 잡음 데이터에 대해 민감하게 분류하지 않음
- 과대적합(Overfitting)에 취약함
- 설명력이 좋음

3-67. 설명 변수 선택 방법

step 함수를 사용한 후진제거법

```
> step(lm(y~x1+x2+x3+x4, df), direction='backward')
Start:  AIC=26.94
y ~ x1 + x2 + x3 + x4

       Df Sum of Sq    RSS    AIC
- x3    1    0.1091 47.973 24.974
- x4    1    0.2470 48.111 25.011
- x2    1    2.9725 50.836 25.728
<none>              47.864 26.944
- x1    1   25.9509 73.815 30.576

Step:  AIC=24.97
y ~ x1 + x2 + x4

       Df Sum of Sq    RSS    AIC
<none>              47.97 24.974
- x4    1      9.93  57.90 25.420
- x2    1     26.79  74.76 28.742
- x1    1    820.91 868.88 60.629
```

- 후진제거법 : direction = 'backward'
- 전진선택법 : direction = 'forward'
- 단계선택법 : direction = 'both'

최종 선택 설명 변수 : x1, x2, x4

```
Call:
lm(formula = y ~ x1 + x2 + x4, data = df)

Coefficients:
(Intercept)           x1           x2           x4
    71.6483       1.4519       0.4161      -0.2365
```

-x3 … 24.974 : x3을 제거했을 때 AIC 가 24.974가 된다는 것입니다 그래서 제거 1순위가 되는 것입니다! (그것을 제거해야 AIC가 작은 값이 되어 좋아지니까요!).

3-83. 앙상블(Ensemble) 모형

출 : 21, 25, 27, 30*2, 34

앙상블 모형

내용 추가

- 여러 개의 분류 모형에 의한 결과를 종합하여 분류의 정확도를 높이는 방법
- 각 모형의 상호연관성이 높을수록 정확도가 떨어짐
- 적절한 표본추출법으로 데이터에서 여러 개의 훈련용 데이터 집합을 만들어 각 데이터 집합에 하나의 분류기를 만들어 결합하는 방법
- 약하게 학습 된 여러 모델들을 결합하여 사용
- 성능을 분산시키기 때문에 과적합(overfitting) 감소 효과가 있음

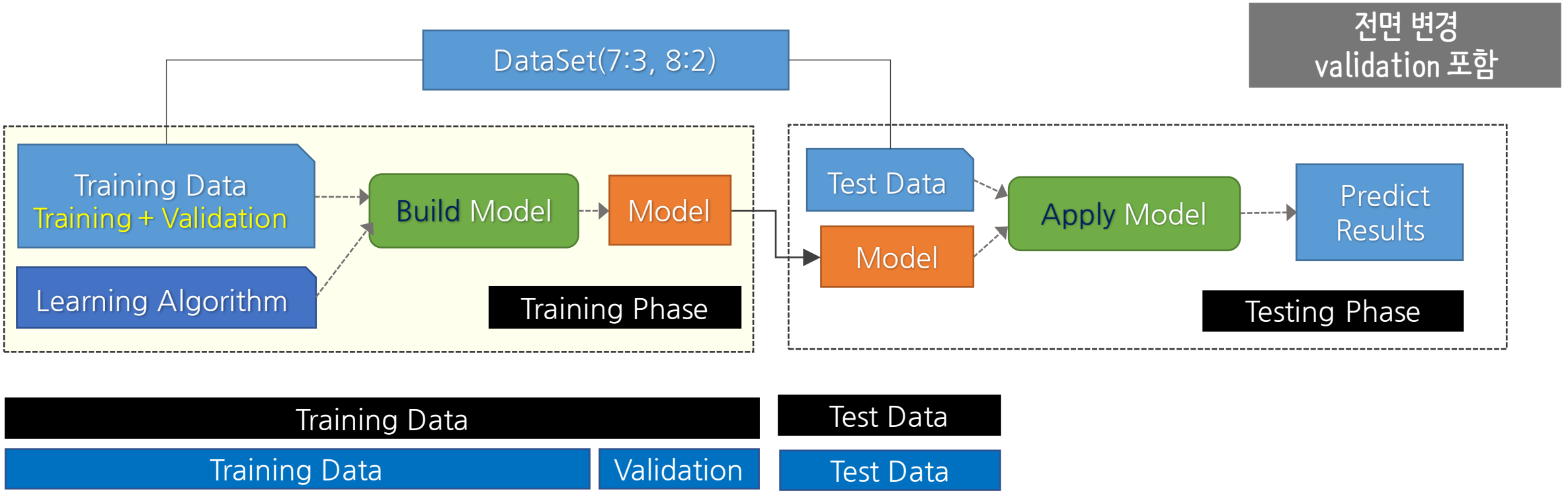
앙상블 모형의 종류

- Voting
- Bagging
- Boosting
- Random Forest

3-90. 모형 평가

출 : 17*2, 20, 22, 29, 35, 36

홀드아웃(Hold Out) | 교차검증(Cross Validation) | 붓스트랩(Bootstrap)



- Training Data : 학습용 데이터, Test Data : 학습 종료 후 성능 확인용 데이터
- Validation Data : 학습 중 성능 확인용 데이터 (Overfitting 여부 확인, Early Stopping 등을 위해 사용)

3-90. 모형 평가

홀드아웃(Hold Out) | 교차검증(Cross Validation) | 붓스트랩(Bootstrap)

- 원천 데이터를 랜덤하게 두 개로 분리하여 성능을 평가하는 방법
- 하나는 모형 학습 및 구축을 위한 훈련용 자료, 다른 하나는 성과평가를 위한 테스트용 자료로 사용
- 과대적합(overfitting) 발생 여부 확인을 위해 train, test에 대해 평가 후 비교 (train >> test면 과대적합)
- 7:3 또는 8:2 등의 비율로 학습 데이터와 테스트 데이터로 분할함
- 훈련용 데이터는 다시 훈련과 검증(Validation) 자료로 나눠 사용할 수 있음
- 즉, 데이터를 학습 세트, 검증(validataion) 세트, 테스트(test) 세트 세 가지로 분할하여 사용할 수 있음
- 일반적으로 랜덤추출 방법을 활용하여 데이터가 편향되지 않도록 함
- 학습 데이터와 테스트 데이터는 겹쳐서는 안됨, 검증 데이터와 테스트 데이터는 일치할 수 있음
- 학습데이터가 부족하면 알고리즘 학습이 어려워질 수 있음
- 불균형 데이터의 경우 불균형 범주의 비율을 유지하도록 분할하기도 함
- 잘못된 가설을 가정하게 되는 2종 오류의 발생 방지

전면 변경
validation 포함

- idx <- sample(2, nrow(iris), replace=TRUE, prob=c(0.7, 0,3))

```
trainData <- iris[idx==1,]
testData <- iris[idx==2,]
```

  - iris 데이터를 7:3 비율로 나누어 Training에서 70%, Testing에 30% 사용하도록 하는 것

20. 데이터마이닝을 위한 데이터 분할과 관련된 설명 중 알맞지 않은 것은?

1. 데이터는 학습용, 검증용, 평가용 데이터로 분할하여 사용할 수 있다.
2. 검증용 데이터(validation data)는 학습과정에서 사용되지 않는다.
3. 검증용 데이터는 훈련에 사용되지 않는다.
4. 데이터 수가 적을 때는 교차 검증을 사용한다

Hold Out
- 데이터를 학습 세트, 검증 세트, 시험(테스트) 세트 세 가지로 분할하여 사용할 수 있음
- Training Data : 학습용 데이터, Test Data : 학습 종료 후 성능 확인용 데이터
- Validation Data : 학습 중 성능 확인용 데이터 (Overfitting 여부 확인, Early Stopping 등을 위해 사용)
- 데이터 수가 적을 때는 Hold Out 보다 교차 검증을 사용함

3-95. 비계층적 군집(Non-Hierarchical Clustering)

Non-Hierarchical!

비계층적 군집 - 분할적 군집 방법 | k-중심 군집

출 : 17, 20, 22, 26, 28, 30, 31, 32, 33

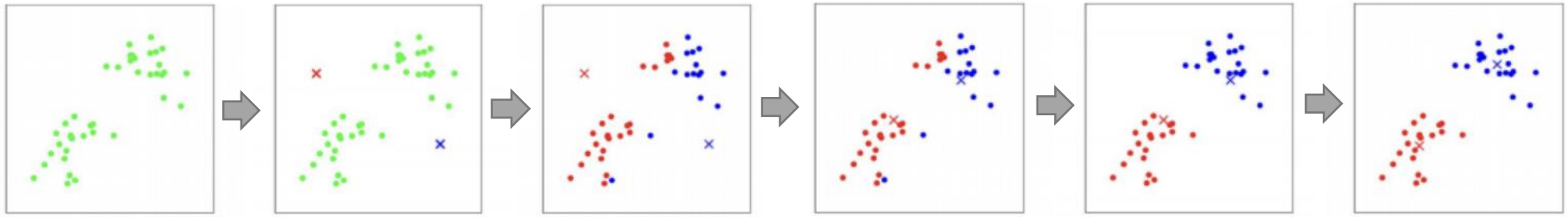
k-means

- k-mean 방법은 사전에 군집의 수 k를 정해 주어야 함 (k : hyper-parameter)
- 군집수 k가 원데이터 구조에 적합하지 않으면 좋은 결과를 얻을 수 없음
- 알고리즘이 단순하며 빠르게 수행되어 계층적 군집보다 많은 양의 자료를 처리
- k-means 군집은 잡음이나 이상값에 영향을 받기 쉬움
- k-means 분석 전에 이상값을 제거하는 것도 좋은 방법
- 평균 대신 중앙값을 사용하는 k-medoids 군집을 사용할 수 있음

Nbclust 패키지를 통해 군집 수에 대한 정보 참고

k-means 절차

1. 초기 군집의 중심으로 k개의 객체를 임의로 선택한다
2. 각 자료를 가장 가까운 군집의 중심에 할당한다
3. 각 군집 내의 자료들의 평균을 계산하여 군집의 중심을 갱신한다
4. 군집 중심의 변화가 거의 없을 때까지 2, 3을 반복한다



출처 : http://stanford.edu/~cpiech/cs221/img/kmeansViz.png

3-96. 군집 분석의 안정성, 타당성

군집 분석의 안정성(stability)과 타당성(validation)

- 군집 분석의 안정성과 타당성은 모두 중요한 요소임
  - 두 가지를 모두 고려하여 최적의 결과를 도출하여야 함 (분석의 목적과 우선 순위에 따라 달라질 수 있음)
  - 안정성, 타당성을 점검하기는 어려움
- 안정성: 군집 분석의 결과가 변동 없이 일관되게 나타나는지 나타냄
  - 안정성이 높을수록 동일 데이터에 대해 반복 수행 시 결과가 일관되어 신뢰성이 높음
  - 집단별 특성이 유사할 경우 안정성이 높을 수 있지만, 데이터 복잡성, 다양성에 따라 결과가 달라질 수 있음
- 타당성: 군집 분석 결과가 실제 데이터의 특성과 부합하는지 나타냄
  - 타당성이 높을수록 군집분석 결과가 실제로 의미 있는 구조를 잘 나타내며, 결과를 신뢰할 수 있음
- 군집 결과에 대한 타당성과 안정성에 대한 검정으로 교차타당성(cross-validation)을 이용할 수 있음
  - 데이터를 A, B 두 개 부분으로 랜덤하게 분류해 놓은 다음 각 부분에서 따로 군집분석을 한 후, 합쳐서 군집 분석한 결과와 비교하여 비슷하면 결과에 대한 안정성이 있다고 판단함
  - 지도 학습의 교차타당성과 동일한 방법은 아님 (지도학습은 결과를 평균!)

내용 추가

14. 다음 군집분석(Cluster analysis) 관한 설명 중 올바르지 않은 것은?

1. 비계층적 군집분석 기법의 경우 사용자가 사전 지식 없이 그룹의 수를 정해주는 일이 많기 때문에 결과가 잘 나오지 않을 수 있다.
2. 군집분석은 신뢰성과 타당성을 점검하기 어렵다
3. 군집 결과에 대한 안정성을 검토하는 방법으로 지도학습과 동일한 교차타당성을 이용한다.
4. 계층적 군집분석은 이상치에 민감하다.

- 군집 결과에 대한 안정성 검토는 교차타당성을 이용하지만 지도학습의 교차타당성과 다르다

2022년 34회 기출 복원 3-22번

22. 군집분석에 대한 설명으로 잘못된 것은?

1. 형성된 군집에 대해 논리성보다 안정성이 더 중요하다
2. 비지도학습으로 군집간 분산 최대화, 군집내 분산을 최소화 한다
3. 집단별 특성이 유사할 경우 안정성이 높을 수 있다
4. 유사성을 이용하여 몇 개의 집단으로 그룹화하는 분석이다

군집분석
- 유사성을 이용하여 몇 개의 집단으로 그룹화하는 분석이다
- 안정성은 일부 입력값이 변경되었을 때 군집의 변화가 유의하게 변하는지에 대한 개념이다
- 집단별 특성이 유사할 경우 안정성이 높을 수 있다
- 군집분석에 있어 군집 타당성 검증을 위해 논리성과 안정성 모두가 중요한 부분이다.

보기 변경

3-95. 비계층적 군집(Non-Hierarchical Clustering)

비계층적 군집 - 분할적 군집 방법 | k-중심 군집

출 : 17, 20, 22, 26, 28, 30, 31, 32, 33

k-means

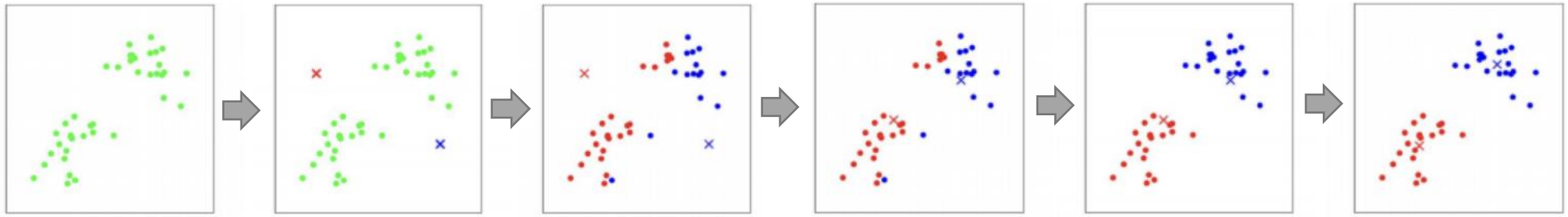
Nbclust 패키지를 통해 군집 수에 대한 정보 참고

- k-mean 방법은 사전에 군집의 수 k를 정해 주어야 함 (k : hyper-parameter)
- 군집수 k가 원데이터 구조에 적합하지 않으면 좋은 결과를 얻을 수 없음
- 알고리즘이 단순하며 빠르게 수행되어 계층적 군집보다 많은 양의 자료를 처리
- k-means 군집은 잡음이나 이상값에 영향을 받기 쉬움
- 볼록한 형태(non-convex)가 아닌 군집이 존재하면 성능이 떨어짐 예) U자 형태의 군집
- 평균 대신 중앙값을 사용하는 k-medoids 군집을 사용할 수 있음

k-means 절차

1. 초기 군집의 중심으로 k개의 객체를 임의로 선택한다
2. 각 자료를 가장 가까운 군집의 중심에 할당한다
3. 각 군집 내의 자료들의 평균을 계산하여 군집의 중심을 갱신한다
4. 군집 중심의 변화가 거의 없을 때까지 2, 3을 반복한다

내용 추가

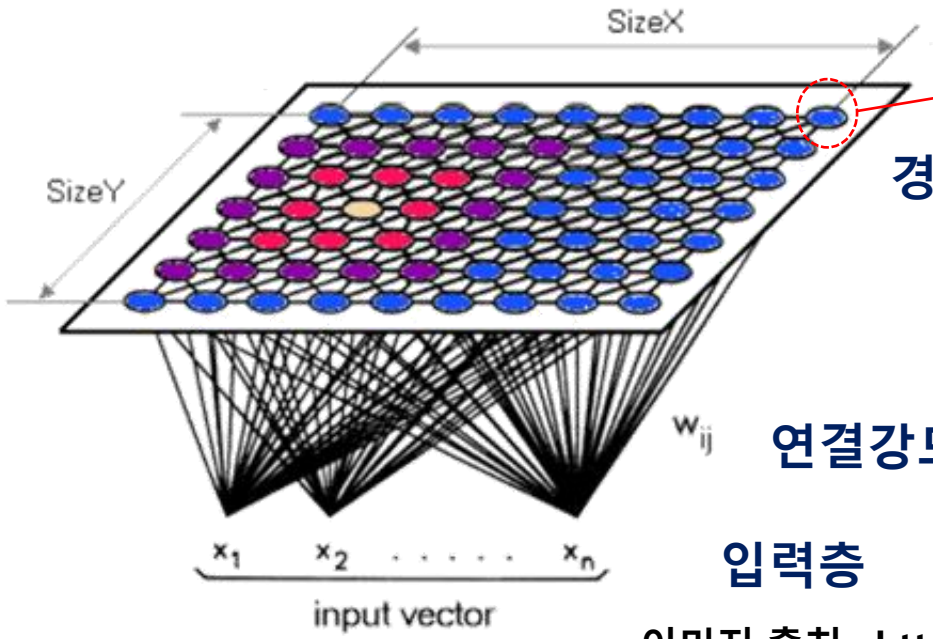


출처 : http://stanford.edu/~cpiech/cs221/img/kmeansViz.png

출 : 18, 21, 26, 28, 30, 33, 35

SOM 이란?

- 자기조직화지도
- 인공신경망의 한 종류로, 차원축소와 군집화를 동시에 수행하는 기법
- 비지도 학습(Unsupervised Learning)의 한 가지 방법
- 고차원으로 표현된 데이터를 저차원으로 변환해서 보는데 유용함
- 입력층과 2차원의 격자 형태의 경쟁층으로 이루어져 있음(2개의 층으로 구성)



노드

경쟁층

연결강도(weight)

입력층

경쟁층(=출력층)이라는 표현 제외

이미지 출처 : https://m.blog.naver.com/pmw9440/221588292503

계층적 군집의 예

- 아래는 학생들의 키와 몸무게를 정규화 한 데이터이다. **최단연결법**을 통해 학생들을 3개의 군집으로 나누면 어떻게 나누어 지는가? (Euclidean 거리 사용)

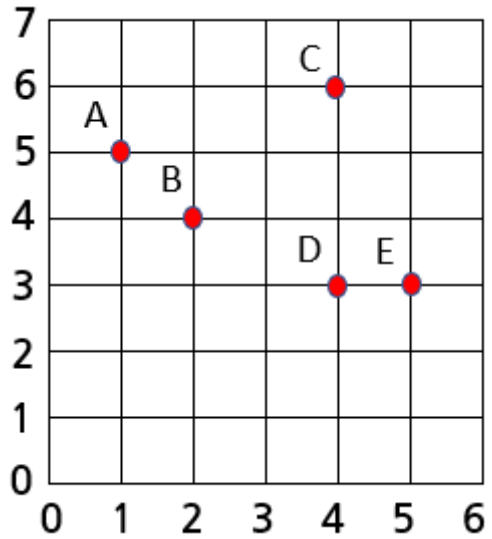
	A	B	C	D
B	2			
C	10	8		
D	13	5	9	
E	18	10	10	1

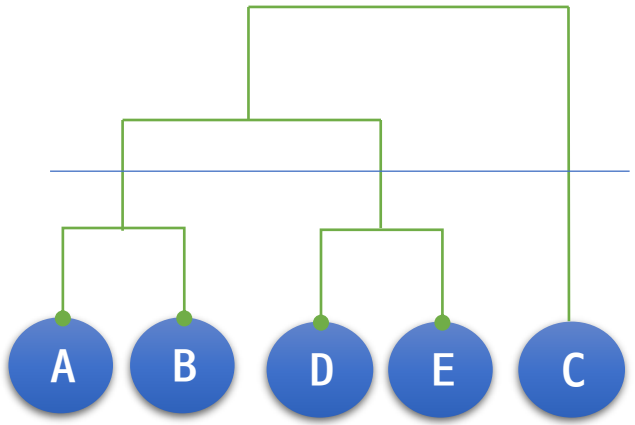
18 → 20 수정

	A	B	C
B	2		
C	10	8	
DE	13	5	9

	AB	C
C	8	
DE	5	9

	C
ABDE	8





1. 각 학생 사이의 거리를 Euclidean 거리의 제곱으로 표시한 거리표를 작성한다 (표기, 연산의 간략화)
2. 가장 작은 숫자를 찾아 가장 먼저 군집을 형성하는 것을 찾고, 최단거리표를 작성한다 (최단연결법)
3. 그 다음 작은 값들을 찾아가며 계속 군집을 만들고 최단거리표를 다시 작성한다.

2-10. 데이터 분석 기획 - 문제 10

C

10. 아래 내용은 **빅데이터 분석 방법론의** 데이터 분석 단계의 어떤 Task에 대한 설명인가?

분석용 데이터를 이용한 가설 설정을 통하여 통계모델을 만들거나 기계학습을 이용한 데이터의 분류, 예측, 군집 등의 기능을 수행하는 과정을 의미한다.

2-10. 빅데이터 분석 방법론
분석기획 - 데이터 준비 - 데이터 분석 - 시스템 구현 - 평가 및 전개

2-10-3. 데이터 분석 단계
분석용 데이터 준비 - 텍스트 분석 - 탐색적 분석 - 모델링 - 모델 평가 및 검증

문제에 '빅데이터 분석 방법론의' 추가

2과목 키워드 살펴보기

분석 방법론의 모델 3가지

폭포수 모델 - 하향식 접근 방법
나선형 모델 - 점증적 개발
프로토타입 모델 - 상향식 접근 방법

KDD 분석 방법론 (5단계)

데이터셋 선택 - 데이터 전처리 - 데이터 변환 - 데이터 마이닝 - 데이터 마이닝 결과평가

잡음, 이상치, 결측치 식별/제거

변수 선택, 차원 축소, 데이터셋 변경 작업

CRISP-DM 분석 방법론 (6단계) - 영문도!

업무(Business) 이해 - 데이터 이해 - 데이터 준비 - 모델링 - 평가 - 전개

모델 평가

분석 결과 평가
모델링 과정 평가
모델 적용성 평가

KDD	CRISP-DM
데이터셋 선택	데이터 이해
데이터 전처리 데이터 변환	데이터 준비
데이터 마이닝	모델링

CRISP-DM : Business Understanding - Data Understanding - Data Preparation - Modeling - Evaluation - Deployment

CRISP-DM 분석 방법론 - 업무 이해 순서

업무 목적 파악 - 상황 파악 - 데이터 마이닝 목표 설정 - 프로젝트 계획 수립

KDD/CRISP-DM 매치가
다른 교재와 달라요 ^^”

3과목 3 정형 데이터 마이닝 - 키워드 살펴보기

데이터 마이닝

모든 사용가능한 원천 데이터를 기반으로 감춰진 지식, 기대하지 못했던 경향 또는 새로운 규칙 등을 발견하고 이를 실제 비즈니스 의사결정 등에 유용한 정보로 활용하는 일련의 작업

데이터 마이닝 기법

- 분류, 추정, 연관분석, 예측, 군집, 기술에 대한 정의/예시
- 분류 : 답이 있는 상태(기존의 분류, 정의된 집합에 배정)
- 군집 : 미리 정의된 기준,예시 없음, 유사성에 의해 그룹화되고 이질성에 의해 세분화
- 연관 : 카탈로그 배열 및 교차판매, 마케팅 계획
- 기술 : 데이터가 가진 특징 및 의미를 단순하게 설명

로지스틱 회귀

- 종속변수 범주형(=이산형)
- 최대우도법, 가중최소자승법, χ^2 test
- Sigmoid 함수 : log odds값을 연속형 0 ~ 1의 비선형 값으로 바꾸는 함수
- 승산비(odds ratio) = 관심있는 사건이 발생할 상대 비율

분류분석의 종류

로지스틱 회귀, 의사결정나무, 앙상블, 신경망 모형, kNN, 나이브베이즈, SVM, 유전 알고리즘

의사결정 나무

- 나무 구조로 나타내 전체 자료를 몇 개의 소집단으로 분류하거나 예측을 수행하는 분석 방법
- 순수도가 높아지고 불확실성이 낮아지는 방향 분리
- 분류 : 지니, 엔트로피, 카이제곱 통계량의 p-value 작은 것
- 회귀 : 분산 감소량이 큰 것, F 통계량의 p-value 작은 것
- Gini(T) = 1 - ∑(각 범주별수/전체수)^2 1- $\sum_{i=1}^{k} P_i^2$
- Entropy(T) = - $\sum_{i=1}^{k} P_i \log_2 P_i$
- CART : 지니지수, 분산 감소량 C5.0 : 엔트로피 지수
- CHAID : 카이제곱 통계량의 p-value, ANOVA F-통계량 p-value