



2과목.빅데이터 탐색

(Ch_01. 데이터 전처리 - SEC 01. 데이터 정제
SEC 02. 분석 변수 처리)

빅데이터 분석 기사(2과목. 빅데이터 탐색)

CHAPTER 1. 데이터 전처리

CHAPTER 2. 데이터 탐색

CHAPTER 3. 통계 기법 이해

데이터 전처리

데이터 전처리 챕터는 총 2개의 작은 섹션으로 구성된다.

1. 데이터 정제
2. 분석 변수 처리

2. 데이터 전처리 – 데이터 정제

01 데이터의 정제

1) 데이터 전처리(Data pre-processing)

- 데이터 분석 업무 중 가장 많은 시간이 소요되는 단계가 데이터 수집과 전처리 단계이다.
- 분석가는 업무 시간 중 70~80% 정도를 데이터 수집 및 전처리 과정에 사용한다.
- 데이터 전처리는 여러 번 수행될 수 있다.

[전처리 과정]

데이터 정제 -> 결측값 처리 -> 이상값 처리 -> 분석 변수 처리

결측값(Missing Value) : 필수 데이터가 입력되지 않고 누락된 값

이상값(Outlier) : 데이터 범위에서 많이 벗어난 매우 크거나 작은 값

2. 데이터 전처리 – 데이터 정제

2) 데이터 정제(Data Cleansing)

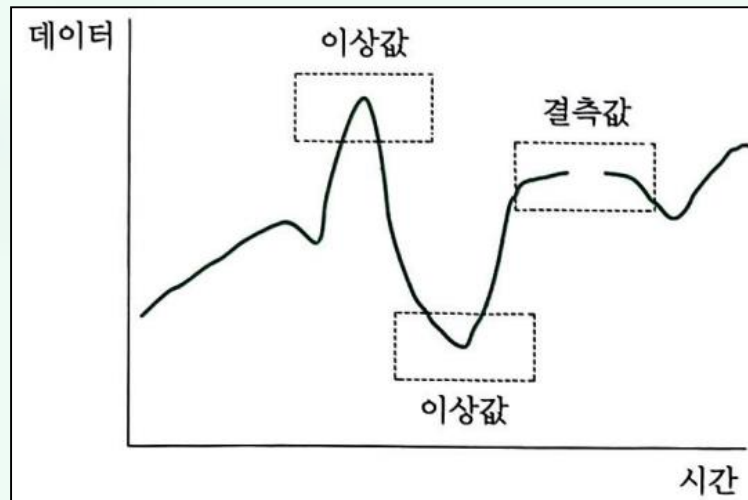
; 데이터 정제는 데이터를 깨끗하게 다듬어서 데이터의 신뢰도를 높이는 작업이라고 할 수 있다.

[정제 과정]

데이터 오류 원인 분석 -> 데이터 정제 대상 선정 -> 데이터 정제 방법 결정

① 데이터 오류 원인

- ㉠ 결측값(Missing Value) : 필수 데이터가 입력되지 않고 누락된 값(NA(Not Available), 999999, Null)
- ㉡ 노이즈(Noise) : 실제로 입력되지 않았으나 입력되었다고 잘못 판단된 값
- ㉢ 이상값(Outlier) : 데이터 범위에서 많이 벗어난 매우 크거나 작은 값



2. 데이터 전처리 – 데이터 정제

2) 데이터 정제(Data Cleansing)

- ② 데이터 정제 방법 결정 : 삭제, 대체(최빈값, 중앙값, 평균값 활용), 예측값 삽입
- ③ 데이터 일관성 유지를 위한 정제 기법
 - ㉠ 변환(Transform) : 다양한 형태로 표현된 데이터를 일관된 형태로 변환하는 작업
 - ㉡ 파싱(Parsing) : 데이터를 유의미한 최소 단위로 분할하는 작업
 - ㉢ 보강(Enhancement) : 변환, 파싱, 표준화 등을 통한 추가적인 정보를 반영하는 작업

최빈값 : 통계학 용어로, 가장 많이 관측되는 수, 즉 주어진 값 중에서 가장 자주 나오는 값이다.
중앙값: 모든 데이터를 크기 순으로 정렬해서 가운데에 있는 데이터를 선택한다.

2. 데이터 전처리 – 데이터 정제

개념 체크

1. 다음 중 데이터 전처리에 대한 설명 중 틀린 것은?

- ① 데이터 분석 업무 중에 데이터 수집 및 전처리 과정에 가장 많은 시간이 소요된다.
- ② 데이터 전처리 과정은 데이터 정제, 결측값 처리, 이상값 처리, 분석변수 처리이다.
- ③ 데이터 전처리 여부에 따라 분석 결과가 달라질 수 있다.
- ④ 데이터 전처리는 최초에 한 번만 수행된다.

데이터 전처리는 데이터 상태 및 분석 상황에 따라서 여러 번 수행될 수 있다.

2. 데이터 오류의 원인 중 하나로 필수 데이터가 입력되지 않고 누락된 값을 의미하는 것은?

- ① 결측값
- ② 노이즈
- ③ 이상값
- ④ 파생변수

데이터 오류 원인

1. 결측값(Missing Value) : 필수 데이터가 입력되지 않고

3. 다음 중 결측값을 의미하는 말이 아닌 것은?

- ① NA ② 999999
- ③ NO ④ Null

결측값을 의미하는 것은 NA, 999999, Null이다.

2. 데이터 전처리 – 데이터 정제

02 데이터 결측값 처리

; 결측값(Missing Value)이란 입력되어야 할 데이터가 입력되지 않아 누락된 값을 의미한다.

1) 데이터 결측값 종류

① 완전 무작위 결측(MCAR, Missing Completely At Random)

▶ 발생한 결측값이 다른 변수들과 아무런 연관이 없는 경우

② 무작위 결측(MAR, Missing At Random)

▶ 누락된 자료가 특정 변수와 관련되지만, 그 변수의 결과와는 관계가 없는 경우

③ 비무작위 결측(MNAR, Missing Not At Random)

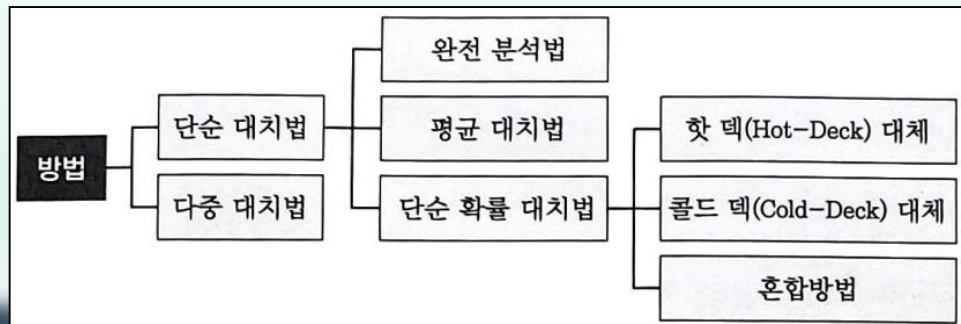
▶ 결측값이 다른 변수와 연관이 있는 경우

2) 데이터 결측값 처리 절차

결측값 식별 -> 결측값 부호화 -> 결측값 대체

2. 데이터 전처리 – 데이터 정제

3) 데이터 결측값 처리 방법



① 단순 대치법(Single Imputation)

▶ 결측값을 그럴듯한 값으로 대치하는 통계적 기법이다.

- ㉠ **완전 분석법(Completes Analysis)** : 불완전 자료는 모두 무시하고, 완전하게 관측된 자료만 사용하여 분석하는 방법이다.
- ㉡ **평균 대치법(Mean Imputation)** : 관측되어 얻어진 자료의 평균값으로 결측값을 대치하는 방법이다.
- ㉢ **단순 확률대치법(Single Stochastic Imputation)** : 적절한 확률값을 부여한 후 이를 결측값으로 대치하는 방법이다.
 - **핫덱(Hot-Deck) 대체** : 진행 중인 연구 내에서 비슷한 성향의 자료로 결측값을 대체하는 방법
 - **콜드덱(Cold-Deck) 대체** : 진행 중 연구 내부가 아닌 외부 출처 또는 이전의 비슷한 연구에서 대체 값을 가져오는 방법
 - **혼합방법** : 다양한 방법을 혼합하는 방법

2. 데이터 전처리 – 데이터 정제

3) 데이터 결측값 처리 방법

② 다중 대치법

- ▶ 단순 대치법을 한 번 하지 않고, n 번 대치를 통해 1개의 완전한 자료를 만들어 분석하는 방법이다.
대치 → 분석 → 결합의 3단계로 구성된다.

1. 데이터 수집 및 저장 계획 – 데이터 수집 및 전환

개념 체크

1. 다음 중 데이터 결측값 종류가 아닌 것은?

- ① 반무작위 결측
- ② 완전 무작위 결측
- ③ 비무작위 결측
- ④ 무작위 결측

데이터 결측값 종류

1. 완전 무작위 결측(MCAR, Missing Completely At Random)

발생한 결측값이 다른 변수들과 아무런 연관이 없는 경우

2. 무작위 결측(MAR, Missing At Random)

누락된 자료가 특정 변수와 관련되지만, 그 변수의 결과와는 관계가 없는 경우

3. 비무작위 결측(MNAR, Missing Not At Random)

결측값이 다른 변수와 연관이 있는 경우

2. 다음에 설명하는 통계적 기법은?

불완전 자료는 모두 무시하고, 완전하게 관측된 자료만 사용하여 분석하는 방법

- ① 다중 대체법
- ② 단순 확률 대체법
- ③ 평균 대체법
- ④ 완전 분석법

1. 다중 대체법(Multiple Imputation)

단순 대체법을 한 번 하지 않고, n번 대체를 통해 1개의 완전한 자료를 만들어 분석하는 방법

대치 -> 분석 -> 결합의 3단계로 구성이 된다.

2. 단순 확률 대체법

적절한 확률값을 부여한 후 이를 결측값으로 대체하는 방법

3. 평균 대체법

관측되어 얻어진 자료의 평균값으로 결측값을 대체하는 방법이다.

2. 데이터 전처리 – 데이터 정제

03 데이터 이상값 처리

; **이상값(Outlier)**이란 일반적인 데이터 범위를 많이 벗어난 아주 작은 값 또는 큰 값을 의미한다.

1) 데이터 이상값 발생원인 7가지

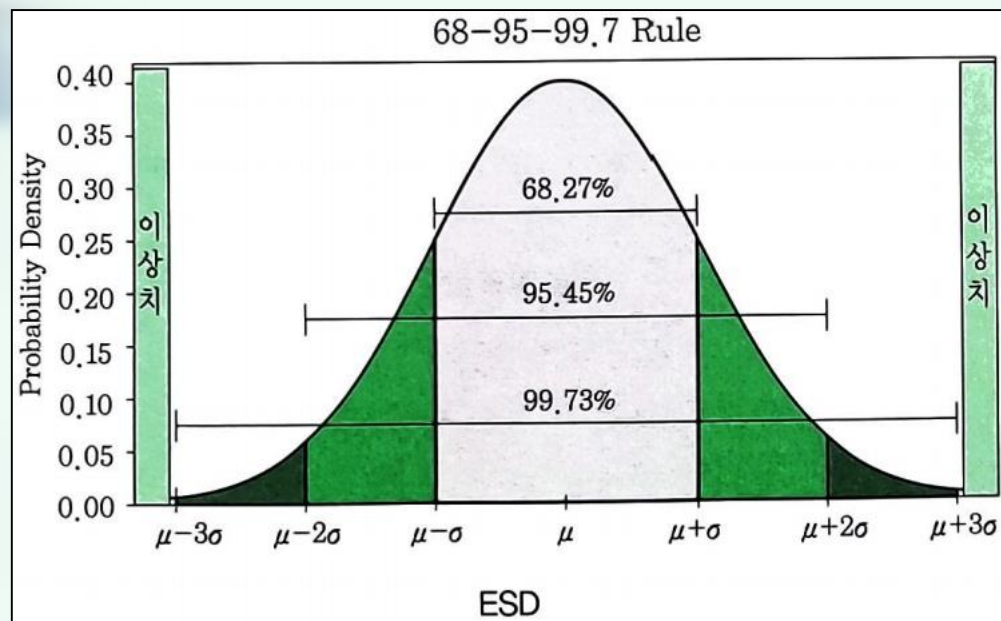
- ① **표본 추출 오류** : 데이터 표집이 잘못된 경우(이상값 포함)
- ② **고의적인 이상값** : 자기 보고식 측정(Self Reported Measures)의 경우 피실험자들이 고의적인 이상값을 기입한 경우
- ③ **데이터 입력 오류** : 데이터 수집 과정에서 발생한 입력 오류
- ④ **실험 오류** : 동일하지 않은 실험조건에서 발생하는 오류
- ⑤ **측정 오류** : 데이터 측정 과정에서 발생하는 오류
- ⑥ **데이터 처리 오류** : 여러 데이터를 처리하는 과정에서 발생하는 오류
- ⑦ **자연 오류** : 자연적으로 발생하는 오류

2. 데이터 전처리 – 데이터 정제

2) 이상값 검출 방법

① 통계 기법을 이용한 데이터 이상값 검출 방법

㉠ **ESD(Extreme Studentized Deviation)** : 평균(μ)으로부터 3시그마(σ , 표준편차) 떨어진 값을 이상치로 인식하는 방법. 3표준편차에 해당하는 값이 **99.7%**이므로 **양쪽 0.15%**에 해당하는 값을 이상치로 인식한다.



표준 편차(standard deviation) : 분산을 제곱근한 것이다. 편차들(deviations)의 제곱합(SS, sum of square)에서 얻어진 값의 평균치인 분산의 성질로부터 다시 제곱근해서 원래 단위로 만들어줌으로써 얻게 된다.

분산(variance, Var) : 확률변수가 기댓값으로부터 얼마나 떨어진 곳에 분포하는지를 가늠하는 숫자이다

확률변수 : 확률적인 결과에 따라 결과값이 바뀌는 변수를 묘사하는 통계학 및 확률론의 개념. 일정한 확률을 갖고 일어나는 사건에 수치가 부여된 것으로 해석할 수 있다.

2. 데이터 전처리 – 데이터 정제

2) 이상값 검출 방법

① 통계 기법을 이용한 데이터 이상값 검출 방법

- ㉠ **기하평균을 활용한 방법** : 기하평균으로부터 2.5시그마(σ) 떨어진 값을 이상값으로 판단하는 방법이다.
- ㉡ **사분위수를 활용한 방법** : 제1사분위(Q1), 제3사분위(Q3)를 기준으로 사분위 간 범위(IQR, $Q3-Q1$)의 1.5배 이상 떨어진 값을 이상값으로 판단하는 방법이다.
- ㉢ **표준화 점수(Z-Score)를 활용한 이상값 검출** : 평균이 μ 이고, 표준편차가 σ 인 정규분포에서 관측된 자료들이 중심인 평균에서 얼마나 떨어져 있는지에 따라 이상값으로 판단하는 방법이다.
- ㉣ **딕슨의 Q 검정 (Dixon Q-Test)** : 오름차순으로 정렬된 데이터에서 범위에 대한 관측치 간 차이의 비율을 활용하여 이상값을 검출하는 방법이다. 데이터 수가 30개 미만인 경우가 적합하다.
- ㉤ **그립스 T 검정(Grubbs T-Test)** : 정규분포의 단변량 자료에서 이상값을 검출하는 방법이다.
- ㉥ **카이제곱 검정(Chi-Square Test)** : 데이터가 정규분포 형태이지만 자료의 수가 적은 경우 이상값을 검출하는 방법이다.

Z-Score(Z-점수) : 자료가 평균으로부터 표준편차의 몇 배만큼 떨어져 있는지 보여주는 수치로 (데이터 값(X) - 평균(μ))/표준편차(σ) 로 연산한다.

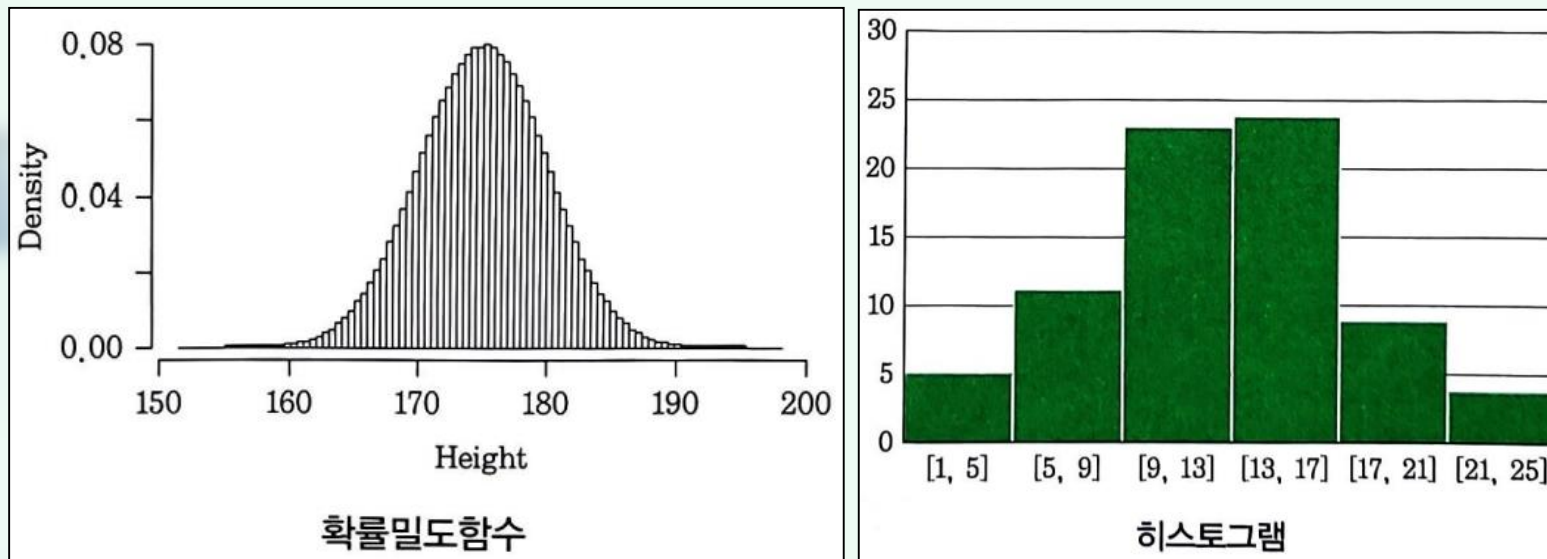
단변량 자료 : 단위에 대해 하나의 속성만 측정하여 얻게 되는 변수에 대한 자료

2. 데이터 전처리 – 데이터 정제

2) 이상값 검출 방법

② 시각화를 이용한 데이터 이상값 검출 방법

- ▶ 확률밀도함수, 히스토그램, 시계열차트, 상자수염그림 등이 있다.

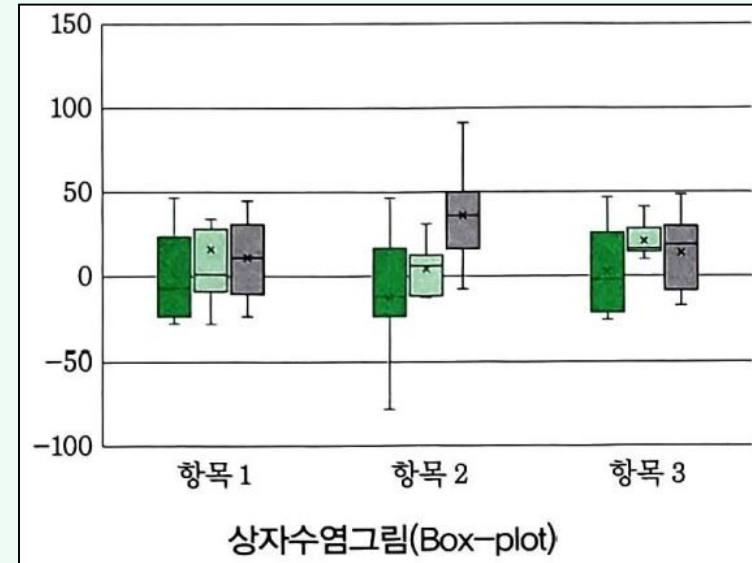
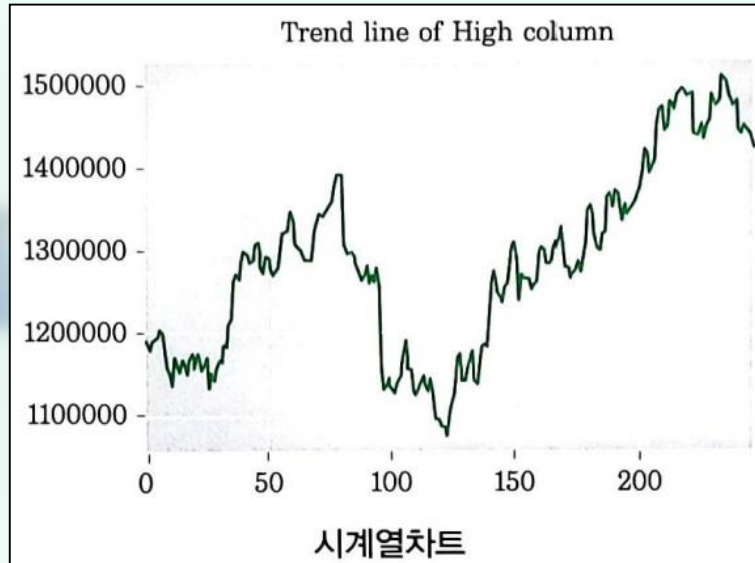


확률밀도함수 : 어떤 값이 자주 발생되는지 확인하기 위해 확률 변수가 나올 수 있는 전체 구간 ($-\infty \sim \infty$)을 아주 작은 폭을 갖는 구간들로 세분화 하여 나눈 다음 각 구간의 확률을 살펴보는 함수

2. 데이터 전처리 - 데이터 정제

2) 이상값 검출 방법

② 시각화를 이용한 데이터 이상값 검출 방법



2. 데이터 전처리 – 데이터 정제

2) 이상값 검출 방법

- ③ 머신러닝 기법(비지도 학습) 활용 : 머신러닝 기법 중 군집화 기술을 활용하여 이상값을 검출하는 방법이다.
- ④ 마할라노비스거리(Mahalanobis Distance) 활용 : 관측된 값이 평균으로부터 벗어난 정도를 측정하는 방법이다.
- ⑤ LOF(Local Outlier Factor) : 전체 데이터 분포에서 지역적인 밀집도(density)를 고려하여 이상값을 확인하는 방법이다.
- ⑥ iForest(Isolation Forest) : 거리 또는 밀도를 활용하지 않고 의사결정나무를 이용하여 이상값을 확인하는 방법이다.

2. 데이터 전처리 – 데이터 정제

3) 이상값 처리 방법

; 삭제, 대체, 변환을 통해 이상값을 제거한다.

- ① **삭제(Deleting Observations)** : 이상값으로 확인된 값을 삭제하는 방법이다.
- ② **대체(Imputation)** : 이상값을 평균 또는 중위수로 대체하는 방법이다.
- ③ **변환(Transformation)** : 극단적인 값으로 인해 발생한 이상값의 경우 데이터에 자연로그를 취해서 값을 감소시키는 방법이다.

자연로그는 기호 e 로 표기되는 특정 상수를 밑으로 하는 로그다.

2. 데이터 전처리 – 데이터 정제

개념 체크

1. 다음 중 이상값이 발생하는 원인이 아닌 것은?

- ① 자동 연산 오류
- ② 표본 추출 오류
- ③ 고의적 이상값
- ④ 데이터 입력 오류

데이터 이상값 발생원인 7가지

- 1. 표본 추출 오류 : 데이터 표집이 잘못된 경우(이상값 포함)
- 2. 고의적인 이상값 : 자기 보고식 측정의 경우 피실험자들이 고의적인 이상값을 기입한 경우
- 3. 데이터 입력 오류 : 데이터 수집 과정에서 발생한 입력 오류
- 4. 실험 오류 : 동일하지 않은 실험조건에서 발생하는 오류
- 5. 측정 오류 : 데이터 측정 과정에서 발생하는 오류
- 6. 데이터 처리 오류 : 여러 데이터를 처리하는 과정에서 발생하는 오류
- 7. 자연 오류 : 자연적으로 발생하는 오류

2. 다음 중 통계기법을 이용한 이상값 검출 방법이 아닌 것은?

3. 다음 중 시각화를 이용한 데이터 이상값 검출 방법에 사용될 수 없는 방법은?

- ① 확률밀도함수
- ② 히트맵
- ③ 히스토그램
- ④ 상자수염그림

시각화를 이용한 데이터 이상값 검출 방법에는 확률밀도함수, 히스토그램, 시계열 차트, 상자수염그림(Box-Plot) 이 있다. 히트맵(Heatmap)은 색상으로 표현할 수 있는 다양한 정보를 일정한 이미지 위에 열분포 형태의 비주얼 그래픽으로 출력한 차트이다.

4. 이상값 검출 방법 중 하나로 관측된 값이 평균으로부터 벗어난 정도를 측정하는 방법은?

- ① 카이제곱 검정 방법
- ② iForest 활용 방법
- ③ 유클리디안 거리 활용 방법
- ④ 마할라노비스 거리 활용 방법

카이제곱 검정 방법 : 통계 기법 중 하나로 데이터가 정규

2. 데이터 전처리 – 데이터 정제

5. 다음 설명의 빈칸에 알맞은 것은?

ESD(Extreme Studentized Deviation)는 평균으로부터 3표준편차 떨어진 값을 이상치로 인식하는 방법으로 3표준편차에 해당하는 값은 ()% 이다.

- ① 80.5 ② 99.7
- ③ 68.2 ④ 35.4

6. 다음 중 제3사분위(Q3)에서 제1사분위(Q1)를 뺀 범위를 나타내는 명칭은?

- ① LOF ② Outlier
- ③ IQR ④ Ridge

LOF(Local Outlier Factor) : 전체 데이터 분포에서 지역적인 밀집도(density)를 고려하여 이상값을 확인하는 방법이다.

7. 다음 중 극단적인 값으로 인해 발생한 이상값의 경우 데이터에 자연로그를 취해서 값을 감소시키는 방법은?

- ① 삭제 ② 대체
- ③ 변환 ④ 치환

이상값 처리 방법

1. 삭제 : 이상값으로 확인된 값을 삭제하는 방법이다.
2. 대체 : 이상값을 평균 또 중위수로 대체하는 방법이다.
3. 변환 : 극단적인 값으로 인해 발생한 이상값의 경우 데이터에 자연로그를 취해서 값을 감소시키는 방법이다.

2. 데이터 전처리 – 분석 변수 처리

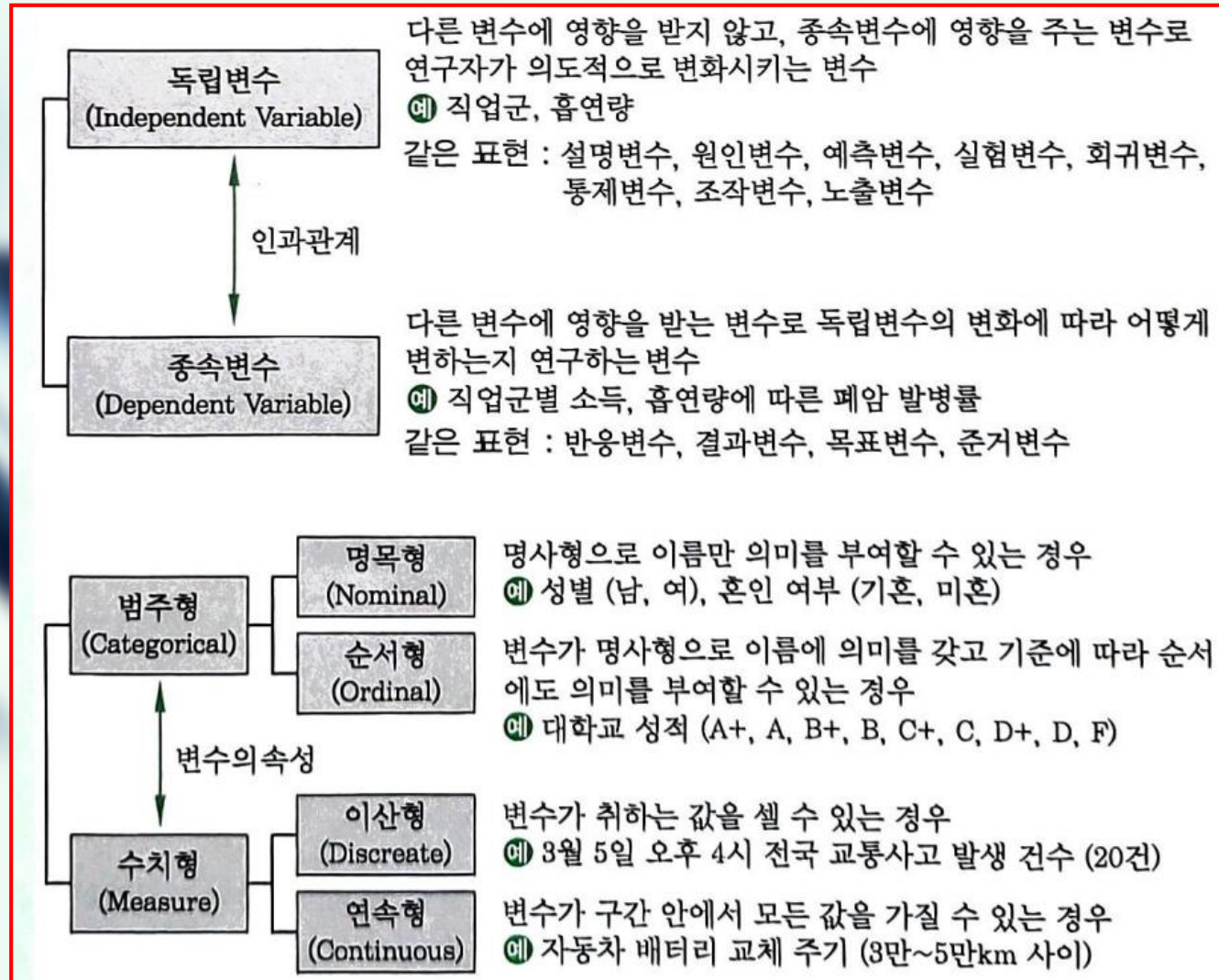
01 변수 선택

1) 변수(Feature)의 개념

- 변수란 데이터 모델에서 사용하는 예측을 수행하는 데 사용되는 **입력 변수**이다.
- 데이터 분석에서의 변수는 **Variable(변할 수 있는)**이 아닌 **Feature(특성)**를 의미한다.

2. 데이터 전처리 – 분석 변수 처리

2) 변수(Feature)의 유형



2. 데이터 전처리 – 분석 변수 처리

3) 변수 선택(Feature Selection)

- 변수 선택은 독립변수(X) 중 종속변수(Y)와 가장 관련이 깊은 변수(Feature)를 선택하는 것이다.
- 어떤 변수를 선택하느냐에 따라 분석 결과가 달라질 수 있다.

4) 변수 선택 기법

; 변수 선택 기법에는 필터 기법, 래퍼 기법, 임베디드 기법이 있다.

① 필터 기법(Filter Method)

▶ 데이터의 통계적 특성으로부터 변수를 선택하는 기법이다.

예) 정보소득, 카이제곱검정, 피셔스코어, 상관계수

정보 소득(Information Gain)	가장 정보 소득이 높은 속성을 선택하여 데이터를 더 잘 구분하게 되는 것
카이제곱 검정(Chi-Square Test)	카이제곱 분포에 기초한 통계적 방법으로 관찰된 빈도가 기대되는 빈도와 의미있게 다른지 여부를 검증하기 위해 사용되는 검증 방법
피셔 스코어(Fisher Score)	최대 가능성 방정식을 풀기 위해 통계에 사용되는 뉴턴(Newton)의 방법
상관계수(Correlation Coefficient)	두 변수 사이의 통계적 관계를 표현하기 위해 특정한 상관관계의 정도를 수치적으로 나타낸 계수

뉴턴법(Newton's method) : 뉴턴-랩슨법(Newton-Raphson method)이라고도 불리는데, 방정식 $f(x) = 0$ 의 해를 근사적으로 찾을 때 유용하게 사용되는 방법이다.

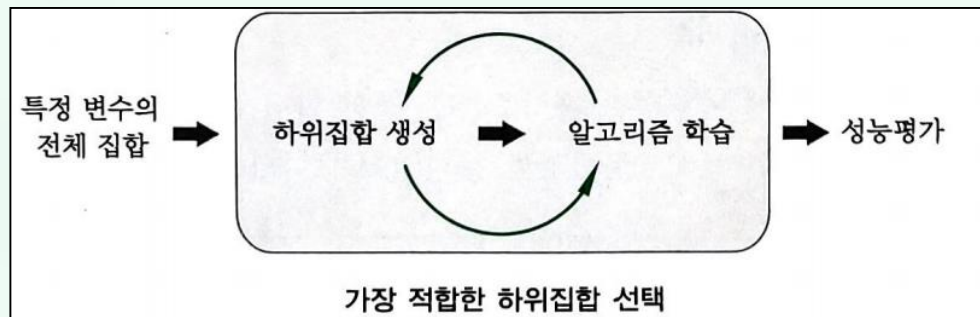
2. 데이터 전처리 – 분석 변수 처리

4) 변수 선택 기법

② 래퍼 기법(Wrapper Method)

- ▶ 변수의 일부분만 모델링에 사용하고, 그 결과를 확인하는 작업을 반복하면서 변수를 선택해가는 기법으로, 예측정확도 측면에서 가장 좋은 성능을 보이는 하위집합을 선택하는 기법이다.
- ▶ 변수 선택을 위한 알고리즘 유형 - 전진 선택법, 후진 소거법, 단계적 방법
 - 전진 선택법(Forward Selection) : 가장 큰 영향을 주는 변수를 하나씩 추가하는 방법
 - 후진 소거법 (Backward Elimination) : 가장 적은 영향을 주는 변수를 하나씩 제거하는 방법
 - 단계적 방법(Stepwise Method) : 전진 선택과 후진 소거 방법을 함께 사용하는 방법

래퍼 기법 사례 : RFE, SFS, 유전알고리즘, 단변량 선택, mRMR



RFE(Recursive Feature Elimination) : 재귀적으로 제거

SFS(Sequential Feature Selection) : 빈 모델에 하나씩 추가

유전 알고리즘(Genetic Algorithm) : 자연세계 진화과정에 기초한 전역 최적화 기법 (존 홀랜드, 1975)

단변량 선택(Univariate Selection) : 각 변수를 개별 검사 → 변수와 반응변수간 관계 강도 결정

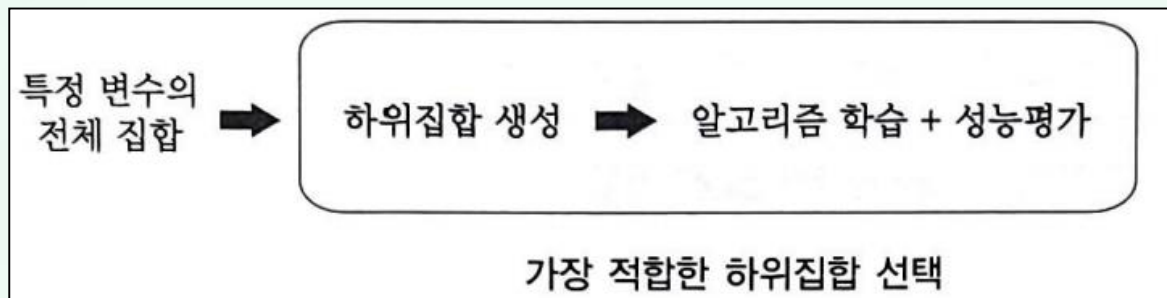
mRMR(Minimum Redundancy Maximum Relevance) : 특성변수의 중복성 최소화하는 기법

2. 데이터 전처리 – 분석 변수 처리

4) 변수 선택 기법

③ 임베디드 기법(Embedded Method)

- ▶ 모델 자체에 변수 선택이 포함된 기법으로 모델의 학습 또는 생성 과정에서 최적의 변수를 선택하는 기법이다.
- ▶ 임베디드 기법 사례 - 라쏘, 릿지, 엘라스틱넷, Select From Model
 - 라쏘(LASSO) : L1-norm을 통해 제약, 가중치 절댓값의 합을 최소화하는 방법
 - 릿지(Ridge) : L2-norm을 통해 제약, 가중치들의 제곱합을 최소화하는 방법
 - 엘라스틱넷(Elastic Net) : 라쏘(LASSO)와 릿지(Ridge) 두 기법을 선형 결합한 방법
 - Select From Model : 의사결정나무 기반 알고리즘에서 변수를 선택하는 방법



L1-norm : 벡터 p, q 각 원소간 차이의 절댓값의 합

L2-norm : 유클리디안 거리(직선 거리)

2. 데이터 전처리 – 분석 변수 처리

개념 체크

1. 다음 중 변수에 대한 설명으로 옳지 않은 것은?

- ① 데이터 분석에서 변수는 Variable을 의미한다.
- ② 변수는 인과관계에 따라 독립변수와 종속 변수로 나뉜다.
- ③ 독립변수는 연구자가 의도적으로 변화시키는 변수이다.
- ④ 독립변수는 X로, 종속변수는 Y로 표시한다.

변수란 데이터 모델에서 사용하는 예측을 수행하는데 사용되는 입력 변수이다.

데이터 분석에서의 변수는 Variable(변할 수 있는)이 아닌 **Feature**(특성, 속성, 특징, 컬럼, Attribute)를 의미한다.

독립변수 : 다른 변수에 영향을 받지 않고, 종속변수에 영향을 주는 변수로 연구자가 의도적으로 변화시키는 변수

종속변수 : 다른 변수에 영향을 받는 변수로 독립변수의 변화에 따라 어떻게 변하는지 연구하는 변수

독립변수와 종속변수는 인과관계이다.

변수 선택은 독립변수(X) 중 종속변수(Y)와 가장 관련이 깊은 변수를 선택하는 것이다. 어떤 변수를 선택하느냐에 따라 분석 결과가 달라질 수 있다.

3. 다음 중 변수 선택 기법이 아닌 것은?

- ① 필터 기법
- ② 피쳐 기법
- ③ 임베디드 기법
- ④ 래퍼 기법

1. 필터 기법 : 데이터의 통계적 특성으로부터 변수를 선택하는 기법이다.

예) 정보 소득, 카이제곱 검정, 피셔스코어, 상관계수

2. 래퍼 기법 : 변수의 일부분만 모델링에 사용하고, 그 결과를 확인하는 작업을 반복하면서 변수를 선택해가는 기법

전진 선택법 : 가장 큰 영향을 주는 변수를 하나씩 추가 하는 방법

후진 소거법 : 가장 적은 영향을 주는 변수를 하나씩 제거 하는 방법

단계적 방법 : 전진 선택과 후진 소거 방법을 함께 사용하는 방법

3. 임베디드 기법 : 모델 자체에 변수 선택이 포함된 기법으로 모델의 학습 또는 생성 과정에서 최적의 변수를 선택

2. 데이터 전처리 – 분석 변수 처리

개념 체크

1. 다음 중 변수에 대한 설명으로 옳지 않은 것은?

- ① 데이터 분석에서 변수는 Variable을 의미한다.
- ② 변수는 인과관계에 따라 독립변수와 종속 변수로 나뉜다.
- ③ 독립변수는 연구자가 의도적으로 변화시키는 변수이다.
- ④ 독립변수는 X로, 종속변수는 Y로 표시한다.

변수란 데이터 모델에서 사용하는 예측을 수행하는데 사용되는 입력 변수이다.

데이터 분석에서의 변수는 Variable(변할 수 있는)이 아닌 **Feature**(특성, 속성, 특징, 컬럼, Attribute)를 의미한다.

독립변수 : 다른 변수에 영향을 받지 않고, 종속변수에 영향을 주는 변수로 연구자가 의도적으로 변화시키는 변수

종속변수 : 다른 변수에 영향을 받는 변수로 독립변수의 변화에 따라 어떻게 변하는지 연구하는 변수

독립변수와 종속변수는 인과관계이다.

변수 선택은 독립변수(X) 중 종속변수(Y)와 가장 관련이 깊은 변수를 선택하는 것이다. 어떤 변수를 선택하느냐에 따라 분석 결과가 달라질 수 있다.

3. 다음 중 변수 선택 기법이 아닌 것은?

- ① 필터 기법
- ② 피쳐 기법
- ③ 임베디드 기법
- ④ 래퍼 기법

1. 필터 기법 : 데이터의 통계적 특성으로부터 변수를 선택하는 기법이다.

예) 정보 소득, 카이제곱 검정, 피셔스코어, 상관계수

2. 래퍼 기법 : 변수의 일부분만 모델링에 사용하고, 그 결과를 확인하는 작업을 반복하면서 변수를 선택해가는 기법

전진 선택법 : 가장 큰 영향을 주는 변수를 하나씩 추가 하는 방법

후진 소거법 : 가장 적은 영향을 주는 변수를 하나씩 제거 하는 방법

단계적 방법 : 전진 선택과 후진 소거 방법을 함께 사용하는 방법

3. 임베디드 기법 : 모델 자체에 변수 선택이 포함된 기법으로 모델의 학습 또는 생성 과정에서 최적의 변수를 선택

2. 데이터 전처리 – 분석 변수 처리

5. 다음 중 모델 자체에 변수 선택이 포함된 기법으로 모델의 학습 또는 생성 과정에서 최적의 변수를 선택하는 기법을 의미하는 것은?

- ① 임베디드 기법
- ② 필터 기법
- ③ 래퍼 기법
- ④ 결합 기법

6. 다음 중, 임베디드 기법 사례가 아닌 것은?

- ① 라쏘 ② 릿지
- ④ 엘라스틱넷 ④ SFS

SFS(Sequential Feature Selection) : 래퍼 기법 사례 중 하나로써 빈 모델에 하나씩 추가

라쏘(LASSO) : L1-norm을 통해 제약, 가중치 절댓값의 합을 최소화하는 방법

릿지(Ridge) : L2-norm을 통해 제약, 가중치들의 제곱합을 최소화하는 방법

엘라스틱넷(Elastic Net) : 라쏘와 릿지 두 기법을 선형적

2. 데이터 전처리 – 분석 변수 처리

02 차원 축소

1) 차원축소(Dimensionality Reduction)의 정의

- 막연히 데이터의 개수가 많다고 하여 정확한 분석 결과를 얻을 수 있는 것은 아니기 때문에 원활한 데이터 분석 작업을 위해 차원축소 기법을 사용한다.
- 차원축소는 분석에 활용되는 데이터의 변수 정보는 최대한 유지하면서 데이터 세트 변수의 개수를 줄이는 데이터 분석 기법이다.

2. 데이터 전처리 – 분석 변수 처리

2) 차원 축소 기법

; 차원 축소 기법에는 주성분 분석(PCA), 선형 판별 분석(LDA), 특이값 분해(SVD), 요인 분석, 독립성분 분석(ICA), 다차원 척도법(MDS)이 있다.

① 주성분 분석(PCA : Principal Component Analysis)

- ▶ 가장 보편적으로 사용되는 차원 축소 기법 중 하나로 원본 데이터를 최대한 보존하면서 고차원 공간의 데이터를 저차원 공간 데이터로 변환하는 기법이다.
- ▶ 기존 변수들을 조합하여 서로 연관성이 없는 새로운 변수(주성분 PC, Principal Component)를 생성한다.
- ▶ 행과 열의 크기가 같은 정방행렬에서만 사용한다.
- ▶ 첫 번째 주성분(PC1)은 원 데이터의 분포를 가장 많이 보존하고, 두 번째 주성분(PC2)이 그 다음으로 원 데이터의 분포를 많이 보존한다.

② 선형 판별 분석(LDA : Linear Discriminant Analysis)

- ▶ 데이터를 특정한 직선(축)에 사영(projection)하여 두 범주를 잘 구분할 수 있는 직선을 찾는 기법이다. 사영 또는 투영은 어떤 집합을 부분집합으로 특정한 조건을 만족시키면서 옮기는 작용이다.

2. 데이터 전처리 – 분석 변수 처리

2) 차원축소 기법

③ 특이값 분해(SVD: Singular Value Decomposition)

- ▶ 주성분 분석과 유사하나 행과 열의 크기가 다른 임의의 $M \times N$ 차원의 행렬에서 특이값을 추출하여 효율적으로 차원을 축소하는 기법이다.

④ 요인 분석(Factor Analysis)

- ▶ 변수들 간의 상관관계를 고려하여 유사한 변수끼리 묶어서 변수의 요인(Factor)을 축소시키는 차원 축소 기법이다.
- ▶ 실제 결과를 초래하게 되는 잠재 요인을 찾아냄으로써 데이터 안의 구조를 확인하는 기법이다.

⑤ 독립성분 분석(ICA: Independent Component Analysis)

- ▶ 데이터를 가장 잘 설명할 수 있는 축을 찾는 주성분 분석(PCA)과 다르게 가장 독립적인 축을 찾는 기법이다.
- ▶ 다변량의 신호를 통계적으로 독립적인 하부 성분으로 분리하여 차원을 축소하는 기법이다.
- ▶ 비정규 분포를 따르는 데이터들의 관계를 독립적으로 변환시키는 방법이다.

다변량 분석(Multivariate analysis) : 여러 현상이나 사건에 대한 측정치를 개별적으로 분석하지 않고 동시에 한번에 분석하는 통계적 기법을 말한다. 즉 여러 변인들 간의 관계성을 동시에 고려해 그 효과를 밝히는 것이다. 이때 여러 변인을 동시에 고려 하려다 보니 다변량 분포는 평면상의 면적이 아니라 공간상의 입체적 표현이 필요하게 된다.

2. 데이터 전처리 – 분석 변수 처리

2) 차원축소 기법

⑥ 다차원척도법(MDS: Multi-Dimensional Scaling)

- ▶ 군집분석과 유사하게 개체들 사이의 유사성과 비유사성을 측정하여 개체들을 2차원 혹은 3차원 공간상에 점으로 표현하여 개체들 간의 근접성(Proximity)을 시각적으로 표현할 수 있는 차원 축소 기법이다.

군집분석(cluster analysis) : 동일한 성격을 가진 여러 개의 그룹을 대상을 분류하는 것을 말한다. 여기서 나뉜 부분집단을 군집이라 칭한다. 유사한 성격을 가지는 몇 개의 군집으로 집단화 한 후, 형성된 군집들의 특성을 파악하여 군집들 사이의 관계를 분석하고 데이터 전체의 구조에 대한 이해를 돕고자 하는 탐색적 분석방법이다.

2. 데이터 전처리 – 분석 변수 처리

개념 체크

1. 다음 중 차원축소(Dimensionality Reduction)에 대한 설명으로 옳지 않은 것은?

- ① 데이터의 개수가 많으면 정확한 분석 결과를 얻을 수 있다.
- ② 원활한 데이터 분석 작업을 위해 차원축소 기법을 사용한다.
- ③ 차원축소는 분석에 활용되는 데이터의 변수 정보는 최대한 유지한다.
- ④ 데이터 세트 변수의 개수를 줄이는 데이터 분석 기법이다.

차원축소

막연히 데이터의 개수가 많다고 하여 정확한 분석 결과를 얻을 수 있는 것은 아니기 때문에 원활한 데이터 분석 작업을 위해 차원축소 기법을 이용한다.

차원축소는 분석에 활용되는 데이터의 변수 정보는 최대한 유지하면서 데이터 세트 변수의 개수를 줄이는 데이터 분석 기법이다.

2. 다음 설명은 차원축소 기법 중 무엇을 의미하는 것인가?
가장 보편적으로 사용되는 차원 축소 기법 중 하나로

3. 다음 설명은 차원축소 기법 중 무엇을 의미하는 것인가?
군집분석과 유사하게 개체들 사이의 유사성과 비유사성을 측정하여 개체들을 2차원 혹은 3차원 공간상에 점으로 표현하여 개체들 간의 근접성 (Proximity)을 시각적으로 표현할 수 있는 차원 축소 기법이다.

- ① 특이값 분해(SVD: Singular Value Decomposition)
- ② 요인 분석(Factor Analysis)
- ③ 다차원척도법(MDS: Multi-Dimensional Scaling)
- ④ 독립성분 분석(ICA: Independent Component Analysis)

독립성분 분석(ICA: Independent Component Analysis)

- ▶ 데이터를 가장 잘 설명할 수 있는 축을 찾는 주성분 분석(PCA)과 다르게 가장 독립적인 축을 찾는 기법이다.
- ▶ 다변량의 신호를 통계적으로 독립적인 하부 성분으로 분리하여 차원을 축소하는 기법이다.
- ▶ 비정규 분포를 따르는 데이터들의 관계를 독립적으로 변환시키는 방법이다.

2. 데이터 전처리 – 분석 변수 처리

03 파생변수의 생성

1) 파생변수(Derived Variable)

; 파생변수(유도변수)란 기존 변수에 특정 조건 혹은 함수 등을 사용하여 새롭게 재정의한 변수를 의미한다. 변수를 생성할 때는 논리적 타당성 및 명확한 기준을 갖도록 한다.

2) 파생변수(Derived Variable) 생성 방법

방법	설명
단위 변환	주어진 변수의 단위를 새로운 단위로 변환하여 데이터를 표현하는 방법 예) 온도를 100℃에서 50℃로 변환
표현방식 변환	표현방식을 단순화하는 변환 방법 예) 미혼, 기혼 데이터를 0, 1로 변환
요약 통계량 변환	요약 통계량을 활용하는 변환 방법 예) 성별에 따른 A제품 재구매율 확인
정보 추출	하나의 정보에서 새로운 정보를 추출하는 방법 예) 차량 번호판에서 개인소유 혹은 렌터카 여부 확인
변수 결합	변수를 결합하여 새로운 변수를 정의하는 방법 예) 타이타닉 생존자 데이터에서 형제, 부모 데이터를 가족 데이터로 결합
조건문 이용	조건문을 활용하여 파생변수를 생성하는 방법 예) 백화점 구매 데이터에서 고객의 1년 구매액이 8,000만원 이상인 경우 A 그룹, 아닌 경우 B 그룹의 변수로 분류

2. 데이터 전처리 – 분석 변수 처리

3) 인코딩(Encoding)

- 인코딩은 데이터의 형태나 형식을 변환하는 처리 방법으로 데이터 분석에서는 문자열 데이터를 숫자형 데이터로 변환하는 기술을 의미한다.
- 인코딩의 종류에는 원-핫 인코딩, 레이블 인코딩, 카운트 인코딩, 대상 인코딩이 있다.

① 원-핫 인코딩(One-Hot Encoding)

- ▶ 원-핫 인코딩은 표현하고자 하는 데이터를 1값으로, 그렇지 않은 데이터를 0값으로 표현하는 방식이다.

분류	서울	경기	강원	제주
서울	1	0	0	0
경기	0	1	0	0
강원	0	0	1	0
강원	0	0	1	0
서울	1	0	0	0
경기	0	1	0	0
제주	0	0	0	1


원-핫 인코딩 예시

2. 데이터 전처리 - 분석 변수 처리

3) 인코딩(Encoding)

② 레이블 인코딩(Labeled-Encoding)

▶ 레이블 인코딩은 범주형 변수의 문자열 데이터를 수치형으로 변환하는 방식이다.

분류		분류
서울		0
경기		1
강원		2
강원		2
서울		0
경기		1
제주		3

레이블 인코딩 예시

2. 데이터 전처리 – 분석 변수 처리

3) 인코딩(Encoding)

③ 카운트 인코딩(Count Encoding)

▶ 카운트 인코딩은 각 범주의 데이터 개수를 총합하여 그 개수의 수치값을 인코딩하는 방식이다.

분류		분류	횟수
서울		서울	2
경기		경기	2
강원		강원	2
강원		강원	2
서울		서울	2
경기		경기	2
제주		제주	1
카운트 인코딩 예시			

2. 데이터 전처리 - 분석 변수 처리

3) 인코딩(Encoding)

④ 대상 인코딩(Target Encoding)

- ▶ 대상 인코딩은 범주형 데이터의 값들을 목표하는 데이터 값으로 바꿔주는 방식이다.
- ▶ 대상 인코딩은 원-핫 인코딩에서 변수의 값이 많아지는 문제를 해결해 준다.

분류	전체 지역 업체 수		분류	단위 지역당 업체 수(평균)
서울	120		서울	50
경기	80		경기	30
강원	40		강원	20
강원	40		강원	20
서울	120		서울	50
경기	80		경기	30
제주	30		제주	15

대상 인코딩 예시

2. 데이터 전처리 – 분석 변수 처리

개념 체크

1. 다음 설명의 파생변수(Derived Variable) 생성 방법 중 어느 것인가?

백화점 구매 데이터에서 고객의 1년 구매액이 8,000만원 이상인 경우 A 그룹, 아닌 경우 B 그룹의 변수로 분류

- ① 조건문 이용
- ② 변수 결합
- ③ 단위 변환
- ④ 요약 통계량 변환

파생변수 생성 방법

- 1. 단위 변환 : 주어진 변수의 단위를 새로운 단위로 변환하여 데이터를 표현하는 방법
- 2. 표현방식 변환 : 표현방식을 단순화하는 변환 방법
- 3. 요약 통계량 변환 : 요약 통계량을 활용하는 변환 방법
- 4. 정보 추출 : 하나의 정보에서 새로운 정보를 추출하는 방법
- 5. 변수 결합 : 변수를 결합하여 새로운 변수를 정의하는 방법
- 6. 조건문 이용 : 조건문을 활용하여 파생변수를 생성하는

2. 다음 중 인코딩의 종류로 틀린 것은?

- ① 원-핫 인코딩(One-Hot Encoding)
- ② 레이블 인코딩(Labeled-Encoding)
- ③ 카운트 인코딩(Count Encoding)
- ④ 테이블 인코딩(Table Encoding)

원-핫 인코딩 : 표현하고자 하는 데이터를 1값으로, 그렇지 않은 데이터를 0값으로 표현하는 방식이다.

레이블 인코딩 : 범주형 변수의 문자열 데이터를 수치형으로 변환하는 방식이다.

카운트 인코딩 : 각 범주의 데이터 개수를 총합하여 그 개수의 수치값을 인코딩하는 방식이다.

대상 인코딩 : 범주형 데이터의 값들을 목표하는 데이터 값으로 바꿔주는 방식이다.

2. 데이터 전처리 – 분석 변수 처리

04 변수 변환

1) 변수 변환의 정의

; 변수 변환(Variable Transformation)이란 데이터 분석을 위해 불필요한 변수를 제거하고, 변수를 변환하여 새로운 변수를 생성시키는 작업이다.

2) 변수 변환의 방법

; 변수 변환 방법에는 단순 기능 변환, 비닝, 정규화, 표준화, 박스-콕스 변환이 있다.

① 단순 기능 변환(Simple Functions)

▶ 한쪽으로 치우친 변수를 변환하여 분석 모형을 적합하게 만드는 방법이다.

예) 지수/로그 변환, 루트 변환

② 비닝(구간화, Binning)

▶ 데이터 값을 몇 개의 Bin으로 분할하여 계산하는 방법이다.

▶ 구조화, 연속형 변수를 특정 구간으로 나누어 범주형 또는 순위형 변수로 변환하는 방법이다.

예) 연령별 데이터를 40대, 50대, 60대 이상의 범주로 나누기

2. 데이터 전처리 – 분석 변수 처리

2) 변수 변환의 방법

③ 스케일링(Scaling)

▶ 데이터의 성질은 유지한 채 데이터의 범위를 조정하는 방법이다.

예) 정규화, 표준화

④ 정규화(Normalization)

▶ 데이터의 값을 0~1 사이의 값으로 변환하는 방법이다.

예) 최소-최대 정규화, Z-점수 정규화

정규화 : 데이터의 값을 공통 척도 또는 비슷한 값 분포에 맞추는 데이터 변환 프로세스이다. 즉, 정규화의 목적은 Dataset의 범위의 차이를 왜곡하지 않고 비슷한 정도의 Scale로 반영되도록 변경하는 것이다.

최소-최대 정규화 : 데이터를 정규화하는 가장 일반적인 방법이다. 모든 feature에 대해 각각의 최소값 0, 최대값 1로, 그리고 다른 값들은 0과 1 사이의 값으로 변환하는 거다.

z-점수 정규화(z-score Normalization) : 전체 데이터의 평균을 0, 표준편차를 1로 만드는 정규화 방법이다.

2. 데이터 전처리 – 분석 변수 처리

2) 변수 변환의 방법

⑤ 표준화(Standardization)

- ▶ 입력된 데이터를 평균이 0이고, 분산이 1인 표준 정규 분포로 변환하는 방법이다.
- ▶ 평균 0을 중심으로 양쪽으로 데이터를 분포시키는 방법이다.

예) StandardScaler, RobustScaler

⑥ 박스-콕스 변환(Box-Cox Transformation)

- ▶ 정규성에 맞지 않은 변수를 정규분포에 가깝게 로그/지수 변환하는 방법으로 데이터의 분산을 안정화하는 기법이다.

StandardScaler : Sklearn(사이킷런)에서 제공하는 표준화를 위한 클래스이며, 개별 변수를 평균이 0이고 분산이 1인 가우시안 분포를 가질 수 있도록 값을 변환해준다.

가우시안 분포는 정규 분포와 같은 의미로 연속 확률 분포의 하나이다.

RobustScaler : 데이터의 중앙값 = 0, IQR = 1이 되도록 스케일링하는 기법이다.

2. 데이터 전처리 – 분석 변수 처리

개념 체크

1. 다음 중 변수 변환 방법에 속하지 않는 것은?

- ① 단순 기능 변환
- ② 비닝
- ③ 스케일링

④ 과소 표집

과소표집(Under-Sampling)은 불균형 데이터 처리 기법 중 하나이다.

▶ 다수 클래스의 데이터 중 일부만 선택하여 데이터 비율을 맞추는 방법이다.

▶ 데이터 소실 가능성과 중요한 데이터를 잃을 가능성이 높다.

변수 변환 방법

1. 단순 기능 변환

▶ 한쪽으로 치우친 변수를 변환하여 분석 모형을 적합하게 만드는 방법이다.

2. 비닝(구간화)

▶ 데이터 값을 몇 개의 Bin으로 분할하여 계산하는 방법

2. 다음 설명은 어떤 변수 변환 방법인가?

정규성에 맞지 않은 변수를 정규분포에 가깝게

로그/지수 변환하는 방법으로 데이터의 분산을 안정화하는 기법이다.

① 정규화

② 비닝

③ 표준화

④ 박스-콕스 변환

2. 데이터 전처리 – 분석 변수 처리

05 불균형 데이터 처리

데이터의 클래스별 불균형이 심한 경우 정확한 분석 결과를 도출하기 어렵기 때문에 불균형 데이터를 전처리한 뒤 데이터를 분석한다.

1) 불균형 데이터 처리 기법

● 불균형 데이터 처리 기법으로는 과소표집, 과대표집, 임계값 이동, 앙상블 기법, 가중치 균형이 있다.

① 과소표집(Under-Sampling)

- ▶ 다수 클래스의 데이터 중 일부만 선택하여 데이터 비율을 맞추는 방법이다.
- ▶ 데이터 소실 가능성과 중요한 데이터를 잃을 가능성이 높다.

[기법] 랜덤 과소표집, ENN(Edited Nearest Neighbor), 토맥 링크 방법, CNN(Condensed Nearest Neighbor), OSS(One Sided Selection)

② 과대표집(Over-Sampling)

- ▶ 소수 클래스의 데이터를 복제 또는 생성하여 데이터 비율을 맞추는 방법이다.
- ▶ 과적합 가능성 있다.
- ▶ 알고리즘 성능은 높지만, 검증 성능은 나빠질 수 있다.

[기법] 랜덤 과대표집, SMOTE, Borderline-SMOTE, ADASYN

2. 데이터 전처리 – 분석 변수 처리

1) 불균형 데이터 처리 기법

③ 임계값 이동(Cut-off Value Moving)

- ▶ 임계값을 데이터가 많은 쪽으로 이동시키는 방법이다.
- ▶ 학습 단계에서는 변화 없이 학습하고, 테스트 단계에서 임계값을 이동한다.

④ 앙상블 기법(Ensemble Technique)

- ▶ 같거나 서로 다른 여러 가지 모형들의 예측 및 분류 결과를 종합하여 최종적인 의사결정에 활용하는 방법이다.

⑤ 가중치 균형(Weight balancing)

- ▶ 학습 데이터셋의 각 데이터에서 손실(loss)을 계산할 때 특정 클래스의 데이터에 더 큰 손실(loss)값을 갖도록 하는 방법이다.
- ▶ 클래스 비율에 가중치를 두기도 한다.

예) 높은 비율의 데이터에 높은 가중치 적용

2. 데이터 전처리 – 분석 변수 처리

개념 체크

1. 다음 중, 불균형 데이터 처리 기법이 아닌 것은?

- ① 과소표집 ② 과대표집
- ③ 임계값 이동 ④ 비닝

비닝(구간화)

▶ 데이터 값을 몇 개의 Bin으로 분할하여 계산하는 방법이며, 변수 변환 방법에 속한다.

과소표집(Under-Sampling)

- ▶ 다수 클래스의 데이터 중 일부만 선택하여 데이터 비율을 맞추는 방법이다.
- ▶ 데이터 소실 가능성과 중요한 데이터를 잃을 가능성이 높다.

과대표집(Over-Sampling)

- ▶ 소수 클래스의 데이터를 복제 또는 생성하여 데이터 비율을 맞추는 방법이다.
- ▶ 과적합 가능성이 있다.
- ▶ 알고리즘 성능은 높지만, 검증 성능은 나빠질 수가 있다.

임계값 이동(Cut-Off Value Moving)

2. 다음 설명에 해당하는 불균형 데이터 처리 기법은?

학습 데이터셋의 각 데이터에서 손실(loss)을 계산할 때 특정 클래스의 데이터에 더 큰 손실(loss)값을 갖도록 하는 방법이다.

클래스 비율에 가중치를 두기도 한다.

- ① 가중치 균형(Weight balancing)
- ② 과대표집(Over-Sampling)
- ③ 앙상블 기법(Ensemble Technique)
- ④ 임계값 이동(Cut-off Value Moving)

앙상블 기법(Ensemble Technique)

▶ 같거나 서로 다른 여러 가지 모형들의 예측 및 분류 결과를 종합하여 최종적인 의사결정에 활용하는 방법이다.

2. 데이터 전처리 예상문제

예상 문제

1. 다음은 데이터 정제의 과정과 요소를 나열한 것이다. 괄호 안의 요소로 알맞은 것은?

다양한 매체로부터 데이터를 수집, 원하는 형태로 변환, 원하는 장소에 저장, (), 필요한 시기와 목적에 따라 사용이 원활하도록 관리의 과정이 필요하다.

- ① 비정형 데이터의 경우 기본적으로 구조화된 정형 데이터로의 변환을 수행
- ② 결측치의 처리, 이상치 처리, 노이즈 처리
- ③ 저장된 데이터의 활용 가능성을 타진하기 위한 품질확인
- ④ 데이터 분석이 용이하도록 기존 또는 유사 데이터와의 연계 통합

데이터 정제 과정 : 수집 -> 변환 -> 저장 -> 품질확인 -> 관리 의 과정이다.

비정형 데이터의 경우 기본적으로 구조화된 정형 데이터로의 변환을 수행(변환의 세부 수행)

결측치의 처리, 이상치 처리, 노이즈 처리(데이터 변환 과정의

2. 데이터 집합에서 다른 측정값들과 비교하여 현저한 차이를 보이는 샘플 또는 변수 값은?

- ① 결측값
- ② 잡음
- ③ 이상값
- ④ 데이터 정제

3. 나이대별 성별과 체중에 대해서 조사를 하고자 한다. 이때 발생 가능한 결측치에 대해서 분류를 다음 아래와 같이 구분하였다. 옳은 것은?

- ① 데이터의 누락 : 비무작위 결측
- ② 여성은 체중 공개를 꺼림 : 무작위 결측
- ③ 젊은 여성은 체중 공개를 꺼림 : 비무작위 결측
- ④ 무거운 사람은 체중 공개를 꺼림 : 무작위 결측

나이대별(X), 성별(Y), 체중(Z) 분석에 대한 모델링을 가정해보면

X, Y, Z가 관계없이 Z가 없는 경우 : 데이터의 누락(응답 없음) -> 완전 무작위 결측(MCAR)

2. 데이터 전처리 예상문제

4. 다음 중 데이터 전처리에 대한 설명 중 틀린 것은?

- ① 데이터 분석 업무 중에 데이터 수집 및 전처리 과정에 가장 많은 시간이 소요된다.
- ② 데이터 전처리 과정은 데이터 정제, 결측값 처리, 이상값 처리, 분석변수 처리이다.
- ③ 데이터 전처리 여부에 따라 분석 결과가 달라질 수 있다.
- ④ 데이터 전처리는 최초에 한 번만 수행된다.

5. 데이터 오류의 원인 중 하나로 필수 데이터가 입력되지 않고 누락된 값을 의미하는 것은?

- ① 결측값
- ② 노이즈
- ③ 이상값
- ④ 파생변수

데이터 오류 원인

- 1. 결측값 : 필수 데이터가 입력되지 않고 누락된 값
- 2. 노이즈 : 실제 입력되지 않았으나 입력되었다고 잘못 판단한 값

6. 다음 중 결측값을 의미하는 말이 아닌 것은?

- ① NA ② 999999
- ③ NO ④ Null

7. 다음에 설명하는 통계적 기법은?

불완전 자료는 모두 무시하고, 완전하게 관측된 자료만 사용하여 분석하는 방법

- ① 다중 대치법
- ② 단순 확률 대치법
- ③ 평균 대치법
- ④ 완전 분석법

다중 대치법 : 단순 대치법을 한 번 하지 않고, n번 대치를 통해 1개의 완전한 자료를 만들어 분석하는 방법

단순 확률 대치법 : 적절한 확률값을 부여한 후 이를 결측값으로 대치하는 방법

평균 대치법 : 관측되어 얻어진 자료의 평균값으로 결측값을 대치하는 방법

2. 데이터 전처리 예상문제

8. 다음 중 시각화를 이용한 데이터 이상값 검출 방법에 사용될 수 없는 방법은?

- ① 확률밀도함수
- ② 히트맵
- ③ 히스토그램
- ④ 상자수염그림

시각화를 이용한 데이터 이상값 검출 방법에는 확률밀도함수, 히스토그램, 시계열 차트, 상자수염그림이 있다.

히트맵은 색상으로 표현할 수 있는 다양한 정보를 일정한 이미지 위에 열분포 형태의 비주얼 그래픽으로 출력한 차트

9. 이상값 검출 방법 중 하나로 관측된 값이 평균으로부터 벗어난 정도를 측정하는 방법은?

- ① 카이제곱 검정 방법
- ② iForest 활용 방법
- ③ 유클리디안 거리 활용 방법
- ④ 마할라노비스 거리 활용 방법

10. 다음 중 차원축소(Dimensionality Reduction)에 대한 설명으로 옳지 않은 것은?

- ① 데이터의 개수가 많으면 정확한 분석 결과를 얻을 수 있다.
 - ② 원활한 데이터 분석 작업을 위해 차원축소 기법을 사용한다.
 - ③ 차원축소는 분석에 활용되는 데이터의 변수 정보는 최대한 유지한다.
 - ④ 데이터 세트 변수의 개수를 줄이는 데이터 분석 기법 이다.
11. 다음 설명은 차원축소 기법 중 무엇을 의미하는 것인가?

가장 보편적으로 사용되는 차원 축소 기법 중 하나로 원본 데이터를 최대한 보존하면서 고차원 공간의 데이터를 저차원 공간 데이터로 변환하는 기법이다. 행과 열의 크기가 같은 정방행렬에서만 사용한다.

- ① 주성분 분석(PCA : Principal Component Analysis)
- ② 선형 판별 분석(LDA : Linear Discriminant Analysis)
- ③ 특이값 분해(SVD: Singular Value Decomposition)
- ④ 요인 분석(Factor Analysis)

2. 데이터 전처리 예상문제

12. 다음 설명의 파생변수(Derived Feature)생성 방법 중 어느 것인가?

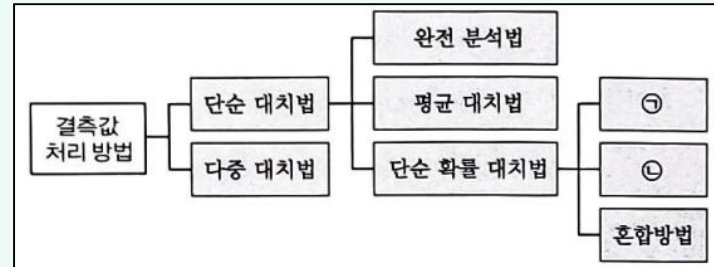
주어진 변수의 단위를 새로운 단위로 변환하여 데이터를 표현하는 방법

- ① 조건문 이용
- ② 변수 결합
- ③ 단위 변환
- ④ 요약 통계량 변환

파생변수 생성 방법

1. 단위 변환 : 주어진 변수의 단위를 새로운 단위로 변환하여 데이터를 표현하는 방법
2. 표현방식 변환 : 표현방식을 단순화하는 변환 방법
3. 정보 추출 : 하나의 정보에서 새로운 정보를 추출하는 방법
4. 변수 결합 : 변수를 결합하여 새로운 변수를 정의하는 방법
5. 요약 통계량 변환 : 요약 통계량을 활용하는 변환 방법
6. 조건문 이용 : 조건문을 활용하여 파생변수를 생성하는 방법

14. 다음의 결측값 처리 방법에 대한 표에서 ㉠과 ㉡에 알맞은 명칭은?



- ① ㉠ : 핫-덱 대체 ㉡ : SMOTE
- ② ㉠ : 핫-덱 대체 ㉡ : 콜드-덱 대체
- ③ ㉠ : 콜드-덱 대체 ㉡ : LOF
- ④ ㉠ : 파생변수 대체 ㉡ : ESD

핫덱(Hot-Deck) 대체 : 진행 중인 연구 내에서 비슷한 성향의 자료로 결측값을 대체하는 방법

콜드덱(Cold-Deck) 대체 : 진행 중 연구 내부가 아닌 외부 출처 또는 이전의 비슷한 연구에서 대체 값을 가져오는 방법이다.

혼합 방법 : 2가지 이상 다양한 방법을 혼합하는 방법

15. 다음 중 데이터를 가장 잘 설명할 수 있는 축을 찾는 주성분 분석(PCA)과 다르게 가장 독립적인 축을 찾는 기법은?



감사합니다.