1. 데이터 전처리

데이터 정제 → 결측값 처리 → 이상값 처리 → 분석 변수 처리

2. 데이터 정제

결측값을 채우거나 이상값을 제거하는 과정을 통해 데이터의 신뢰도를 높이는 작업

절차

1) 데이터 오류 원인 분석

원인) 결측값, 노이즈, 이상값

2) 데이터 정제 대상 선정

3) 데이터 정제 방법 결정

방법) 삭제, 대체, 예측값 삽입

3. 결측값 처리

결측값 → 입력이 누락된 값

결측값 종류

완전 무작위 결측 (MCAR)

변수상에서 발생한 결측값이 다른 변수들과 상관이 없는 경우

무작위 결측(MAR)

누락된 자료가 특정 변수와 관련되어 일어나지만, 그 변수의 결과는 관계가 없는 경우

비 무작위 결측 (MNAR)

비무작위 결측 (MNAR)

누락된 값이 다른 변수와 연관 있는 경우

처리 절차

결측값 식별 → 결측값 부호화 → 결측값 대체

처리 방법

단순 대치법

결측값을 그럴듯한 값으로 대체하는 통계적 기법

완전 분석법, 평균 대치법, 단순 확률 대치법

다중 대치법

단순 대치법을 m번 대치를 통해 m개의 가상적 완전한 자료를 만들어서 분석하는 방법

대치, 분석, 결합

4. 이상값 처리

↳ 관측된 데이터 범위에서 벗어난 작은 값이나 큰 값

이상값 발생 원인

데이터 입력 오류

데이터 수집 과정에서 발생할 수 있는 에러

측정 오류

데이터 측정 과정에서 발생하는 에러

실험 오류

실험 조건이 통일하지 않을 경우 발생

실험 조건이 충분하지 않은 경우 발생

고의적인 이상값

자기 보고식 측정에서 나타나는 에러

표본 추출 에러

데이터 샘플링하는 과정에서 나타나는 에러

검출 방법

개별 데이터 관찰, 통계값, 시각화, 머신러닝, 마할라노비스 거리, LOF, iForest

처리 방법

삭제

이상값으로 판단되는 관측값을 제외하고 분석하는 방법

대체법

이상값을 평균이나 중앙값 등으로 대체

변환

극단적인 값으로 인해 이상값이 발생하면 자연로그를 취해서 값을 감소시키는 방법

박스 플롯 해석

사분위 수를 이용해서 제거하는 방법

분류

이상값이 많을 경우 사용하는 방법으로, 서로 다른 그룹으로 통계적인 분석을 실행하여 처리

5. 분석 변수 처리

↳ Feature) 데이터 모델에서 사용하는 예측을 수행하는 데 사용되는 입력 변수

변수 유형

독립 변수

다른 변수에 영향을 받지 않고 종속 변수에 영향을 주는 변수

연구자가 의도적으로 변화시키는 변수

종속 변수

다른 변수로부터 영향을 받는 변수

독립 변수의 변화에 따라 어떻게 변화하는지 연구하는 변수

차원 축소

여러 변수의 정보를 최대한 유지하면서 데이터 세트 변수의 수를 줄이는 탐색적 분석 기법

주성분 분석(PCA), 특이값 분해(SVD), 요인 분석, 독립성분분석(ICA), 다차원척도법(MDS)

파생 변수

기존 변수에 특정 조건이나 함수 등을 사용하여 재정의한 변수

단위 변환, 표현 형식 변환, 요약 통계량 변환, 변수 결합

변수 변환

불필요한 변수를 제거하고, 변수를 반환하며, 새로운 변수를 생성시키는 작업

단순 기능 변환, 비닝, 정규화, 표준화