데이터 전처리

데이터 전처리 개요

- 데이터 전처리
  - 데이터를 분석에 사용할 때 성능이 더 좋게 나오도록 데이터를 수정하거나 형태를 변형하는 것
- 수집한 데이터를 머신 러닝 등에 바로 사용할 수 있는 경우는 거의 없다.
- 수집한 데이터가 너무 크면 한번에 분석하기 어려우므로 적절한 크기로 줄여야 한다.
- 데이터가 비정형이라면 이를 정형 데이터로 바꾸어야 한다.
  - 이미지나 텍스트와 같은 비정형 데이터의 의미를 컴퓨터가 바로 다룰 수는 없다.
- 분석 목적에 맞게 데이터의 품질을 확인, 필요하면 품질을 높이는 작업
- 데이터 품질
  - 신뢰성
  - 정확성
  - 적시성 (최신성) 등

데이터 전처리 – 전처리 유형

- 전처리의 유형
  - 중간에 데이터가 빠진 경우
  - 틀린 값이 들어 있는 경우
  - 데이터의 단위가 틀리는 경우 (m/inch, kg/파운드 등)
  - 범주형 데이터의 경우 카테고리를 나타내는 표현으로 변경 필요
    - 예) 월요일: 1, 화요일: 2, 등

- 데이터 변환
  - 로그 변환
  - 역수 변환
  - 정규분포로 변환
    - 예) 10점 만점으로 처리한 것과 100점 만점으로 처리한 데이터를 같이 활용하려면 동일한 분포로 변환할 필요

*** 변환의 목적 : 선형시스템이어야 정확한 예측이 가능

데이터 전처리 – 전처리 유형

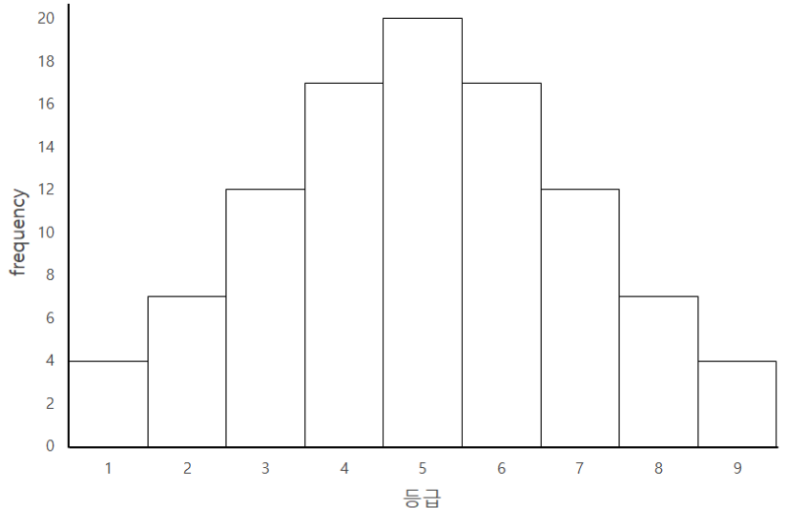
구분	처리 방법
결측치 (missing) 처리	• 결측치가 포함된 항목을 모두 버리는 방법 – 버리는 항목의 비중이 크면 무시하기 어려움 • 결측치를 적절한 값으로 대체 – 0, 평균값, 최소값, 특정 상수, 인접 값으로 추정 등 • 분석 단계로 결측치 처리를 넘김 (NA로 표기) • 별도의 범주형 변수를 정의하여 추적 가능하게 관리
틀린 값 (invalid) 처리	• 틀린 값이 포함된 항목을 모두 버리는 방법 • 틀린 값을 다른 적절한 값으로 대체 • 분석 단계로 틀린 값의 처리를 넘김
이상치 (outlier) 검출	• 값이 일반적인 범위를 벗어나 특별한 값을 갖는 경우 • 데이터 분석 과정의 활동이므로 분석 단계로 넘김 • 도난 카드의 사용, 불법 보험료 청구 등의 탐지

데이터 전처리 – 데이터 변환

- 데이터를 주어진 그대로 사용하지 않고 다른 형태로 바꾸어 사용한 것이 필요한 경우가 많다
  - 같은 성적을 나타내는데 A, B, C 등 학점으로 표현하거나 100점 만점으로 환산하기도 한다 (97, 94, 91 등)
  - 변환의 종류
    - 범주형 변환
    - 로그 변환
    - 역수 변환

데이터 전처리 – 범주형으로 변환

- 수치형 데이터의 **개별 값** 구분이 **오히려 혼란**스러울 경우
  - 나이 : 10대, 20대, 30대, 40대 등
  - 연간 소득 : 고소득층, 중간층, 저소득층 등
- 수치형 데이터를 범주형으로 변환할 때, **각 구간의 범위**를 **균등**하게 정할 수도 있고 **서로 다른 범위**를 정할 수도 있다.
  - 고교 내신 성적 : 1등급/9등급(각 4%), 2등급/8등급(15%) 등



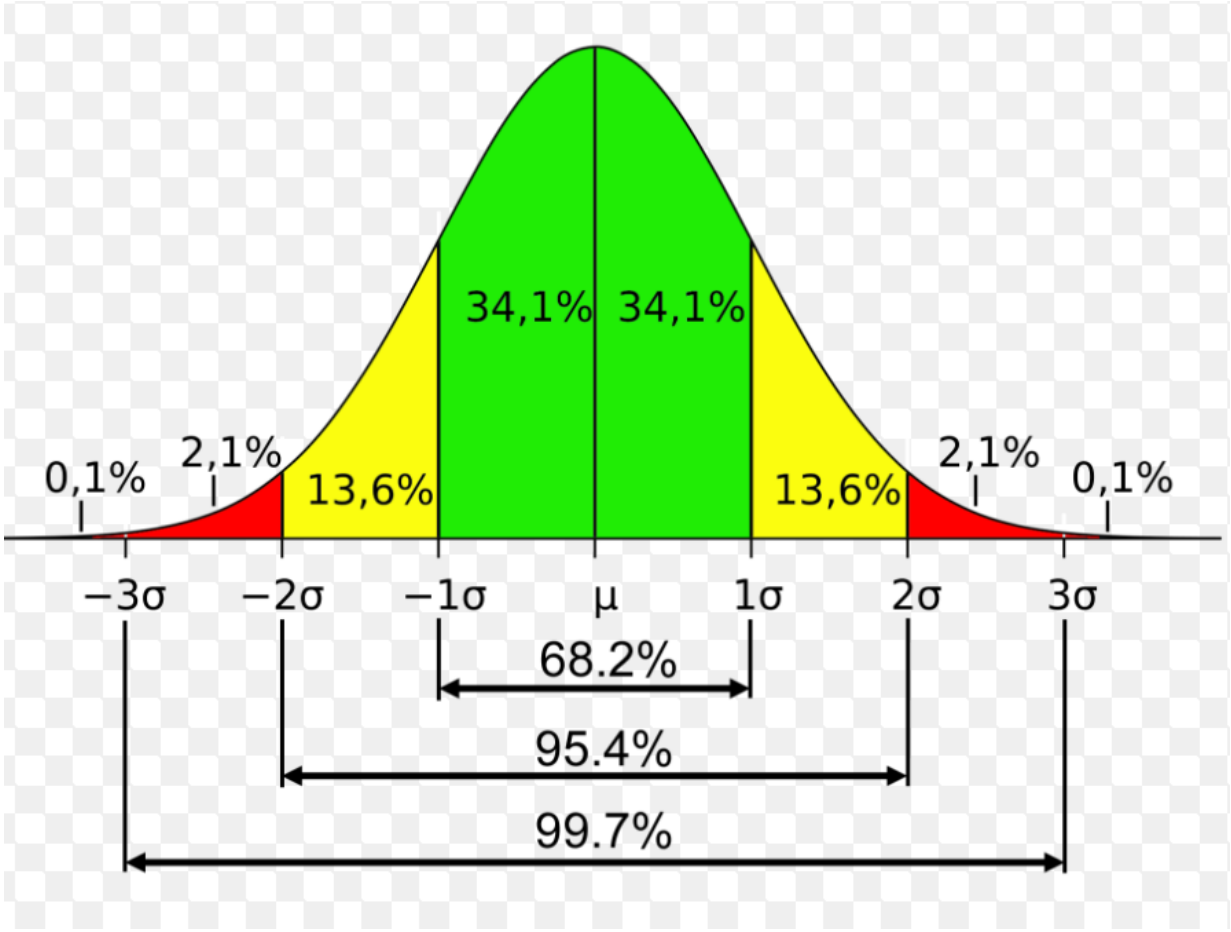
데이터 전처리 – 범주형 변수 코딩

- 요일을 1, 2, 3, 4, 5, 6, 7 등으로 표시한 경우 이 변수를 컴퓨터가 연산(덧셈이나 곱셈)을 할 수 있는 숫자로 인식해서는 안 된다.
  - 이 숫자를 범주형(Category형)으로 분명하게 처리되어야 한다.
  - 컴퓨터가 범주형 변수를 분명히 인식하게 하는 방법이 필요하다
- one hot encoding
  - 하나의 특성(컬럼)만 1이 될 수 있고, 다른 특성은 모두 0으로 코딩하는 방법
    - 월요일 : (1, 0, 0, 0, 0, 0, 0)
    - 수요일 : (0, 0, 1, 0, 0, 0, 0)
  - get_dummies() : pandas
    - 범주형 변수들을 one hot encoding으로 만들어준다

데이터 전처리 – 스케일링(Scaling)

- 원래 데이터가 갖는 값의 범위를 다르게 조정하는 작업
- 스케일링을 하는 이유
  - 여러 특성 변수의 중요도를 같도록 맞추기 위해서
  - 예) 모든 시험은 100점 만점으로 환산해야 동일한 비중으로 취급된다
    - 어떤 과목은 50점 만점, 어떤 과목을 80점 만점이면 동일한 조건으로 특성이 반영되지 않는다.
- 최소-최대 스케일링
  - 주어진 값을 [최소값=0, 최대값=1]로 재조정하는 것
  - MinMaxScaler() 함수 (sklearn)
- 표준 스케일링
  - 데이터 분포를 표준 정규분포(평균 0, 표준 편차 1)이 되도록 정규화하는 방법
  - 표준화(standardization), z-score 정규화
  - StandardScaler() 함수 (sklearn)

데이터 전처리 – 정규 분포(Normal Distribution)

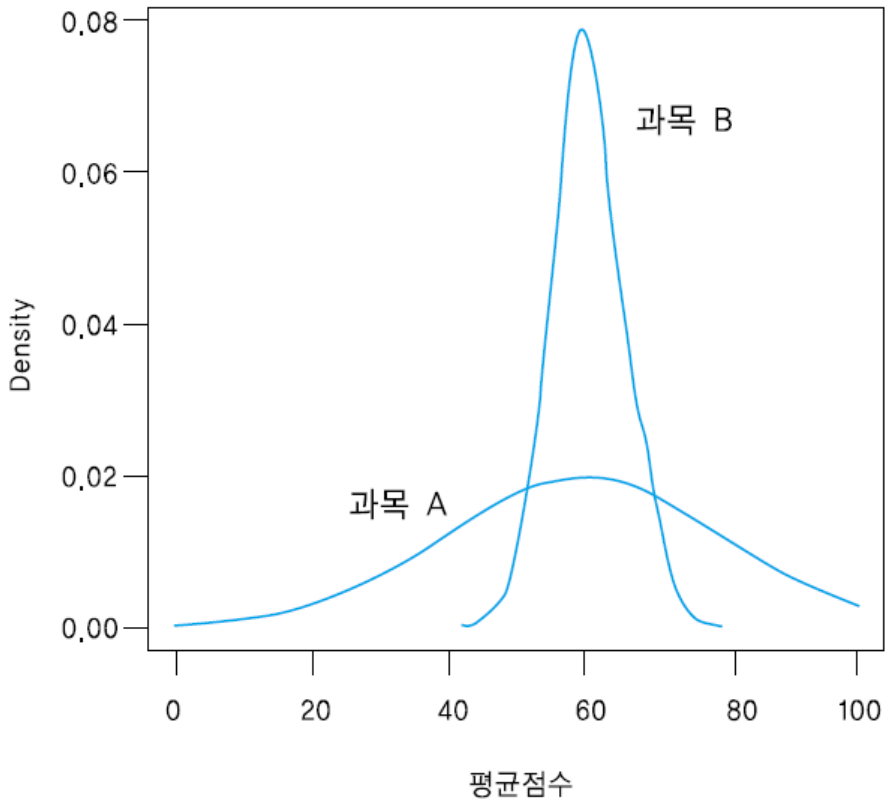


데이터 전처리 – 표준 스케일링

- 모든 과목 (동일)
  - 100점 만점
  - 학급 평균 : 60점

- 과목의 표준편차
  - 과목 A : 20점
  - 과목 B : 5점

학생	과목 A	과목 B	평균
갑	90	80	85
을	80	90	85



- 누가 더 공부를 잘 하는 학생일까?

데이터 전처리 – 표준 스케일링

- 과목 B는 편차가 작은데 이것의 의미는
  - 대부분의 학생이 60점 근처에 모여 있다는 것이고 따라서 고득점을 받기가 매우 어려운 과목인 것을 의미
- 이러한 과목의 점수 분포 특성을 고려하면 학생 을이 어려운 과목에서 90점을 받았으므로 더 우수한 학생이라는 생각이 든다.
  - 이러한 문제를 정확히 해결하려면 원 점수가 아니라 표준편차를 고려한 점수를 사용해야 한다.
  - 표준 변환
    - 각 점수가 평균에서 얼마나 떨어져 있는 지를 표준편차를 기준으로 나누어 비교

$$z = \frac{x - u}{\sigma}$$

데이터 전처리 – 표준 스케일링

- 각 점수가 평균에서 얼마나 떨어져 있는 지를 표준편차를 기준으로 나누어 비교

$$z = \frac{x - u}{\sigma}$$

- **표준(z) 변환 후**의 데이터 분포 : 평균 0, 표준편차 1

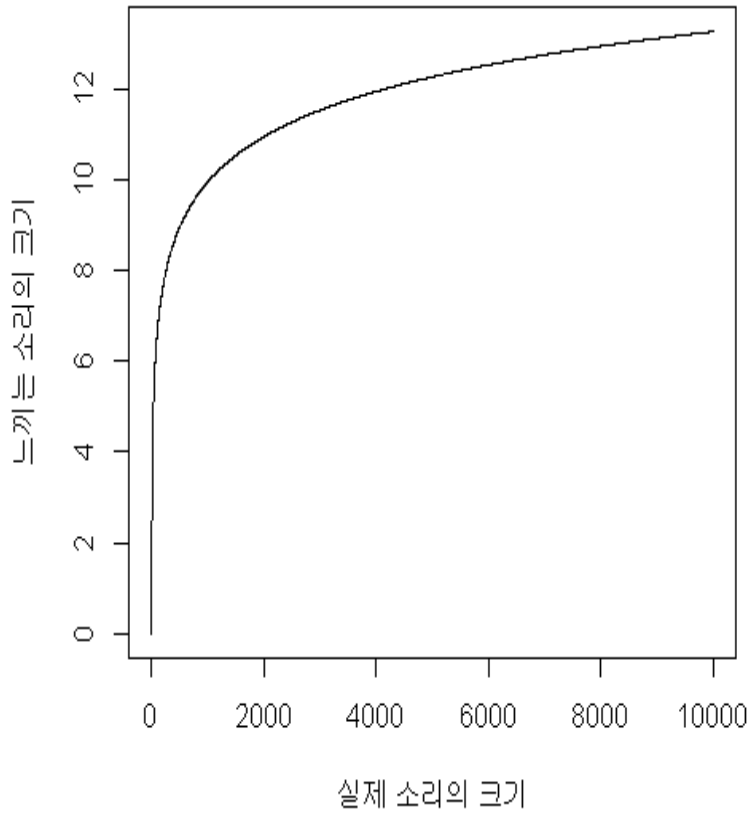
	학생	과목 A	과목 B	평균
변환 전	갑	90	80	85
	을	80	90	85
변환 후	갑	(90 - 60) / 20 = 1.5	(80 - 60) / 5 = 4	2.75
	을	(80 - 60) / 20 = 1	(90 - 60) / 5 = 6	3.50

데이터 전처리 – 로그 변환

- 체감형 수치를 선형적으로 표현할 때 사용
  - 사람이 자연적으로 느끼는 느낌의 양을 수학적 모델로 설명할 때 사용
  - 돈, 소리, 빛, 압력, 냄새 등 생물학적인 자극을 주는 경우

- 같은 자극 정도를 느끼려면 현재 보유한 양이 많을수록 이에 비례한 더 강한 자극이 필요하다.
  - 느끼는 자극 정도 : 현재 보유량에 반비례
  - 이를 수학적으로 표현하면 로그 함수가 됨
    - 현재 보유한 양 x, 이의 변화량(미분값) $1/x$
    - 로그 형태로 변화하는 신호의 기울기(입력의 변화대 출력의 변화량)은 현재의 양에 반비례
- 로그를 취한 후의 값에 대해서 사람들이 변화량을 느끼는 것이 선형적이라는 특성

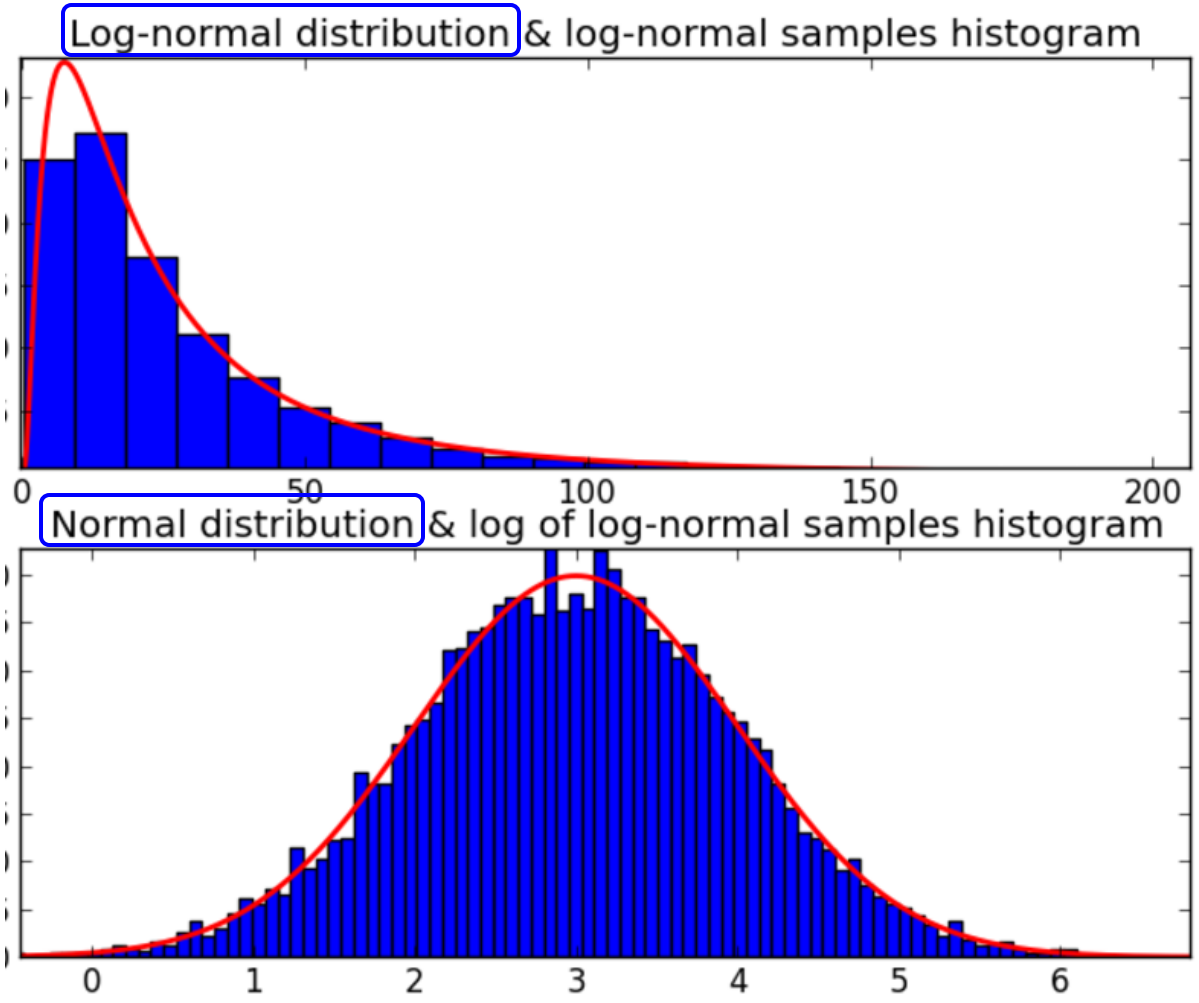
데이터 전처리 – 로그 변환



실제 소리의 크기와 느낌의 차이는 로그 관계

- 스피커 소리를 들을 때 소리가 2배, 3배, 4배 크게 들리게 하려면 실제 소리의 크기는 지수 함수로 키워야 한다.

데이터 전처리 – 로그 변환


Log-normal distribution & log-normal samples histogram
0
50
100
150
200
Normal distribution & log of log-normal samples histogram
0
1
2
3
4
5
6

데이터 전처리 – 역수 변환

- **역수**를 사용하면 **선형적인 특성**을 가져 **분석의 정확도**가 높아지는 경우

- 역수 관계 : 자동차의 성능 지표
  - 자동차 마일리지 (연료 1l로 가는 거리 km)
  - 연비 (100 km 주행하는데 필요한 연료 l)

- 측정 목적
  - 같은 비용을 얼마나 멀리 갈 수 있는가?
  - 같은 거리를 여행하는데 비용이 얼마가 드는가?

데이터 전처리 – 데이터 변환

구분	내용
범주형으로 변환	• 수치 데이터가 아닌 것을 명시
min-max 정규화	• 수치 데이터의 범위가 다를 때
z-score 정규화	• 일반 정규화에 표준 편차를 고려한 변환
로그 변환	• 로그를 취하면 선형 특성을 가질 때 (또는 로그 정규 분포를 가질 때)
역수 변환	• 역수를 사용하면 선형적인 특성을 가질 때

- 적절히 잘 선택을 해야 한다