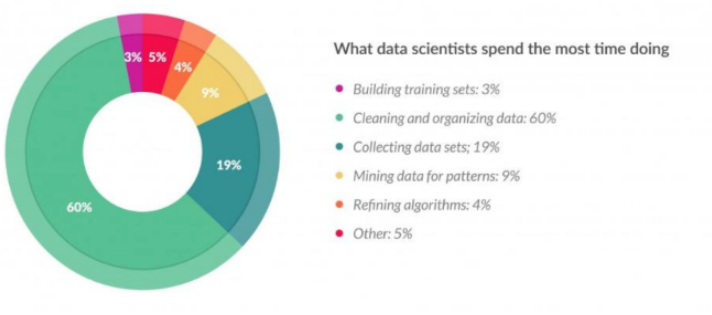
Preprocessing

Preprocessing

1. 전처리란?

: 데이터를 **모델링에 적합한 형태**로 처리하는 절차



2. 데이터 전처리의 필요성

- 데이터 분석과정에서 가장 많은 비중 차지
- 모든 수집된 데이터가 데이터 분석에 최적화된 상태가 아님
- 각 분석 방향에 따라 데이터의 적절한 변환을 통해 신뢰할 수 있는 결과를 도출할 수 있어야함

ex) - 나이를 수집하였는데 음수 데이터
- 남자 1 여자 0 으로 표기하였는데 3의 데이터

Preprocessing

3. 각 데이터 분석 별 전처리 과정

* 표, 그래프 작성용

- 지표의 계산, 표나 그래프로 쉽게 변환할 수 있는 데이터를 준비하기 위한 전처리 과정
- 필요한 열이 모두 있고, 다루기 쉬운 단위로 집약된 행이 필요한 범위만큼 존재하도록 데이터를 작성

* 지도학습용 전처리

- 지도학습 모델이 이용할 학습데이터, 테스트데이터, 적용데이트 세 종류 준비
- 머신러닝 모델의 종류에 따라 다루기 쉬운 데이터로 변환

* 비지도학습용 전처리

- 비지도학습 모델이 이용할 설명 변수를 가진 데이터를 준비하는 전처리 과정
- 충분한 열과 행, 머신러닝 모델의 종류에 따라 다루기 쉬운 자료형으로 변환

Data cleansing

Data cleansing

1. Missing Imputation(결측치)

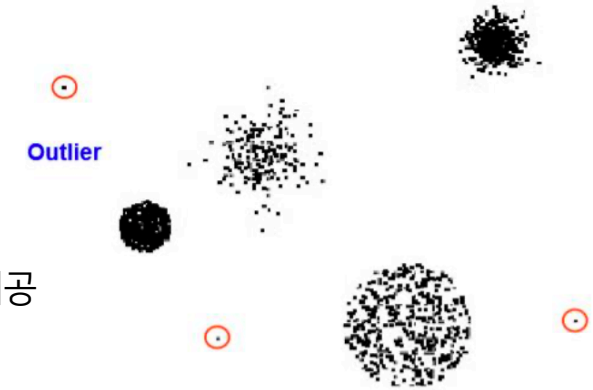
- 정확한 분석을 방해하는 결측값이 있는 데이터는 분석 가능한 형태로 만들어줘야 함
  → 결측값의 유형, 규모, 다른 변수와의 관련성 등을 고려하여 결측값을 처리할 방법을 정해야함

- 행/ 열 무시하기
- 중심 경향 측정값 (평균값, 중간값, 최빈값) 글로벌 상수값으로 대체
- 가장 가능성이 높은 값으로 결측치 채우기
  (회기분석, 베이지안 공식, 의사결정나무, KNN 등 다른 모델에서 추론하여 대체값 정함)

Data cleansing

2. Outlier(이상치)

: 변수의 분포에서 비정상적으로 분포를 벗어난 값
각 변수의 분포에서 비정상적인 극단값을 갖는 경우나
자료에 타당도가 없는 경우, 비현실적인 변수값들이 해당됨



** 노이즈와는 다른 개념

이상치는 나머지 데이터와 생성 메커니즘이 다를 수 있다는 의혹 제공
따라서 왜 다른 메커니즘을 따르는지 밝혀내는 것이 중요함

Data cleansing

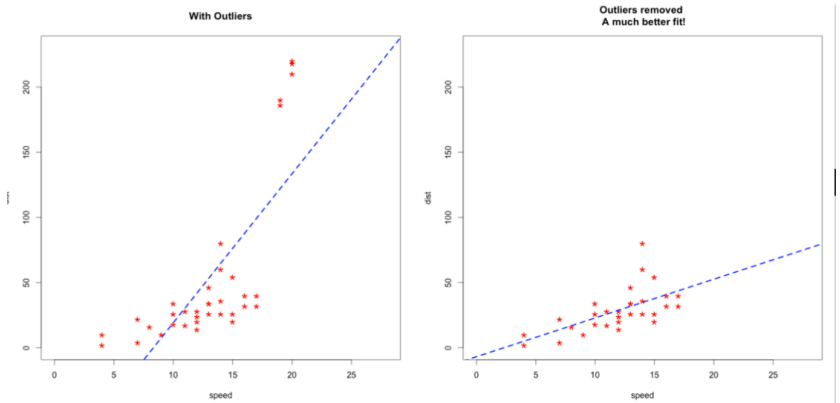
2. Outlier(이상치)

검출

내면 스튜던트화 잔차 확인
Leverage
Cooks'D
1.5*IQR

처리

삭제
상,하한선 제한
케이스 분리분석



Data cleansing

3. Noise(노이즈)

: 측정 변수의 랜덤 오류나 분산

Smoothing(평활화): removing noise from a data set
in order to make a pattern more visible

1. **비닝** : 근접한 다른값을 참고하여 정렬한 데이터 값을 평활화하는 법
2. **회귀분석**: 두 개 이상의 속성으로 다른 속성을 예측해 최선의 직선을 찾음

Data reduction

Data reduction

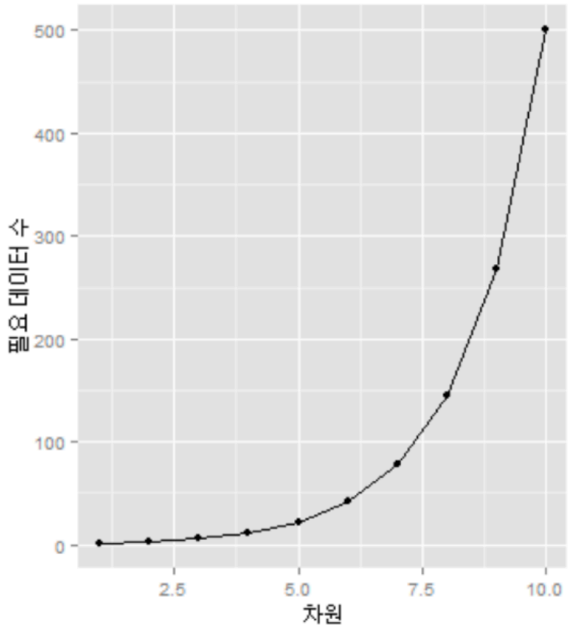
1. Dimension Reduction

: 많은 변수로 인해 야기되는 비효율성을 개선하기 위해
샘플링 등으로 데이터 양을 줄이거나 차원을 줄이는 작업

* 차원의 저주 (Curse of Dimensionality):

: 훈련 샘플 각각이 수천, 수백만 개의 특성을 가지고 있을 때
학습이 느리고 좋은 솔루션에 방해가 되는 상황

- 분석에 영향을 미치지 않는 데이터를 배제하거나
타 변수와 중복적 성격을 띠는 변수들을 통합/제거하여 효율성 제고



Data reduction

2. PCA(principal component analysis | 주성분 분석)

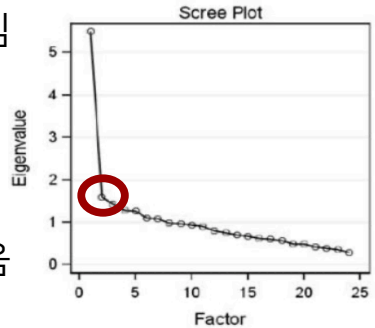
: 가장 인기있는 차원 축소 알고리즘으로, 변수들의 선형 결합을 통해 새로운 변수를 만들며 차원을 줄임
샘플링 등으로 데이터 양을 줄이거나 차원을 줄이는 작업

장점

- 간단하면서 직관적인 방법으로 원데이터 정보를 최대한 보존하면서 변수의 수를 줄임
- 특정 기준을 통해 유의한 주성분의 개수를 구하면
  정보의 중첩이 없는 데이터의 순수한 차원을 구할 수 있음

단점

- 새롭게 구한 주성분을 해석하기 매우 모호해지므로 추가적인 분석에 활용하기 어려움
- 과도한 차원 축소는 데이터의 본질을 흐림



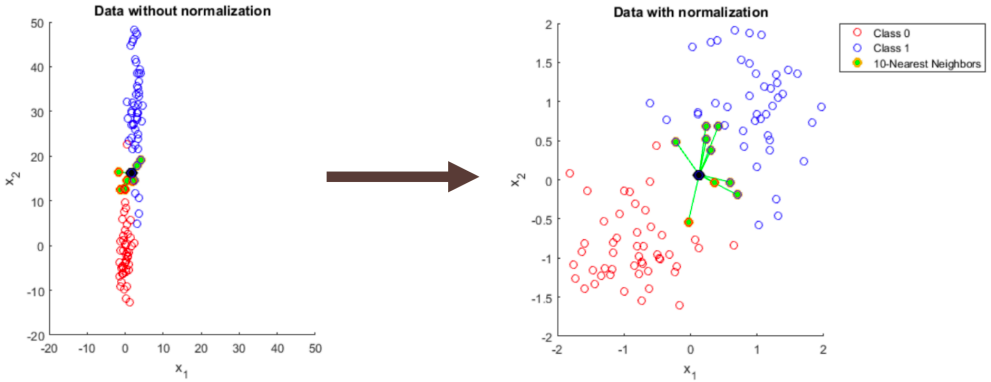
Data Transformation

Data transformation

1. Data transformation

: 데이터 분석시 적절한 측정단위를 사용하기 위해 정규화 또는 표준화를 통해 작은 범위 안에 위치하도록 함

- Min-Max Normalization
  (최소-최대 정규화)
- Standardization
  (Z스코어 정규화)
- Decimal scaling
  (십진스케일 정규화)



Thank you