데이터 전처리 둘러보기

데이터 전처리 과정

- 데이터 평가Access Data
  - Dirty Data 판별 -> Data Cleaning
  - Data Mining 성능 -> Data Preprocessing

- 데이터 정제Cleaning Data
  - 누락 데이터 처리
  - 이상치 처리

- 데이터 전처리Data Preprocessing
  - Feature Scaling
  - Dimensionality Reduction

데이터 평가

연구와 경험을 통해 자신만의 판별법을 구축한다.

- **Dirty Data (Low Quality Data) 판별 사례**
  - Step1. Completeness
  - Step2. Validity
  - Step3. Accuracy
  - Step4. Consistency

- **계속되는 방법론 연구**
  - Messy Data (Untidy Data) 판별 : Dirty Data는 데이터 자체에 문제가 있는 것이고, Messy Data는 구조적으로 문제가 있는 데이터
  - Tool(Coding, Library등)을 사용하여 판별할 수 있는가?

Dirty Data 판별

- **Step1. Completeness**
  - 모든 데이터가 채워져 있는가?
  - 행, 열에 Null값이 없는지 확인
  - DataFrame의 Library로 판별 가능

- **Step2. Validity**
  - 미리 정해놓은 제한사항constraint을 충족하는지 확인
  - 예를 들어, 사람의 키에 대한 데이터를 수집한다고 할 때, height>=0라는 제한사항을 정해놓는다. 그러면 -10이 데이터에 있으면 Null은 아니므로 Step1의 Completeness에서는 문제가 없지만, Validity는 충족하지 못해 Dirty Data로 판별 -> 알고리즘 coding으로 구현
  - cf. 이상치Outlier

Dirty Data 판별

- **Step3. Accuracy**
  - 부정확한 데이터가 없는가?
  - 조사기관의 실수로 사람의 키가 180인데 170은 넣은 경우, Completeness, Validity는 모두 충족하지만 Accuracy를 충족 못함
  - 재조사를 통해 해결해야 할 문제 -> 툴 사용이 어려운 경우

- **Step4. Consistency**
  - 테이블간, 테이블 내 표준 포맷을 지켰는가?
  - 예를 들어, 사람의 키에 대한 데이터를 수집한다고 할 때, A파일에서는 cm단위로, B파일에서는 inch단위를 사용했다면, 앞의 3개의 Step은 모두 충족하지만 Consistency는 충족 못함.
  - 데이터 형식을 통일

데이터 정제

- **누락 데이터 처리** Handling missing values
  - 행 또는 열에서 누락된 데이터는 머신러닝의 결과에 악영향을 끼칠 수 있다.
  - 누락데이터는 삭제하거나 새로운 값(?)을 삽입하여 처리
  - 새로운 값은 코딩으로 평균값 등 또는 재조사

- **이상치 처리** Dealing with outliers and erroneous data
  - 이상치는 오류 성 데이터로 분석에 오류를 줄 가능성이 높음
  - 이상치를 포함한 행은 삭제하는 것이 좋음

Feature Scaling

- **Feature Scaling**
  - 서로 다른 변수variable [feature]의 값 범위를 일정한 수준으로 맞추는 방법
  - 비교해야 할 데이터의 기준이 서로 다른 경우에 같은 기준으로 만들어서 비교
  - 머신 러닝 등의 분석 성능 향상을 위해

- **표준화** standardization **및 정규화**normalization
  - 표준화 : 일정 기준 안으로 ex) z-score와 같은 수식 사용
  - 정규화 : 최소 0 ~ 최대 1의 값으로 변환

용어정리

- **Observation**
  - 개별 관측대상에 대한 다양한 속성 데이터들의 모음인 레코드
  - Observation은 보통 데이터프레임의 행이 된다

- **Feature**
  - 공통의 속성을 갖는 일련의 데이터로 관측대상에 대한 정보
  - Feature는 보통 데이터프레임의 열이 된다.
  - 데이터 마이닝Data Mining 분야에서는 feature라는 용어를 사용하지만, 통계 등의 분야에서 variable과 같은 의미

설명예제

- 용어 설명예제
  - Observation은 각각의 행에 있는 학생으로 관측대상이 된다
  - 각각의 열에 있는 학교나 나이가 feature가 된다 즉, 관측대상인 학생의 정보들이 feature가 된다

	나이	성별	학교	수학	과학
준서	15	남	덕영중	80	90
예은	17	여	수리중	80	80
길동	16	여	연수중	80	70
희수	15	남	반포중	100	95

Dimensionality Reduction

- **차원** Dimension
  - 데이터 마이닝 관점에서는 Observation을 표현하기 위한 Feature의 개수가 차원이 된다
  - 즉, 변수가 늘어나면 차원이 커지면서 문제 발생 -> 차원 축소 필요

- **차원축소** Dimensionality Reduction
  - Feature Selection (특성선택, 변수선택)

    : 가지고 있는 특성 중에서 훈련에 가장 유용한 특성을 선택하는 것을 말한다. (다른 표혀으로는 원본 데이터의 불필요한 특징을 제거한다)
  - Feature Extraction (특성추출, 변수추출)

    : 원본 데이터의 특징들의 조합으로 새로운 특징을 생성한다

참고도서

- **reference**

[1] 파이썬을 활용한 데이터길들이기

- 프로그래밍인사이트

[2] 모두의 데이터분석

- 길벗

[3] 파이썬머신러닝 판다스데이터분석

- 정보문화사

End