Ch.2 데이터 정제(Data Cleaning)

실무에서의 데이터는 비어 있거나(missing value), 오류값이 들어 있거나 정합성이 맞는 않는 경우가 많다. 불완전한 데이터로 분석 작업을 수행했을 때는 분석 결과 또한 신뢰성이 반감된다. 그러므로, 본격적인 분석 작업을 수행하기 전에 데이터의 불완전성을 최대한 제거하는 것이 필요하다.

2.1 결측값(missing value)의 처리

- 결측값(missing value)
  - 존재하지 않고 비어있는 상태
  - DB에서의 NULL값
- 결측값을 채우는 방법
  1. 해당 투플을 무시한다 (row-wise deletion)
  2. 결측값을 수동으로 채워넣는다
  3. 전역상수(global constant)를 사용하여 결측값을 채워 넣는다
  4. 속성의 평균값을 사용하여 결측값을 채워 넣는다
  5. 주어진 투플과 같은 클래스(분류)에 속하는 투플 들의 속성 평균값을 사용한다
  6. 가장 가능성이 높은 값(예측)으로 결측값을 채워 넣는다 (회귀분석, 베이지안기법, 의사결정트리 기법 등)

2.1 결측값(missing value)의 처리

결측값 처리 예제

A	B
A01	10
A01	
A01	20
A02	30

① 해당 투플을 무시 (row-wise deletion)

A	B
A01	10
A01	20
A02	30

③ 전역상수(global constant) 사용

A	B
A01	10
A01	**0**
A01	20
A02	30

④ 속성의 평균값 사용

A	B
A01	10
A01	**20**
A01	20
A02	30

2.1 결측값(missing value)의 처리

결측값 처리 예제

A	B
A01	10
A01	
A01	20
A02	30

⑤ 주어진 투플과 같은 클래스(분류)에 속하는 투플 들의 속성 평균값 사용

A	B
A01	10
A01	15
A01	20
A02	30

⑥ 가장 가능성이 높은 값(예측)으로 결측값을 채워 넣는다

A	B
A01	10
A01	?
A01	20
A02	30

2.1 결측값(missing value)의 처리 - 실무예제

다음은 부산광역시 사상구 약수터 수질 현황 표(검사일: 2015년 11월 5일)이다. 표상에 나타난 결측 속성값(일반세균, 질산성질소)을 채우시오.

⊙ 데이터 파일: ch2-1(약수터 수질 현황).csv

⊙ 원본 투플 수: 24개

연번	약수터명	동명	총대장균군	일반세균	질산성질소	적합
1	백수	모라	양 성	10	6.7	부적합
2	이칠	모라	음 성	20	0.9	적합
3	운수사	모라	음 성	10	1.1	적합

출처: 공공데이터포털(www.data.go.kr)

☞ **해답은 ch2-1.ipynb 참고**

2.2 잡음(noisy data) 제거

- 잡음(noise)
  - 변수(속성)에서의 오류나 오차 값
  - 오류나 오차에 의한 값의 경향성 훼손을 줄이기 위해서 데이터 평활화 기법(smoothing technique)을 적용
- 데이터 평활화 기법
  - 구간화(Binning)

    정렬된 데이터 값들을 몇 개의 빈(혹은 버킷)으로 분할하여 평활화하는 방법

    ① 평균값 평활화(smoothing by bin means)

    ② 중앙값 평활화(smoothing by bin medians)

    ③ 경계값 평활화(smoothing by bin boundaries)
  - 구간화(Binning) 방식

    ① 동일 너비 방식

    ② 동일 높이 방식

2.2 잡음(noisy data) 제거

평활화 예제 [동일 너비(범위) 방식]

2.2 잡음(noisy data) 제거

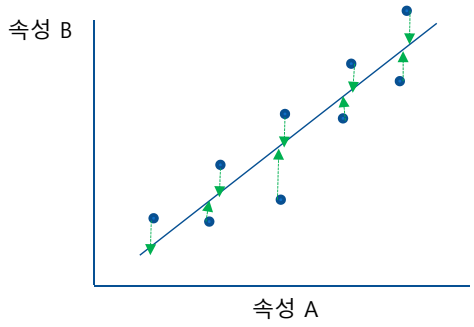
평활화 예제 [동일 높이(개수) 방식]

2.2 잡음(noisy data) 제거

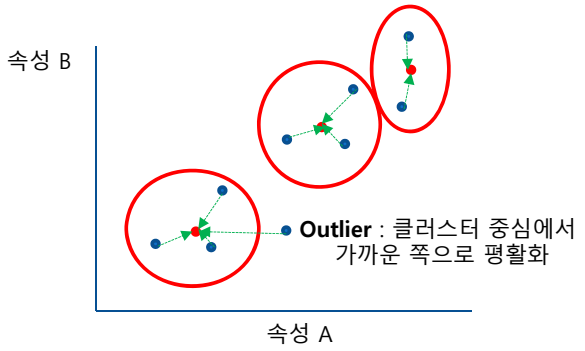
- 데이터 평활화 기법
  - 회귀(Regression)

    회귀 함수에 의한 데이터 평활화 기법
    - ✓ 선형회귀분석 : 하나의 속성값으로 다른 하나의 속성값 예측
    - ✓ 다중회귀분석 : 두 개 이상의 속성값으로 다른 속성값을 예측
  - 군집화(Clustering)

    유사한 값들끼지 그룹화하는 기법 (outlier는 평활화 대상)



< 회귀에 의한 평활화 >



< 군집화에 의한 평활화 >

2.2 잡음(noisy data) 제거 - 실무예제

실무 예제 2-2

다음은 2016년 항만별 선박 입출항 현황이다. 선박 수 속성은 구간화(Binning) 기법으로 평활화하고, 선박 톤수 속성은 선박 수 속성과의 회귀(regression) 분석을 통하여 평활화시키시오.

- ⊙ 데이터 파일 : ch2-2(선박입출항).csv
- ⊙ 원본 투플 수 : 30개

항만	입항선박수	입항선박톤수	출항선박수	출항선박톤수
부산	7,301	105,138,280	7,409	103,857,903
인천	2,715	30,716,710	2,716	30,779,186
평택.당진	1,558	23,153,226	1,536	22,778,109
경인항	28	126,236	27	128,344
동해.묵호	552	4,202,603	546	4,039,929

출처 공공데이터포털(www.data.go.kr)

☞ **해답은 ch2-2.ipynb 참고**

[여기서 잠깐!] PyCharm 실습환경 구축

- 왜 갑자기 PyCharm?
  - PyCharm : Python의 통합개발환경(IDE) 제공
  - PyCharm의 장점 : 디버깅(Debugging)
    - ✓ Python 코딩 시, 변수값의 변화를 보거나 오류를 잡을 때 유용
    - ✓ 복잡한 Python 로직(logic) 이해 시 도움
  - PyCharm 다운로드 및 설치
    - ① https://www.jetbrains.com/pycharm/ 사이트 접속
    - ② PyCharm Community ver2018.3.5 다운로드 (무료)
    - ③ PyCharm 설치
  - 데이터전처리 실습환경 구축
    - ① PyCharm 실행
    - ② 프로젝트(Project) 생성
    - ③ 예제 소스코드(*.py)와 데이터 파일(*.csv)을 프로젝트에 삽입
    - ④ 패키지(Package) 설치