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