노트북: 첫 번째 노트북
만든 날짜: 2020-11-15 오후 10:31
URL: https://continuous-development.tistory.com/157?category=736685

ML,DL

[ML/DL]결측치의 종류와 결측치 처리 가이드라인

2020. 10. 26. 01:37 수정 삭제 공개

결측치의 종류

- 완전 무작위 결측(MCAR : Missing completely at random)

변수 상에서 발생한 결측치가 다른 변수들과 아무런 상관이 없는 경우 우리는 완전 무작위 결측(MCAR)이라고 부릅니다. 대부분의 결측치 처리 패키지가 MCAR을 가정으로 하고 있고 보통 우리가 생각하는 결측치라고 생각하시면 됩니다. 예를 들어, 데이터를 입력하는 사람이 깜빡하고 입력을 안 했다든지 전산오류로 누락된 경우 등입니다. 이러한 결측치는 보통 제거하거나 대규모 데이터 셋에서 단순 무작위 표본추출을 통해서 완벽한 데이터셋으로 만들 수 있습니다.

- 무작위 결측(MAR : Missing at random)

누락된 자료가 특정 변수와 관련되어 일어나지만, 그 변수의 결과는 관계가 없는 경우를 의미합니다. 그리고 누락이 전체 정보가 있는 변수로 설명이 될 수 있음을 의미합니다.(누락이 완전히 설명 될 수 있는 경우 발생) 예를 들어, 남성은 우울증 설문 조사에 기입 할 확률이 낮지만 우울함의 정도와는 상관이 없는 경우입니다.

- 비 무작위 결측(MNAR : Missing at not random)

위의 두가지 유형이 아닌 경우를 MNAR이라고 합니다. MNAR은 누락된 값(변수의 결과)이 다른 변수와 연관 있는 경우를 의미합니다. 위의 예시를 확장해서, 만약 남성이 우울증 설문 조사에 기입하는 게 우울증의 정도와 관련이 있다면 이것은 MNAR입니다.

결측치의 종류는 이렇게 3가지로 나눠져 있고 결측치의 종류에 따라 해야되는 결측치 방법이 달라진다.

#결측치 가이드라인

두번째로는 결측치 가이드라인이다.

> **결측값(결측치) 처리 가이드라인**

- `10% 미만` : `삭제` or `대치`
- `10 ~ 20%` : `Hot deck` (매년자료->해당년자료 추정) or `regression` or `model based imputation`
- `20 ~ 50% 이상` : `regression` or `model based imputation`
- `50% 이상` : 해당 `칼럼(변수)자체 제거`

이러한 형태로 결측치를 처리한다고 나와있긴 하다.

하지만 다른 사람들의 자료나 kaggle을 보면 결측치 제거에 대한 부분은 원본을 훼손할 가능성이 있어 삭제하지 않는 게 좋다고 하는 글도 있었고 20프로 이상일 때 삭제하는 경우도 있었다.

또한 칼럼자체가 필요 없다고 생각 될 때는 그 컬럼 자체를 삭제하는 경우도 있었다.

이 부분은 공부를 더해야 될 것 같다. 아직 답은 없지만 일반적으로는 이렇게 한다라고 나와 있었다.

'ML,DL' 카테고리의 다른 글□

[ML/DL] 데이터 인코딩 - Label Encoding / One-hot Encoding/ dummies□

[ML/DL] 파이썬(python)을 이용한 분류(Classification)하기□

[ML/DL] 대체법의 종류와 다중 대체법 사용법□

[ML,DL] 머신러닝에 대한 간단한 개념들과 사용 하는 주요 패키지□

[ML/DL]결측치의 종류와 결측치 처리 가이드라인□

[ML,DL] 머신러닝(Machine lerning)과 딥러닝(Deep lerning)의 정의와 차이점□

결측치 가이드라인 결측치 종류 결측치 처리

나아무늘보

혼자 끄적끄적하는 블로그 입니다.