기계학습

질문 / 응답

2021840216
서동현
2019040255
강성욱

① 데이터를 나누는 다른방법이 있는가?

서동현: data를 자동적이게 하기위해 train_test, test_set을 7:3으로 나누면 어느 상황이든 자동적으로 할 수 있는 것 같다.

7:3

김성욱: 보통 data를 train_set, test_set을 8:2로 나누는데 좀더 정확한 확률을 하기위해 Train_set의 8에서 Train_test 할 부분을 나눈다.

∴ (Train_data-, Train_test) : Test_data

= 8:2

② 무작위 샘플링 VS 계층적 샘플링 ?

서동현: 계층적 샘플링이 좋을 것 같다. 편향 발생률이 적어 더 좋은 예측을 할 수 있을 것 같다.

김성욱: 무작위샘플링도 괜찮을 수도 있는거 정해진 틀에서 데이터를 나누면 무작위적인 데이터가 들어왔을때 오히려 예측을 못 할것 같다.

질문?

강성욱: 지도, 비지도, 강화학습을
고르는 기준이 뭘까?
무엇보고 결정하지?

서동현 = 지도는 경험O, 비지도는
경험 X, 강화학습은 보정이
필요할때

강성욱: Label 이 무슨 이야기인지?

서동현 = ㅇㅇㅋ

서동현: 데이터 전처리에서 null
채워지면 null값을 왜
고려해야하는가?

강성욱 = null값으로 채워진 데이터는
학습에 아무런 영향을 주지
않는 데이터이기에 학습에
도움을주는 방법으로

바꿔주는게 학습한 모델의
정확도를 올려준다.

더 오래 기억하기 위한 부분

** 전처리 **
데이터 전처리 : 모델 학습을 효율적으로 진행하기 위해 주어진 데이터를 변환시키는 것

1. 수치형 데이터 전처리 과정 :
*null 값 처리 방법
1) 해당 구역 제거 - dropna
2) 전체 특성 제거 - drop
3) 특정 값으로 채우기 - 0, 평균, 중간값 등

2. 텍스트와 범주형 :
* 원-핫 인코딩 : 더미 특성을 추가하여 활용
1) 해당 카테고리의 특성 값 : 1
2) 나머지 카테고리의 특성 값 : 2
- OneHotEncoder 클래스 제공

3. 나만의 변환

4. 특성 스케일 : 수치 데이터 전처리 과정