John

Follow 53 Followers About

데이터 전처리 : 레이블 인코딩과 원핫 인코딩

John Feb 3, 2020 · 4 min read

기본적으로 사이킷런의 머신러닝 알고리즘은 문자열 값을 입력 값으로 허락하지 않는다.

그렇기 때문에 모든 문자열 값들을 숫자 형으로 인코딩하는 전처리 작업 후에 머신러닝 모델에 학습을 시켜야한다.
이렇게 인코딩 하는 방식에는 크게 레이블 인코딩**(Lable encoding)**과 원-핫 인코딩**(One Hot Encoding)**이 있다.

레이블 인코딩

```
from sklearn.preprocessing import LabelEncoder

items=['트와이스','BTS','레드벨벳','신화','GOD','GOD']

# LabelEncoder를 객체로 생성한 후 , fit( ) 과 transform( ) 으로 label 인코딩 수행.

encoder = LabelEncoder()
encoder.fit(items)

labels = encoder.transform(items)
print('인코딩 변환값:',labels)
```

```
인코딩 변환값: [4 0 2 3 1 1]
```

- 여기서 각 속성의 인코딩 전 원래의 값을 확인할 수 있는 방법이 있다.

```
print('인코딩 클래스:',encoder.classes_)
```

인코딩 클래스: ['BTS' 'GOD' '레드벨벳' '신화' '트와이스']

0번 부터 순서대로 속성 값을 알 수 있다.

- 레이블 인코딩의 문제점

1. 일괄적인 숫자 값으로 변환되면서 예측 성능이 떨어질 수 있다.

-> 숫자의 크고 작음에 대한 특성이 작용

2. 선형 회귀와 같은 ML 알고리즘에는 적용하지 않아야 함 (트리 계열의 ML알고리즘은 숫자의 이러한 특성을 반영하지 않으므로 괜찮음)

원 — 핫 인코딩

간단하게 피처 값의 유형에 따라 새로운 피처를 추가해 고유 값에 해당하는 칼럼에만 1을 표시하고 나머지 칼럼에는 0을 표시하는 방법이다.

Label Encoding

Food Name	Categorical #	Calories
Apple	1	95
Chicken	2	231
Broccoli	3	50

→

One Hot Encoding

Apple	Chicken	Broccoli	Calories
1	0	0	95
0	1	0	231
0	0	1	50

예시

- SKlearn (사이킷런 사용 — 상대적으로 복잡)

-> 레이블 인코딩을 한번 거쳐야함.

```
# 먼저 숫자값으로 변환을 위해 LabelEncoder로 변환합니다.
encoder = LabelEncoder()
encoder.fit(items)
labels = encoder.transform(items)
# 2차원 데이터로 변환합니다.
labels = labels.reshape(-1,1)

# 원-핫 인코딩을 적용합니다.
oh_encoder = OneHotEncoder()
oh_encoder.fit(labels)
oh_labels = oh_encoder.transform(labels)
print('원-핫 인코딩 데이터')
print(oh_labels.toarray())
print('원-핫 인코딩 데이터 차원')
print(oh_labels.shape)
```

- pandas (get_dummies() 함수 사용)

```
import pandas as pd

df = pd.DataFrame({'item':['트와이스','BTS','레드벨벳','신
화','GOD','GOD'] })
df

# pd.get_dummies(df)# 원핫인코딩 실행
```

	item
0	트와이스
1	BTS
2	레드벨벳
3	신화
4	GOD
5	GOD

	item_BTS	item_GOD	item_레드벨벳	item_신화	item_트와이스
0	0	0	0	0	1
1	1	0	0	0	0
2	0	0	1	0	0
3	0	0	0	1	0
4	0	1	0	0	0
5	0	1	0	0	0

전(왼) / 후(오)

결론적으로는 pandas의 get_dummies 함수가 가장 좋은 방법으로 생각된다.

One Hot Encoding