[ML/DL] 데이터 인코딩 - Label Encoding / One-hot Encoding/ dummies — 나무늘보의 개발 블로그

노트북: 첫 번째 노트북
만든 날짜: 2020-12-28 오후 1:28
URL: https://continuous-development.tistory.com/164?category=736685

ML,DL

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

2020. 10. 28. 11:27 수정 삭제 공개

데이터 인코딩이란?

머신러닝 알고리즘은 문자열 데이터 속성을 입력받지 않으며 모든 데이터는 숫자형으로 표현되어야 한다.
그래서 문자형 카테고리형 속성은 모두 숫자 값으로 변환/인코딩 되어야 한다.

인코딩의 종류

label Encoding - 범주형 변수의 문자열을 수치형으로 변환

One-hot Encoding - 피처값의 유형에 따라 새로운 피처를 추가해 고윳값에 해당하는 칼럼에만 1을 표시하고 나머지 칼럼에는 0을 표시한다.

get_dummies() - pandas에서 제공해주는 함수로서 더미의 가변수를 만들어준다.

예제)

label Encoding

```
#[실습] breast_cancer
from sklearn.datasets import load_iris, load_breast_cancer
from sklearn.tree     import DecisionTreeClassifier
from sklearn.model_selection import GridSearchCV,train_test_split
from sklearn.metrics import accuracy_score

from sklearn.preprocessing import LabelEncoder

import pandas as pd
import numpy  as np
```

```
item_label = ['TV','냉장고','전자렌지','컴퓨터','선풍기','선풍기','믹서','믹서']
encoder = LabelEncoder() # labelEncoder함수를 가져온다.
encoder.fit(item_label) # 이걸 내가 가지고 있는 데이터에 fit한다.
digit_label = encoder.transform(item_label) # transform으로 변환한다.
print('encoder',encoder)
print('encoder 결과',digit_label)
print("*"*50)
print('decoder 결과',encoder.inverse_transform(digit_label)) # 변환 했던것을 다시 변환한다.
```

```
encoder LabelEncoder()
encoder 결과 [0 1 4 5 3 3 2 2]
**************************************************
decoder 결과 ['TV' '냉장고' '전자렌지' '컴퓨터' '선풍기' '선풍기' '믹서' '믹서']
```

One-hot encoding

```
from sklearn.preprocessing import OneHotEncoder
item_label = ['TV','냉장고','전자렌지','컴퓨터','선풍기','선풍기','믹서','믹서']
encoder = LabelEncoder() # labelEncoder함수를 가져온다.
encoder.fit(item_label) # 이걸 내가 가지고 있는 데이터에 fit한다.
digit_label = encoder.transform(item_label) # transform으로 변환한다.

print('type', type(digit_label))

# 2차원 데이터로 변환
digit_label = digit_label.reshape(-1,1)
print(digit_label)
print(digit_label.shape)

# One-Hot 인코딩
one_hot_encoder = OneHotEncoder()
one_hot_encoder.fit(digit_label)
one_hot_label = one_hot_encoder.transform(digit_label)
print(one_hot_label.toarray())
print(one_hot_label.shape)
```

```
type <class 'numpy.ndarray'>
[[0]
 [1]
 [4]
 [5]
 [3]
 [3]
 [2]
 [2]]
(8, 1)
[[1. 0. 0. 0. 0. 0.]
 [0. 1. 0. 0. 0. 0.]
 [0. 0. 0. 0. 1. 0.]
 [0. 0. 0. 0. 0. 1.]
 [0. 0. 0. 1. 0. 0.]
 [0. 0. 0. 1. 0. 0.]
 [0. 0. 1. 0. 0. 0.]
 [0. 0. 1. 0. 0. 0.]]
(8, 6)
```

get_dummies()

```
one_hot_df = pd.DataFrame({'item':['TV', '냉장고', '전자렌지', '컴퓨터', '선풍기', '선풍기', '믹서','믹서']})
pd.get_dummies(one_hot_df)
```

	item_TV	item_냉장고	item_믹서	item_선풍기	item_전자렌지	item_컴퓨터
0	1	0	0	0	0	0
1	0	1	0	0	0	0
2	0	0	0	0	1	0
3	0	0	0	0	0	1
4	0	0	0	1	0	0
5	0	0	0	1	0	0
6	0	0	1	0	0	0
7	0	0	1	0	0	0

'ML,DL' 카테고리의 다른 글

[ML/DL] python 을 통한 교차검증 (k -Fold , stratifiedkFold)

[ML/DL] python 을 통한 결측값 확인 및 결측치 처리 방법

[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

data encoding get_dummies() label encoding One-Hot Encoding 데이터 인코딩 라벨 인코딩 원핫 인코딩

나아무늘보

혼자 끄적끄적하는 블로그 입니다.