

더미변수

#01. 더미변수의 이해

#02. 작업 준비

패키지 참조

샘플 데이터 가져오기

#03. 더미변수 생성

성별에 대한 처리

더미변수 만들기

생성된 변수의 이름만 추출

값만 추출

데이터프레임으로 변환

성별에 대해 모든 경우의 수에
대한 더미변수를 생성하는 경
우더미변수 생성 및 데이터프레임
구성

비만도에 대한 더미변수 생성

N-1개의 변수 생성

N개의 변수 생성

성별 + 비만도

N-1개

더미변수

#01. 더미변수의 이해

명목형 변수를 연속형 변수 "스럽게" 변환한 것.

카테고리 형태의 데이터를 `0, 1, 2` 등의 연속형 숫자로 변환한 형태이다.

ex) 남자, 여자 --> 0, 1

기존의 범주형 변수를 이진 변수로 대체하여 모델에 적용할 수 있다.

일반적으로 머신 러닝 알고리즘들은 연속적인 숫자를 다루는 데 더 효과적이기 때문에 더미 변수 변환이 필요함.

#02. 작업 준비

패키지 참조

`patsy` 패키지의 설치가 필요하다.

```
from pandas import read_excel, DataFrame, get_dummies
from patsy import dmatrix
import numpy as np
```

샘플 데이터 가져오기

더미변수

#01. 더미변수의 이해

#02. 작업 준비

패키지 참조

샘플 데이터 가져오기

#03. 더미변수 생성

성별에 대한 처리

더미변수 만들기

생성된 변수의 이름만 추출

값만 추출

데이터프레임으로 변환

성별에 대해 모든 경우의 수에
대한 더미변수를 생성하는 경
우

더미변수 생성 및 데이터프레임
구성

비만도에 대한 더미변수 생성

N-1개의 변수 생성

N개의 변수 생성

성별 + 비만도

N-1개

```
df = read_excel("https://data.hossam.kr/C02/dum.xlsx")
df.head()
```

	성별	비만도
0	남자	정상
1	여자	경도
2	여자	정상
3	남자	고도
4	남자	정상

#03. 더미변수 생성

성별에 대한 처리

더미변수 만들기

남자 와 여자 에 대해서 더미변수를 생성한 결과 여자인 경우 1 로 표시되는 하나의 더미변수가 생성된다.

```
dv = dmatrix('성별', df)
dv
```

더미변수

#01. 더미변수의 이해

#02. 작업 준비

패키지 참조

샘플 데이터 가져오기

#03. 더미변수 생성

성별에 대한 처리

더미변수 만들기

생성된 변수의 이름만 추출

값만 추출

데이터프레임으로 변환

성별에 대해 모든 경우의 수에
대한 더미변수를 생성하는 경
우더미변수 생성 및 데이터프
레임 구성

비만도에 대한 더미변수 생성

N-1개의 변수 생성

N개의 변수 생성

성별 + 비만도

N-1개

DesignMatrix with shape (20, 2)

Intercept 성별[T.여자]

1 0

1 1

1 1

1 0

1 0

1 0

1 0

1 1

1 1

1 0

1 1

1 1

1 1

1 0

1 1

1 1

1 0

1 1

1 1

1 1

Terms:

'Intercept' (column 0)

'성별' (column 1)

생성된 변수의 이름만 추출

더미변수

#01. 더미변수의 이해

#02. 작업 준비

패키지 참조

샘플 데이터 가져오기

#03. 더미변수 생성

성별에 대한 처리

더미변수 만들기

생성된 변수의 이름만 추출

값만 추출

데이터프레임으로 변환

성별에 대해 모든 경우의 수에
대한 더미변수를 생성하는 경
우더미변수 생성 및 데이터프레임
구성

비만도에 대한 더미변수 생성

N-1개의 변수 생성

N개의 변수 생성

성별 + 비만도

N-1개

`dv.design_info.column_names``['Intercept', '성별[T.여자]']`

값만 추출

`dmarray = np.asarray(dv)`
`dmarray`

```
array([[1., 0.],
       [1., 1.],
       [1., 1.],
       [1., 0.],
       [1., 0.],
       [1., 0.],
       [1., 1.],
       [1., 1.],
       [1., 0.],
       [1., 1.],
       [1., 1.],
       [1., 1.],
       [1., 0.],
       [1., 1.],
       [1., 1.],
       [1., 0.],
       [1., 1.],
       [1., 1.],
       [1., 0.]])
```

더미변수

#01. 더미변수의 이해

#02. 작업 준비

패키지 참조

샘플 데이터 가져오기

#03. 더미변수 생성

성별에 대한 처리

더미변수 만들기

생성된 변수의 이름만 추출

값만 추출

데이터프레임으로 변환

성별에 대해 모든 경우의 수에
대한 더미변수를 생성하는 경
우더미변수 생성 및 데이터프레임
구성

비만도에 대한 더미변수 생성

N-1개의 변수 생성

N개의 변수 생성

성별 + 비만도

N-1개

```
[1., 1.],
[1., 1.],
[1., 1.]])
```

데이터프레임으로 변환

데이터프레임 생성 후 절편인 Intercept 필드는 제거한다.

```
dummy_df = DataFrame(np.asarray(dv), columns=dv.design_info.column_names)
dummy_df.drop('Intercept', axis=1, inplace=True)
dummy_df.head()
```

	성별[T.여자]
0	0.0
1	1.0
2	1.0
3	0.0
4	0.0

성별에 대해 모든 경우의 수에 대한 더미변수를 생성하는 경우

더미 변수들은 서로 상관관계가 있으므로, 다중공선성이 발생할 수 있다.

이러한 문제를 피하기 위해 일반적으로 N-1개의 더미 변수를 생성한다.

더미변수

#01. 더미변수의 이해

#02. 작업 준비

패키지 참조

샘플 데이터 가져오기

#03. 더미변수 생성

성별에 대한 처리

더미변수 만들기

생성된 변수의 이름만 추출

값만 추출

데이터프레임으로 변환

성별에 대해 모든 경우의 수에
대한 더미변수를 생성하는 경
우더미변수 생성 및 데이터프레임
구성

비만도에 대한 더미변수 생성

N-1개의 변수 생성

N개의 변수 생성

성별 + 비만도

N-1개

예를 들어, 성별의 경우 "남성"과 "여성" 두 가지 범주만 있으므로 한 개의 더미 변수만 생성하고, 다른 하나는 자동으로 포함시키지 않는다.

그러므로 이 방법은 필요한 경우가 아닌 이상 사용하지 않는것이 좋다.

더미변수 생성 및 데이터프레임 구성

컬럼이름을 지정하는 문자열에 항상 `+ 0` 을 추가해야 한다

```
dv = dmatrix('성별 + 0', df)
dummy_df = DataFrame(np.asarray(dv), columns=dv.design_info.column_names)
dummy_df.head()
```

	성별[남자]	성별[여자]
0	1.0	0.0
1	0.0	1.0
2	0.0	1.0
3	1.0	0.0
4	1.0	0.0

비만도에 대한 더미변수 생성

N-1 개의 변수 생성

데이터프레임 생성 후 `Intercept` 필드 삭제 필요

더미변수

#01. 더미변수의 이해

#02. 작업 준비

패키지 참조

샘플 데이터 가져오기

#03. 더미변수 생성

성별에 대한 처리

더미변수 만들기

생성된 변수의 이름만 추출

값만 추출

데이터프레임으로 변환

성별에 대해 모든 경우의 수에
대한 더미변수를 생성하는 경
우더미변수 생성 및 데이터프레임
구성

비만도에 대한 더미변수 생성

N-1개의 변수 생성

N개의 변수 생성

성별 + 비만도

N-1개

```
dv = dmatrix('비만도', df)
dummy_df = DataFrame(np.asarray(dv), columns=dv.design_info.column_names)
dummy_df.drop('Intercept', axis=1, inplace=True)
dummy_df.head()
```

	비만도[T.고도]	비만도[T.정상]
0	0.0	1.0
1	0.0	0.0
2	0.0	1.0
3	1.0	0.0
4	0.0	1.0

N 개의 변수 생성

표현식에 + 0 추가

데이터프레임 생성 후 Intercept 필드 삭제 안함

```
dv = dmatrix('비만도 + 0', df)
dummy_df = DataFrame(np.asarray(dv), columns=dv.design_info.column_names)
dummy_df.head()
```

더미변수

#01. 더미변수의 이해

#02. 작업 준비

패키지 참조

샘플 데이터 가져오기

#03. 더미변수 생성

성별에 대한 처리

더미변수 만들기

생성된 변수의 이름만 추출

값만 추출

데이터프레임으로 변환

성별에 대해 모든 경우의 수에
대한 더미변수를 생성하는 경
우더미변수 생성 및 데이터프레임
구성

비만도에 대한 더미변수 생성

N-1개의 변수 생성

N개의 변수 생성

성별 + 비만도

N-1개

	비만도[경도]	비만도[고도]	비만도[정상]
0	0.0	0.0	1.0
1	1.0	0.0	0.0
2	0.0	0.0	1.0
3	0.0	1.0	0.0
4	0.0	0.0	1.0

성별 + 비만도

N-1 개

```
dv = dmatrix('성별:비만도', df)
dummy_df = DataFrame(np.asarray(dv), columns=dv.design_info.column_names)
dummy_df.drop('Intercept', axis=1, inplace=True)
dummy_df.head()
```

	비만도[T. 고도]	비만도[T. 정상]	성별[T.여자]:비만 도[경도]	성별[T.여자]:비만 도[고도]	성별[T.여자]:비만 도[정상]
0	0.0	1.0	0.0	0.0	0.0
1	0.0	0.0	1.0	0.0	0.0
2	0.0	1.0	0.0	0.0	1.0
3	1.0	0.0	0.0	0.0	0.0

더미변수

#01. 더미변수의 이해

#02. 작업 준비

패키지 참조

샘플 데이터 가져오기

#03. 더미변수 생성

성별에 대한 처리

더미변수 만들기

생성된 변수의 이름만 추출

값만 추출

데이터프레임으로 변환

성별에 대해 모든 경우의 수에
대한 더미변수를 생성하는 경우더미변수 생성 및 데이터프레임
구성

비만도에 대한 더미변수 생성

N-1개의 변수 생성

N개의 변수 생성

성별 + 비만도

N-1개

	비만도[T. 고도]	비만도[T. 정상]	성별[T.여자]:비만 도[경도]	성별[T.여자]:비만 도[고도]	성별[T.여자]:비만 도[정상]
4	0.0	1.0	0.0	0.0	0.0

N 개

```
dv = dmatrix('성별:비만도+0', df)
dummy_df = DataFrame(np.asarray(dv), columns=dv.design_info.column_names)
dummy_df.head()
```

	성별[남자]: 비만도[경 도]	성별[여자]: 비만도[경 도]	성별[남자]: 비만도[고 도]	성별[여자]: 비만도[고 도]	성별[남자]: 비만도[정 상]	성별[여자]: 비만도[정 상]
0	0.0	0.0	0.0	0.0	1.0	0.0
1	0.0	1.0	0.0	0.0	0.0	0.0
2	0.0	0.0	0.0	0.0	0.0	1.0
3	0.0	0.0	1.0	0.0	0.0	0.0
4	0.0	0.0	0.0	0.0	1.0	0.0

#04. 원본 데이터가 라벨링 되어 있는 경우

예제를 위해 원본 데이터 라벨링 수행

더미변수

#01. 더미변수의 이해

#02. 작업 준비

패키지 참조

샘플 데이터 가져오기

#03. 더미변수 생성

성별에 대한 처리

더미변수 만들기

생성된 변수의 이름만 추출

값만 추출

데이터프레임으로 변환

성별에 대해 모든 경우의 수에
대한 더미변수를 생성하는 경
우더미변수 생성 및 데이터프레임
구성

비만도에 대한 더미변수 생성

N-1개의 변수 생성

N개의 변수 생성

성별 + 비만도

N-1개

df2 = df.copy()

컬럼을 구분하지 않고 모든 값을 변경

df2.replace("남자", 0, inplace=True)

df2.replace("여자", 1, inplace=True)

성별 컬럼에서만 변경

df2.replace({"비만도": "정상"}, 0, inplace=True)

df2.replace({"비만도": "경도"}, 1, inplace=True)

df2.replace({"비만도": "고도"}, 2, inplace=True)

df2.head()

	성별	비만도
0	0	0
1	1	1
2	1	0
3	0	2
4	0	0

라벨링 된 데이터의 더미 변수화

각 변수별로 더미변수 변환

각 필드이름에 범주형(Category)임을 의미하는 **C**를 표기한다.각각 필드를 **+**로 연결한다.

더미변수

#01. 더미변수의 이해

#02. 작업 준비

패키지 참조

샘플 데이터 가져오기

#03. 더미변수 생성

성별에 대한 처리

더미변수 만들기

생성된 변수의 이름만 추출

값만 추출

데이터프레임으로 변환

성별에 대해 모든 경우의 수에
대한 더미변수를 생성하는 경
우

더미변수 생성 및 데이터프레임
구성

비만도에 대한 더미변수 생성

N-1개의 변수 생성

N개의 변수 생성

성별 + 비만도

N-1개

```
dm = dmatrix('C(성별)+C(비만도)', df2)
dummy_df = DataFrame(np.asarray(dm), columns=dm.design_info.column_names)
dummy_df.drop('Intercept', axis=1, inplace=True)
dummy_df.head()
```

	C(성별)[T.1]	C(비만도)[T.1]	C(비만도)[T.2]
0	0.0	0.0	0.0
1	1.0	1.0	0.0
2	1.0	0.0	0.0
3	0.0	0.0	1.0
4	0.0	0.0	0.0

각 변수를 교차로 조합한 형태로 더미변수 생성

남자:정상, 남자:경도, 남자:고도, 여자:정상, 여자:경도, 여자:고도 형태로 모든 경우의 수를 조합하여 더미변수를 생성한다.

각 필드를 : 으로 연결한다.

```
dm = dmatrix('C(성별):C(비만도)', df2)
dummy_df = DataFrame(np.asarray(dm), columns=dm.design_info.column_names)
dummy_df.drop('Intercept', axis=1, inplace=True)
dummy_df.head()
```

더미변수

#01. 더미변수의 이해

#02. 작업 준비

패키지 참조

샘플 데이터 가져오기

#03. 더미변수 생성

성별에 대한 처리

더미변수 만들기

생성된 변수의 이름만 추출

값만 추출

데이터프레임으로 변환

성별에 대해 모든 경우의 수에
대한 더미변수를 생성하는 경
우더미변수 생성 및 데이터프
레임 구성

비만도에 대한 더미변수 생성

N-1개의 변수 생성

N개의 변수 생성

성별 + 비만도

N-1개

	C(비만도) [T.1]	C(비만도) [T.2]	C(성별)[T.1]:C(비 만도)[0]	C(성별)[T.1]:C(비 만도)[1]	C(성별)[T.1]:C(비 만도)[2]
0	0.0	0.0	0.0	0.0	0.0
1	1.0	0.0	0.0	1.0	0.0
2	0.0	0.0	1.0	0.0	0.0
3	0.0	1.0	0.0	0.0	0.0
4	0.0	0.0	0.0	0.0	0.0

#05. Pandas의 함수 사용하기

모든 필드를 더미변수로 변환 (N개 생성)

원본 데이터프레임과 더미변수로 변환할 컬럼 이름을 파라미터로 전달한다.

원본 데이터프레임에서 기존의 컬럼은 제거되고 더미변수로 변경된 컬럼들이 추가된다.

```
dummy_df = get_dummies(df, columns=['성별', '비만도'])
dummy_df.head()
```

	성별_남자	성별_여자	비만도_경도	비만도_고도	비만도_정상
0	True	False	False	False	True
1	False	True	True	False	False
2	False	True	False	False	True

더미변수

#01. 더미변수의 이해

#02. 작업 준비

패키지 참조

샘플 데이터 가져오기

#03. 더미변수 생성

성별에 대한 처리

더미변수 만들기

생성된 변수의 이름만 추출

값만 추출

데이터프레임으로 변환

성별에 대해 모든 경우의 수에
대한 더미변수를 생성하는 경
우더미변수 생성 및 데이터프레임
구성

비만도에 대한 더미변수 생성

N-1개의 변수 생성

N개의 변수 생성

성별 + 비만도

N-1개

	성별_남자	성별_여자	비만도_경도	비만도_고도	비만도_정상
3	True	False	False	True	False
4	True	False	False	False	True

N-1 개의 더미변수 생성

drop_first 파라미터를 True 로 설정한다.

```
dummy_df = get_dummies(df, columns=['성별', '비만도'], drop_first=True)
dummy_df.head()
```

	성별_여자	비만도_고도	비만도_정상
0	False	False	True
1	True	False	False
2	True	False	True
3	False	True	False
4	False	False	True

변환 데이터 타입 지정

```
dummy_df = get_dummies(df, columns=['성별', '비만도'], drop_first=True, dtype=int)
dummy_df.head()
```

더미변수

#01. 더미변수의 이해

#02. 작업 준비

패키지 참조

샘플 데이터 가져오기

#03. 더미변수 생성

성별에 대한 처리

더미변수 만들기

생성된 변수의 이름만 추출

값만 추출

데이터프레임으로 변환

성별에 대해 모든 경우의 수에
대한 더미변수를 생성하는 경
우더미변수 생성 및 데이터프
레임 구성

비만도에 대한 더미변수 생성

N-1개의 변수 생성

N개의 변수 생성

성별 + 비만도

N-1개

	성별_여자	비만도_고도	비만도_정상
0	0	0	1
1	1	0	0
2	1	0	1
3	0	1	0
4	0	0	1