원핫 인코딩 실습

학습 목표

01 pd.get_dummy를 이용한 원핫 인코딩 실습
02 성인 인구조사 소득 데이터 셋(adult.data)을 활용한 onehot encoding 실습

목차

01. 원핫 인코딩 실습

목차로 이동하기

- 데이터 셋을 불러와 원핫 인코딩 실습
- hello world 원핫 인코딩 실습

In [1]:
```
import mglearn
import pandas as pd
import os
```

In [2]:
```
demo_df = pd.DataFrame({"Product":['양말', '여우', '양말', '상자']})
display(demo_df)
```

	Product
0	양말
1	여우
2	양말
3	상자

In [3]:
```
onehot = pd.get_dummies(demo_df)
onehot
```

Out[3]:

	Product_상자	Product_양말	Product_여우
0	0	1	0
1	0	0	1
2	0	1	0
3	1	0	0

In [4]:
```
df = pd.concat([demo_df, onehot], axis=1)
df
```

Out[4]:

	Product	Product_상자	Product_양말	Product_여우
0	양말	0	1	0
1	여우	0	0	1
2	양말	0	1	0
3	상자	1	0	0

02. adult.data 셋을 활용한 onehot encoding 실습

목차로 이동하기

In [5]:
```
path = os.path.join(mglearn.datasets.DATA_PATH, 'adult.data')
print(path)
```

```
C:₩Users₩totofriend₩anaconda3₩lib₩site-packages₩mglearn₩data₩adult.data
```

In [6]:
```
data = pd.read_csv(path,
                   header=None,
                   index_col=False,
                   names=['age', 'workclass', 'fnlwgt', 'education',
                          'education-num', 'marital-status', 'occupation', 'relationship
                          'race', 'gender', 'capital-gain', 'capital-loss',
                          'hours-per-week', 'native-country', 'income'])
data.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 32561 entries, 0 to 32560
Data columns (total 15 columns):
 #   Column          Non-Null Count  Dtype
---  ------          --------------  -----
 0   age             32561 non-null  int64
 1   workclass       32561 non-null  object
 2   fnlwgt          32561 non-null  int64
 3   education       32561 non-null  object
 4   education-num   32561 non-null  int64
 5   marital-status  32561 non-null  object
 6   occupation      32561 non-null  object
 7   relationship    32561 non-null  object
 8   race            32561 non-null  object
 9   gender          32561 non-null  object
 10  capital-gain    32561 non-null  int64
 11  capital-loss    32561 non-null  int64
 12  hours-per-week  32561 non-null  int64
 13  native-country  32561 non-null  object
 14  income          32561 non-null  object
dtypes: int64(6), object(9)
memory usage: 3.7+ MB
```

In [8]:
```
data.columns
```

Out[8]:
```
Index(['age', 'workclass', 'fnlwgt', 'education', 'education-num',
       'marital-status', 'occupation', 'relationship', 'race', 'gender',
       'capital-gain', 'capital-loss', 'hours-per-week', 'native-country',
       'income'],
      dtype='object')
```

일부 변수 선택 후, 진행

In [ ]:
```

```

In [9]:
```
sel = ['age', 'workclass','education','gender','hours-per-week',
       'occupation','income']
data = data[sel]
data.head()
```

Out[9]:

	age	workclass	education	gender	hours-per-week	occupation	income
0	39	State-gov	Bachelors	Male	40	Adm-clerical	<=50K
1	50	Self-emp-not-inc	Bachelors	Male	13	Exec-managerial	<=50K
2	38	Private	HS-grad	Male	40	Handlers-cleaners	<=50K
3	53	Private	11th	Male	40	Handlers-cleaners	<=50K
4	28	Private	Bachelors	Female	40	Prof-specialty	<=50K

의미 있는 범주형 데이터 있는지 확인

In [10]:
```
print(data.gender.value_counts())
```

```
 Male      21790
 Female    10771
Name: gender, dtype: int64
```

pandas에서 get_dummies 함수를 이용하여 인코딩

In [11]:
```
print("원본 특성 :₩n", list(data.columns), "₩n")
data_dummies = pd.get_dummies(data)
print("get_dummies 후 특성 : ₩n", list(data_dummies.columns))
```

```
원본 특성 :
 ['age', 'workclass', 'education', 'gender', 'hours-per-week', 'occupation', 'incom
e']

get_dummies 후 특성 :
 ['age', 'hours-per-week', 'workclass_ ?', 'workclass_ Federal-gov', 'workclass_ Loc
al-gov', 'workclass_ Never-worked', 'workclass_ Private', 'workclass_ Self-emp-inc',
'workclass_ Self-emp-not-inc', 'workclass_ State-gov', 'workclass_ Without-pay', 'ed
ucation_ 10th', 'education_ 11th', 'education_ 12th', 'education_ 1st-4th', 'educati
on_ 5th-6th', 'education_ 7th-8th', 'education_ 9th', 'education_ Assoc-acdm', 'educ
ation_ Assoc-voc', 'education_ Bachelors', 'education_ Doctorate', 'education_ HS-gr
ad', 'education_ Masters', 'education_ Preschool', 'education_ Prof-school', 'educat
ion_ Some-college', 'gender_ Female', 'gender_ Male', 'occupation_ ?', 'occupation_
Adm-clerical', 'occupation_ Armed-Forces', 'occupation_ Craft-repair', 'occupation_
Exec-managerial', 'occupation_ Farming-fishing', 'occupation_ Handlers-cleaners', 'o
ccupation_ Machine-op-inspct', 'occupation_ Other-service', 'occupation_ Priv-house-
serv', 'occupation_ Prof-specialty', 'occupation_ Protective-serv', 'occupation_ Sal
es', 'occupation_ Tech-support', 'occupation_ Transport-moving', 'income_ <=50K', 'i
ncome_ >50K']
```

- age와 hours-per-week는 그대로이지만 범주형 특성은 새로운 특성으로 확장

특성을 포함한 열 'age'~'occupation_ Transport-moving' 모두 추출

In [12]:
```
features = data_dummies.loc[:, "age":"occupation_ Transport-moving"]
X = features.values
y = data_dummies['income_ >50K'].values
```

In [13]:
```
print("X.shape : {}, y.shape : {}".format(X.shape, y.shape))
```

```
X.shape : (32561, 44), y.shape : (32561,)
```

실습 1

- 로지스틱 모델을 만들어보기
  - (1) 데이터를 나누어준다.
  - (2) 모델을 만든다.
  - (3) 모델을 학습한다.(학습 데이터를 이용해서)
  - (4) score를 확인(테스트 데이터를 이용해서)

로지스틱 모델 사용해 보기

In [14]:
```
from sklearn.linear_model import LogisticRegression
from sklearn.model_selection import train_test_split

X_train, X_test, y_train, y_test = train_test_split(X, y, random_state=0)
logreg = LogisticRegression()
logreg.fit(X_train, y_train)
```

```
C:₩Users₩totofriend₩anaconda3₩lib₩site-packages₩sklearn₩linear_model₩_logistic.py:814: ConvergenceWarning: lbfgs failed to converge (status=1):
STOP: TOTAL NO. of ITERATIONS REACHED LIMIT.

Increase the number of iterations (max_iter) or scale the data as shown in:
    https://scikit-learn.org/stable/modules/preprocessing.html
Please also refer to the documentation for alternative solver options:
    https://scikit-learn.org/stable/modules/linear_model.html#logistic-regression
  n_iter_i = _check_optimize_result(
```

Out[14]:
```
LogisticRegression()
```

In [15]:
```
print("학습용 점수 {:.2f}".format(logreg.score(X_train, y_train)))
print("테스트 점수 {:.2f}".format(logreg.score(X_test, y_test)))
```

```
학습용 점수 0.81
테스트 점수 0.81
```

In [16]:
```
from sklearn.ensemble import RandomForestClassifier
from sklearn.neighbors import KNeighborsClassifier
```

랜덤 포레스트를 활용한 모델 구축

In [19]:
```
model = RandomForestClassifier().fit(X_train, y_train)

print("학습용 점수 {:.2f}".format(model.score(X_train, y_train)))
print("테스트 점수 {:.2f}".format(model.score(X_test, y_test)))
```

```
학습용 점수 0.94
테스트 점수 0.79
```

knn모델을 활용한 모델 구축

In [18]:
```
model = KNeighborsClassifier().fit(X_train, y_train)

print("학습용 점수 {:.2f}".format(model.score(X_train, y_train)))
print("테스트 점수 {:.2f}".format(model.score(X_test, y_test)))
```

학습용 점수 0.84
테스트 점수 0.78