2장 머신러닝 프로젝트 처음부터 끝까지 (2부)

감사의 글

자료를 공개한 저자 오렐리앙 제롱과 강의자료를 지원한 한빛아카데미에게 진심어린 감사를 전합니다.

2.5 머신러닝 알고리즘을 위한 데이터 준비

데이터 준비 자동화

- 모든 전처리 과정의 자동화. 언제든지 재활용 가능.
- 자동화는 **파이프라인**(pipeline)으로 구현
- 훈련 세트 준비: 훈련에 사용되는 특성과 타깃 특성(레이블) 구분하여 복사본 생성

```
housing = strat_train_set.drop("median_house_value", axis=1)
housing_labels = strat_train_set["median_house_value"].copy()
```

- 테스트 세트는 훈련이 완성된 후에 성능 측정 용도로만 사용.

데이터 전처리

- 데이터 전처리(data preprocessing): 효율적인 모델 훈련을 위한 데이터 변환
- 수치형 특성과 범주형 특성에 대해 다른 변환과정을 사용
- 수치형 특성 전처리 과정
    - 데이터 정제
    - 조합 특성 추가
    - 특성 스케일링
- 범주형 특성 전처리 과정
    - 원-핫-인코딩(one-hot-encoding)

변환 파이프라인

- **파이프라인**(pipeline)
    - 여러 과정을 한 번에 수행하는 기능을 지원하는 도구
    - 여러 사이킷런 API를 묶어 순차적으로 처리하는 사이킷런 API
- 파이프라인 적용 과정
    - 수치형 특성 전처리 과정에 사용된 세 가지 변환 과정의 자동화 파이프라인 구현
    - 수치형 특성 파이프라인과 범주형 특성 전처리 과정을 결합한 파이프라인 구현

사이킷런 API 활용

- '조합 특성 추가' 과정을 제외한 나머지 변환은 사이킷런에서 제공하는 관련 API 직접 활용 가능
- '조합 특성 추가' 과정도 다른 사이킷런 API와 호환이 되는 방식으로 사용자가 직접 구현 가능
- 사이킷런에서 제공하는 API는 일관되고 단순한 인터페이스를 제공

사이킷런 API의 세 가지 유형

추정기(estimator)

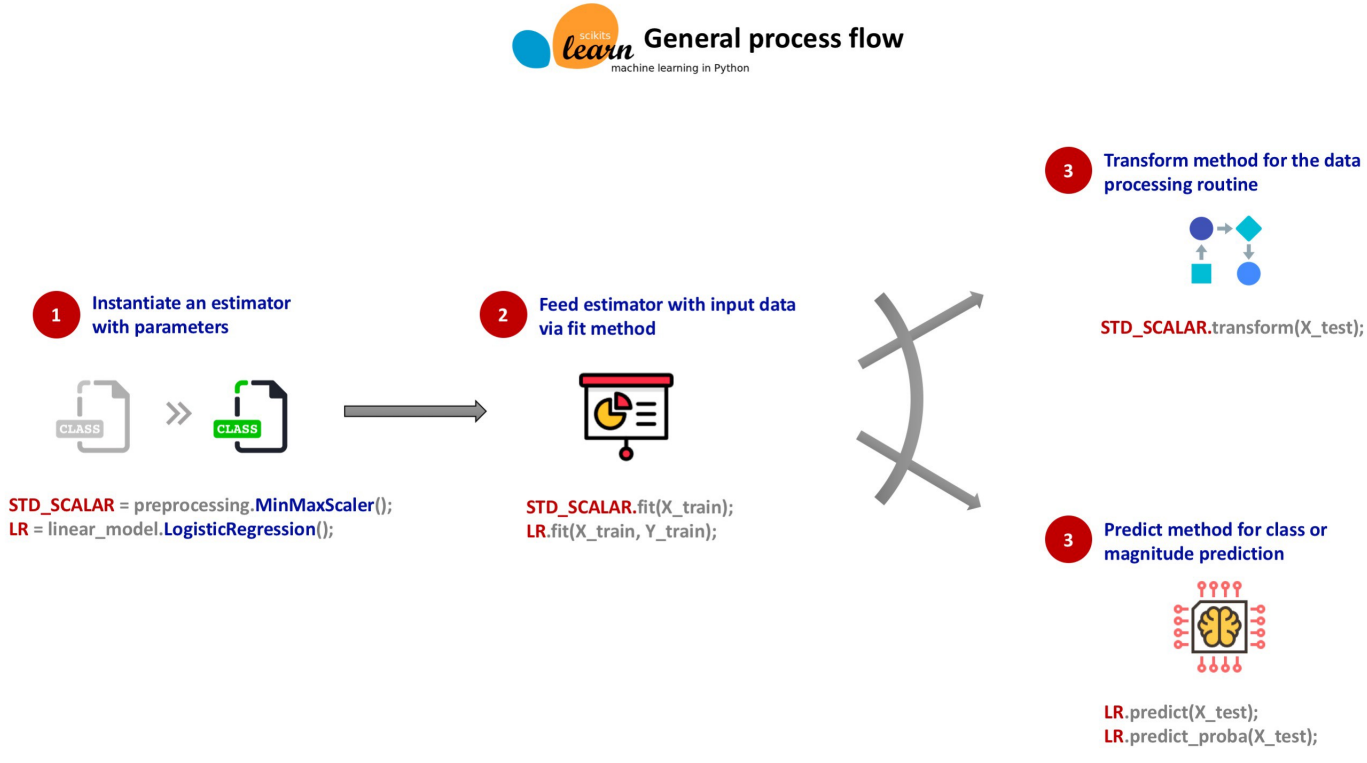
- 주어진 데이터셋과 관련된 특정 파라미터 값들을 추정하는 객체
- `fit()` 메서드 활용: 특정 파라미터 값을 저장한 속성이 업데이트된 객체 자신 반환

변환기(transformer):

- fit() 메서드에 의해 학습된 파라미터를 이용하여 주어진 데이터셋 변환
- `transform()` 메서드 활용
- `fit()` 메서드와 `transform()` 메서드를 연속해서 호출하는 `fit_transform()` 메서드 활용 가능

예측기(predictor)

- 주어진 데이터셋과 관련된 값을 예측하는 기능을 제공하는 추정기
- `predict()` 메서드 활용
- `fit()`과 `predict()` 메서드가 포함되어 있어야 함
- `predict()` 메서드가 추정한 값의 성능을 측정하는 `score()` 메서드도 포함
- 일부 예측기는 추정치의 신뢰도를 평가하는 기능도 제공



<그림 출처: Scikit-Learn: A silver bullet for basic machine learning>

2.5.1 데이터 정제: 수치형 특성 전치러 과정 1

- 누락된 특성값이 존재 경우, 해당 값 또는 특성을 먼저 처리해야 함.
- `total_bedrooms` 특성에 207개 구역에 대한 값이 null로 채워져 있음, 즉, 일부 구역에 대한 정보가 누락됨.

	longitude	latitude	housing_median_age	total_rooms	total_bedrooms	population	households	median_income	ocean_p
4629	-118.30	34.07	18.0	3759.0	NaN	3296.0	1462.0	2.2708	<1H
6068	-117.86	34.01	16.0	4632.0	NaN	3038.0	727.0	5.1762	<1H
17923	-121.97	37.35	30.0	1955.0	NaN	999.0	386.0	4.6328	<1H
13656	-117.30	34.05	6.0	2155.0	NaN	1039.0	391.0	1.6675	
19252	-122.79	38.48	7.0	6837.0	NaN	3468.0	1405.0	3.1662	<1H

null 값 처리 옵션

- 옵션 1: 해당 구역 제거
- 옵션 2: 전체 특성 삭제
- 옵션 3: 평균값, 중앙값, 0, 주변에 위치한 값 등 특정 값으로 채우기. 책에서는 중앙값으로 채움.

옵션	코드
옵션 1	`housing.dropna(subset=["total_bedrooms"])`
옵션 2	`housing.drop("total_bedrooms", axis=1)`
옵션 3	`median = housing["total_bedrooms"].median()`
	`housing["total_bedrooms"].fillna(median, inplace=True)`

<옵션 3 활용>

	longitude	latitude	housing_median_age	total_rooms	total_bedrooms	population	households	median_income	ocean_p
4629	-118.30	34.07	18.0	3759.0	433.0	3296.0	1462.0	2.2708	<1H
6068	-117.86	34.01	16.0	4632.0	433.0	3038.0	727.0	5.1762	<1H
17923	-121.97	37.35	30.0	1955.0	433.0	999.0	386.0	4.6328	<1H
13656	-117.30	34.05	6.0	2155.0	433.0	1039.0	391.0	1.6675	
19252	-122.79	38.48	7.0	6837.0	433.0	3468.0	1405.0	3.1662	<1H

SimpleImputer 변환기

- 옵션 3를 지원하는 사이킷런 변환기
- 중앙값 등 통계 요소를 활용하여 누락 데이터를 특정 값으로 채움

```
from sklearn.impute import SimpleImputer
imputer = SimpleImputer(strategy="median")
housing_num = housing.drop("ocean_proximity", axis=1)
imputer.fit(housing_num)
```

2.5.2 텍스트와 범주형 특성 다루기: 원-핫 인코딩

- 범주형 특성인 해안 근접도(ocean_proximity)에 사용된 5개의 범주를 수치형 특성으로 변환해야 함.

단순 수치화의 문제점

- 단순 수치화 적용 가능

범주	숫자
<1H OCEAN	0
INLAND	1
ISLAND	2
NEAR BAY	3
NEAR OCEAN	4

- 하지만 여기서는 다음 문제 발생
    - 해안 근접도는 단순히 구분을 위해 사용. 해안에 근접하고 있다 해서 주택 가격이 기본적으로 더 비싸지 않음.
    - 반면에 수치화된 값들은 크기를 비교할 수 있는 숫자
    - 따라서 모델 학습 과정에서 숫자들의 크기 때문에 잘못된 학습이 이루어질 수 있음.

원-핫 인코딩(one-hot encoding)

- 수치화된 범주들 사이의 크기 비교를 피하기 위해 더미(dummy) 특성을 추가하여 활용
    - 범주 수 만큼의 더미 특성 추가
- 예를 들어, 해안 근접도 특성 대신에 다섯 개의 범주 전부를 새로운 특성으로 추가한 후 각각의 특성 값을 아래처럼 지정
    - 해당 카테고리의 특성값: 1
    - 나머지 카테고리의 특성값: 0
- 더미 특성 별로 1 또는 0의 값을 취하도록 모델 훈련 유도.

OneHotEncoder 변환기

- 원-핫 인코딩 지원
- `sparse` 키워드 인자
    - 기본값은 `True`.
    - 1의 위치만 기억하는 희소 행렬로 처리. 대용량 행렬 처리에 효과적임.
    - `False`로 지정할 경우 일반 행렬로 처리.

```
In [68]: cat_encoder = OneHotEncoder(sparse=False)
         housing_cat_1hot = cat_encoder.fit_transform(housing_cat)
         housing_cat_1hot
```

```
Out[68]: array([[1., 0., 0., 0., 0.],
                [1., 0., 0., 0., 0.],
                [0., 0., 0., 0., 1.],
                ...,
                [0., 1., 0., 0., 0.],
                [1., 0., 0., 0., 0.],
                [0., 0., 0., 1., 0.]])
```

2.5.3 나만의 변환기: 수치형 특성 전처리 과정 2 (조합 특성 추가)

- 아래 특성 추가 용도 변환기 클래스 직접 선언하기
    - 가구당 방 개수(rooms for household)
    - 방 하나당 침실 개수(bedrooms for room)
    - 가구당 인원(population per household)
- 변환기 클래스: `fit()`, `transform()` 메서드를 구현하면 됨.
    - 주의: fit() 메서드의 리턴값은 self

예제: CombinedAttributesAdder 변환기 클래스 선언

- `__init__()` 메서드: 생성되는 모델의 **하이퍼파라미터** 지정 용도
    - 모델에 대한 적절한 하이퍼파라미터를 튜닝할 때 유용하게 활용됨.
    - 예제: 방 하나당 침실 개수 속성 추가 여부
- `fit()` 메서드: 계산해야 하는 파라미터가 없음. 바로 `self` 리턴
- `transform()` 메서드: 넘파이 어레이를 입력받아 속성을 추가한 어레이를 반환

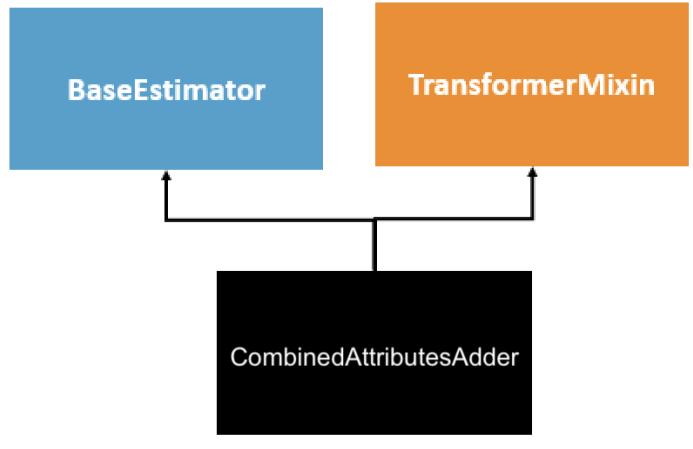
```
class CombinedAttributesAdder(BaseEstimator, TransformerMixin):
    def __init__(self, add_bedrooms_per_room = True):
        ...

    def fit(self, X, y=None):
        return self

    def transform(self, X):
        ...
```

상속하면 좋은 클래스

- `BaseEstimator` 상속: 하이퍼파라미터 튜닝 자동화에 필요한 `get_params()`, `set_params()` 메서드 제공
- `TransformerMixin` 상속: `fit_transform()` 자동 생성



<그림 아이디어 출처: Get the Most out of scikit-learn with Object-Oriented Programming>

2.5.4 특성 스케일링: 수치형 특성 전처리 과정 3

- 머신러닝 알고리즘은 입력 데이터셋의 특성값들의 스케일(범위)이 다르면 제대로 작동하지 않음
- 특성에 따라 다루는 숫자의 크기가 다를 때 통일된 스케일링이 필요
- 아래 두 가지 방식이 일반적으로 사용됨.
    - min-max 스케일링
    - 표준화 (책에서 사용)
- **주의**: 타깃(레이블)에 대한 스케일링은 하지 않음

min-max 스케일링

- **정규화**(normalization)라고도 불림
- 특성값 x를 $\frac{x-min}{max-min}$로 변환
- 변환 결과: **0에서 1** 사이
- 이상치에 매우 민감
    - 이상치가 매우 **크면 분모가 매우 커져서** 변환된 값이 **0 근처**에 몰림

표준화(standardization)

- 특성값 x 를 $\frac{x-\mu}{\sigma}$로 변환
    - μ: 특성값들의 **평균**값
    - σ: 특성값들의 **표준편차**
- 결과: 변환된 데이터들이 **표준정규분포**를 이룸
    - 이상치에 상대적으로 영향을 덜 받음.
- 사이킷런의 `StandardScaler` 변환기 활용 가능 (책에서 사용)

변환기 관련 주의사항

- `fit()` 메서드: 훈련 세트에 대해서만 적용. 테스트 세트는 활용하지 않음.
- `transform()` 메서드: 테스트 세트 포함 모든 데이터에 적용
    - 훈련 세트를 이용하여 필요한 파라미터를 확인한 후 그 값들을 이용하여 전체 데이터셋트를 변환

2.5.5 변환 파이프라인

- 모든 전처리 단계가 정확한 순서대로 진행되어야 함
- 사이킷런의 `Pipeline` 클래스를 이용하여 파이프라인 변환기 객체 생성 가능

수치형 특성 변환 파이프라인

```
num_pipeline = Pipeline([
        ('imputer', SimpleImputer(strategy="median")),
        ('attribs_adder', CombinedAttributesAdder()),
        ('std_scaler', StandardScaler()),
    ])
```

- 인자: 이름과 추정기로 이루어진 쌍들의 리스트
- 마지막 추정기 제외 나머지 추정기는 모두 변환기이어야 함.
    - `fit_transform()` 메서드 지원

- 파이프라인으로 정의된 추정기의 유형은 마지막 추정기의 유형과 동일
    - `num_pipeline`는 변환기. 이유는 `std_scaler`가 변환기이기 때문임.
- `num_pipeline.fit()` 호출:
    - 마지막 단계 이전 추정기: `fit_transform()` 메소드 연속 호출. 즉, 변환기가 실행될 때마다 변환도 동시에 진행.
    - 마지막 추정기: `fit()` 메서드 호출

수치형 / 범주형 특성 전처리 과정 통합 파이프라인

- 사이킷런의 `ColumnTransformer` 클래스를 이용하여 특성별로 지정된 전처리를 처리할 수 있도록 지정 가능
- 인자: (이름, 추정기, 적용 대상 열(column) 리스트) 튜플로 이루어진 리스트
- `fit()` 메서드에 pandas의 데이터프레임을 직접 인자로 사용 가능

- 수치형 특성: `num_pipeline` 변환기
    - 적용 대상 열(columns): `list(housing_num)`
- 범주형 특성: `OneHotEncoder` 변환기
    - 적용 대상 열(columns): `["ocean_proximity"]`

```
num_attribs = list(housing_num)
cat_attribs = ["ocean_proximity"]

full_pipeline = ColumnTransformer([
        ("num", num_pipeline, num_attribs),
        ("cat", OneHotEncoder(), cat_attribs),
    ])

housing_prepared = full_pipeline.fit_transform(housing)
```