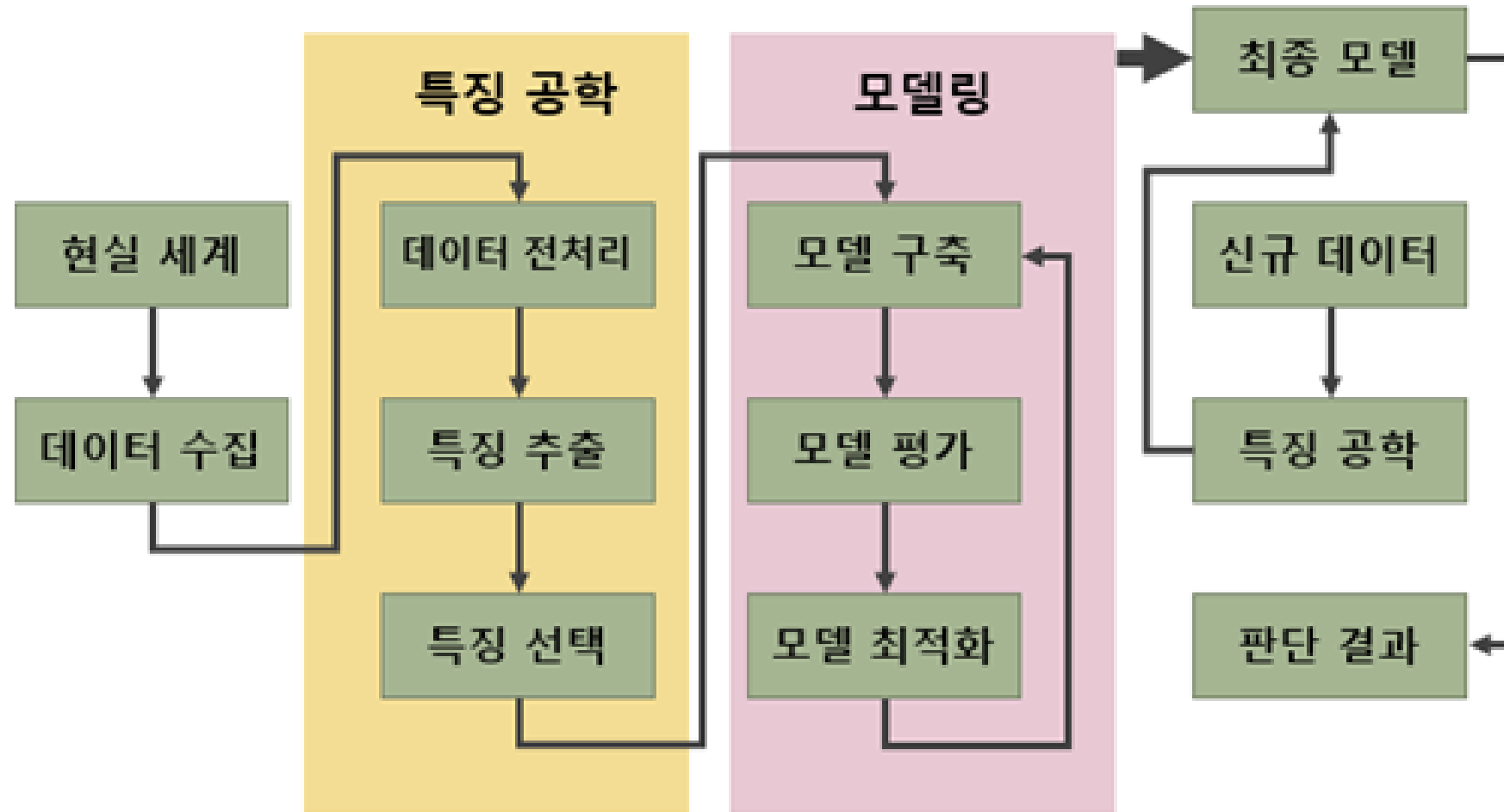



사이킷런으로 시작하는 머신러닝

류영표

머신러닝 프로세스



사이킷런 이란?

 [Install](#) [User Guide](#) [API](#) [Examples](#) [More ▾](#)

scikit-learn

Machine Learning in Python

[Getting Started](#) [Release Highlights for 0.24](#) [GitHub](#)

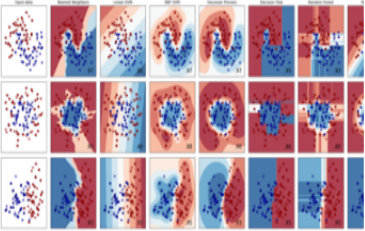
- Simple and efficient tools for predictive data analysis
- Accessible to everybody, and reusable in various contexts
- Built on NumPy, SciPy, and matplotlib
- Open source, commercially usable - BSD license

Classification

Identifying which category an object belongs to.

Applications: Spam detection, image recognition.

Algorithms: SVM, nearest neighbors, random forest, and more...



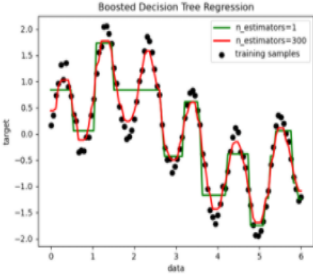
Examples

Regression

Predicting a continuous-valued attribute associated with an object.

Applications: Drug response, Stock prices.

Algorithms: SVR, nearest neighbors, random forest, and more...



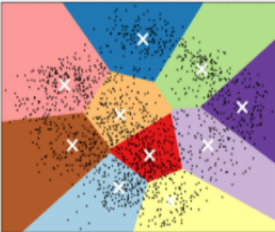
Examples

Clustering

Automatic grouping of similar objects into sets.

Applications: Customer segmentation, Grouping experiment outcomes

Algorithms: k-Means, spectral clustering, mean-shift, and more...



Examples

Dimensionality reduction

Reducing the number of random variables to consider.

Applications: Visualization, Increased efficiency

Algorithms: k-Means, feature selection, non-negative matrix factorization, and more...

Model selection

Comparing, validating and choosing parameters and models.

Applications: Improved accuracy via parameter tuning

Algorithms: grid search, cross validation, metrics,

Preprocessing

Feature extraction and normalization.

Applications: Transforming input data such as text for use with machine learning algorithms.

Algorithms: preprocessing, feature extraction, and more...

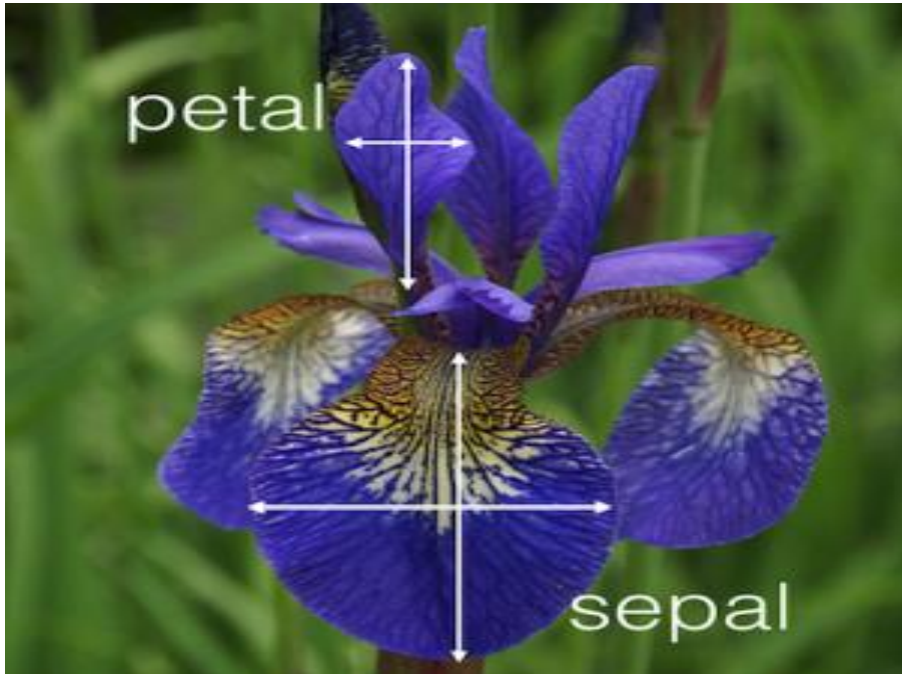
- 파이썬 기반의 다른 머신러닝 패키지도 사이킷런 스타일의 API를 지향할 정도로 쉽고 파이썬스러운 API를 제공
- 머신러닝을 위한 매우 다양한 알고리즘과 개발을 위한 편리한 프레임워크와 API를 제공
- 오랜 기간 실전 ghksruid에서 검증됐으며, 매우 많은 환경에서 성숙한 라이브러리입니다.
- 주로 Numpy와 Scipy 기반 위에서 구축된 라이브러리.

Scipy

- 과학, 분석 그리고 엔지니어링을 위한 과학적 영역의 여러 기본적인 작업을 위한 라이브러리
- 수치 적분과 미분방정식 해석기, 방정식의 근을 구하는 알고리즘, 표준 연속/이산 확률분포와 다양한 통계관련 도구 등을 제공

사이킷런을 이용한 붓꽃 데이터 분류

- 첫 번째의 머신러닝 모델 : 붓꽃의 품종을 예측
 - 붓꽃 데이터 세트로 붓꽃의 품종을 분류(Classification) 하는 것.
 - 붓꽃 데이터 세트는 꽃잎의 길이와 너비, 꽃받침의 길이와 너비 피쳐 (Feature)를 기반으로 꽃의 품종



iris setosa



petal sepal

iris versicolor



petal sepal

iris virginica



petal sepal

머신러닝을 위한 용어정리

- 피쳐(Feature)? 속성(attribute)?
 - 피쳐는 데이터 세트의 일반 속성
 - 머신러닝은 2차원 이상의 다차원 데이터에서도 많이 사용되므로 타겟값을 제외한 나머지 속성을 모두 피쳐로 지칭.
- 레이블, 클래스, 타겟(값), 결정(값)
 - 타겟값 또는 결정값은 지도 학습 시 데이터의 학습을 위해 주어지는 정답 데이터
 - 지도 학습 중 분류의 경우에는 이 결정값을 레이블 또는 클래스로 지칭

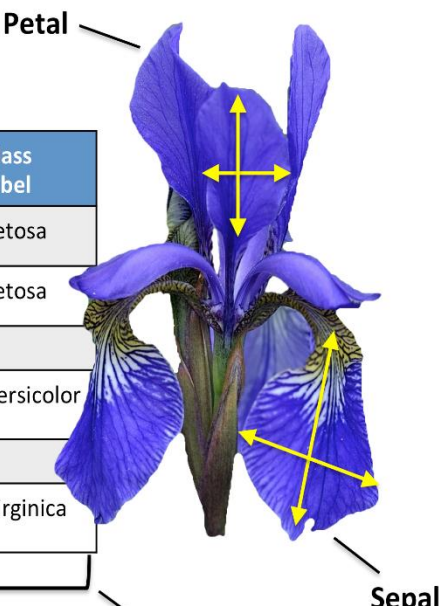
지도학습-분류(Classification)

Samples
(instances, observations)

	Sepal length	Sepal width	Petal length	Petal width	Class label
1	5.1	3.5	1.4	0.2	Setosa
2	4.9	3.0	1.4	0.2	Setosa
...					
50	6.4	3.5	4.5	1.2	Versicolor
...					
150	5.9	3.0	5.0	1.8	Virginica

Features
(attributes, measurements, dimensions)

Class labels
(targets)



Petal

Sepal

분류(Classification)는 대표적인 지도학습(Supervised learning) 방법의 하나. 지도학습은 학습을 위한 다양한 피쳐와 결정값인 레이블(Label) 데이터로 모델을 학습한 뒤, 별도의 데이터 세트에서 미지의 레이블을 예측

즉, 지도학습은 명확한 정답이 주어진 데이터를 먼저 학습한 뒤 미지의 정답을 예측하는 방식. 이때 학습을 위해 주어진 데이터 세트를 학습 데이터 세트, 머신러닝 모델의 예측 성능을 평가하기 위해 별도로 주어진 데이터 세트를 테스트 데이터 세트로 지칭

붓꽃 데이터 분류 예측 프로세스

데이터 세트 분리

데이터를 학습 데이터와 테스트 데이터로 분리

모델 학습

학습 데이터를 기반으로 ML 알고리즘을 적용해 모델을 학습

예측 수행

학습된 ML 모델을 이용해 테스트 데이터의 분류(즉, 붓꽃 종류)를 예측

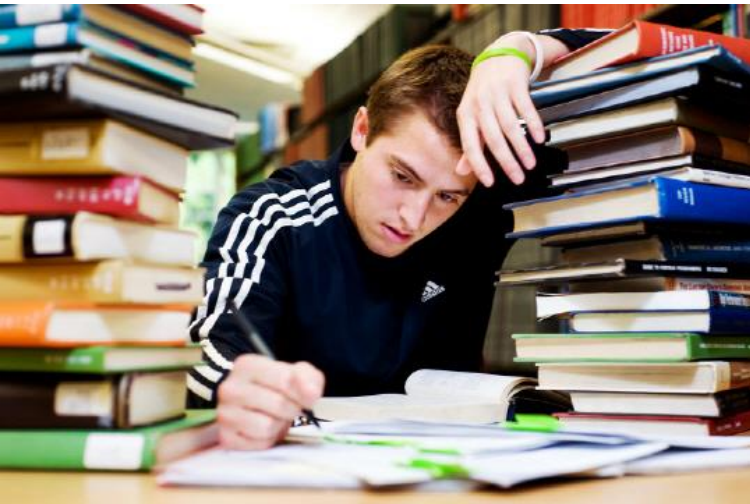
평가

이렇게 예측된 결과값과 테스트 데이터의 실제 결과값을 비교해 ML 모델 성능을 평가.

Model Selection_학습/테스트 데이터

학습 데이터 세트

- 머신러닝 알고리즘의 학습을 위해 사용
- 데이터의 속성들과 결정값(레이블) 값 모두를 가지고 있음
- 학습 데이터를 기반으로 머신러닝 알고리즘이 데이터 속성과 결정값의 패턴을 인지하고 학습

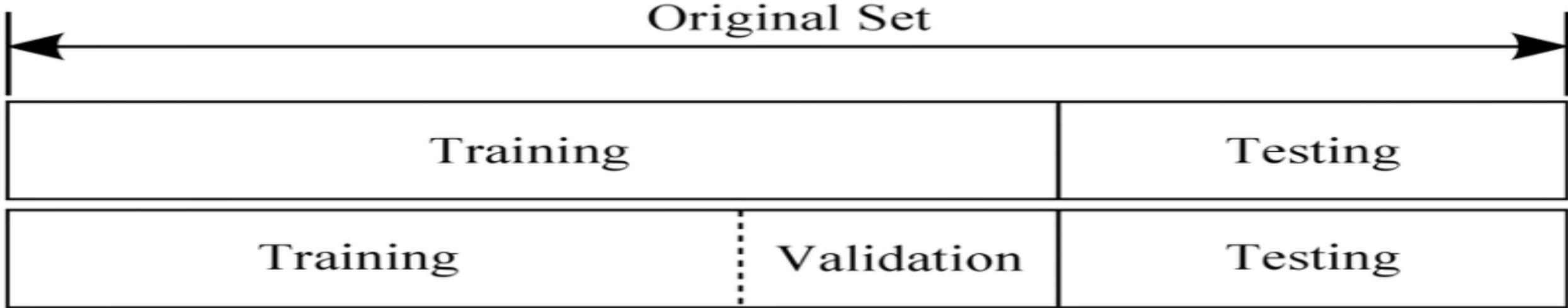


테스트 데이터 세트

- 테스트 데이터 세트에서 학습된 머신러닝 알고리즘을 테스트
- 테스트 데이터는 속성 데이터만 머신러닝 알고리즘에 제공하며, 머신러닝 알고리즘은 제공된 데이터를 기반으로 결정값을 예측
- 테스트 데이터는 학습 데이터와 별도의 데이터 세트로 제공되어야 함.

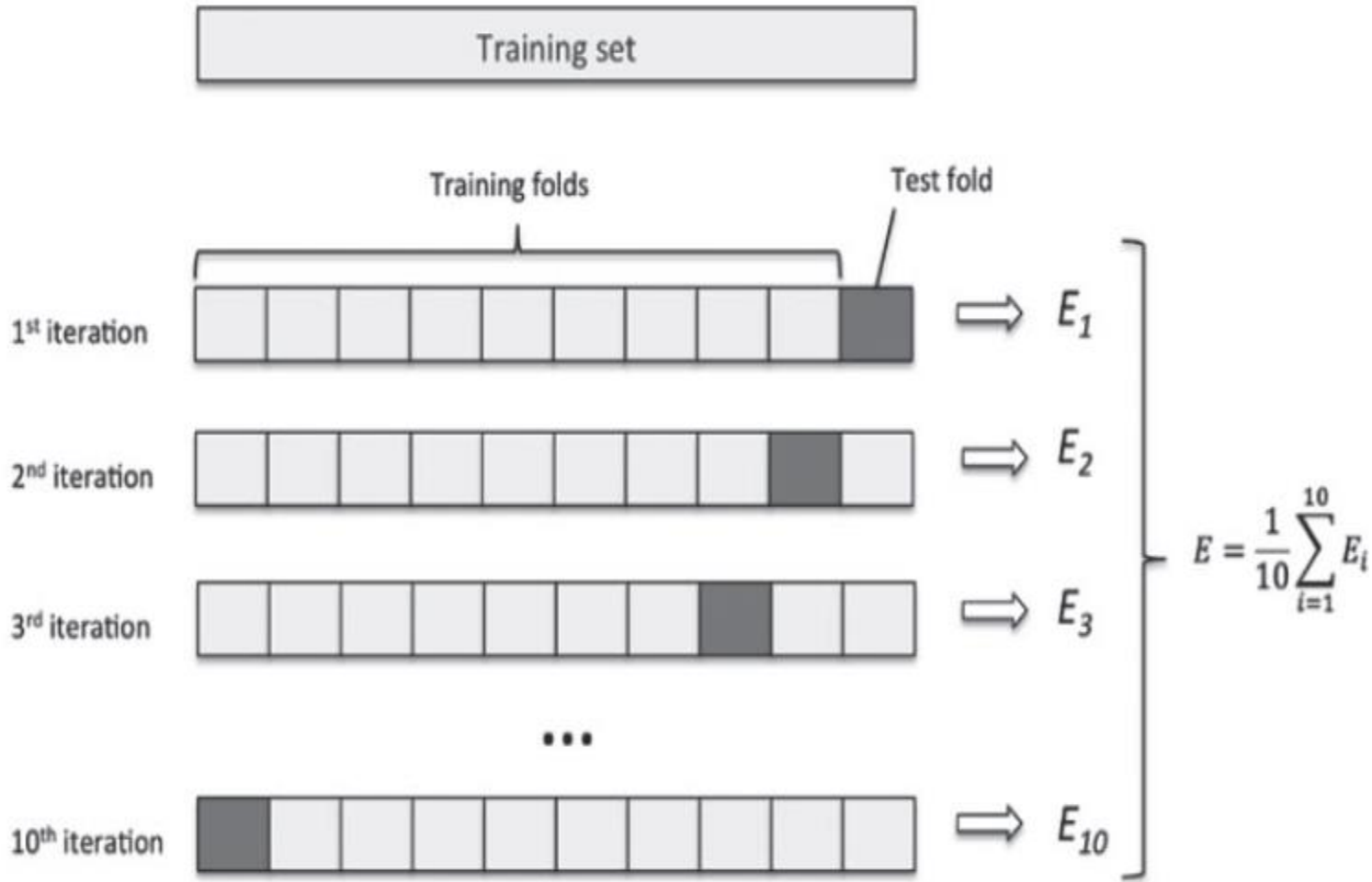


Model Selection_학습/테스트 데이터



- Training을 한 후에 만들어진 모형이 잘 예측을 하는지 그 성능을 평가
- Training Set의 일부를 모델의 성능을 평가하기 위해서 희생함.
- 이 희생을 감수하지 못 할만큼 data set의 크기가 작다면 cross-validation이라는 방법을 쓰기도 함.

K-폴드 교차검증



- 모델 평가를 위해서 데이터를 훈련 세트와 검증 세트로 나눌 때 데이터의 편향을 방지하기 위해 사용
- 데이터를 K개로 나누어 K-1개를 분할하고 나머지는 평가에 사용됩니다.
- 모델의 검증 점수는 K개의 검증 점수 평균이 됨.

K-폴드 교차 검증

K- 폴드 교차 검증

- 일반 K 폴드

- Stratified K 폴드

- 불균형한(imbalanced) 분포도를 가진 레이블(결정 클래스) 데이터 집합을 위한 K 폴드 방식
- 학습데이터와 검증 데이터 세트가 가지는 레이블 분포도가 유사하도록 검증 데이터 추출

Standard VS Stratified Cross-Validation

