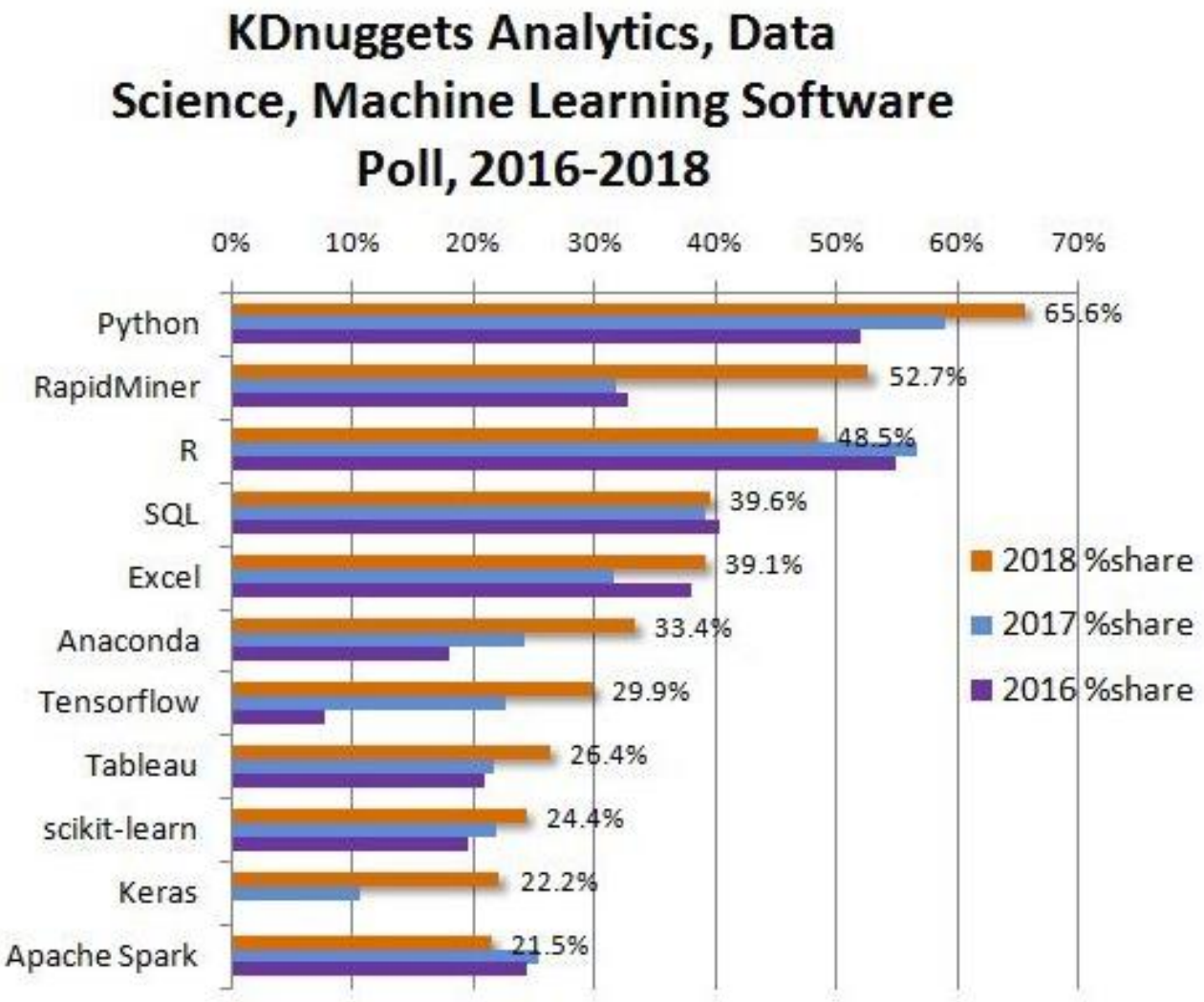
파이썬 기반 머신러닝 라이브러리

Scikit-learn 소개

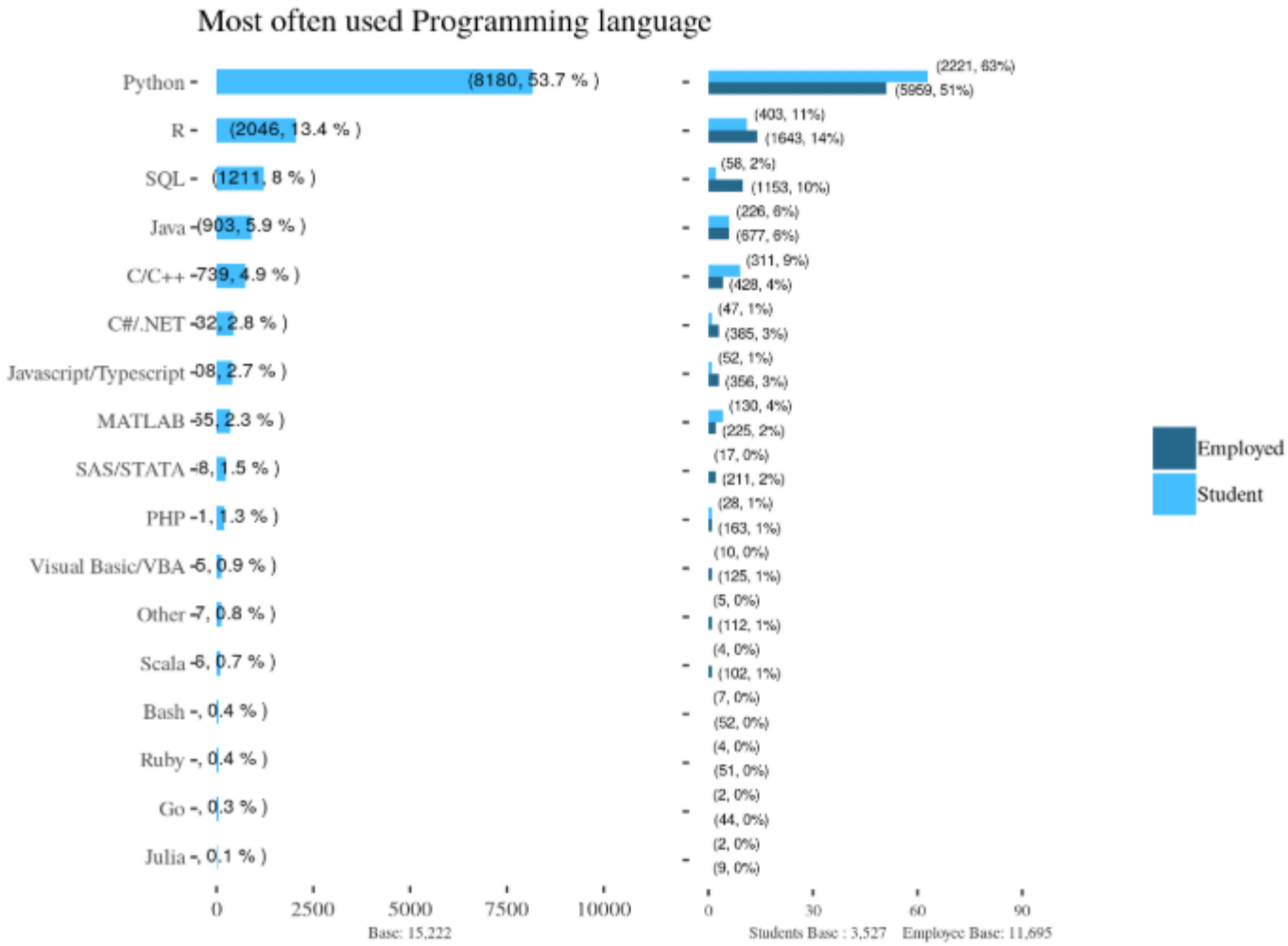
Python for Data science, Machine learning

동양미래대학교



State of Data Science & Machine Learning-2018 (By Kaggle)

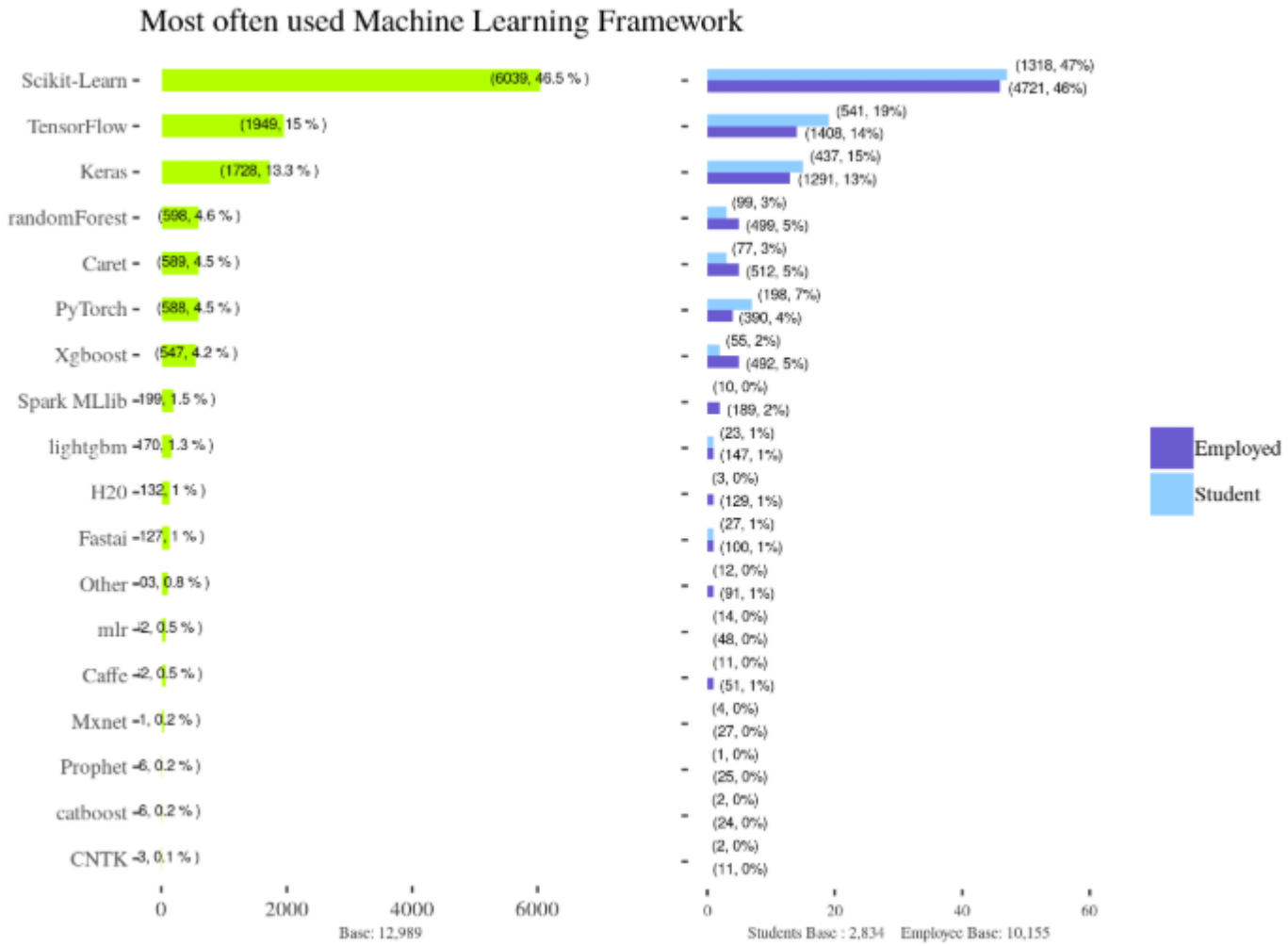
동양미래대학교



https://www.kaggle.com/rajeshcv/state-of-data-science-machine-learning-2018

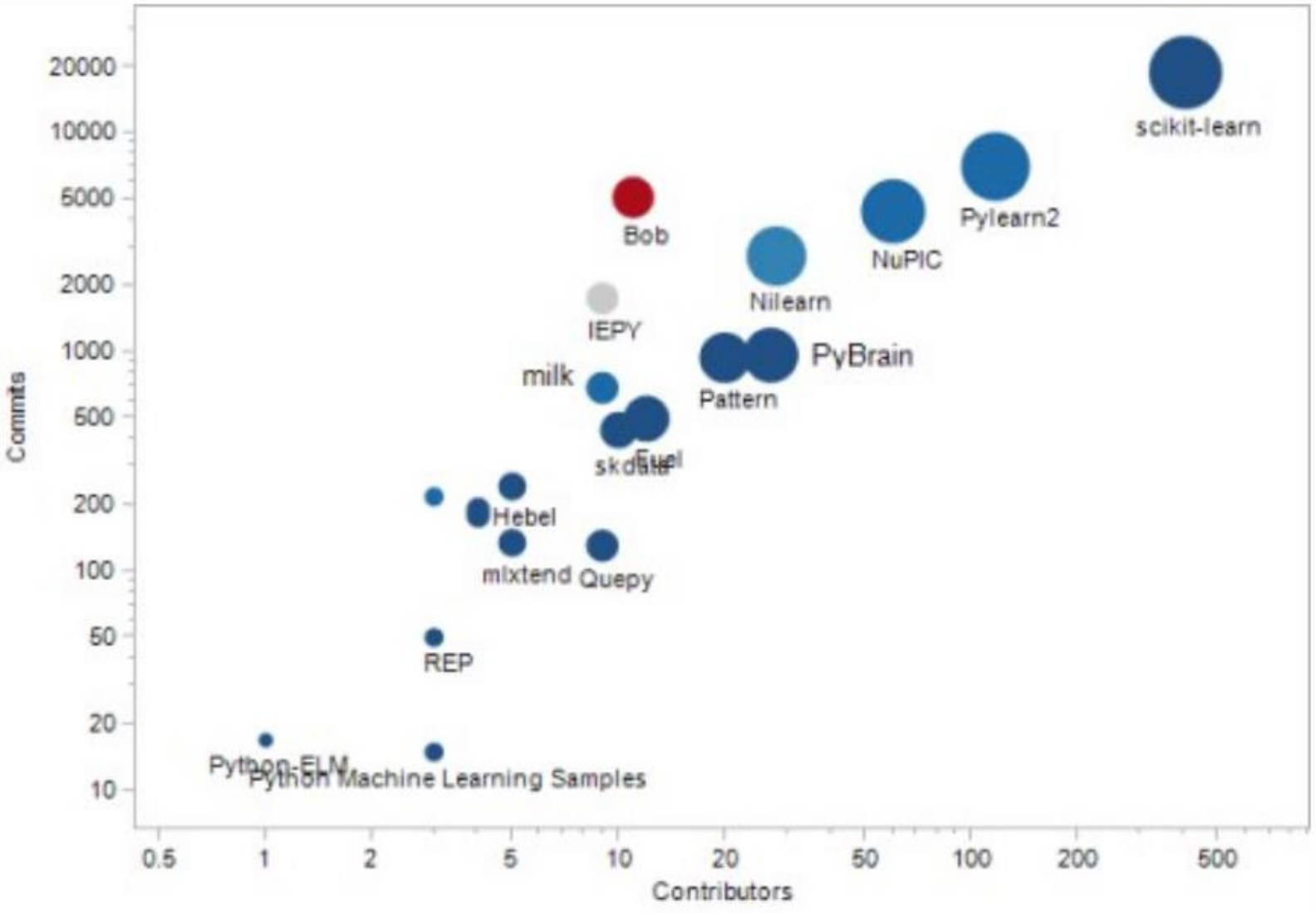
State of Data Science & Machine Learning-2018 (By Kaggle)

동양미래대학교



Python Library for Machine Learning

동양미래대학교



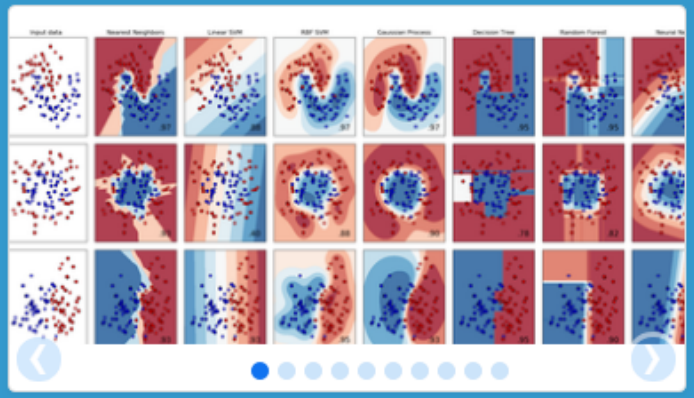
Deep Learning Frameworks

동양미래대학교

	Languages	Tutorials and training materials	CNN modeling capability	RNN modeling capability	Architecture: easy-to-use and modular front end	Speed	Multiple GPU support	Keras compatible
Theano	Python, C++	++	++	++	+	++	+	+
Tensor-Flow	Python	+++	+++	++	+++	++	++	+
Torch	Lua, Python (new)	+	+++	++	++	+++	++	
Caffe	C++	+	++		+	+	+	
MXNet	R, Python, Julia, Scala	++	++	+	++	++	+++	
Neon	Python	+	++	+	+	++	+	
CNTK	C++	+	+	+++	+	++	+	

Scikit-learn

동양미래대학교



scikit-learn

Machine Learning in Python

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

Classification

Identifying to which category an object belongs to.

Applications: Spam detection, Image recognition.
Algorithms: SVM, nearest neighbors, random forest, ... — Examples

Regression

Predicting a continuous-valued attribute associated with an object.

Applications: Drug response, Stock prices.
Algorithms: SVR, ridge regression, Lasso, ... — Examples

Clustering

Automatic grouping of similar objects into sets.

Applications: Customer segmentation, Grouping experiment outcomes
Algorithms: k-Means, spectral clustering, mean-shift, ... — Examples

Dimensionality reduction

Reducing the number of random variables to consider.

Applications: Visualization, Increased efficiency
Algorithms: PCA, feature selection, non-negative matrix factorization. — Examples

Model selection

Comparing, validating and choosing parameters and models.

Goal: Improved accuracy via parameter tuning
Modules: grid search, cross validation, metrics. — Examples

Preprocessing

Feature extraction and normalization.

Application: Transforming input data such as text for use with machine learning algorithms.
Modules: preprocessing, feature extraction.

— Examples

http://scikit-learn.org/stable/

분류	모듈명	설명
예제 데이터	sklearn.datasets	Scikit-learn에 내장된 예제 데이터셋
피쳐 처리	sklearn.preprocessing	데이터 전처리에 필요한 다양한 기능 제공(인코딩, 정규화, 스케일링 등)
	sklearn.feature_selection	알고리즘에 영향을 미치는 특성을 우선순위대로 선택 작업을 수행하는 다양한 기능 제공
	sklearn.feature_extraction	텍스트 데이터나 이미지 데이터의 벡터화된 피쳐를 추출하는데 사용됨.
	sklearn.decomposition	차원 축소와 관련된 알고리즘 지원 (PCA, NMF 등)
데이터 분리, 검증 & 파라미터 튜닝	sklearn.model_selection	교차 검증을 위한 훈련용/테스트용 데이터 분리, GridSearch 기능 등 제공
평가	sklearn.metrics	각종 머신 러닝 알고리즘(회귀, 분류, 클러스터링 등)의 다양한 성능 측정 방법 제공 (accuracy, precision, recall, ROC-AUC 곡선 등)
머신 러닝 알고리즘	sklearn.ensemble	앙상블 알고리즘 제공(Random Forest, AdaBoost, Gradient boosting, 등)
	sklearn.linear_model	선형회귀, 릿지, 라쏘 및 로지스틱 회귀 등 회기ㅜ 관련 알고리즘 지원
	sklearn.naïve_bayes	나이브 베이즈 알고리즘 제공. 가우시안 NB, 다항 분포 NB 등
	sklearn.neighbors	KNN 알고리즘 제공
	sklearn.svm	SVM 알고리즘 제공
	sklearn.tree	의사 결정 트리 알고리즘 제공
	sklearn.cluster	클러스터링 알고리즘 제공 (k-means, DBScan 등)
유틸리티 (지원 기능)	sklearn.pipeline	피쳐 처리 등의 변환과 ML 알고리즘 학습, 예측 등을 함께 묶어서 실행할 수 있는 유틸리티 제공

datasets 모듈

동양미래대학교

별도의 외부 웹사이트에서 데이터 세트를 내려받을 필요 없이 예제로 활용가능한 데이터셋 혹은 데이터 생성 기능 제공

datasets.clear_data_home([data_home])	Delete all the content of the data home cache.
datasets.dump_svmlight_file(X, y, f[, ...])	Dump the dataset in svmlight / libsvm file format.
datasets.fetch_20newsgroups([data_home, ...])	Load the filenames and data from the 20 newsgroups dataset (classification).
datasets.fetch_20newsgroups_vectorized([...])	Load the 20 newsgroups dataset and vectorize it into token counts (classification).
datasets.fetch_california_housing([...])	Load the California housing dataset (regression).
datasets.fetch_covtype([data_home, ...])	Load the covertype dataset (classification).
datasets.fetch_kddcup99([subset, data_home, ...])	Load the kddcup99 dataset (classification).
datasets.fetch_lfw_pairs([subset, ...])	Load the Labeled Faces in the Wild (LFW) pairs dataset (classification).
datasets.fetch_lfw_people([data_home, ...])	Load the Labeled Faces in the Wild (LFW) people dataset (classification).
datasets.fetch_olivetti_faces([data_home, ...])	Load the Olivetti faces data-set from AT&T (classification).
datasets.fetch_openml([name, version, ...])	Fetch dataset from openml by name or dataset id.
datasets.fetch_rcv1([data_home, subset, ...])	Load the RCV1 multilabel dataset (classification).
datasets.fetch_species_distributions([...])	Loader for species distribution dataset from Phillips et
datasets.get_data_home([data_home])	Return the path of the scikit-learn data dir.
datasets.load_boston([return_X_y])	Load and return the boston house-prices dataset (regression).
datasets.load_breast_cancer([return_X_y])	Load and return the breast cancer wisconsin dataset (classification).
datasets.load_diabetes([return_X_y])	Load and return the diabetes dataset (regression).
datasets.load_digits([n_class, return_X_y])	Load and return the digits dataset (classification).
datasets.load_files(container_path[, ...])	Load text files with categories as subfolder names.
datasets.load_iris([return_X_y])	Load and return the iris dataset (classification).
datasets.load_linnerud([return_X_y])	Load and return the linnerud dataset (multivariate regression).
datasets.load_sample_image(image_name)	Load the numpy array of a single sample image
datasets.load_sample_images()	Load sample images for image manipulation.
datasets.load_svmlight_file(f[, n_features, ...])	Load datasets in the svmlight / libsvm format into sparse CSR matrix
datasets.load_svmlight_files(files[, ...])	Load dataset from multiple files in SVMlight format
datasets.load_wine([return_X_y])	Load and return the wine dataset (classification).

데이터의 크기가 커서 패키지에 저장되어 있는 것이 아니라 인터넷에서 다운로드 받는 함수 (fetch_XXX)

분류/회귀용 샘플 데이터 (load_XXX)

https://scikit-learn.org/stable/modules/classes.html#module-sklearn.datasets

Model_selection 모듈 (1) – 학습/테스트 데이터셋 분리

학습/데이터셋 분리 – 주어진 데이터셋을 학습용 데이터와 테스트용 데이터로 분리

train_test_split(*arrays, test_size, random_state, shuffle*)

- ***arrays***: 데이터
- ***test_size*** : 전체 데이터 중 테스트 데이터셋의 비중 (0과 1사이의 값, 기본값 = 0.25)
- ***random_state***: 호출할 때마다 동일한 학습/테스트용 데이터세트를 생성하기 위해 주어지는 난수 값. 특정 값으로 지정하면 항상 같은 데이터로 나뉨
- ***shuffle***: 데이터를 분리하기 전에 데이터를 미리 섞을 것인지를 결정함. 기본값은 True. 효율적인 데이터 분리를 위해서는 대부분 기본값을 사용함.

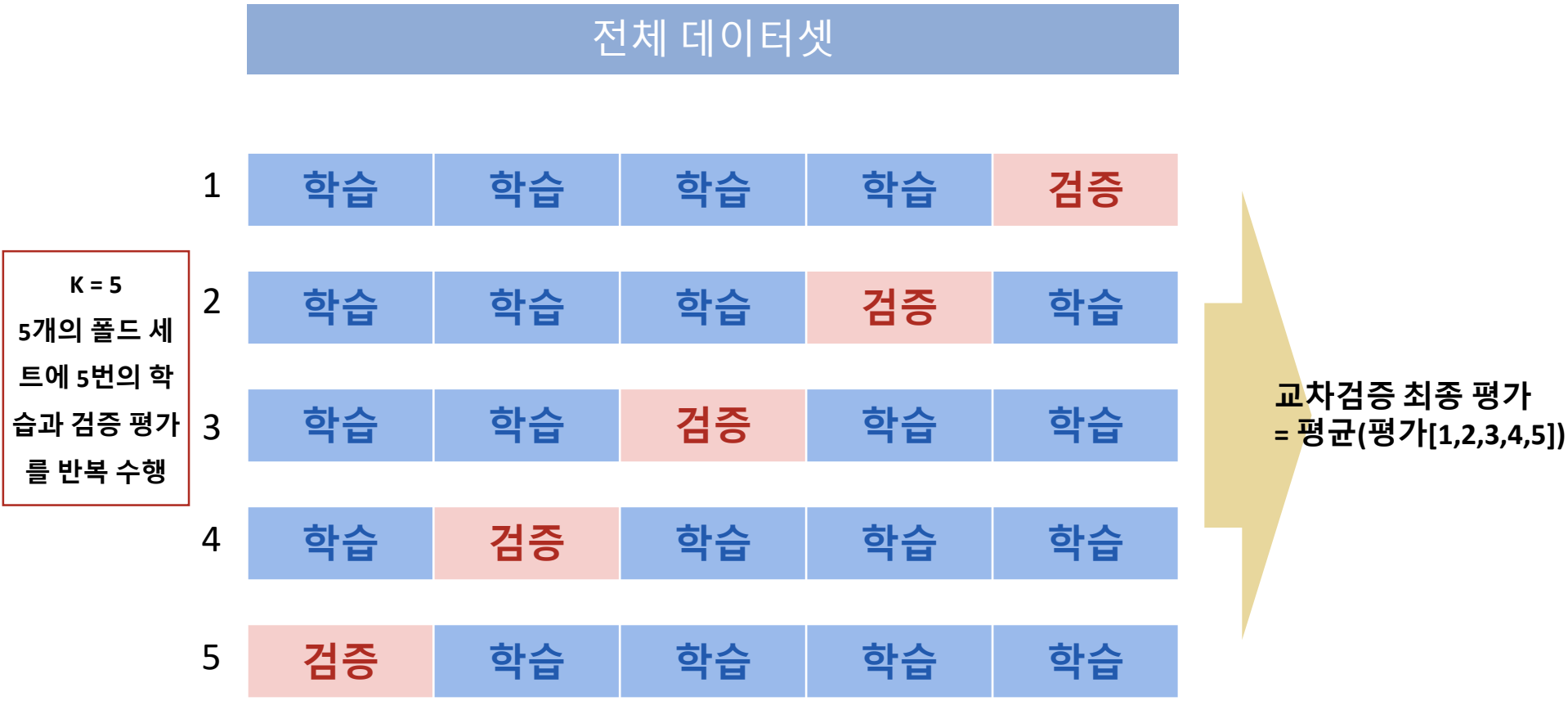
Model_selection 모듈 (2) - 교차검증

동양미래대학교

교차 검증 – 주어진 데이터에 과적합(overfitting)되는 문제점을 개선하기 위하여 사용

→ 학습 데이터를 학습 데이터와 검증 데이터로 구성하여, 학습과 평가를 반복 수행

K 폴드 교차 검증(K-fold Cross Validation)



Stratified K 폴드 교차 검증(stratified K-fold Cross Validation)

K겹 교차검증의 단점

→ 부분데이터집합에 데이터의 클래스가 skew되는 경우에는 정확도가 매우 떨어질 수 있다. (e.g. 대출 사기 데이터)

Stratified k겹 교차검증

- 데이터를 k개로 나눌 때, 각 클래스별로 고르게 분포하도록 나눔.

Scikit-learn에서의 교차 검증

- 분류 모델: Stratified K- Fold cross validation 사용
- 회귀 모델: K-fold cross validation 사용

Model_selection 모듈 (3) – 하이퍼 파라미터 튜닝

동양미래대학교

하이퍼파라미터 튜닝

알고리즘의 성능을 결정하는 파라미터들을 최적화하는 과정

교차 검증과 하이퍼 파라미터 튜닝을 한 번에 할 수 있는 함수를 제공

GridSearchCV(*estimator, param_grid, scoring, n_jobs, cv, refit*)

- ***estimator*** : 머신러닝 학습 객체 (classification, regressor, …)
- ***param_grid***: 사전 타입으로 튜닝에 사용할 파라미터들을 정의 {파라미터명:[값1, 값2, 값3, …]
- ***scoring***: 예측 성능을 측정할 평가 방법 (기본값: accuracy)
- ***n_jobs***: 동시 수행할 job의 개수. 기본값: None(1과 같음). 모든 CPU를 사용하고자 한다면 -1로 설정
- ***cv***: 교차검증 횟수 (현재 기본값은 3이나, 0.22버전부터는 5로 변경)
- **refit**: 기본값이 True. 최적의 파라미터 조합을 찾은 후 estimator 객체를 해당 하이퍼 파라미터로 재학습시킴

https://scikit-learn.org/stable/modules/generated/sklearn.model_selection.GridSearchCV.html#sklearn.model_selection.GridSearchCV

Preprocessing 모듈

동양미래대학교

데이터 품질을 향상 시키기 위한 다양한 기능 제공

- **데이터 인코딩** : 문자열 피처를 숫자형 피처로 변환 (대부분의 ML 알고리즘들은 문자열 피처 제공 X)
- **피처 스케일링/정규화**: 피처들의 값 범위를 일정한 수준으로 맞추는 기능

분류	모듈명	설명
데이터 인코딩	LabelEncoder	레이블 인코딩 (카테고리 피처를 코드형 숫자값으로 변환) (e.g. 성별 피처의 값인 Male, Female을 각각 0과 1로 변환)
	OneHotEncoder	특성(피처)의 모든 값들을 새로운 특성으로 추가해 해당 값에 해당하는 칼럼에만 1을 표시, 나머지는 0으로 표시하는 방식
피처 스케일링/정규화	StandardScaler	피처의 값들이 가우시안 분포(평균 = 0, 분산 = 1)를 따르도록 변환
	MinMaxScaler	피처의 값들을 0과 1사이의 값으로 변환

https://scikit-learn.org/stable/modules/classes.html#module-sklearn.preprocessing

Data

Car
Sonata
Grandeur
Genesis
Grandeur

인코딩

1) 레이블 인코딩

Car_code
0
1
2
1

2) 원-핫 인코딩

Sonata_yn	Grandeur_yn	Genesis_yn
1	0	0
0	1	0
0	0	1
0	1	0

동양미래대학교

IRIS 데이터 분석 실습

Samples
(instances, observations)

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

Petal

Class labels
(targets)

Features
(attributes, measurements, dimensions)

IRIS 데이터 분석

동양미래대학교