Related Projects

Projects implementing the scikit-learn estimator API are encouraged to use the scikit-learn-contrib template which facilitates best practices for testing and documenting estimators. The scikit-learn-contrib GitHub organisation also accepts high-quality contributions of repositories conforming to this template.

Below is a list of sister-projects, extensions and domain specific packages.

Interoperability and framework enhancements

These tools adapt scikit-learn for use with other technologies or otherwise enhance the functionality of scikit-learn's estimators.

Data formats

- Fast svmlight / libsvm file loader Fast and memory-efficient svmlight / libsvm file loader for Python.
- sklearn_pandas bridge for scikit-learn pipelines and pandas data frame with dedicated transformers.
- sklearn_xarray provides compatibility of scikit-learn estimators with xarray data structures.

Auto-ML

- auto-sklearn An automated machine learning toolkit and a drop-in replacement for a scikit-learn estimator
- autoviml Automatically Build Multiple Machine Learning Models with a Single Line of Code. Designed as a faster way to use scikit-learn models without having to preprocess data.
- TPOT An automated machine learning toolkit that optimizes a series of scikit-learn operators to design a machine learning pipeline, including data and feature preprocessors as well as the estimators. Works as a drop-in replacement for a scikit-learn estimator.
- Featuretools A framework to perform automated feature engineering. It can be used for transforming temporal and relational datasets into feature matrices for machine learning.
- Neuraxle A library for building neat pipelines, providing the right abstractions to both ease research, development, and deployment of machine learning applications. Compatible with deep learning frameworks and scikit-learn API, it can stream minibatches, use data checkpoints, build funky pipelines, and serialize models with custom per-step savers.
- EvalML EvalML is an AutoML library which builds, optimizes, and evaluates machine learning pipelines using domain-specific objective functions. It incorporates multiple modeling libraries under one API, and the objects that EvalML creates use an sklearn-compatible API.

Experimentation frameworks

- Neptune Metadata store for MLOps, built for teams that run a lot of experiments.â€Œ It gives you a single place to log, store, display, organize, compare, and query all your model building metadata.
- Sacred Tool to help you configure, organize, log and reproduce experiments
- REP Environment for conducting data-driven research in a consistent and reproducible way
- Scikit-Learn Laboratory A command-line wrapper around scikit-learn that makes it easy to run machine learning experiments with multiple learners and large feature sets.

Model inspection and visualisation

- dtreeviz A python library for decision tree visualization and model interpretation.
- eli5 A library for debugging/inspecting machine learning models and explaining their predictions.
- mlxtend Includes model visualization utilities.
- yellowbrick A suite of custom matplotlib visualizers for scikit-learn estimators to support visual feature analysis, model selection, evaluation, and diagnostics.

Model selection

- scikit-optimize A library to minimize (very) expensive and noisy black-box functions. It implements several methods for sequential model-based optimization, and includes a replacement for `GridSearchCV` or `RandomizedSearchCV` to do cross-validated parameter search using any of these strategies.
- sklearn-deap Use evolutionary algorithms instead of gridsearch in scikit-learn.

Model export for production

- sklearn-onnx Serialization of many Scikit-learn pipelines to ONNX for interchange and prediction.
- sklearn2pmml Serialization of a wide variety of scikit-learn estimators and transformers into PMML with the help of JPMML-SkLearn library.
- sklearn-porter Transpile trained scikit-learn models to C, Java, Javascript and others.
- m2cgen A lightweight library which allows to transpile trained machine learning models including many scikit-learn estimators into a native code of C, Java, Go, R, PHP, Dart, Haskell, Rust and many other programming languages.
- treelite Compiles tree-based ensemble models into C code for minimizing prediction latency.

Other estimators and tasks

Not everything belongs or is mature enough for the central scikit-learn project. The following are projects providing interfaces similar to scikit-learn for additional learning algorithms, infrastructures and tasks.

Structured learning

- tslearn A machine learning library for time series that offers tools for pre-processing and feature extraction as well as dedicated models for clustering, classification and regression.
- sktime A scikit-learn compatible toolbox for machine learning with time series including time series classification/regression and (supervised/panel) forecasting.
- HMMLearn Implementation of hidden markov models that was previously part of scikit-learn.
- PyStruct General conditional random fields and structured prediction.
- pomegranate Probabilistic modelling for Python, with an emphasis on hidden Markov models.
- sklearn-crfsuite Linear-chain conditional random fields (CRFsuite wrapper with sklearn-like API).

Deep neural networks etc.

- nolearn A number of wrappers and abstractions around existing neural network libraries
- Keras High-level API for TensorFlow with a scikit-learn inspired API.
- lasagne A lightweight library to build and train neural networks in Theano.
- skorch A scikit-learn compatible neural network library that wraps PyTorch.
- scikeras provides a wrapper around Keras to interface it with scikit-learn. SciKeras is the successor of *tf.keras.wrappers.scikit_learn*.

Federated Learning

- Flower A friendly federated learning framework with a unified approach that can federate any workload, any ML framework, and any programming language.

Broad scope

- mlxtend Includes a number of additional estimators as well as model visualization utilities.
- scikit-lego A number of scikit-learn compatible custom transformers, models and metrics, focusing on solving practical industry tasks.

Other regression and classification

- xgboost Optimised gradient boosted decision tree library.
- ML-Ensemble Generalized ensemble learning (stacking, blending, subsemble, deep ensembles, etc.).
- lightning Fast state-of-the-art linear model solvers (SDCA, AdaGrad, SVRG, SAG, etc...).
- py-earth Multivariate adaptive regression splines
- Kernel Regression Implementation of Nadaraya-Watson kernel regression with automatic bandwidth selection
- gplearn Genetic Programming for symbolic regression tasks.
- scikit-multilearn Multi-label classification with focus on label space manipulation.
- seglearn Time series and sequence learning using sliding window segmentation.
- libOPF Optimal path forest classifier
- fastFM Fast factorization machine implementation compatible with scikit-learn

Decomposition and clustering

- lda: Fast implementation of latent Dirichlet allocation in Cython which uses Gibbs sampling to sample from the true posterior distribution. (scikit-learn's :class:`~sklearn.decomposition.LatentDirichletAllocation` implementation uses variational inference to sample from a tractable approximation of a topic model's posterior distribution.)

> **System Message: ERROR/3** (`D:\onboarding-resources\sample-onboarding-resources\scikit-learn-main\doc\[scikit-learn-main][doc]related_projects.rst`, **line 231**); ***backlink***
>
> Unknown interpreted text role "class".

- kmodes k-modes clustering algorithm for categorical data, and several of its variations.
- hdbscan HDBSCAN and Robust Single Linkage clustering algorithms for robust variable density clustering.
- spherecluster Spherical K-means and mixture of von Mises Fisher clustering routines for data on the unit hypersphere.

Pre-processing

- categorical-encoding A library of sklearn compatible categorical variable encoders.
- imbalanced-learn Various methods to under- and over-sample datasets.
- Feature-engine A library of sklearn compatible transformers for missing data imputation, categorical encoding, variable transformation, discretization, outlier handling and more. Feature-engine allows the application of preprocessing steps to selected groups of variables and it is fully compatible with the Scikit-learn Pipeline.

Topological Data Analysis

- giotto-tda A library for Topological Data Analysis aiming to provide a scikit-learn compatible API. It offers tools to transform

data inputs (point clouds, graphs, time series, images) into forms suitable for computations of topological summaries, and components dedicated to extracting sets of scalar features of topological origin, which can be used alongside other feature extraction methods in scikit-learn.

Statistical learning with Python

Other packages useful for data analysis and machine learning.

- Pandas Tools for working with heterogeneous and columnar data, relational queries, time series and basic statistics.
- statsmodels Estimating and analysing statistical models. More focused on statistical tests and less on prediction than scikit-learn.
- PyMC Bayesian statistical models and fitting algorithms.
- Seaborn Visualization library based on matplotlib. It provides a high-level interface for drawing attractive statistical graphics.
- scikit-survival A library implementing models to learn from censored time-to-event data (also called survival analysis). Models are fully compatible with scikit-learn.

Recommendation Engine packages

- implicit, Library for implicit feedback datasets.
- lightfm A Python/Cython implementation of a hybrid recommender system.
- OpenRec TensorFlow-based neural-network inspired recommendation algorithms.
- Spotlight Pytorch-based implementation of deep recommender models.
- Surprise Lib Library for explicit feedback datasets.

Domain specific packages

- scikit-image Image processing and computer vision in python.
- Natural language toolkit (nltk) Natural language processing and some machine learning.
- gensim A library for topic modelling, document indexing and similarity retrieval
- NiLearn Machine learning for neuro-imaging.
- AstroML Machine learning for astronomy.
- MSMBuilder Machine learning for protein conformational dynamics time series.

Translations of scikit-learn documentation

Translationâ€™s purpose is to ease reading and understanding in languages other than English. Its aim is to help people who do not understand English or have doubts about its interpretation. Additionally, some people prefer to read documentation in their native language, but please bear in mind that the only official documentation is the English one [1].

Those translation efforts are community initiatives and we have no control on them. If you want to contribute or report an issue with the translation, please contact the authors of the translation. Some available translations are linked here to improve their dissemination and promote community efforts.

- Chinese translation (source)
- Persian translation (source)
- Spanish translation (source)

Footnotes

[1] following linux documentation Disclaimer