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

- sklearn_pandas bridge for scikit-learn pipelines and pandas data frame with dedicated transformers.

Auto-ML

- auto_ml Automated machine learning for production and analytics, built on scikit-learn and related projects. Trains a pipeline wth all the standard machine learning steps. Tuned for prediction speed and ease of transfer to production environments.
- auto-sklearn An automated machine learning toolkit and a drop-in replacement for a scikit-learn estimator
- 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.

Experimentation frameworks

- REP Environment for conducting data-driven research in a consistent and reproducible way
- ML Frontend provides dataset management and SVM fitting/prediction through web-based and programmatic interfaces.
- 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.
- Xcessiv is a notebook-like application for quick, scalable, and automated hyperparameter tuning and stacked ensembling. Provides a framework for keeping track of model-hyperparameter combinations.

Model inspection and visualisation

«

- eli5 A library for debugging/inspecting machine learning models and explaining their predictions.
- mlxtend Includes model visualization utilities.
- scikit-plot A visualization library for quick and easy generation of common plots in data analysis and machine learning.
- yellowbrick A suite of custom matplotlib visualizers for scikit-learn estimators to support visual feature analysis, model selection, evaluation, and diagnostics.

Model export for production

- sklearn-pmml Serialization of (some) scikit-learn estimators into PMML.
- 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.
- sklearn-compiledtrees Generate a C++ implementation of the predict function for decision trees (and ensembles) trained by sklearn. Useful for latency-sensitive production environments.

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

- Seqlearn Sequence classification using HMMs or structured perceptron.

- 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.

- pylearn2 A deep learning and neural network library build on theano with scikit-learn like interface.
- sklearn_theano scikit-learn compatible estimators, transformers, and datasets which use Theano internally
- nolearn A number of wrappers and abstractions around existing neural network libraries
- keras Deep Learning library capable of running on top of either TensorFlow or Theano.
- lasagne A lightweight library to build and train neural networks in Theano.

Broad scope

- mlxtend Includes a number of additional estimators as well as model visualization utilities.
- sparkit-learn Scikit-learn API and functionality for PySpark's distributed modelling.

Other regression and classification

- xgboost Optimised gradient boosted decision tree library.
- 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.
- multiisotonic Isotonic regression on multidimensional features.

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 **sklearn.decomposition.LatentDirichletAllocation** implementation uses variational inference to sample from a tractable approximation of a topic model's posterior distribution.)
- Sparse Filtering Unsupervised feature learning based on sparse-filtering
- 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.

«

Previous

- 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.

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.
- theano A CPU/GPU array processing framework geared towards deep learning research.
- 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.
- Sacred Tool to help you configure, organize, log and reproduce experiments
- Seaborn Visualization library based on matplotlib. It provides a high-level interface for drawing attractive statistical graphics.
- Deep Learning A curated list of deep learning software libraries.

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.

Snippets and tidbits

The wiki has more!

«