Utilities for Developers

Scikit-learn contains a number of utilities to help with development. These are located in `sklearn.utils`, and include tools in a number of categories. All the following functions and classes are in the module `sklearn.utils`.

Warning: These utilities are meant to be used internally within the scikit-learn package. They are not guaranteed to be stable between versions of scikit-learn. Backports, in particular, will be removed as the scikit-learn dependencies evolve.

Validation Tools

These are tools used to check and validate input. When you write a function which accepts arrays, matrices, or sparse matrices as arguments, the following should be used when applicable.

- `assert_all_finite`: Throw an error if array contains NaNs or Infs.
- `as_float_array`: convert input to an array of floats. If a sparse matrix is passed, a sparse matrix will be returned.
- `check_array`: check that input is a 2D array, raise error on sparse matrices. Allowed sparse matrix formats can be given optionally, as well as allowing 1D or N-dimensional arrays. Calls `assert_all_finite` by default.
- `check_X_y`: check that X and y have consistent length, calls check_array on X, and column_or_1d on y. For multilabel classification or multitarget regression, specify multi_output=True, in which case check_array will be called on y.
- `indexable`: check that all input arrays have consistent length and can be sliced or indexed using safe_index. This is used to validate input for cross-validation.
- `validation.check_memory` checks that input is `joblib.Memory`-like, which means that it can be converted into a `sklearn.utils.Memory` instance (typically a str denoting the `cachedir`) or has the same interface.

If your code relies on a random number generator, it should never use functions like `numpy.random.random` or `numpy.random.normal`. This approach can lead to repeatability issues in unit tests. Instead, a `numpy.random.RandomState` object should be used, which is built from a `random_state` argument passed to the class or function. The function `check_random_state`, below, can then be used to create a random number generator object.

- `check_random_state`: create a `np.random.RandomState` object from a parameter `random_state`.
  - If `random_state` is `None` or `np.random`, then a randomly-initialized `RandomState` object is returned.
  - If `random_state` is an integer, then it is used to seed a new `RandomState` object.
  - If `random_state` is a `RandomState` object, then it is passed through.

For example:

```
>>> from sklearn.utils import check_random_state
>>> random_state = 0
>>> random_state = check_random_state(random_state)
>>> random_state.rand(4)
array([0.5488135 , 0.71518937, 0.60276338, 0.54488318])
```

When developing your own scikit-learn compatible estimator, the following helpers are available.

- `validation.check_is_fitted`: check that the estimator has been fitted before calling `transform`, `predict`, or similar methods. This helper allows to raise a standardized error message across estimator.
- `validation.has_fit_parameter`: check that a given parameter is supported in the `fit` method of a given estimator.

Efficient Linear Algebra & Array Operations

- `extmath.randomized_range_finder`: construct an orthonormal matrix whose range approximates the range of the input. This is used in `extmath.randomized_svd`, below.
- `extmath.randomized_svd`: compute the k-truncated randomized SVD. This algorithm finds the exact truncated singular values decomposition using randomization to speed up the computations. It is particularly fast on large matrices on which you wish to extract only a small number of components.
- `arrayfuncs.cholesky_delete`: (used in `sklearn.linear_model.lars_path`) Remove an item from a cholesky factorization.
- `arrayfuncs.min_pos`: (used in `sklearn.linear_model.least_angle`) Find the minimum of the positive values within an array.
- `extmath.fast_logdet`: efficiently compute the log of the determinant of a matrix.
- `extmath.density`: efficiently compute the density of a sparse vector

- `extmath.safe_sparse_dot`: dot product which will correctly handle `scipy.sparse` inputs. If the inputs are dense, it is equivalent to `numpy.dot`.
- `extmath.weighted_mode`: an extension of `scipy.stats.mode` which allows each item to have a real-valued weight.
- `resample`: Resample arrays or sparse matrices in a consistent way. used in `shuffle`, below.
- `shuffle`: Shuffle arrays or sparse matrices in a consistent way. Used in `sklearn.cluster.k_means`.

Efficient Random Sampling

- `random.sample_without_replacement`: implements efficient algorithms for sampling `n_samples` integers from a population of size `n_population` without replacement.

Efficient Routines for Sparse Matrices

The `sklearn.utils.sparsefuncs` cython module hosts compiled extensions to efficiently process `scipy.sparse` data.

- `sparsefuncs.mean_variance_axis`: compute the means and variances along a specified axis of a CSR matrix. Used for normalizing the tolerance stopping criterion in `sklearn.cluster.KMeans`.
- `sparsefuncs_fast.inplace_csr_row_normalize_l1` and `sparsefuncs_fast.inplace_csr_row_normalize_l2`: can be used to normalize individual sparse samples to unit L1 or L2 norm as done in `sklearn.preprocessing.Normalizer`.
- `sparsefuncs.inplace_csr_column_scale`: can be used to multiply the columns of a CSR matrix by a constant scale (one scale per column). Used for scaling features to unit standard deviation in `sklearn.preprocessing.StandardScaler`.

Graph Routines

- `graph.single_source_shortest_path_length`: (not currently used in scikit-learn) Return the shortest path from a single source to all connected nodes on a graph. Code is adapted from networkx. If this is ever needed again, it would be far faster to use a single iteration of Dijkstra's algorithm from `graph_shortest_path`.
- `graph_shortest_path.graph_shortest_path`: (used in `sklearn.manifold.Isomap`) Return the shortest path between all pairs of connected points on a directed or undirected graph. Both the Floyd-Warshall algorithm and Dijkstra's algorithm are available. The algorithm is most efficient when the connectivity matrix is a `scipy.sparse.csr_matrix`.

Testing Functions

- `all_estimators` : returns a list of all estimators in scikit-learn to test for consistent behavior and interfaces.

Multiclass and multilabel utility function

- `multiclass.is_multilabel`: Helper function to check if the task is a multi-label classification one.
- `multiclass.unique_labels`: Helper function to extract an ordered array of unique labels from different formats of target.

Helper Functions

- `gen_even_slices`: generator to create `n`-packs of slices going up to `n`. Used in `sklearn.decomposition.dict_learning` and `sklearn.cluster.k_means`.
- `safe_mask`: Helper function to convert a mask to the format expected by the numpy array or scipy sparse matrix on which to use it (sparse matrices support integer indices only while numpy arrays support both boolean masks and integer indices).
- `safe_sqr`: Helper function for unified squaring (`**2`) of array-likes, matrices and sparse matrices.

Hash Functions

- `murmurhash3_32` provides a python wrapper for the `MurmurHash3_x86_32` C++ non cryptographic hash function. This hash function is suitable for implementing lookup tables, Bloom filters, Count Min Sketch, feature hashing and implicitly defined sparse random projections:

```
>>> from sklearn.utils import murmurhash3_32
>>> murmurhash3_32("some feature", seed=0) == -384616559
True

>>> murmurhash3_32("some feature", seed=0, positive=True) == 3910350737
True
```

The `sklearn.utils.murmurhash` module can also be “cimported” from other cython modules so as to benefit from the high performance of MurmurHash while skipping the overhead of the Python interpreter.

Warnings and Exceptions

- `deprecated`: Decorator to mark a function or class as deprecated.
- `sklearn.exceptions.ConvergenceWarning`: Custom warning to catch convergence problems. Used in `sklearn.covariance.graphical_lasso`.

Toggle Menu

 Show this page source