Configuration options for as-auto-sklearn

The training and testing scripts both accept a yaml configuration file. This document describes the available options and their effects.

The configuration file can contain other keys and values (such as those used by AutoFolio), but they will be ignored.

- `wallclock_limit`. The amount of time (in seconds) to use for training one fold of one solver
- `allowed_feature_groups`. A list of feature groups included in training (and testing). The strings must exactly match the keys of the feature_steps in the description of the aslib scenario.
- `imputer_strategy`. The approach to use for replacing missing values. The same strategy is applied to all features. Valid values are:
  - `median`
  - `mean`
  - `most_frequent`
- `preprocessing_strategy`. The approach to use for preprocessing the data. **N.B.** auto-sklearn already learns "optimal" strategies for preprocessing, so sophisticated methods are not sensible for this option. Valid values are:
  - `scale`. Value for each feature are scaled such that the values of the feature have a mean of 0 and a variance of 1.
  - `null`. No preprocessing is applied.
- `log_performance_data`. If this key is present with any value (even something like "no" or "False"), then the performance data (that is, the solver runtimes) will be transformed with `np.log1p` before training. After testing, the predictions will be transformed back using `np.exp1m`. The predictions reported with `test-as-auto-sklearn` will be in "normal" (not logged) space.
  N.B. auto-sklearn *does not* attempt to optimize taking the logarithm of input features, so it is reasonable to choose these by hand.
- `fields_to_log`. A list of features to transform using `np.log1p` before training. This is implemented as part of an `sklearn.Pipeline`, so the same transformations will be applied during testing.