# Model selection: choosing estimators and their parameters

## Score, and cross-validated scores

As we have seen, every estimator exposes a `score` method that can judge the quality of the fit (or the prediction) on new data. **Bigger is better**.

```
>>> from sklearn import datasets, svm
>>> X_digits, y_digits = datasets.load_digits(return_X_y=True)
>>> svc = svm.SVC(C=1, kernel='linear')
>>> svc.fit(X_digits[:-100], y_digits[:-100]).score(X_digits[-100:], y_digits[-100:])
0.98
```

To get a better measure of prediction accuracy (which we can use as a proxy for goodness of fit of the model), we can successively split the data in *folds* that we use for training and testing:

```
>>> import numpy as np
>>> X_folds = np.array_split(X_digits, 3)
>>> y_folds = np.array_split(y_digits, 3)
>>> scores = list()
>>> for k in range(3):
...     # We use 'list' to copy, in order to 'pop' later on
...     X_train = list(X_folds)
...     X_test = X_train.pop(k)
...     X_train = np.concatenate(X_train)
...     y_train = list(y_folds)
...     y_test = y_train.pop(k)
...     y_train = np.concatenate(y_train)
...     scores.append(svc.fit(X_train, y_train).score(X_test, y_test))
>>> print(scores)
[0.934..., 0.956..., 0.939...]
```

**System Message: ERROR/3** (D:\onboarding-resources\sample-onboarding-resources\scikit-learn-main\doc\tutorial\statistical_inference\[scikit-learn-main][doc][tutorial][statistical_inference]model_selection.rst, line 42)

Unknown directive type "currentmodule".

```
.. currentmodule:: sklearn.model_selection
```

This is called a :class:`KFold` cross-validation.

**System Message: ERROR/3** (D:\onboarding-resources\sample-onboarding-resources\scikit-learn-main\doc\tutorial\statistical_inference\[scikit-learn-main][doc][tutorial][statistical_inference]model_selection.rst, line 44); *backlink*

Unknown interpreted text role "class".

## Cross-validation generators

Scikit-learn has a collection of classes which can be used to generate lists of train/test indices for popular cross-validation strategies.

They expose a `split` method which accepts the input dataset to be split and yields the train/test set indices for each iteration of the chosen cross-validation strategy.

This example shows an example usage of the `split` method.

```
>>> from sklearn.model_selection import KFold, cross_val_score
>>> X = ["a", "a", "a", "b", "b", "c", "c", "c", "c", "c"]
>>> k_fold = KFold(n_splits=5)
>>> for train_indices, test_indices in k_fold.split(X):
...      print('Train: %s | test: %s' % (train_indices, test_indices))
Train: [2 3 4 5 6 7 8 9] | test: [0 1]
Train: [0 1 4 5 6 7 8 9] | test: [2 3]
Train: [0 1 2 3 6 7 8 9] | test: [4 5]
Train: [0 1 2 3 4 5 8 9] | test: [6 7]
Train: [0 1 2 3 4 5 6 7] | test: [8 9]
```

The cross-validation can then be performed easily:

```
>>> [svc.fit(X_digits[train], y_digits[train]).score(X_digits[test], y_digits[test])
...  for train, test in k_fold.split(X_digits)]
[0.963..., 0.922..., 0.963..., 0.963..., 0.930...]
```

The cross-validation score can be directly calculated using the :func:`cross_val_score` helper. Given an estimator, the cross-validation object and the input dataset, the :func:`cross_val_score` splits the data repeatedly into a training and a testing set, trains the estimator using the training set and computes the scores based on the testing set for each iteration of cross-validation.

**System Message: ERROR/3** (D:\onboarding-resources\sample-onboarding-resources\scikit-learn-main\doc\tutorial\statistical_inference\[scikit-learn-main][doc][tutorial][statistical_inference]model_selection.rst, line 77); *backlink*

Unknown interpreted text role "func".

**System Message: ERROR/3** (D:\onboarding-resources\sample-onboarding-resources\scikit-learn-main\doc\tutorial\statistical_inference\[scikit-learn-main][doc][tutorial][statistical_inference]model_selection.rst, line 77); *backlink*

Unknown interpreted text role "func".

By default the estimator's `score` method is used to compute the individual scores.

Refer the :ref`metrics module <metrics>` to learn more on the available scoring methods.

**System Message: ERROR/3** (D:\onboarding-resources\sample-onboarding-resources\scikit-learn-main\doc\tutorial\statistical_inference\[scikit-learn-main][doc][tutorial][statistical_inference]model_selection.rst, line 85); *backlink*

Unknown interpreted text role "ref".

```
>>> cross_val_score(svc, X_digits, y_digits, cv=k_fold, n_jobs=-1)
array([0.96388889, 0.92222222, 0.9637883 , 0.9637883 , 0.93036212])
```

*n_jobs=-1* means that the computation will be dispatched on all the CPUs of the computer.

Alternatively, the `scoring` argument can be provided to specify an alternative scoring method.

```
>>> cross_val_score(svc, X_digits, y_digits, cv=k_fold,
...                 scoring='precision_macro')
array([0.96578289, 0.92708922, 0.96681476, 0.96362897, 0.93192644])
```

**Cross-validation generators**

| :class:`KFold` **(n_splits, shuffle, random_state)** | :class:`StratifiedKFold` **(n_splits, shuffle, random_state)** | :class:`GroupKFold` **(n_splits)** |
|---|---|---|
| System Message: ERROR/3 (D:\onboarding-resources\sample-onboarding-resources\scikit-learn-main\doc\tutorial\statistical_inference\[scikit-learn-main][doc][tutorial][statistical_inference]model_selection.rst, line 108); *backlink*<br>Unknown interpreted text role "class". | System Message: ERROR/3 (D:\onboarding-resources\sample-onboarding-resources\scikit-learn-main\doc\tutorial\statistical_inference\[scikit-learn-main][doc][tutorial][statistical_inference]model_selection.rst, line 110); *backlink*<br>Unknown interpreted text role "class". | System Message: ER<br>resources\sample-o<br>resources\scikit-l<br>main\doc\tutorial\<br>[scikit-learn-main]<br>[statistical_infere<br>line 112); *backlink*<br>Unknown interpreted tex |
| Splits it into K folds, trains on K-1 and then tests on the left-out. | Same as K-Fold but preserves the class distribution within each fold. | Ensures that the same group is n |
| :class:`ShuffleSplit` **(n_splits, test_size, train_size, random_state)** | :class:`StratifiedShuffleSplit` | :class:`GroupShuffleSplit` |
| System Message: ERROR/3 (D:\onboarding-resources\sample-onboarding-resources\scikit-learn-main\doc\tutorial\statistical_inference\[scikit-learn-main][doc][tutorial][statistical_inference]model_selection.rst, line 128); *backlink*<br>Unknown interpreted text role "class". | System Message: ERROR/3 (D:\onboarding-resources\sample-onboarding-resources\scikit-learn-main\doc\tutorial\statistical_inference\[scikit-learn-main][doc][tutorial][statistical_inference]model_selection.rst, line 130); *backlink*<br>Unknown interpreted text role "class". | System Message: ER<br>resources\sample-o<br>resources\scikit-l<br>main\doc\tutorial\<br>[scikit-learn-main]<br>[statistical_infere<br>line 132); *backlink*<br>Unknown interpreted tex |
| Generates train/test indices based on random permutation. | Same as shuffle split but preserves the class distribution within each iteration. | Ensures that the same group is n |
| :class:`LeaveOneGroupOut` **()** | :class:`LeavePGroupsOut` **(n_groups)** | :class:`LeaveOneOut` **()** |
| System Message: ERROR/3 (D:\onboarding-resources\sample-onboarding-resources\scikit-learn-main\doc\tutorial\statistical_inference\[scikit-learn-main][doc][tutorial][statistical_inference]model_selection.rst, line 147); *backlink*<br>Unknown interpreted text role "class". | System Message: ERROR/3 (D:\onboarding-resources\sample-onboarding-resources\scikit-learn-main\doc\tutorial\statistical_inference\[scikit-learn-main][doc][tutorial][statistical_inference]model_selection.rst, line 149); *backlink*<br>Unknown interpreted text role "class". | System Message: ER<br>resources\sample-o<br>resources\scikit-l<br>main\doc\tutorial\<br>[scikit-learn-main]<br>[statistical_infere<br>line 151); *backlink*<br>Unknown interpreted tex |
| Takes a group array to group observations. | Leave P groups out. | Leave one observation out. |
| :class:`LeavePOut` **(p)** | :class:`PredefinedSplit` | |
| System Message: ERROR/3 (D:\onboarding-resources\sample-onboarding-resources\scikit-learn-main\doc\tutorial\statistical_inference\[scikit-learn-main][doc][tutorial][statistical_inference]model_selection.rst, line 169); *backlink*<br>Unknown interpreted text role "class". | System Message: ERROR/3 (D:\onboarding-resources\sample-onboarding-resources\scikit-learn-main\doc\tutorial\statistical_inference\[scikit-learn-main][doc][tutorial][statistical_inference]model_selection.rst, line 171); *backlink*<br>Unknown interpreted text role "class". | |
| Leave P observations out. | Generates train/test indices based on predefined splits. | |

System Message: ERROR/3 (D:\onboarding-resources\sample-onboarding-resources\scikit-learn-main\doc\tutorial\statistical_inference\[scikit-learn-main][doc][tutorial][statistical_inference]model_selection.rst, line 180)

Unknown directive type "currentmodule".

```
.. currentmodule:: sklearn.svm
```

**Exercise**

On the digits dataset, plot the cross-validation score of a :class:`SVC` estimator with an linear kernel as a function of parameter `C` (use a logarithmic grid of points, from 1 to 10).

System Message: ERROR/3 (D:\onboarding-resources\sample-onboarding-resources\scikit-learn-main\doc\tutorial\statistical_inference\[scikit-learn-main][doc][tutorial][statistical_inference]model_selection.rst, line 184); *backlink*

Unknown interpreted text role "class".

System Message: ERROR/3 (D:\onboarding-resources\sample-onboarding-resources\scikit-learn-main\doc\tutorial\statistical_inference\[scikit-learn-main][doc][tutorial][statistical_inference]model_selection.rst, line 188)

Unknown directive type "literalinclude".

```
.. literalinclude:: ../../auto_examples/exercises/plot_cv_digits.py
    :lines: 13-23
```

**Solution:** :ref:`sphx_glr_auto_examples_exercises_plot_cv_digits.py`

System Message: ERROR/3 (D:\onboarding-resources\sample-onboarding-resources\scikit-learn-main\doc\tutorial\statistical_inference\[scikit-learn-main][doc][tutorial][statistical_inference]model_selection.rst, line 196); *backlink*

Unknown interpreted text role "ref".

## Grid-search and cross-validated estimators

### Grid-search

System Message: ERROR/3 (D:\onboarding-resources\sample-onboarding-resources\scikit-

System Message: ERROR/3 (D:\onboarding-resources\sample-onboarding-resources\scikit-learn-main\doc\tutorial\statistical_inference\[scikit-learn-main][doc][tutorial][statistical_inference]model_selection.rst, line 204)

Unknown directive type "currentmodule".

```
.. currentmodule:: sklearn.model_selection
```

scikit-learn provides an object that, given data, computes the score during the fit of an estimator on a parameter grid and chooses the parameters to maximize the cross-validation score. This object takes an estimator during the construction and exposes an estimator API:

```
>>> from sklearn.model_selection import GridSearchCV, cross_val_score
>>> Cs = np.logspace(-6, -1, 10)
>>> clf = GridSearchCV(estimator=svc, param_grid=dict(C=Cs),
...                    n_jobs=-1)
>>> clf.fit(X_digits[:1000], y_digits[:1000])        # doctest: +SKIP
GridSearchCV(cv=None,...
>>> clf.best_score_                                  # doctest: +SKIP
0.925...
>>> clf.best_estimator_.C                            # doctest: +SKIP
0.0077...

>>> # Prediction performance on test set is not as good as on train set
>>> clf.score(X_digits[1000:], y_digits[1000:])      # doctest: +SKIP
0.943...
```

By default, the :class:`GridSearchCV` uses a 5-fold cross-validation. However, if it detects that a classifier is passed, rather than a regressor, it uses a stratified 5-fold.

System Message: ERROR/3 (D:\onboarding-resources\sample-onboarding-resources\scikit-learn-main\doc\tutorial\statistical_inference\[scikit-learn-main][doc][tutorial][statistical_inference]model_selection.rst, line 227); *backlink*

Unknown interpreted text role "class".

**Nested cross-validation**

```
>>> cross_val_score(clf, X_digits, y_digits) # doctest: +SKIP
array([0.938..., 0.963..., 0.944...])
```

Two cross-validation loops are performed in parallel: one by the :class:`GridSearchCV` estimator to set `gamma` and the other one by `cross_val_score` to measure the prediction performance of the estimator. The resulting scores are unbiased estimates of the prediction score on new data.

System Message: ERROR/3 (D:\onboarding-resources\sample-onboarding-resources\scikit-learn-main\doc\tutorial\statistical_inference\[scikit-learn-main][doc][tutorial][statistical_inference]model_selection.rst, line 238); *backlink*

Unknown interpreted text role "class".

**Warning**

You cannot nest objects with parallel computing (`n_jobs` different than 1).

### Cross-validated estimators

Cross-validation to set a parameter can be done more efficiently on an algorithm-by-algorithm basis. This is why, for certain estimators, scikit-learn exposes :ref:`cross_validation` estimators that set their parameter automatically by cross-validation:

System Message: ERROR/3 (D:\onboarding-resources\sample-onboarding-resources\scikit-learn-main\doc\tutorial\statistical_inference\[scikit-learn-main][doc][tutorial][statistical_inference]model_selection.rst, line 254); *backlink*

Unknown interpreted text role "ref".

```
>>> from sklearn import linear_model, datasets
>>> lasso = linear_model.LassoCV()
>>> X_diabetes, y_diabetes = datasets.load_diabetes(return_X_y=True)
>>> lasso.fit(X_diabetes, y_diabetes)
LassoCV()
>>> # The estimator chose automatically its lambda:
>>> lasso.alpha_
0.00375...
```

These estimators are called similarly to their counterparts, with 'CV' appended to their name.

**Exercise**

On the diabetes dataset, find the optimal regularization parameter alpha.

**Bonus**: How much can you trust the selection of alpha?

System Message: ERROR/3 (D:\onboarding-resources\sample-onboarding-resources\scikit-learn-main\doc\tutorial\statistical_inference\[scikit-learn-main][doc][tutorial][statistical_inference]model_selection.rst, line 278)

Unknown directive type "literalinclude".

```
.. literalinclude:: ../../auto_examples/exercises/plot_cv_diabetes.py
    :lines: 17-24
```

**Solution:** :ref:`sphx_glr_auto_examples_exercises_plot_cv_diabetes.py`

System Message: ERROR/3 (D:\onboarding-resources\sample-onboarding-resources\scikit-learn-main\doc\tutorial\statistical_inference\[scikit-learn-main][doc][tutorial][statistical_inference]model_selection.rst, line 281); *backlink*

Unknown interpreted text role "ref".