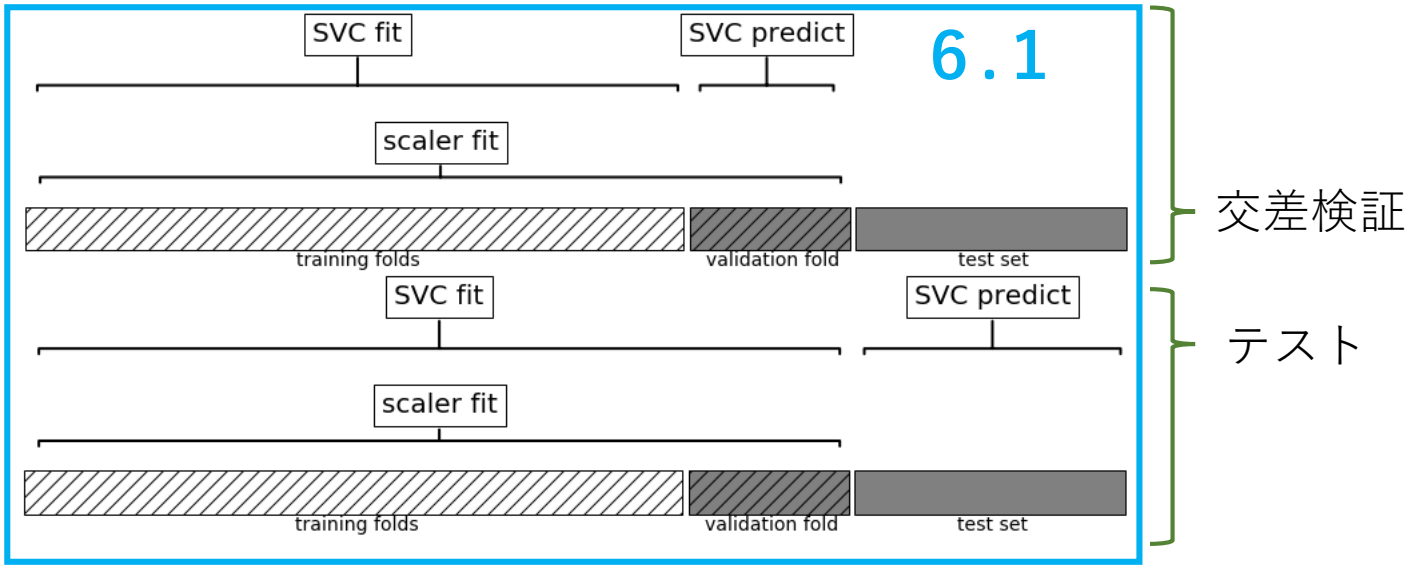
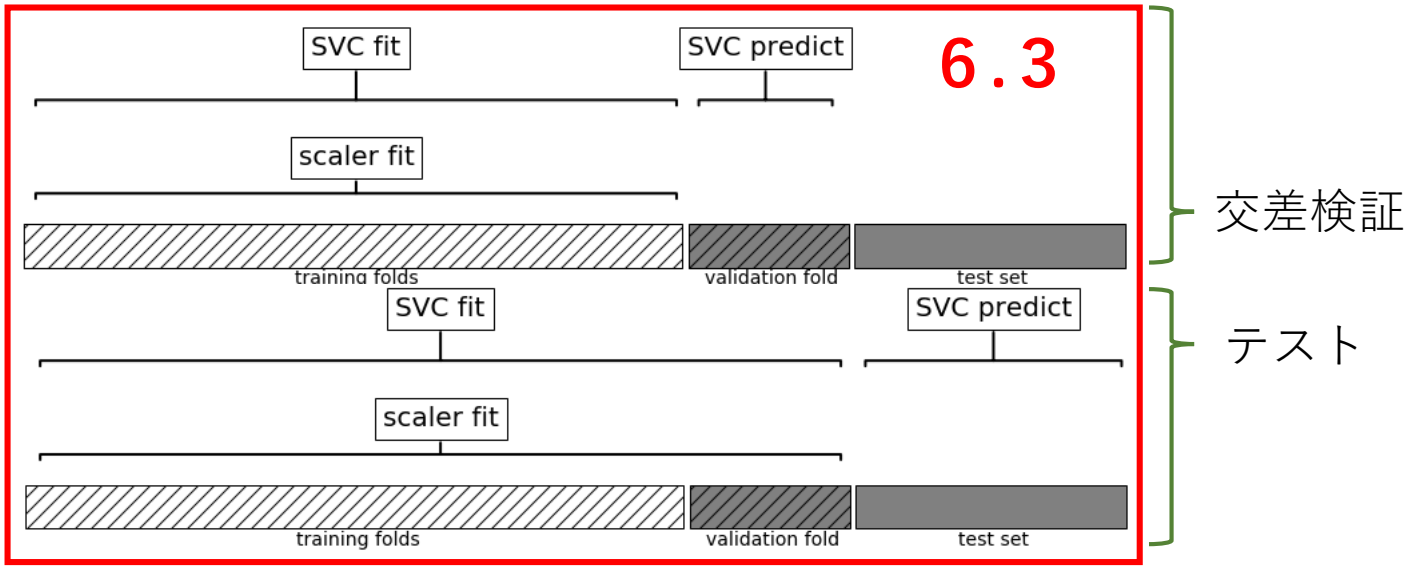
Python ではじめる機械学習

6.3 パイプラインを用いたグリッドサーチ

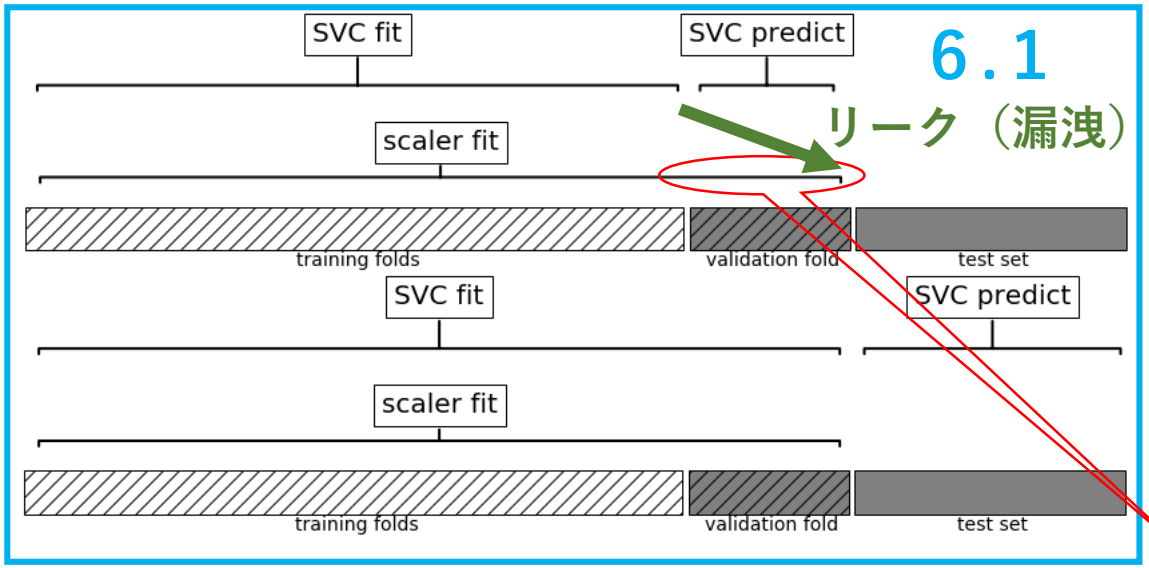
15T4057F　藤井　真

6.3 パイプラインを用いたグリッドサーチ





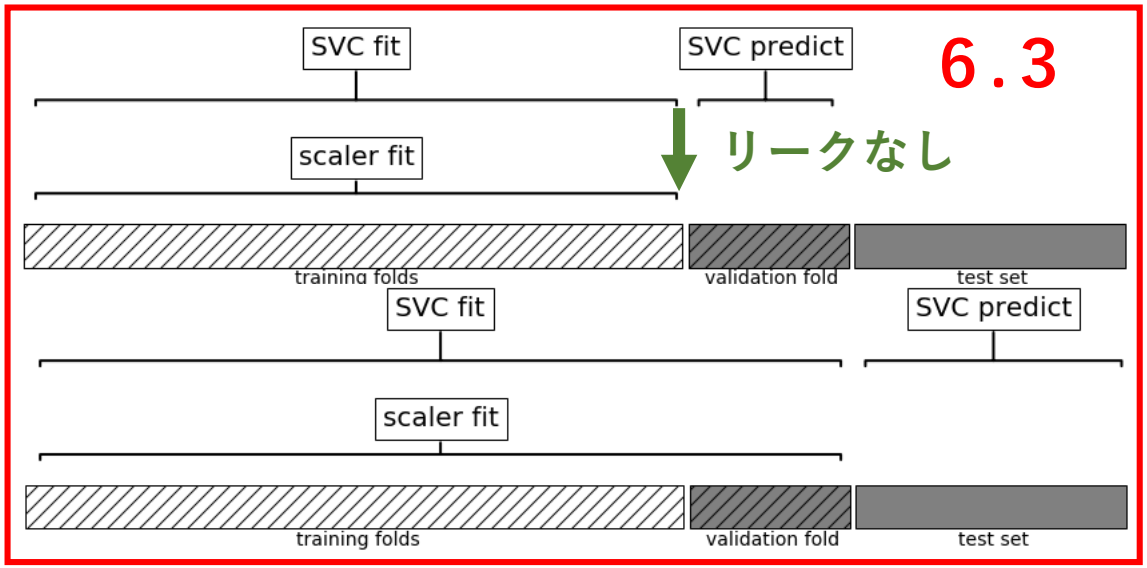
6.3 パイプラインを用いたグリッドサーチ



交差検証

テスト

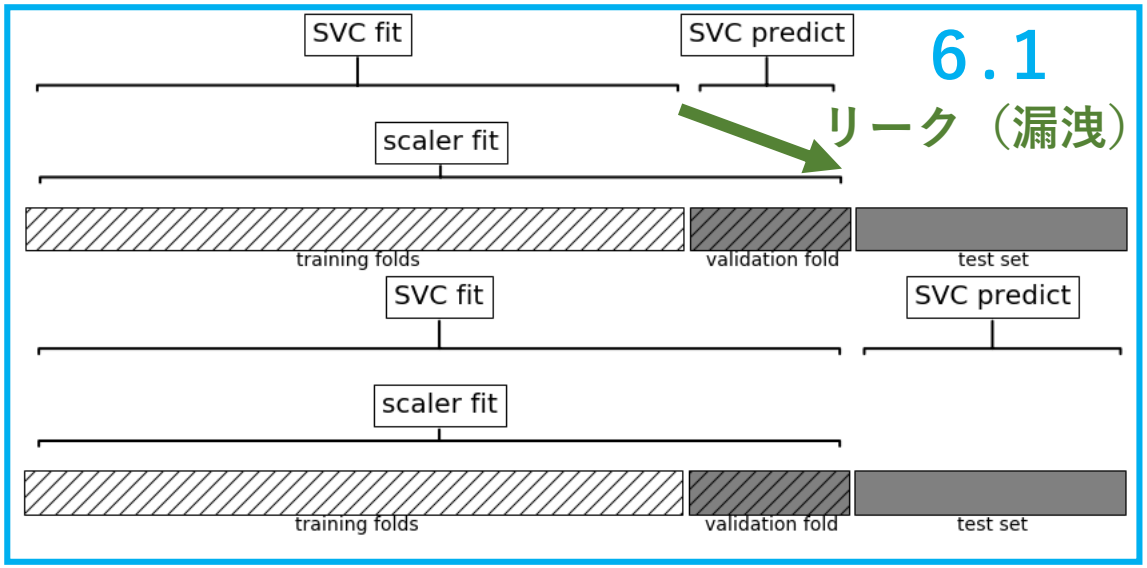
6.1では交差検証時に
準テストデータとして
扱うべきデータも
スケール変換に
含んでいる

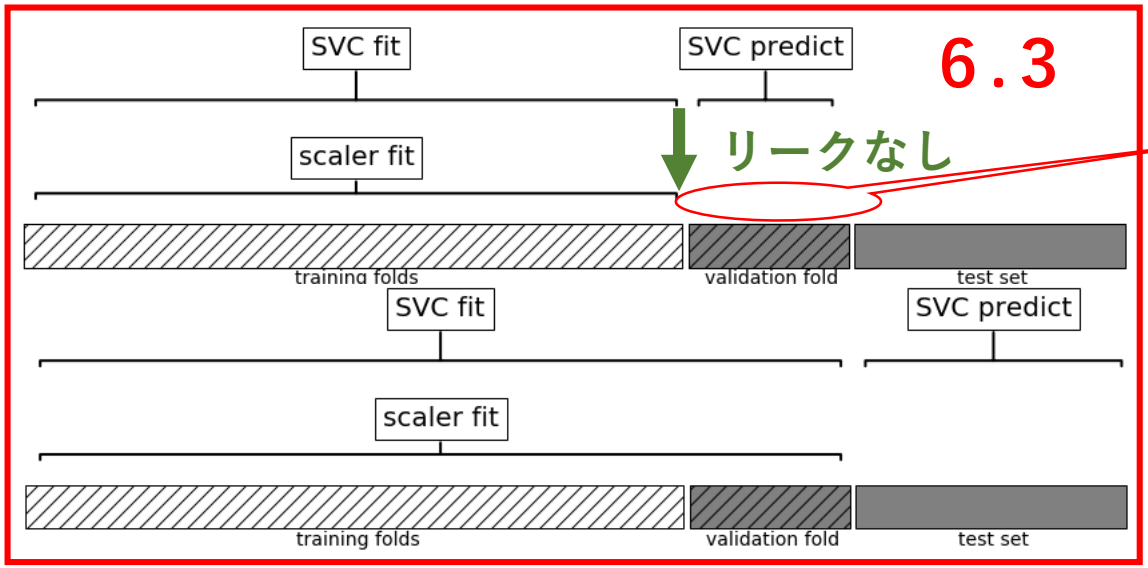


交差検証

テスト

6.3 パイプラインを用いたグリッドサーチ





6.3はパイプラインを用いて、このリークを無くす方法の紹介。

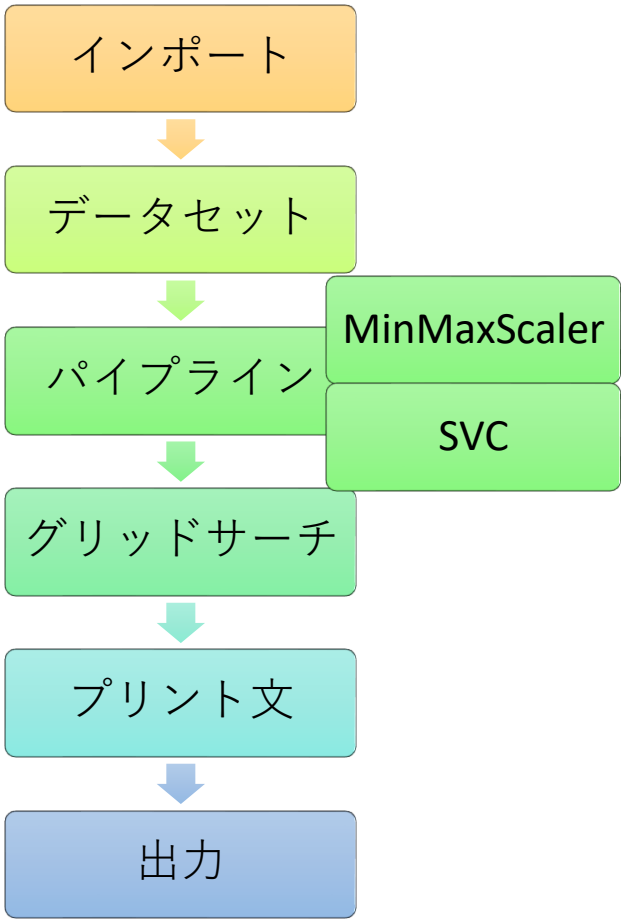
6.3 パイプラインを用いたグリッドサーチ

ソース

```
from sklearn.svm import SVC
from sklearn.datasets import load_breast_cancer
from sklearn.model_selection import train_test_split
from sklearn.preprocessing import MinMaxScaler
from sklearn.model_selection import GridSearchCV
from sklearn.pipeline import Pipeline

cancer = load_breast_cancer()
X_train, X_test, y_train, y_test = train_test_split(
    cancer.data, cancer.target, random_state=0)

pipe = Pipeline([("scaler", MinMaxScaler()), ("svm", SVC())])
pipe.fit(X_train, y_train)

param_grid = {'svm__C': [0.001, 0.01, 0.1, 1, 10, 100],
              'svm__gamma': [0.001, 0.01, 0.1, 1, 10, 100]}

grid = GridSearchCV(pipe, param_grid=param_grid, cv=5)
grid.fit(X_train, y_train)

print("Best cross-validation accuracy: {:.2f}".format(grid.best_score_))
print("Test set score: {:.2f}".format(grid.score(X_test, y_test)))
print("Best parameters: {}".format(grid.best_params_))
```

```
Best cross-validation accuracy: 0.98
Test set score: 0.97
Best parameters: {'svm__C': 1, 'svm__gamma': 1}
```

概要図



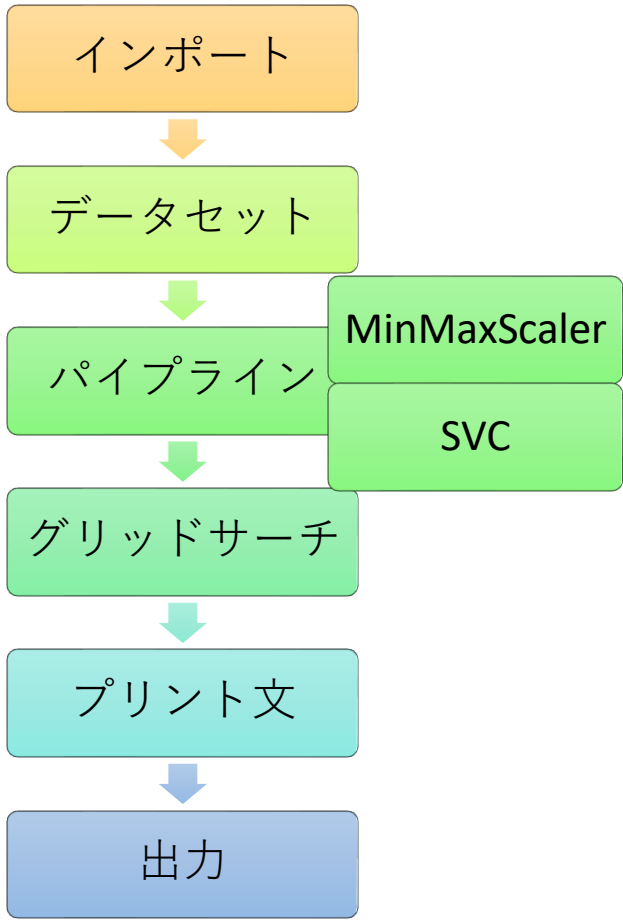
6.3 パイプラインを用いたグリッドサーチ

ソース

```
from sklearn.svm import SVC
from sklearn.datasets import load_breast_cancer
from sklearn.model_selection import train_test_split
from sklearn.preprocessing import MinMaxScaler
from sklearn.model_selection import GridSearchCV
from sklearn.pipeline import Pipeline

cancer = load_breast_cancer()
X_train, X_test, y_train, y_test = train_test_split(
    cancer.data, cancer.target, random_state=0)

pipe = Pipeline([("scaler", MinMaxScaler()), ("svm", SVC())])
pipe.fit(X_train, y_train)

param_grid = {'svm__C': [0.001, 0.01, 0.1, 1, 10, 100],
              'svm__gamma': [0.001, 0.01, 0.1, 1, 10, 100]}

grid = GridSearchCV(pipe, param_grid=param_grid, cv=5)
grid.fit(X_train, y_train)

print("Best cross-validation accuracy: {:.2f}".format(grid.best_score_))
print("Test set score: {:.2f}".format(grid.score(X_test, y_test)))
print("Best parameters: {}".format(grid.best_params_))
```

```
Best cross-validation accuracy: 0.98
Test set score: 0.97
Best parameters: {'svm__C': 1, 'svm__gamma': 1}
```

概要図



変更点 パイプラインで2つの処理（scalerとsvm）を行っているため、
グリッドサーチパラメータがどちらのものか明示する必要がある。
今回はsvmなので‘svm__C’として（_）アンダースコア2つを挟む。

6.3 パイプラインを用いたグリッドサーチ

- 情報リーク（漏洩）の影響
  ⇒前処理の性質による

  - スケール変換（今回）　⇒影響は限定的
  - 特徴量抽出や特徴量選択⇒影響が大きい

  テストデータや検証用データの押さえるべき特徴を
  前処理時に参照してしまうため。