Analytics Vidhya
Learn Everything About Analytics

# Cheatsheet:Scikit Learn

Scikit-Learn is the most popular and widely used library for machine learning in Python.

## Pre-Processing

| | Function | Description |
|---|---|---|
| 1 | sklearn.preprocessing.StandardScaler | Standardize features by removing the mean and scaling to unit variance |
| 2 | sklearn.preprocessing.Imputer | Imputation transformer for completing missing values. |
| 3 | sklearn.preprocessing.LabelBinarizer | Binarize labels in a one-vs-all fashion |
| 4 | sklearn.preprocessing.OneHot Encoder | Encode categorical integer features using a one-hot a.k.a one-of-K scheme. |
| 5 | sklearn.preprocessing.Polynomial Features | Generate polynomial and interaction features. |

## Regression

| | Function | Description |
|---|---|---|
| 1 | sklearn.tree.DecisionTreeRegressor | A decision tree regressor |
| 2 | sklearn.svm.SVR | Epsilon-Support Vector Regression |
| 3 | sklearn.linear_model.Linear Regression | Ordinary least squares Linear Regression |
| 4 | sklearn.linear_model.Lasso | Linear Model trained with L1 prior as regularizer (a.k.a the Lasso) |
| 5 | sklearn.linear_model.SGDRegressor | Linear model fitted by minimizing a regularized empirical loss with SGD |
| 6 | sklearn.linear_model.ElasticNet | Linear regression with combined L1 and L2 priors as regularizer |
| 7 | sklearn.ensemble.RandomForest Regressor | A random forest regressor |
| 8 | sklearn.ensemble.GradientBoosting Regressor | Gradient Boosting for regression |
| 9 | sklearn.neural_network. MLPRegressor | Multi-layer Perceptron regressor |

## Classification

| | Function | Description |
|---|---|---|
| 1 | sklearn.neural_network.MLP Classifier | Multi-layer Perceptron classifier |
| 2 | sklearn.tree.DecisionTreeClassifier | A decision tree classifier |
| 3 | sklearn.svm.SVC | C-Support Vector Classification |
| 4 | sklearn.linear_model.Logistic Regression | Logistic Regression (at.k.a logit, Max Ent) classifier |
| 5 | sklearn.linear_model.SGDClassifier | Linear classifiers (SVM, logistic regression, a.o.) with SGD training |
| 6 | sklearn.naive_bayes.GaussianNB | Gaussian Naive Bayes |
| 7 | sklearn.neighbors.KNeighbors Classifier | Classifier implementing the k-nearest neighbors vote |
| 8 | sklearn.ensemble.RandomForest Classifier | A random forest classifier |
| 9 | sklearn.ensemble.GradientBoosting Classifier | Gradient Boosting for classification |

## Clustering

| | Function | Description |
|---|---|---|
| 1 | sklearn.cluster.KMeans | K-Means clustering |
| 2 | sklearn.cluster.DBSCAN | Perform DBSCAN clustering from vector array or distance matrix |
| 3 | sklearn.cluster.Agglomerative Clustering | Agglomerative Clustering |
| 4 | sklearn.cluster.SpectralBiclustering | Spectral bi-clustering |

## DimensionalityReduction

| | Function | Description |
|---|---|---|
| 1 | sklearn.decomposition.PCA | Principal component analysis (PCA) |
| 2 | sklearn.decomposition.Latent DirichletAllocation | Latent Dirichlet Allocation with online variational Bayes algorithm |
| 3 | sklearn.decomposition.SparseCoder | Sparse coding |
| 4 | sklearn.decomposition.Dictionary Learning | Dictionary learning |

## Model Selection

| | Function | Description |
|---|---|---|
| 1 | sklearn.model_selection.KFold | K-Folds cross-validator |
| 2 | sklearn.model_selection.Stratified KFold | Stratified K-Folds cross-validator |
| 3 | sklearn.model_selection.TimeSeries Split | Time Series cross-validator |
| 4 | sklearn.model_selection.train _test_split | Split arrays or matrices into random train and test subsets |
| 5 | sklearn.model_selection.GridSearch CV | Exhaustive search over specified parameter values for an estimator. |
| 6 | sklearn.model_selection.Randomized SearchCV | Randomized search on hyper parameters. |
| 7 | sklearn.model_selection.cross_val_ score | Evaluate a score by cross-validation |

## Metric

| | Function | Description |
|---|---|---|
| 1 | sklearn.metrics.accuracy_score | Classification Metric: Accuracy classification score |
| 2 | sklearn.metrics.log_loss | Classification Metric: Log loss, a.k.a logistic loss or cross-entropy loss |
| 3 | sklearn.metrics.roc_auc_score | Classification Metric: Compute Rece iver operating characteristic (ROC) |
| 4 | sklearn.metrics.mean_absolute_error | Regression Metric: Mean absolute error regression loss |
| 5 | sklearn.metrics.r2_score | Regression Metric: R^2 (coefficient of determination)regression score function. |
| 6 | sklearn.metrics.label_ranking_loss | Ranking Metric: Compute Ranking loss measure |
| 7 | sklearn.metrics.mutual_info_score | Clustering Metric: Mutual Information between two clusterings. |

## Miscellaneous

| | Function | Description |
|---|---|---|
| 1 | sklearn.datasets.load_boston | Load and return the boston house-prices dataset (regression) |
| 2 | sklearn.datasets.make_classification | Generate a random n-class classifi-cation problem |
| 3 | sklearn.feature_extraction.Feature Hasher | Implements feature hashing, a.k.a the hashing trick |
| 4 | sklearn.feature_selection.SelectK Best | Select features according to the k highest scores |
| 5 | sklearn.pipeline.Pipeline | Pipeline of transforms with a final estimator |
| 6 | sklearn.semi_supervised.Label Propagation | Label Propagation classifier for semi-supervised learning |