Creating your first model

Using SKLearn's datasets

Terry McCann | @SQLShark

Microsoft Partner
Microsoft
Gold Data Analytics
Gold Data Platform
Gold Cloud Platform

MVP Microsoft® Most Valuable Professional

adatis

Sample Data

`sklearn.datasets` : Datasets

The `sklearn.datasets` module includes utilities to load datasets, including methods to load and fetch popular reference datasets. It also features some artificial data generators.

User guide: See the Dataset loading utilities section for further details.

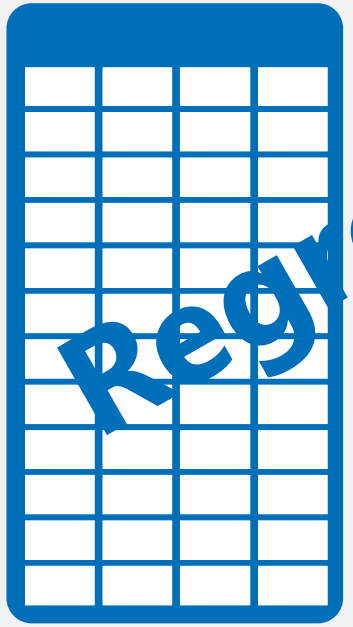
Loaders

`datasets.clear_data_home` ([data_home])	Delete all the content of the data home cache.
`datasets.dump_svmlight_file` (X, y, f[, ...])	Dump the dataset in svmlight / libsvm file format.
`datasets.fetch_20newsgroups` ([data_home, ...])	Load the filenames and data from the 20 newsgroups dataset (classification).
`datasets.fetch_20newsgroups_vectorized` ([...])	Load the 20 newsgroups dataset and vectorize it into token counts (classification).
`datasets.fetch_california_housing` ([...])	Load the California housing dataset (regression).
`datasets.fetch_covtype` ([data_home, ...])	Load the covertype dataset (classification).
`datasets.fetch_kddcup99` ([subset, data_home, ...])	Load the kddcup99 dataset (classification).
`datasets.fetch_lfw_pairs` ([subset, ...])	Load the Labeled Faces in the Wild (LFW) pairs dataset (classification).
`datasets.fetch_lfw_people` ([data_home, ...])	Load the Labeled Faces in the Wild (LFW) people dataset (classification).
`datasets.fetch_olivetti_faces` ([data_home, ...])	Load the Olivetti faces data-set from AT&T (classification).
`datasets.fetch_openml` ([name, version, ...])	Fetch dataset from openml by name or dataset id.
`datasets.fetch_rcv1` ([data_home, subset, ...])	Load the RCV1 multilabel dataset (classification).
`datasets.fetch_species_distributions` ([...])	Loader for species distribution dataset from Phillips et.
`datasets.get_data_home` ([data_home])	Return the path of the scikit-learn data dir.
`datasets.load_boston` ([return_X_y])	Load and return the boston house-prices dataset (regression).
`datasets.load_breast_cancer` ([return_X_y])	Load and return the breast cancer wisconsin dataset (classification).
`datasets.load_diabetes` ([return_X_y])	Load and return the diabetes dataset (regression).
`datasets.load_digits` ([n_class, return_X_y])	Load and return the digits dataset (classification).
`datasets.load_files` (container_path[, ...])	Load text files with categories as subfolder names.
`datasets.load_iris` ([return_X_y])	Load and return the iris dataset (classification).
`datasets.load_linnerud` ([return_X_y])	Load and return the linnerud dataset (multivariate regression).
`datasets.load_sample_image` (image_name)	Load the numpy array of a single sample image
`datasets.load_sample_images` ()	Load sample images for image manipulation.
`datasets.load_svmlight_file` (f[, n_features, ...])	Load datasets in the svmlight / libsvm format into sparse CSR matrix
`datasets.load_svmlight_files` (files[, ...])	Load dataset from multiple files in SVMlight format
`datasets.load_wine` ([return_X_y])	Load and return the wine dataset (classification).
`datasets.mldata_filename` (dataname)	DEPRECATED: mldata_filename was deprecated in version 0.20 and will be removed in version 0.22

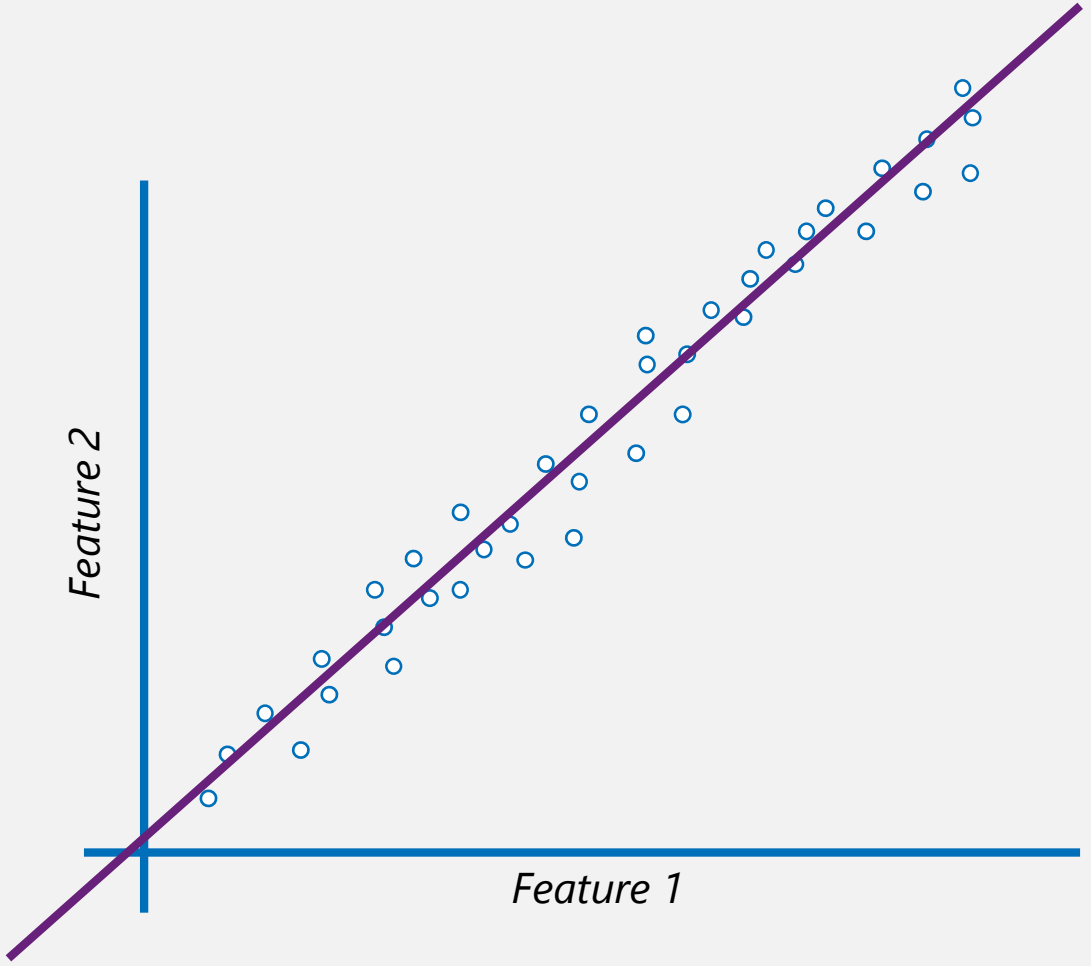
Samples generator

Diabetes data

data



Regression



5.2.3. Diabetes dataset

Ten baseline variables, age, sex, body mass index, average blood pressure, and six blood serum measurements were obtained for each of n = 442 diabetes patients, as well as the response of interest, a quantitative measure of disease progression one year after baseline.

Data Set Characteristics:

Number of Instances:	
	442
Number of Attributes:	
	First 10 columns are numeric predictive values
Target:	Column 11 is a quantitative measure of disease progression one year after baseline
Attribute Information:	
	• Age • Sex • Body mass index • Average blood pressure • S1 • S2 • S3 • S4 • S5 • S6

Note: Each of these 10 feature variables have been mean centered and scaled by the standard deviation times n_samples (i.e. the sum of squares of each column totals 1).

Source URL: http://www4.stat.ncsu.edu/~boos/var.select/diabetes.html

For more information see: Bradley Efron, Trevor Hastie, Iain Johnstone and Robert Tibshirani (2004) “Least Angle Regression,” Annals of Statistics (with discussion), 407-499. (http://web.stanford.edu/~hastie/Papers/LARS/LeastAngle_2002.pdf)

Demo – Exploring Diabetes Data

Lab 01 – Creating a model

15 minutes