Working with Specialized Datasets

Janani Ravi

CO-FOUNDER, LOONYCORN

www.loonycorn.com

Overview

Internal sample datasets in scikit-learn

Real world data for common models

Numeric, text and image data

Artificial datasets in scikit-learn

Generate data for classification, regression and clustering

Generate data for dimensionality reduction

Datasets in scikit-learn

Internal datasets

Artificial datasets

External datasets

Datasets in scikit-learn

Internal datasets

Artificial datasets

External datasets

Internal Datasets

Common datasets in easy-to-use form

Tailored for classification, regression

Image, text, and numeric datasets

Boston home prices

California housing

Olivetti faces

20 newsgroups

...

Datasets in scikit-learn

Internal datasets

Artificial datasets

External datasets

Artificial Datasets

Datasets with specific properties

Useful in training, visualizing models

Classification and regression datasets

S-curves, Swiss Rolls - Manifold learning

Blobs - Clustering

Very handy utilities

Artificial Datasets

sklearn.datasets.make_regression

sklearn.datasets.make_classification

sklearn.datasets.make_blobs

sklearn.datasets.make_circles

sklearn.datasets.make_low_rank_matrix

sklearn.datasets.make_s_curve

sklearn.datasets.make_swiss_roll

Datasets in scikit-learn

Internal datasets

Artificial datasets

External datasets

External Datasets

Usually loaded using Pandas

Can then be processed using scikit-learn, numpy, matplotlib, etc.

Demo

Using the internal real world datasets available in scikit-learn

Demo

Generating artificial datasets to use with regression, classification, clustering and dimensionality reduction models

Summary

Internal sample datasets in scikit-learn

Real world data for common models

Numeric, text and image data

Artificial datasets in scikit-learn

Generate data for classification, regression and clustering

Generate data for dimensionality reduction