scikit-learn toy datasets

scikit-learn provides some built-in datasets that can be used for testing purposes. They're all available in the package `sklearn.datasets` and have a common structure: the data instance variable contains the whole input set `X` while target contains the labels for classification or target values for regression. For example, considering the Boston house pricing dataset (used for regression), we have:

Copy

```
from sklearn.datasets import load_boston

>>> boston = load_boston()
>>> X = boston.data
>>> Y = boston.target

>>> X.shape
(506, 13)
>>> Y.shape
(506,)
```

In this case, we have 506 samples with 13 features and a single target value. In this book, we're going to use it for regressions and the MNIST handwritten digit dataset (`load_digits()`) for classification tasks. scikit-learn also provides functions for creating dummy datasets from scratch: `make_classification()` , `make_regression()` , and `make_blobs()` (particularly useful for testing...

Continue reading with a 10 day free trial

With a Packt Subscription, you can keep track of your learning and progress your skills with 7,000+ eBooks and Videos.

Continue learning now (/checkout/packt-subscription-monthly-launch-offer?freeTrial)

❮ Previous Section (/book/big_data_and_business_intelligence/9781785889622/3)

Next Section ❯ (/book/big_data_and_business_intelligence/9781785889622/3/ch03lvl1sec2