Layout of Datasets

The scikit-learn deals with learning information from one or more datasets that are represented as 2D arrays. They can be understood as a list of multi-dimensional observations. We say that the first axis of these arrays is the samples axis, while the second is the features axis.
When the data is not intially in the (`n_samples, n_features`) shape, it needs to be preprocessed to be used by the scikit.

Packaged Datasets

The scikit-learn library is packaged with datasets. These datasets are useful for getting a handle on a given machine learning algorithm or library feature before using it in your own work.

A simple example shipped with the scikit: iris dataset

```
>>> from scikits.learn import datasets
>>> iris = datasets.load_iris()
>>> data = iris.data
>>> data.shape
(150, 4)
```

Iris is made of 150 observations of irises, each described by 4 features: their sepal and petal length and width, as detailed in `iris.DESCR`.

scikit-learn embeds a copy of the iris CSV file along with a helper function to load it into numpy arrays:

```
from sklearn.datasets import load_iris

## -  Load the packaged iris flowers dataset
## - Iris flower dataset
## - (4x150, reals, multi-label classification)

iris = load_iris()
print(iris)
iris.keys()
```

Load from CSV

- In most of the Scikit-learn algorithms, the data must be loaded as a `Bunch` Object.
- However there are many example in the tutorial where `load_files()` or other functions are used to populate the bunch object.
- Function like `load_files()` expect data to be present in certain format. Suppose we have a different format in which data is stored.
- It is very common to have a dataset as a CSV file on the local workstation or on a remote server.
- You load a CSV file from a URL, in this case the Pima Indians diabetes classification dataset from the UCI Machine Learning Repository.
- From the prepared `X` and `y` variables, you can train a machine learning model.

```
# Pima Indians diabetes
# Load the  dataset from CSV URL

import numpy as np
import urllib

# URL for the Pima Indians Diabetes dataset
# (UCI Machine Learning Repository)

url = "http://goo.gl/j0Rvxq"

# download the file
raw_data = urllib.urlopen(url)

# load the CSV file as a numpy matrix

dataset = np.loadtxt(raw_data, delimiter=",")
print(dataset.shape)

# separate the data from the target attributes
X = dataset[:,0:7]
y = dataset[:,8]
```