

深度學習 Pytorch手把手實作-資料庫

黃志勝

義隆電子人工智慧研發部

國立陽明交通大學AI學院 合聘助理教授

國立台北科技大學電資學院 合聘助理教授



資料庫

- 1. 公開資料庫
UCI database,
MNIST,
ImageNet,
MS-COCO,... etc.
- 2. 私有資料庫



機器學習的資料庫

• 1. UCI database



Welcome to the UC Irvine Machine Learning Repository!

We currently maintain 559 data sets as a service to the machine learning community. You may [view all data sets](#) through our searchable interface. For a general overview of sets in publications, please read our [citation policy](#). If you wish to donate a data set, please consult our [donation policy](#). For any other questions, feel free to [contact the I](#)

Supported By:  In Collaboration With 

Latest News:

09-24-2018: Welcome to the new Repository admins Dheeru Dua and ES Karra Taniskidou!
04-04-2013: Welcome to the new Repository admins Kevin Bache and Moshe Lichman!
03-01-2010: Note from donor regarding Netflix data
10-16-2009: Two new data sets have been added.
09-14-2009: Several data sets have been added.
03-24-2008: New data sets have been added!
06-25-2007: Two new data sets have been added: UCI Pen Characters, MAGIC Gamma Telescope.

Featured Data Set: MAGIC Gamma Telescope



Task: Classification
Data Type: Multivariate
Attributes: 11
Instances: 19020

Newest Data Sets:

10-03-2020:  Codon usage
09-03-2020:  Intelligent Media Accelerometer and Gyroscope (IM-AccGyro) Dataset
07-22-2020:  Facebook Large Page Page Network
07-17-2020:  Amshabians
07-12-2020:  Early stage diabetes risk prediction dataset.
06-28-2020:  Taiwanese Bankruptcy Prediction
06-20-2020:  South German Credit (UPDATE)


Most Popular Data Sets (hits since 2007):

3825761:  Iris

2070996:  Adult

1601121:  Wine

1442822:  Heart Disease

1435469:  Breast Cancer Wisconsin (Diagnostic)

1434425:  Wine Quality

1397755:  Bank Marketing

1326195:  Car Evaluation

Link

Iris Data Set

Download: [Data Folder](#), [Data Set Description](#)

Abstract: Famous database; from Fisher, 1936



Data Set Characteristics:	Multivariate	Number of Instances:	150	Area:	Life
Attribute Characteristics:	Real	Number of Attributes:	4	Date Donated	1988-07-01
Associated Tasks:	Classification	Missing Values?	No	Number of Web Hits:	3825764

Link

Wine Data Set

Download: [Data Folder](#), [Data Set Description](#)

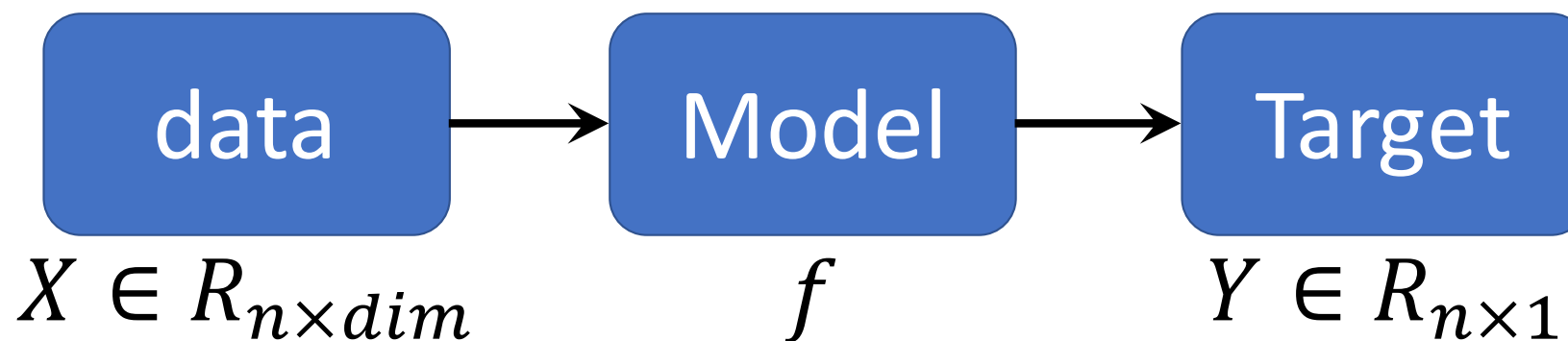
Abstract: Using chemical analysis determine the origin of wines



Data Set Characteristics:	Multivariate	Number of Instances:	178	Area:	Physical
Attribute Characteristics:	Integer, Real	Number of Attributes:	13	Date Donated	1991-07-01
Associated Tasks:	Classification	Missing Values?	No	Number of Web Hits:	1601130

Learning Model

結構資料-learning model



$$Y = f(X)$$



機器學習的資料庫

結構資料

IRIS

1	5.1,3.5,1.4,0.2,Iris-setosa
2	4.9,3.0,1.4,0.2,Iris-setosa
3	4.7,3.2,1.3,0.2,Iris-setosa
4	4.6,3.1,1.5,0.2,Iris-setosa
5	5.0,3.6,1.4,0.2,Iris-setosa
6	5.4,3.9,1.7,0.4,Iris-setosa
7	4.6,3.4,1.4,0.3,Iris-setosa
8	5.0,3.4,1.5,0.2,Iris-setosa
9	4.4,2.9,1.4,0.2,Iris-setosa
10	4.9,3.1,1.5,0.1,Iris-setosa
11	5.4,3.7,1.5,0.2,Iris-setosa
12	4.8,3.4,1.6,0.2,Iris-setosa
13	4.8,3.0,1.4,0.1,Iris-setosa
14	4.3,3.0,1.1,0.1,Iris-setosa
15	5.8,4.0,1.2,0.2,Iris-setosa
16	5.7,4.4,1.5,0.4,Iris-setosa
17	5.4,3.9,1.3,0.4,Iris-setosa
18	5.1,3.5,1.4,0.3,Iris-setosa
19	5.7,3.8,1.7,0.3,Iris-setosa
20	5.1,3.8,1.5,0.3,Iris-setosa
21	5.4,3.4,1.7,0.2,Iris-setosa
22	5.1,3.7,1.5,0.4,Iris-setosa
23	4.6,3.6,1.0,0.2,Iris-setosa

$$X \in R_{n \times dim} = R_{150 \times 4}$$

$$Y \in R_{n \times 1} = R_{150 \times 1}$$

WINE

1	1,14.23,1.71,2.43,15.6,127,2.8,3.06,.28,2.29,5.64,1.04,3.92,1065
2	1,13.2,1.78,2.14,11.2,100,2.65,2.76,.26,1.28,4.38,1.05,3.4,1050
3	1,13.16,2.36,2.67,18.6,101,2.8,3.24,.3,2.81,5.68,1.03,3.17,1185
4	1,14.37,1.95,2.5,16.8,113,3.85,3.49,.24,2.18,7.8,.86,3.45,1480
5	1,13.24,2.59,2.87,21,118,2.8,2.69,.39,1.82,4.32,1.04,2.93,735
6	1,14.2,1.76,2.45,15.2,112,3.27,3.39,.34,1.97,6.75,1.05,2.85,1450
7	1,14.39,1.87,2.45,14.6,96,2.5,2.52,.3,1.98,5.25,1.02,3.58,1290
8	1,14.06,2.15,2.61,17.6,121,2.6,2.51,.31,1.25,5.05,1.06,3.58,1295
9	1,14.83,1.64,2.17,14,97,2.8,2.98,.29,1.98,5.2,1.08,2.85,1045
10	1,13.86,1.35,2.27,16,98,2.98,3.15,.22,1.85,7.22,1.01,3.55,1045
11	1,14.1,2.16,2.3,18,105,2.95,3.32,.22,2.38,5.75,1.25,3.17,1510
12	1,14.12,1.48,2.32,16.8,95,2.2,2.43,.26,1.57,5,1.17,2.82,1280
13	1,13.75,1.73,2.41,16,89,2.6,2.76,.29,1.81,5.6,1.15,2.9,1320
14	1,14.75,1.73,2.39,11.4,91,3.1,3.69,.43,2.81,5.4,1.25,2.73,1150
15	1,14.38,1.87,2.38,12,102,3.3,3.64,.29,2.96,7.5,1.2,3,1547
16	1,13.63,1.81,2.7,17.2,112,2.85,2.91,.3,1.46,7.3,1.28,2.88,1310
17	1,14.3,1.92,2.72,20,120,2.8,3.14,.33,1.97,6.2,1.07,2.65,1280
18	1,13.83,1.57,2.62,20,115,2.95,3.4,.4,1.72,6.6,1.13,2.57,1130
19	1,14.19,1.59,2.48,16.5,108,3.3,3.93,.32,1.86,8.7,1.23,2.82,1680
20	1,13.64,3.1,2.56,15.2,116,2.7,3.03,.17,1.66,5.1,.96,3.36,845
21	1,14.06,1.63,2.28,16,126,3,3.17,.24,2.1,5.65,1.09,3.71,780
22	1,12.93,3.8,2.65,18.6,102,2.41,2.41,.25,1.98,4.5,1.03,3.52,770
23	1,13.71,1.86,2.36,16.6,101,2.61,2.88,.27,1.69,3.8,1.11,4,1035

$$X \in R_{n \times dim} = R_{178 \times 13}$$

$$Y \in R_{n \times 1} = R_{178 \times 1}$$





機器學習的資料庫

- 如果覺得要一個一個去下載和理解太麻煩。
- [Sklearn database](#)

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

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

機器學習的資料庫

實際操作



Pytorch資料庫

- 一般CV或是deep learning課程都以MNIST為課程範例資料庫。
- Pytorch已經提供很好的協助資料庫下載的模組torchvision。
- 範例:
- MNIST
- CIFAR10

TORCHVISION.DATASETS

All datasets are subclasses of `torch.utils.data.Dataset` i.e, they have `__getitem__` and `__len__` methods implemented. Hence, they can all be passed to a `torch.utils.data.DataLoader` which can load multiple samples parallelly using `torch multiprocessing workers`. For example:

```
imagenet_data = torchvision.datasets.ImageNet('path/to/imagenet_root/')
data_loader = torch.utils.data.DataLoader(imagenet_data,
                                          batch_size=4,
                                          shuffle=True,
                                          num_workers=args.nThreads)
```

The following datasets are available:

Datasets

- CelebA
- CIFAR
- Cityscapes
- COCO
 - Captions
 - Detection
- DatasetFolder
- EMNIST
- FakeData
- Fashion-MNIST
- Flickr
- HMDB51
- ImageFolder
- ImageNet
- Kinetics-400
- KMNIST
- LSUN
- MNIST
- Omniglot
- PhotoTour
- Places365
- QMNIST
- SBD
- SBU
- STL10
- SVHN
- UCF101
- USPS
- VOC

MNIST

- MNIST: <http://yann.lecun.com/exdb/mnist/>

THE MNIST DATABASE

of handwritten digits

[Yann LeCun](#), Courant Institute, NYU

[Corinna Cortes](#), Google Labs, New York

[Christopher J.C. Burges](#), Microsoft Research, Redmond

Training set: 60,000 (10,000 images per class)

Testing set: 10,000 (1000 images per class)



Kaggle: <https://www.kaggle.com/c/digit-recognizer>



CIFAR-10

- CIFAR10: <https://www.cs.toronto.edu/~kriz/cifar.html>
- Training set: 50,000 with 10 classes (5000 images per class)
- Testing set: 10,000 (1000 images per class)

[< Back to Alex Krizhevsky's home page](#)

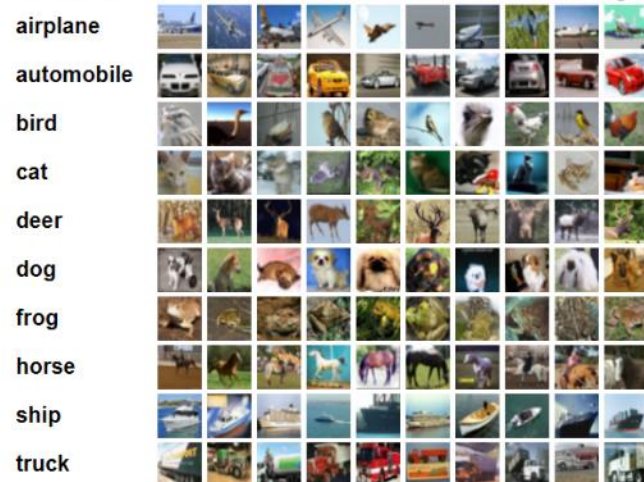
The CIFAR-10 and CIFAR-100 are labeled subsets of the [80 million tiny images](#) dataset. They were collected by Alex Krizhevsky, Vinod Nair, and Geoffrey Hinton.

The CIFAR-10 dataset

The CIFAR-10 dataset consists of 60000 32x32 colour images in 10 classes, with 6000 images per class. There are 50000 training images and 10000 test images.

The dataset is divided into five training batches and one test batch, each with 10000 images. The test batch contains exactly 1000 randomly-selected images from remaining images in random order, but some training batches may contain more images from one class than another. Between them, the training batches contain 6000 images from each class.

Here are the classes in the dataset, as well as 10 random images from each:



CIFAR-100

- **The CIFAR-100 dataset**
- This dataset is just like the CIFAR-10, except it has 100 classes containing 600 images each.
- There are 500 training images and 100 testing images per class. The 100 classes in the CIFAR-100 are grouped into 20 superclasses.
- Each image comes with a "fine" label (the class to which it belongs) and a "coarse" label (the superclass to which it belongs).

Superclass

aquatic mammals
fish
flowers
food containers
fruit and vegetables
household electrical devices
household furniture
insects
large carnivores
large man-made outdoor things
large natural outdoor scenes
large omnivores and herbivores
medium-sized mammals
non-insect invertebrates
people
reptiles
small mammals
trees
vehicles 1
vehicles 2

Classes

beaver, dolphin, otter, seal, whale
aquarium fish, flatfish, ray, shark, trout
orchids, poppies, roses, sunflowers, tulips
bottles, bowls, cans, cups, plates
apples, mushrooms, oranges, pears, sweet peppers
clock, computer keyboard, lamp, telephone, television
bed, chair, couch, table, wardrobe
bee, beetle, butterfly, caterpillar, cockroach
bear, leopard, lion, tiger, wolf
bridge, castle, house, road, skyscraper
cloud, forest, mountain, plain, sea
camel, cattle, chimpanzee, elephant, kangaroo
fox, porcupine, possum, raccoon, skunk
crab, lobster, snail, spider, worm
baby, boy, girl, man, woman
crocodile, dinosaur, lizard, snake, turtle
hamster, mouse, rabbit, shrew, squirrel
maple, oak, palm, pine, willow
bicycle, bus, motorcycle, pickup truck, train
lawn-mower, rocket, streetcar, tank, tractor



Pytorch資料庫

實際操作



其他公開資料庫

Kaggle: <https://www.kaggle.com>

PhysioNet: <https://www.physionet.org/>




私有資料庫

- 私有資料庫如何在pytorch內讀取。
- 我們將用kaggle內Image相關的圖庫進行實際操作
<https://www.kaggle.com/datasets?topic=imageDataset>
- **Car Brands Images:** <https://www.kaggle.com/yamaerenay/100-images-of-top-50-car-brands>



Car Brands Images


- Build a Computer Vision model to predict whether a car image belongs to a luxury or mass-market car brand

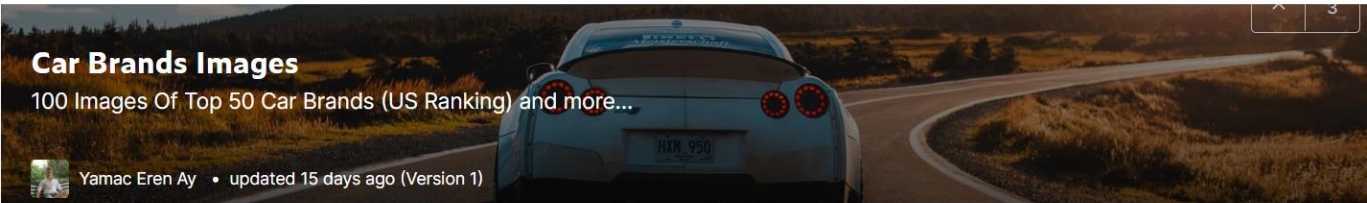


- Home
- Compete
- Data**
- Notebooks
- Communities
- Courses
- More


Recently Viewed

- US Cars Dataset
- Car Image Classificatio...
- Car Brand Images Data...
- Fruits 360
- 99% Acc: Fruits Rec...

 View Active Events





Car Brands Images
100 Images Of Top 50 Car Brands (US Ranking) and more...


 Yamac Eren Ay • updated 15 days ago (Version 1)

[Data](#)
[Tasks \(1\)](#)
[Notebooks \(2\)](#)
[Discussion](#)
[Activity](#)
[Metadata](#)

[Download \(47 MB\)](#)
[New Notebook](#)

 **Usability** 10.0

 **License** Community Data License Agreement - Sharing - Version 1.0

 **Tags** automobiles and vehicles, image data, computer vision, united states

Description

Content

The dataset contains a companies.csv file and a zipped img directory.
companies.csv: General knowledge about the top 50 most popular car brands in US ranking, such as:

- *origin*: The origin of car brand
- *segment*: Whether it is a luxury or a mass-market brand or not
- *logo_link*: Link of the brand logo

imgs:

