

Day 67 Keras Dataset

Keras embedded dataset 的介紹與應用




本日知識點目標



目標
知識點

了解Keras 內建的dataset



獲得
知識點

完成今日課程後你應該可以了解

- 了解Keras內建的dataset
- 如何使用CIFAR10 做類別預測

Keras Dataset – CIFAR10

- **CIFAR10 small image classification**
- Dataset of 50,000 32x32 color training images, labeled over 10 categories, and 10,000 test images.

Usage:

```
from keras.datasets import cifar10  
  
(x_train, y_train), (x_test, y_test) = cifar10.load_data()
```

Keras Dataset – CIFAR100

- **CIFAR100 small image classification**
- Dataset of 50,000 32x32 color training images, labeled over 100 categories, and 10,000 test images.

Usage:

```
from keras.datasets import cifar100  
  
(x_train, y_train), (x_test, y_test) = cifar100.load_data(label_mode='fine')
```

Keras Dataset – MNIST database

- **MNIST database of handwritten digits**
- Dataset of 60,000 28x28 grayscale images of the 10 digits, along with a test set of 10,000 images.

Usage:

```
from keras.datasets import mnist  
(x_train, y_train), (x_test, y_test) = mnsit.load_data()
```


Keras Dataset - Fashion-MNIST

- **Fashion-MNIST database of fashion articles**
- Dataset of 60,000 28x28 grayscale images of 10 fashion categories, along with a test set of 10,000 images. This dataset can be used as a drop-in replacement for MNIST.

Usage:

```
from keras.datasets import fashion_mnsit  
  
(x_train, y_train), (x_test, y_test) = fashion_mnsit.load_data()
```

class labels

| Label | Description |
|-------|-------------|
| 0 | T-shirt/top |
| 1 | Trouser |
| 2 | Pullover |
| 3 | Dress |
| 4 | Coat |
| 5 | Sandal |
| 6 | Shirt |
| 7 | Sneaker |
| 8 | Bag |
| 9 | Ankle boot |

Keras Dataset - Boston housing price

- **Boston housing price regression dataset**
- Dataset taken from the StatLib library which is maintained at Carnegie Mellon University.
- Samples contain 13 attributes of houses at different locations around the Boston suburbs in the late 1970s. Targets are the median values of the houses at a location (in k\$).

Usage:

```
from keras.datasets import boston_housing  
  
(x_train, y_train), (x_test, y_test) = boston_housing.load_data()
```

class labels

| Label | Description |
|-------|-------------|
| 0 | T-shirt/top |
| 1 | Trouser |
| 2 | Pullover |
| 3 | Dress |
| 4 | Coat |
| 5 | Sandal |
| 6 | Shirt |
| 7 | Sneaker |
| 8 | Bag |
| 9 | Ankle boot |

Keras Dataset – IMDB電影評論情緒分類

- 來自IMDB的25,000部電影評論的數據集，標有情緒（正面/負面）。評論已經過預處理，每個評論都被編碼為一系列單詞索引（整數）。
- 單詞由數據集中的整體頻率索引
 - 整數“3”編碼數據中第3個最頻繁的單詞。
 - “0”不代表特定單詞，而是用於編碼任何未知單詞

```
from keras.datasets import imdb
```

```
(x_train, y_train), (x_test, y_test) = imdb.load_data(path="imdb.npz", num_words=  
None, skip_top=0, maxlen=None, seed=113, start_char=1, oov_char=2, index_from=3)
```


Keras Dataset – IMDB電影評論情緒分類

- **path**：如果您沒有本地數據（at '~/.keras/datasets/' + path），它將被下載到此位置。
- **num_words**：整數或無。最常見的詞彙需要考慮。任何不太頻繁的單詞將oov_char在序列數據中顯示為值。
- **skip_top**：整數。最常被忽略的詞（它們將oov_char在序列數據中顯示為值）。
- **maxlen**：int。最大序列長度。任何更長的序列都將被截斷。
- **種子**：int。用於可重複數據改組的種子。
- **start_char**：int。序列的開頭將標有此字符。設置為1，因為0通常是填充字符。
- **oov_char**：int。這是因為切出字num_words 或skip_top限制將這個字符替換。
- **index_from**：int。使用此索引和更高的索引實際單詞。

Keras Dataset –路透社新聞專題主題分類

- 來自路透社的11,228條新聞專線的數據集，標註了46個主題。與IMDB數據集一樣，每條線都被編碼為一系列字索引

```
from keras.datasets import reuters
```

```
(x_train, y_train), (x_test, y_test) = reuters.load_data(path="reuters npz", num_words=
```

```
None, skip_top=0, maxlen=None,
```

```
test_split=0.2, seed=113, start_char=1, oov_char=2, index_from=3)
```

如何使用Keras dataset 做學習

- 適用於文本分析與情緒分類
 - IMDB電影評論情緒分類
 - 路透社新聞專題主題分類
- 適用於Data/Numerical 學習
 - Boston housing price regression dataset

如何使用Keras dataset 做學習

- 適用於影像分類與識別學習
 - CIFAR10/CIFAR100
 - MNIST/ Fashion-MNIST
- 針對小數據集的深度學習
 - 數據預處理與數據提升

前述流程 / python程式 對照

資料準備

```
In [2]: (x_img_train,y_label_train), \
        (x_img_test, y_label_test)=cifar10.load_data()
```

```
In [3]: print('train:',len(x_img_train))
        print('test :',len(x_img_test))
```

```
train: 50000
test : 10000
```

```
In [4]: x_img_train.shape
```

```
Out[4]: (50000, 32, 32, 3)
```

```
In [5]: y_label_train.shape
```

```
Out[5]: (50000, 1)
```


前述流程 / python程式 對照

Image normalize

```
In [13]: x_img_train[0][0][0]
```

```
Out[13]: array([59, 62, 63], dtype=uint8)
```

```
In [14]: x_img_train_normalize = x_img_train.astype('float32') / 255.0  
x_img_test_normalize = x_img_test.astype('float32') / 255.0
```

```
In [15]: x_img_train_normalize[0][0][0]
```

```
Out[15]: array([ 0.23137255,  0.24313726,  0.24705882], dtype=float32)
```

轉換label 為OneHot Encoding

```
In [16]: y_label_train.shape
```

```
Out[16]: (50000, 1)
```

```
In [17]: y_label_train[:5]
```

```
Out[17]: array([[6],  
                [9],  
                [9],  
                [4],  
                [1]], dtype=uint8)
```

```
In [18]: from keras.utils import np_utils  
y_label_train_OneHot = np_utils.to_categorical(y_label_train)  
y_label_test_OneHot = np_utils.to_categorical(y_label_test)
```

```
In [19]: y_label_train_OneHot.shape
```

```
Out[19]: (50000, 10)
```

```
In [20]: y_label_train_OneHot[:5]
```

重要知識點複習

- Keras 內建多種的dataset 提供網路訓練用
 - CIFAR10/100
 - MNIST database of handwritten digits
 - Fashion-MNIST database of fashion articles
 - Boston house price
 - IMDB Movie reviews sentiment classification
 - Reuters newswire topics classification
- 範例以CIFAR10 dataset 為例

解題時間 It's Your Turn

請跳出PDF至官網Sample Code & 作業
開始解題

