Scikit Learn (sklearn)

Sklearn è un modulo del linguaggio python con funzioni di scikit-learn utili per il machine learning e l'apprendimento automatico.

Qual è la differenza tra sklearn e scikit-learn? Il termine sklearn è semplicemente un'abbreviazione di scikit-learn (science kit learning). E' il nome con cui si importa la libreria scikit-learn in python. Ad esempio, per importare le funzioni di scikit-learn in uno script di python scrivo

```
import sklearn
```

Dataset

Un dataset è una raccolta di dati organizzata per un obiettivo, per esempio il training di un modello nel machine learning secondo l'approccio desiderato.

Può essere strutturato, non strutturato, misto, contenendo dati di entrambe le tipologie. I dati strutturati sono dati organizzati in tabelle in cui le colonne rappresentano una variabile, per esempio il numero di lati del poligono, e le righe il suo valore, per esempio 4. I dati non-strutturati sono dati che non hanno questa associazione diretta tra la variabile e il valore, per esempio i testi, l'audio, i video, le immagini.

Come è composto?

x1	x2	x3	x4	x5	x6	x7	x8	x9	x10	y
1	0	0	1	0	0	1	1	0	0	1
1	0	1	1	0	0	1	1	0	1	0
0	1	1	0	1	0	0	0	0	0	0
1	0	0	0	0	1	1	1	1	1	1
1	0	1	1	1	0	1	0	0	1	0
0	0	1	1	0	1	1	0	1	1	1
0	0	1	1	0	0	0	0	0	0	0
1	1	1	0	1	1	1	0	1	0	1

Nei dataset predisposti per il machine learning supervisionato c'è anche un'ulteriore colonna y detta **target** dove è registrata la risposta corretta.

Quali sono i dataset di scikit-learn

Ci sono diversi dataset didattici tra cui scegliere.

load_boston

dataset con i prezzi delle case

load_iris

dataset con le misure di alcuni fiori

load_wine

dataset con le caratteristiche dei vini

Raccolte online

Nello studiare il dominio di un problema si potrebbe cercare su internet un dataset usando le parole chiave più adatte; magari viene fuori un insieme di dati molto più ampio e interessante rispetto ai casi già in possesso. Tra le offerte principali si segnalano:

Raccolte online

UC Irvine Machine Learning

Open Datasets

DataScienceCentral

set di dati pubblici di Google Cloud

strumenti di ricerca Google

selezione delle migliori 50 raccolte

20 dataset più famosi

ampia raccolta su Kaggle

Open Data: dati aperti forniti dalla pubblica amministrazione, accessibili a tutti con l'obbligo di citare la fonte; fonte europea , strumenti specifici di Google , altri siti per dati italiani ;

DATASET IRIS

IRIS

Il dataset Iris è un dataset multivariato introdotto da Ronald Fisher nel 1936. Consiste in 150 istanze di Iris misurate da Edgar Anderson e classificate secondo tre specie: Iris setosa, Iris virginica e Iris versicolor. Le quattro variabili considerate sono la lunghezza e la larghezza del sepalo e del petalo. A causa di errori, esistono diverse versioni del dataset utilizzate nella letteratura scientifica.

Il dataset Iris viene utilizzato nell'ambito dell'apprendimento automatico come esempio di classificazione statistica

Questo set di dati è costituito da 3 diversi tipi di petali e sepali di iris (Setosa, Versicolour e Virginica), memorizzati in un numpy.ndarray 150x4

Le righe sono i campioni e le colonne sono: Lunghezza del sepale, Larghezza del sepalo, Lunghezza del petalo e Larghezza del petalo.

Il grafico seguente utilizza le prime due caratteristiche. Per ulteriori informazioni su questo set di dati, vedere qui.