

1.5. Phân tích dữ liệu học máy với python

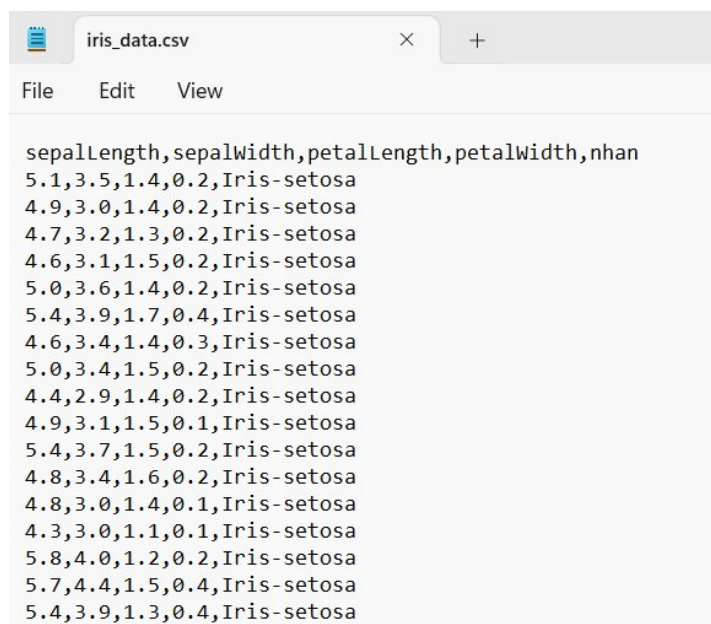
Thường các tập dữ liệu sau khi trích đặc trưng có thể được lưu vào các tập tin có định dạng csv hoặc json tùy vào mục đích lưu trữ của người dùng.

Việc đọc và phân tích tập dữ liệu cũng là việc làm đầu tiên để giải quyết một vấn đề máy học nào đó.

Người ta thường sử dụng thư viện pandas để đọc các tập tin csv hoặc json, dữ liệu trong tập tin sẽ được lưu trữ ở 1 cấu trúc dataframe đã trình bày phía trên.

Ví dụ minh họa sau đây sẽ nói về việc đọc và phân tích 1 tập dữ liệu được lưu trữ ở tập tin có định dạng csv.

Dữ liệu được lưu trữ như hình sau:



```
iris_data.csv
File Edit View
sepalLength,sepalwidth,petalLength,petalwidth,nhan
5.1,3.5,1.4,0.2,Iris-setosa
4.9,3.0,1.4,0.2,Iris-setosa
4.7,3.2,1.3,0.2,Iris-setosa
4.6,3.1,1.5,0.2,Iris-setosa
5.0,3.6,1.4,0.2,Iris-setosa
5.4,3.9,1.7,0.4,Iris-setosa
4.6,3.4,1.4,0.3,Iris-setosa
5.0,3.4,1.5,0.2,Iris-setosa
4.4,2.9,1.4,0.2,Iris-setosa
4.9,3.1,1.5,0.1,Iris-setosa
5.4,3.7,1.5,0.2,Iris-setosa
4.8,3.4,1.6,0.2,Iris-setosa
4.8,3.0,1.4,0.1,Iris-setosa
4.3,3.0,1.1,0.1,Iris-setosa
5.8,4.0,1.2,0.2,Iris-setosa
5.7,4.4,1.5,0.4,Iris-setosa
5.4,3.9,1.3,0.4,Iris-setosa
```

Tiến hành đọc và phân tích tập dữ liệu như sau:

```

import pandas as pd
import numpy as np

df = pd.read_csv("iris_data.csv", sep=',')
print(df)
# Hiện thị thông tin 10 dòng đầu tiên
print(df.head(10))
# Hiện thị thông tin 5 dòng cuối cùng
print(df.tail())
# Hiện thị thông tin về dataset
print(df.info())
# Lấy cột label của dataset
y_data = df.loc[:, "nhân"]
print(y_data)
# Tap các label
labels = np.unique(y_data)
print(labels)
# Lấy dữ liệu của các thuộc tính
x_data = df.iloc[:, 0:4]
print(x_data)
# Lấy dữ liệu của thuộc tính thứ 2
x_data_2 = df.iloc[:, 1]
print(x_data_2)

```

Ví dụ này chỉ minh họa cho việc phân tích dữ liệu học ở mức cơ bản nhằm mục đích cho người học tiếp cận được với dữ liệu, biết cách đọc dữ liệu, phân tích và có thể trích xuất được thông tin từ dữ liệu theo ý muốn. Cho nên, tùy theo mục đích sử dụng, chúng ta sẽ có nhiều cách phân tích và xử lý dữ liệu khác nhau.

Bài tập: Hãy tạo thư mục tên “THB1_Phan1_3”, sau đó tạo tập tin tên là BaiTap_PhanTichData.py để giải quyết yêu cầu sau:

Hãy tải về tập tin dữ liệu theo đường link sau: [\[Liên kết\]](#)

Tiến hành phân tích tập dữ liệu này bằng thư viện pandas.

Một số gợi ý phân tích:

- Đọc dữ liệu và xuất ra thông tin dữ liệu này.
- Hiện thị n dòng đầu tiên cũng như cuối cùng của tập dữ liệu

- Đếm số hàng, số cột
- Lấy được tên các thuộc tính, tên nhãn
- Lấy dữ liệu cột thuộc tính, dữ liệu cột nhãn
- Lấy thông tin thống kê dữ liệu ở mỗi cột thuộc tính và cả tập dataset (sử dụng hàm `describe()`)