Chủ đề: Chuẩn bị dữ liệu

Bài: Các miêu tả thống kê và suy luận

> Làm việc trên tập dữ liệu: adult.data

1.1 Mục tiêu:

- Làm quen với miêu tả Thống kê bao gồm
  - các khái niệm, thuật ngữ, độ đo và công cụ giúp mô tả, hiển thị và tóm tắt dữ liệu một cách có ý nghĩa.
  - Ví dụ, chẳng hạn như giá thuê căn hộ mỗi năm, có thể sử dụng cả mô tả thống kê và suy luận để phân tích kết quả và rút ra một số kết luận.
  - giá trị trung bình, giá trị trung vị, phương sai, tương quan, v.v., để khai thác, mô tả và tóm tắt tập dữ liệu đã cho.

1.1 Giới thiệu

Thống kê dựa trên 2 khái niệm chính:

1. Quần thể là một chủ đề của các đối tượng, các mục (”đơn vị”) về thông tin được tìm kiếm.
2. Một mẫu là một phần của quần thể được quan sát.

- Miêu tả Thống kê.

- 1.1 Lấy dữ liệu
- 1.2 Chuẩn bị dữ liệu
- 1.3 Cải thiện dữ liệu khi là pandas DataFrame
- 1.4 Làm sạch dữ liệu

1.1 Lấy dữ liệu

Tập dữ liệu Adult từ UCI (https://archive.ics.uci.edu/ml/datasets/Adult)

1.2 Chuẩn bị dữ liệu

1. Lấy dữ liệu: Đọc dữ liệu từ tệp hoặc trang web
2. Phân tích cú pháp dữ liệu: theo định dạng của văn bản
3. Làm sạch dữ liệu: Cho dữ liệu không đầy đủ, có sai sót.
4. Xây dựng cấu trúc dữ liệu: lưu trữ dữ liệu theo một cấu trúc

> Chú ý: Đây là tệp văn bản text, không phải là csv, sử dụng hàm mở tệp của python **open()**

- Mở tệp:

```
#file = open('duong dan tới tệp')
file = open('datasets/aduly.data')
```

- Chuyển tệp sang mảng

- Viết hàm kiểm tra kiểm tra một chuỗi nếu là chuỗi thì chuyển thành số

```
def char_int(a):
    if a.isdigit(): return int(a)
    else: return 0
```

- Đọc tệp

```
file = open('datasets/aduly.data')
data =[]
# đọc từng dòng của tệp
for line in file:
    # lấy dữ liệu theo dòng
    data1 = line.split(',')
    # độ dài của
    if len(data1) == 15:
        data.append([char_int(data1[0]), data1[1],
                     char_int(data1[2]), data1[3],
                     char_int(data1[4]), data1[5], data1[6],
                     data1[7],data1[8],data1[9],
                     chr_int(data1[10]), chr_int(data1[11]),
                     chr_int(data1[12]), data1[13], data1[14]
                     ])
```

- In xem dong dau tien

```
print(data[1:2])
```

> Kết quả: [[50, 'Self-emp-not-inc', 83311, 'Bachelors', 13, 'Married-civ-spouse', 'Exec-managerial', 'Husband', 'White', 'Male', 0, 0, 13, 'United-States', '<=50K\n']]

Xuất dữ liệu sang pandas DataFrame

```
df = pd.DataFrame(data)
```

> Chú ý: Tập dữ liệu chưa có nhãn

- Thêm nhãn cho dữ liệu sử dụng thuộc tính columns của DataFrame

```
df.columns = ['age', 'type_employer', 'fnlwgt', 'education',
                "education_num","marital", "occupation", "relationship", "race","sex",
                "capital_gain", "capital_loss", "hr_per_week","country","income"]
# xem dữ liêu
df.head()
df.tail()
df.shape
```

- Nhóm dữ liệu theo quốc gia và lấy kích cỡ (sử dụng groupby() và size())

```
count_countries = df.groupby('country').size()
print(count_countries)
```

- Nhóm dữ liệu theo tuổi và lấy kích cỡ (sử dụng groupby() và size())

```
count_age = df.groupby('age').size()
print(count_age)
```

- Nhóm dữ liệu theo giới tính và lấy kích cỡ, xem dữ liệu

```
male = df[(df.sex == 'Male')]
male
male.shape
```

- Nhóm và xem dữ liệu có thu nhập lớn hơn '50k'

```
df1 = df[(df.income=='>50k'\n)]
```

- Tính phần trăm thu nhập của người có thu nhập cao hơn 50k

```
print('Người có thu nhập cao hơn là:',
int(len(df1)/float(len(df))*100), '%.')
```

Làm sạch dữ liệu

- Dữ liệu trống
- Dữ liệu chứa **NaN**
- Dữ liệu chứa **?**
- Xem dữ liệu

Phân tích, khai thác dữ liệu

Khái quát về dữ liệu

Trung bình mẫu

- Nếu có một mẫu giá trị, x_i , trung bình mẫu là tổng các giá trị chia cho số giá trị:

$\mu = \frac{1}{n}$ $\sum_i {x_i}$

- Cách sử trong python: sử dụng phương thác mean()
- In ra tuổi trung bình của giới tính **nam** trong dữ liệu:

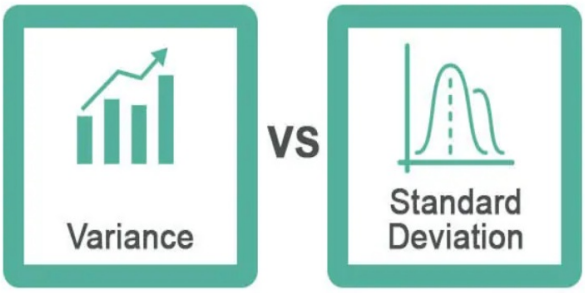
```
# in ra bảng có giới tính nam
male = df[(df.sex == 'Male')]

print('Tuổi trung bình giới tính nam là:', )
male['age'].mean()
```

- In ra tuổi trung bình của giới tính **nữ** trong dữ liệu:

Phương sai

$\sigma^2 = \frac{1}{n} \sum_i (x_i - \mu)^2$



- Sử dụng tính phương sai trong python

```
male['age'].var()
```

Trung vị mẫu

- Đưa ra mẫu ở giữa đã được sắp xếp
- Sử dụng tính trung vị trong python

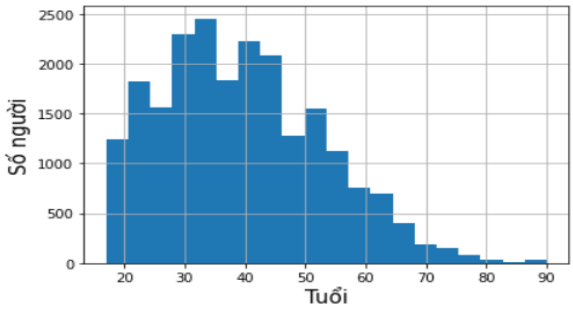
```
male['age'].median()
```

Lược đồ dữ liệu

- Vẽ lược đồ tuổi giới tính nam

```
import matplotlib.pyplot as plt
male_age = male['age']

male_age.hist(density=Fale, histtype = 'stepfilled', bins=20)
# gán nhãn các trục của lược đồ
plt.xlabel('Tuổi', fontsize=15)
plt.ylabel('Số người', fontsize=15)
#hiển thị lược đồ
plt.show()
```



- Vẽ lược đồ tuổi giới tính nữ

```
import matplotlib.pyplot as plt
Female_age = Female['age']

male_age.hist(density=Fale, histtype = 'stepfilled', bins=20)
# gán nhãn các trục của lược đồ
plt.xlabel('Tuổi', fontsize=15)
plt.ylabel('Số người', fontsize=15)
#hiển thị lược đồ
plt.show()
```