

<div style="display: flex; flex-direction: column; align-items: center;"> <div style="display: flex; justify-content: space-between; width: 100%;"> <span>A</span><span>D</span></div> <div style="display: flex; justify-content: space-between; width: 100%;"> <span>A</span><span>I</span></div> </div>	<b>PHÒNG THÍ NGHIỆM THỰC HÀNH ỨNG DỤNG</b> <b>KHOA HỌC DỮ LIỆU VÀ TRÍ TUỆ NHÂN TẠO</b> Applied Data Science and Artificial Intelligence Lab
--	---



## LẬP TRÌNH PYTHON CHO KHOA HỌC DỮ LIỆU

### PYTHON FOR DATA SCIENCE

Hà Nội ★ 11.2020

Nothing is faster than Innovation together

1

<div style="display: flex; flex-direction: column; align-items: center;"> <div style="display: flex; justify-content: space-between; width: 100%;"> <span>A</span><span>D</span></div> <div style="display: flex; justify-content: space-between; width: 100%;"> <span>A</span><span>I</span></div> </div>	<h2>Thông tin giảng viên</h2>	
--	-------------------------------	--



**TS. TẠ QUANG CHIẾU**

**Vị trí & Đào tạo**

- Trưởng bộ môn Hệ thống thông tin và Tri thức, Khoa Công nghệ Thông tin
- PhD & Post-doc in Polytech Tours, France

**Lĩnh vực giảng dạy, nghiên cứu**

- Big data, Data science, machine learning, AI
- Các phương pháp cho bài toán lập kế hoạch

**Contact:**

(E): [quangchieu.ta@gmail.com](mailto:quangchieu.ta@gmail.com)

(M): 0913 522 275

Dr. Tạ Quang Chiếu – (E): [quangchieu.ta@gmail.com](mailto:quangchieu.ta@gmail.com) – (M): 0913 522 275

2/39

2

<div> <div>A</div> <div>D</div> <div>A</div> <div>I</div> </div>	<h2>Objectives</h2>
<ul style="list-style-type: none"> <li>• Hiểu được tầm quan trọng của Khoa học dữ liệu</li> <li>• Vận dụng được các bước trong quy trình thực hiện một dự án về khoa học dữ liệu.</li> <li>• Kiến thức, kỹ năng cần thiết để trở thành một nhà khoa học dữ liệu</li> <li>• Áp dụng được Python và các thư viện phổ biến trong giải quyết một số bài toán cơ bản của Khoa học dữ liệu.</li> </ul>	
<div> <div>Dr. Tạ Quang Chiếu – [E]: <a href="mailto:quangchieu.ta@gmail.com">quangchieu.ta@gmail.com</a> – [M]: 0913 522 275</div> <div>3/39</div> </div>	

3

<div> <div>A</div> <div>D</div> <div>A</div> <div>I</div> </div>	<h2>References</h2>
<ul style="list-style-type: none"> <li>[1]. Introducing Data Science</li> <li>[2]. Python Data Science Handbook</li> <li>[3]. Data Science from Scratch</li> <li>[4]. Python crash course</li> <li>[5]. Bài giảng Khoa học dữ liệu</li> </ul>	
<div> <div>Dr. Tạ Quang Chiếu – [E]: <a href="mailto:quangchieu.ta@gmail.com">quangchieu.ta@gmail.com</a> – [M]: 0913 522 275</div> <div>4/39</div> </div>	

4

A A	D I	<b>PHẦN 2: NHẬP MÔN KHOA HỌC DỮ LIỆU</b>	
		FUNDAMENTALS OF DATA SCIENCE	
<p><b>CONTENT</b></p> <p><b>1</b> CHAPTER 1: INTRODUCTION</p> <p><b>2</b> CHAPTER 2: PYTHON LIBRARIES FOR DATA SCIENCE</p>			
Dr. Tà Quang Chiếu - [E]: <a href="mailto:quangchieu.ta@gmail.com">quangchieu.ta@gmail.com</a> - [M]: 0913 522 275			5/39

5

A A	D I	<b>Content of chapter 2</b>	
		<p>2.1 Python Library for Data Science</p> <p>2.2 Numpy library</p> <p>2.3 Pandas library</p> <p>2.4 Matplotlib library</p> <p><b>2.5 Scikit-learn library</b></p>	
Dr. Tà Quang Chiếu - [E]: <a href="mailto:quangchieu.ta@gmail.com">quangchieu.ta@gmail.com</a> - [M]: 0913 522 275			6/39

6

<div><div>A</div><div>D</div><div>A</div><div>I</div></div>	<h2>NỘI DUNG</h2> <ol style="list-style-type: none"><li>1. Mối quan hệ giữa Khoa học Dữ liệu và Học máy</li><li>2. Một số loại bài toán học máy</li><li>3. Thư viện học máy scikit-learn</li><li>4. Bài tập</li></ol> <p>Dr. Tạ Quang Chiếu – [E]: <a href="mailto:quangchieu.ta@gmail.com">quangchieu.ta@gmail.com</a> – [M]: 0913 522 275</p>
	7/39

7

<div><div>A</div><div>D</div><div>A</div><div>I</div></div>	<h2>PHẦN 1</h2> <h1>Mối quan hệ giữa Khoa học Dữ liệu và Học máy</h1> <p>Dr. Tạ Quang Chiếu – [E]: <a href="mailto:quangchieu.ta@gmail.com">quangchieu.ta@gmail.com</a> – [M]: 0913 522 275</p>
	8/39

8

<div style="display: flex; justify-content: space-between;"> <span>A</span> <span>D</span> </div> <div style="display: flex; justify-content: space-between;"> <span>A</span> <span>I</span> </div>	<b>PHẦN 1: MỐI QUAN HỆ GIỮ KHDL VÀ HỌC MÁY</b>
<p><b>KHDL LÀ GÌ?</b></p> <ul style="list-style-type: none"> <li>▪ Hầu hết các ngành khoa học từ xưa đến nay đều giải quyết vấn đề dựa trên <b>lập luận</b> và <b>tri thức</b> <ul style="list-style-type: none"> <li>▪ Ngành toán: dựa trên các mệnh đề, công thức, lập luận... để chứng minh bài toán</li> <li>▪ Ngành vật lý: dựa trên các quan sát, thực nghiệm, tính toán,... kiểm chứng các giả thiết</li> <li>▪ Ngành hóa học:...</li> <li>▪ ...</li> <li>▪ Ta gọi các ngành khoa học này là “knowledge-driven” (dẫn dắt bởi tri thức)</li> </ul> </li> <li>▪ Có ngành có chút ngoại lệ, ví dụ: <b>ngành xác suất</b></li> </ul>	
<div style="display: flex; justify-content: space-between;"> <span>Dr. Tạ Quang Chiếu – [E]: <a href="mailto:quangchieu.ta@gmail.com">quangchieu.ta@gmail.com</a> – [M]: 0913 522 275</span> <span style="background-color: #000080; color: white; padding: 2px 5px;">9/39</span> </div>	

9

<div style="display: flex; justify-content: space-between;"> <span>A</span> <span>D</span> </div> <div style="display: flex; justify-content: space-between;"> <span>A</span> <span>I</span> </div>	<b>PHẦN 1: MỐI QUAN HỆ GIỮ KHDL VÀ HỌC MÁY</b>
<p><b>KHDL LÀ GÌ?</b></p> <ul style="list-style-type: none"> <li>▪ Với quan điểm như vậy, tất cả những quan sát mà không được chứng minh chặt chẽ thường được cho là “không khoa học” <ul style="list-style-type: none"> <li>▪ Chẳng hạn: chuồn chuồn bay thấp thì mưa</li> </ul> </li> <li>▪ Khoa học dữ liệu ≠ Khoa học thông thường ở quan điểm: <b>tìm tri thức từ dữ liệu</b> (dẫn dắt bởi dữ liệu – “data-driven”) <ul style="list-style-type: none"> <li>▪ Chúng ta rút ra tri thức bằng việc tìm tòi từ dữ liệu chứ không nhất thiết phải chứng minh nó</li> <li>▪ Tất nhiên tri thức tìm ra phải có tính ổn định (luôn có cùng kết quả nếu sử dụng cùng một phương pháp)</li> </ul> </li> </ul>	
<div style="display: flex; justify-content: space-between;"> <span>Dr. Tạ Quang Chiếu – [E]: <a href="mailto:quangchieu.ta@gmail.com">quangchieu.ta@gmail.com</a> – [M]: 0913 522 275</span> <span style="background-color: #000080; color: white; padding: 2px 5px;">10/39</span> </div>	

10

A

D

A

I

PHẦN 1: MỐI QUAN HỆ GIỮ KHDL VÀ HỌC MÁY

- Không có sơ đồ nào minh họa đầy đủ mối quan hệ giữa hai khái niệm này
- Nhiều người (chẳng hạn như Nate Silver) cho rằng ngành khoa học dữ liệu chỉ là một dạng thống kê



Dr. Tạ Quang Chiếu – [E]: [quangchieu.ta@gmail.com](mailto:quangchieu.ta@gmail.com) – [M]: 0913 522 275

11/39

11

A

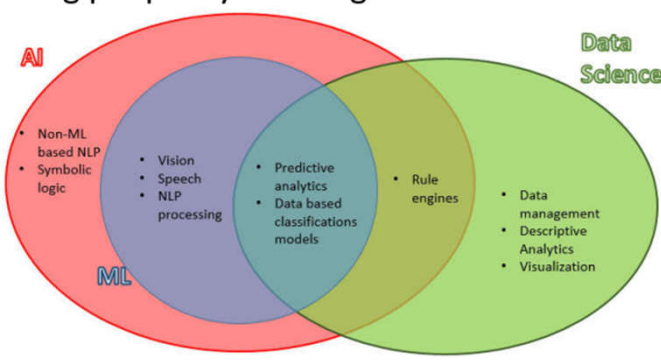
D

A

I

PHẦN 1: MỐI QUAN HỆ GIỮ KHDL VÀ HỌC MÁY

- Học máy là phương pháp quan trọng để xử lý dữ liệu trong ngành data science, bên cạnh những phương pháp truyền thống khác



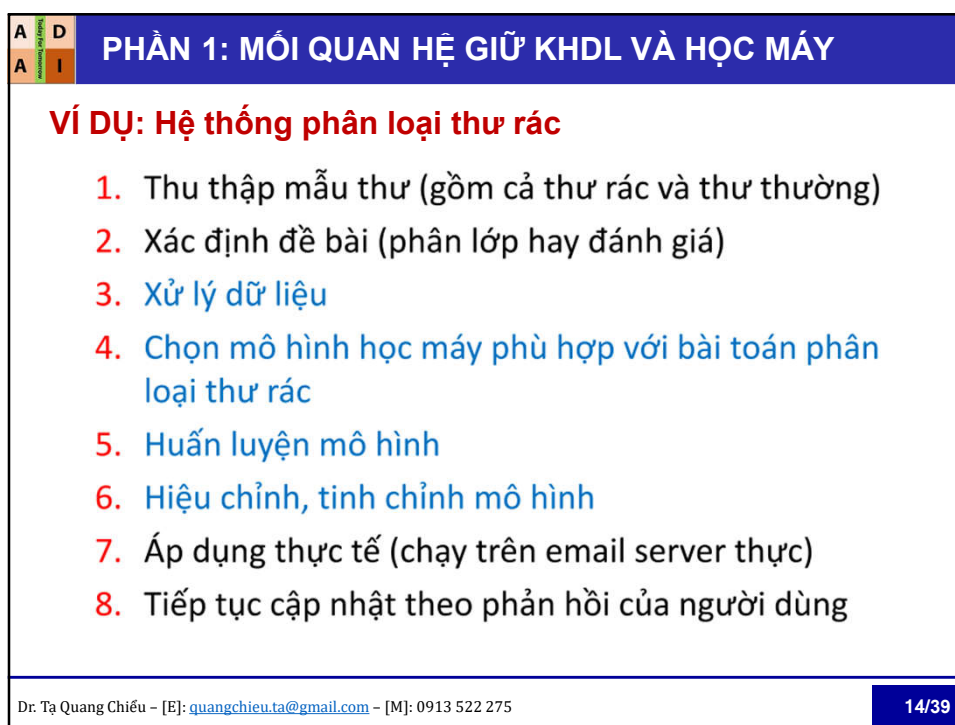
Dr. Tạ Quang Chiếu – [E]: [quangchieu.ta@gmail.com](mailto:quangchieu.ta@gmail.com) – [M]: 0913 522 275

12/39

12



13



14

<div> <div>A</div> <div>D</div> </div> <div> <div>A</div> <div>I</div> </div>	<b>PHẦN 2</b>
<div style="text-align: center; color: red; font-weight: bold; font-size: 1.2em;">MỘT SỐ LOẠI BÀI TOÁN HỌC MÁY</div>	
<div> <div>Dr. Tạ Quang Chiếu – [E]: <a href="mailto:quangchieu.ta@gmail.com">quangchieu.ta@gmail.com</a> – [M]: 0913 522 275</div> <div style="float: right; background-color: #000080; color: white; padding: 2px 5px;">15/39</div> </div>	

15

<div> <div>A</div> <div>D</div> </div> <div> <div>A</div> <div>I</div> </div>	<b>PHẦN 2: MỘT SỐ LOẠI BÀI TOÁN HỌC MÁY</b>
<div style="color: red; font-weight: bold; font-size: 1.1em;">MỘT SỐ BÀI TOÁN THỰC TẾ</div> <ul style="list-style-type: none"> <li>▪ Hệ thống phân loại email</li> <li>▪ Nhận dạng chữ viết từ ảnh</li> <li>▪ Ước lượng giá cả của sản phẩm</li> <li>▪ Dự báo thời tiết</li> <li>▪ Đánh giá trạng thái của người qua ảnh/video</li> <li>▪ Trả lời tự động (chat bot)</li> <li>▪ Gợi ý sản phẩm phù hợp với nhu cầu khách hàng</li> <li>▪ Tự động chơi trò chơi</li> <li>▪ Mô phỏng giọng nói của một người nào đó</li> </ul>	
<div> <div>Dr. Tạ Quang Chiếu – [E]: <a href="mailto:quangchieu.ta@gmail.com">quangchieu.ta@gmail.com</a> – [M]: 0913 522 275</div> <div style="float: right; background-color: #000080; color: white; padding: 2px 5px;">16/39</div> </div>	

16



A	D	PHẦN 2: MỘT SỐ LOẠI BÀI TOÁN HỌC MÁY
A	I	
<p style="color: red; text-align: center;"><b>CÁC LỚP BÀI TOÁN CƠ BẢN</b></p> <ul style="list-style-type: none"> <li>▪ <b>Học có giám sát (supervised learning):</b> học cách tiên đoán đầu ra theo mẫu cho trước           <ul style="list-style-type: none"> <li>▪ Tập mẫu cho trước, cho cả đầu bài và kết quả               <ul style="list-style-type: none"> <li>• Cho email, chỉ rõ trước đâu là spam, đâu không phải spam</li> </ul> </li> <li>▪ Mô hình được huấn luyện trên tập mẫu</li> <li>▪ Thử nghiệm bằng cách cho đầu bài, mô hình tiên đoán kết quả, mô hình đoán càng chính xác càng tốt               <ul style="list-style-type: none"> <li>• Cho một email mới, máy tính đoán xem có phải spam không?</li> </ul> </li> <li>▪ Có 2 loại cơ bản:               <ul style="list-style-type: none"> <li>• Hồi quy (regression): đầu ra là số hoặc vector</li> <li>• Phân lớp (classification): đầu ra thường là xác suất dự báo</li> </ul> </li> </ul> </li> </ul>		
Dr. Tạ Quang Chiếu – [E]: <a href="mailto:quangchieu.ta@gmail.com">quangchieu.ta@gmail.com</a> – [M]: 0913 522 275		17/39

17

A	D	PHẦN 2: MỘT SỐ LOẠI BÀI TOÁN HỌC MÁY
A	I	
<p style="color: red; text-align: center;"><b>CÁC LỚP BÀI TOÁN CƠ BẢN</b></p> <ul style="list-style-type: none"> <li>▪ <b>Học không giám sát (unsupervised learning):</b> tự khai phá các đặc trưng nội tại <b>hợp lý</b> của đầu vào           <ul style="list-style-type: none"> <li>▪ Chỉ cho mẫu vào, không cho biết đầu ra               <ul style="list-style-type: none"> <li>• Cho tập băng ghi âm lời nói của một người</li> </ul> </li> <li>▪ Hệ thống tự học trên các mẫu mà không có định hướng               <ul style="list-style-type: none"> <li>• Tạo ra một đoạn phát âm theo ngữ điệu của người đã cho</li> </ul> </li> <li>▪ Một vài chiến lược cơ bản:               <ul style="list-style-type: none"> <li>• Biến đổi dữ liệu đầu vào có số chiều cao thành dữ liệu có số chiều thấp hơn</li> <li>• Dữ liệu có số chiều cao nhưng các đặc trưng thành phần có tính “kinh tế” (economical) hơn</li> <li>• Gom cụm dữ liệu đầu vào</li> </ul> </li> </ul> </li> </ul>		
Dr. Tạ Quang Chiếu – [E]: <a href="mailto:quangchieu.ta@gmail.com">quangchieu.ta@gmail.com</a> – [M]: 0913 522 275		18/39

18

A

D

A

I

PHẦN 2: MỘT SỐ LOẠI BÀI TOÁN HỌC MÁY

### CÁC LỚP BÀI TOÁN CƠ BẢN

- **Học tăng cường (reinforcement learning):** hiệu chỉnh các siêu tham số (hyperparameter) để cực đại hóa lợi ích trong tương lai
  - Cho bối cảnh và các quy tắc
    - Bàn cờ Vây và các quy tắc của trò chơi cờ Vây
  - Ứng với mỗi hành động (hoặc chuỗi hành động), có một phần thưởng tương ứng
    - Đặt một quân sẽ bị mất điểm, không được hoặc được điểm
  - Hệ thống tự điều chỉnh chuỗi hành động sao cho được phần thưởng lớn nhất
    - Hệ thống học cách chơi để thắng người chơi giỏi nhất

Dr. Tạ Quang Chiếu – [E]: [quangchieu.ta@gmail.com](mailto:quangchieu.ta@gmail.com) – [M]: 0913 522 275

19/39

19

A

D

A

I

PHẦN 2: MỘT SỐ LOẠI BÀI TOÁN HỌC MÁY

### CÁC LỚP BÀI TOÁN CƠ BẢN

MACHINE LEARNING

Dr. Tạ Quang Chiếu – [E]: [quangchieu.ta@gmail.com](mailto:quangchieu.ta@gmail.com) – [M]: 0913 522 275

20/39

20

A

D

A

I

**PHẦN 3: THƯ VIỆN HỌC MÁY SCIKIT-LEARN**

- Scikit-learn xuất phát là một dự án trong một cuộc thi lập trình của Google vào năm 2007, người khởi xướng dự án là David Cournapeau
- Sau đó nhiều viện nghiên cứu và các nhóm ra nhập, đến năm 2010 mới có bản đầu tiên (v0.1 beta)
- Scikit-learn cung cấp gần như tất cả các loại thuật toán học máy cơ bản (khoảng vài chục) và vài trăm biến thể của chúng, cùng với đó là các kỹ thuật xử lý dữ liệu đã được chuẩn hóa
- Cài đặt: **pip install scikit-learn scipy**

Dr. Tạ Quang Chiếu – [E]: [quangchieu.ta@gmail.com](mailto:quangchieu.ta@gmail.com) – [M]: 0913 522 275

21/39

21

A

D

A

I

**PHẦN 3: THƯ VIỆN HỌC MÁY SCIKIT-LEARN**

scikit-learn  
algorithm cheat-sheet

Dr. Tạ Quang Chiếu – [E]: [quangchieu.ta@gmail.com](mailto:quangchieu.ta@gmail.com) – [M]: 0913 522 275

22/39

22

A

D

A

I

**PHẦN 3: THƯ VIỆN HỌC MÁY SCIKIT-LEARN**

**Ví dụ: Dự báo cân nặng của người**

- Tập mẫu quan sát có n người
  - Gồm tên, chiều cao, cân nặng
  - Và nhiều loại chỉ số khác nữa
- Xây dựng một mô hình dự báo về cân nặng người, dựa trên các chỉ số còn lại
  - Trong trường hợp bài toán của ta, chúng ta cố gắng dự báo cân nặng từ chiều cao
  - Thực tế thì cân nặng phụ thuộc vào nhiều thông số khác nữa, như giới tính, vòng eo,...

	A	B	C
1	Tên	Cao	Nang
2	A	147	49
3	B	150	50
4	C	153	51
5	D	155	51
6	E	168	60
7	F	170	62
8	G	173	68
9	H	175	65
10	I	178	66
11	J	180	71
12	K	183	68
13	L	165	59
14	M	163	58
15	N	160	56
16	O	158	54
17	P	169	62
18	Q	172	63
19	S	170	62
20	T	176	62
21	U	180	69

Dr. Tạ Quang Chiếu – [E]: [quangchieu.ta@gmail.com](mailto:quangchieu.ta@gmail.com) – [M]: 0913 522 275

23/39

23

A

D

A

I

**PHẦN 3: THƯ VIỆN HỌC MÁY SCIKIT-LEARN**

**Dự báo hồi quy tuyến tính**

```
import matplotlib.pyplot as plt
import pandas as pd
import numpy as np
from sklearn import linear_model, metrics

# đọc dữ liệu từ file csv
df = pd.read_csv("nguoi.csv", index_col = 0)
print(df)

# vẽ biểu đồ minh họa dataset
plt.plot(df.Cao, df.Nang, 'ro')
plt.xlabel('Chiều cao (cm)')
plt.ylabel('Cân nặng (kg)')
plt.show()
```

Dr. Tạ Quang Chiếu – [E]: [quangchieu.ta@gmail.com](mailto:quangchieu.ta@gmail.com) – [M]: 0913 522 275

24/39

24

A

D

A

I

PHẦN 3: THƯ VIỆN HỌC MÁY SCIKIT-LEARN

### Dự báo hồi quy tuyến tính

```

# sử dụng hồi quy tuyến tính
X = df.loc[:, ['Cao']].values
y = df.Nang.values
model = linear_model.LinearRegression()
model.fit(X, y)

# X là dữ liệu đầu vào
# y là dữ liệu đầu ra
# loại mô hình
# tập huấn trên dữ liệu

# in một số thông tin về mô hình
mse = metrics.mean_squared_error(model.predict(X), y)
print("Tổng bình phương sai số trên tập mẫu:", mse)
print("Hệ số hồi quy:", model.coef_)
print("Sai số:", model.intercept_)
print(f"Công thức: [Nặng] = {model.coef_} x [Cao] + {model.intercept_}")

```

Dr. Tạ Quang Chiếu - [E]: [quangchieu.ta@gmail.com](mailto:quangchieu.ta@gmail.com) - [M]: 0913 522 275

25/39

25

A

D

A

I

PHẦN 3: THƯ VIỆN HỌC MÁY SCIKIT-LEARN

### Dự báo hồi quy tuyến tính

```

# vẽ lại sơ đồ
plt.scatter(X, y, c='b')
plt.plot(X, model.predict(X))
plt.show()

# dự báo một số tình huống
while True:
    x = float(input("Nhập chiều cao (nhập 0 để dừng): "))
    if x <= 0: break
    print("Người cao", x, "cm, dự báo cân nặng",
          model.predict([[x]]))

```

Dr. Tạ Quang Chiếu - [E]: [quangchieu.ta@gmail.com](mailto:quangchieu.ta@gmail.com) - [M]: 0913 522 275

26/39

26

A

D

PHẦN 3: THƯ VIỆN HỌC MÁY SCIKIT-LEARN

**Mở rộng: Thêm cột giới tính**

- Vẫn dữ liệu cũ, bổ sung thêm cột giới tính (Nam/Nu)
- Sử dụng phương pháp cũ, để xem giới tính ảnh hưởng như thế nào đến cân nặng

	A	B	C	D
1	Ten	Gioitinh	Cao	Nang
2	A	Nu	147	49
3	B	Nu	150	50
4	C	Nu	153	51
5	D	Nam	155	51
6	E	Nu	168	60
7	F	Nam	170	62
8	G	Nu	173	68
9	H	Nam	175	65
10	I	Nam	178	66
11	J	Nam	180	71
12	K	Nam	183	68
13	L	Nam	165	59
14	M	Nu	163	58
15	N	Nu	160	56
16	O	Nu	158	54
17	P	Nam	169	62
18	Q	Nam	172	63
19	S	Nu	170	62
20	T	Nam	176	62
21	U	Nam	180	69

Dr. Tạ Quang Chiếu – [E]: [quangchieu.ta@gmail.com](mailto:quangchieu.ta@gmail.com) – [M]: 0913 522 275

27/39

27

A

D

PHẦN 3: THƯ VIỆN HỌC MÁY SCIKIT-LEARN

**Dự báo hồi quy tuyến tính**

```
import matplotlib.pyplot as plt
import pandas as pd
import numpy as np
from sklearn import linear_model, metrics

# đọc dữ liệu từ file csv
df = pd.read_csv("nguoii2.csv", index_col = 0)
print(df)

# thêm cột mới, giới tính Nam = 1, giới tính Nữ = 0
df['GT'] = df.Gioitinh.apply(lambda x: 1 if x=='Nam' else 0)
print(df)
```

Dr. Tạ Quang Chiếu – [E]: [quangchieu.ta@gmail.com](mailto:quangchieu.ta@gmail.com) – [M]: 0913 522 275

28/39

28



A

D

A

I

PHẦN 3: THƯ VIỆN HỌC MÁY SCIKIT-LEARN

### Dự báo hồi quy tuyến tính

```

# sử dụng hồi quy tuyến tính
X = df.loc[:, ['Cao', 'GT']].values      # X là dữ liệu đầu vào
y = df.Nang.values                      # y là dữ liệu đầu ra
model = linear_model.LinearRegression()  # loại mô hình
model.fit(X, y)                         # tập huấn trên dữ liệu

# in một số thông tin về mô hình
mse = metrics.mean_squared_error(model.predict(X), y)
print("Tổng bình phương sai số trên tập mẫu:", mse)
print("Hệ số hồi quy:", model.coef_)
print("Sai số:", model.intercept_)
print(f"Công thức: [Nặng] = {model.coef_} x [Cao, Giới tính] + {model.intercept_}")

```

Dr. Tạ Quang Chiếu – [E]: [quangchieu.ta@gmail.com](mailto:quangchieu.ta@gmail.com) – [M]: 0913 522 275

29/39

29

A

D

A

I

PHẦN 3: THƯ VIỆN HỌC MÁY SCIKIT-LEARN

### Dự báo hồi quy tuyến tính

```

# dự báo một số tình huống
while True:
    x = float(input("Nhập chiều cao (nhập 0 để dừng): "))
    if x <= 0: break
    print("Nam giới cao", x, "cm, dự báo cân nặng",
          model.predict([[x, 1]]))
    print("Nữ giới cao", x, "cm, dự báo cân nặng",
          model.predict([[x, 0]]))

```

Dr. Tạ Quang Chiếu – [E]: [quangchieu.ta@gmail.com](mailto:quangchieu.ta@gmail.com) – [M]: 0913 522 275

30/39

30

A

D

A

I

**PHẦN 4: BÀI TẬP**

**1.** Tải về file **winequality.csv** về các số đo của rượu vang và chất lượng của rượu

- Liên kết: <http://archive.ics.uci.edu/ml/machine-learning-databases/wine-quality/winequality-red.csv>
- Đây là bộ data của đại học California-Berkeley
- Bộ data gồm 1599 mẫu rượu vang, mỗi mẫu gồm 11 loại chỉ số và đánh giá của chuyên gia về chất lượng rượu (cột quality, điểm số từ 0 đến 10)
- Chú ý:
  - Dữ liệu sử dụng dấu chấm phẩy (;) để ngăn giữa các cột
  - Tên các cột có chứa dấu cách

Dr. Tạ Quang Chiếu – [E]: [quangchieu.ta@gmail.com](mailto:quangchieu.ta@gmail.com) – [M]: 0913 522 275

31/39

31

A

D

A

I

**PHẦN 4: BÀI TẬP**

**2.** In ra dữ liệu vừa tải về, ý nghĩa các cột thuộc tính

▪ fixed acidity	Nồng độ axit tartaric
▪ volatile acidity	Tính axit
▪ citric acid	Nồng độ axit Citric
▪ residual sugar	Nồng độ đường dư
▪ chlorides	Nồng độ clo
▪ free sulfur dioxide	Nồng độ acid sulfurus tự do
▪ total sulfur dioxide	Nồng độ acid sulfurus
▪ density	Mật độ (khối lượng/đơn vị thể tích)
▪ pH	Độ pH
▪ sulphates	Nồng độ sunfat
▪ alcohol	Nồng độ chất alcohol

Dr. Tạ Quang Chiếu – [E]: [quangchieu.ta@gmail.com](mailto:quangchieu.ta@gmail.com) – [M]: 0913 522 275

32/39

32



A D		PHẦN 4: BÀI TẬP
A	I	
<ol style="list-style-type: none"> <li>3. Sử dụng các cột thuộc tính “alcohol” để tìm tương quan giữa thuộc tính này và điểm chất lượng rượu</li> <li>4. Sử dụng tất cả 11 thuộc tính để tìm tương quan giữa các thuộc tính với điểm chất lượng rượu</li> <li>5. (*) Loại bỏ những thuộc tính “không quan trọng”, chọn 3 thuộc tính quan trọng nhất và xây dựng tương quan tuyến tính giữa 3 thuộc tính đó với điểm chất lượng rượu</li> </ol>		
Dr. Tà Quang Chiếu – [E]: <a href="mailto:quangchieu.ta@gmail.com">quangchieu.ta@gmail.com</a> – [M]: 0913 522 275		33/39

33

A D		
A	I	
Dr. Tà Quang Chiếu – [E]: <a href="mailto:quangchieu.ta@gmail.com">quangchieu.ta@gmail.com</a> – [M]: 0913 522 275		34/39

34

A

D

A

I

Dr. Tạ Quang Chiếu - [E]: [quangchieu.ta@gmail.com](mailto:quangchieu.ta@gmail.com) - [M]: 0913 522 27535/39

35

A

D

A

I

Dr. Tạ Quang Chiếu - [E]: [quangchieu.ta@gmail.com](mailto:quangchieu.ta@gmail.com) - [M]: 0913 522 27536/39

36

A

D

A

I

Dr. Tạ Quang Chiếu - [E]: [quangchieu.ta@gmail.com](mailto:quangchieu.ta@gmail.com) - [M]: 0913 522 275

37/39

37

A

D

A

I

Dr. Tạ Quang Chiếu - [E]: [quangchieu.ta@gmail.com](mailto:quangchieu.ta@gmail.com) - [M]: 0913 522 275

38/39

38



Dr. Tạ Quang Chiếu - [E]: [quangchieu.ta@gmail.com](mailto:quangchieu.ta@gmail.com) - [M]: 0913 522 275

39/39