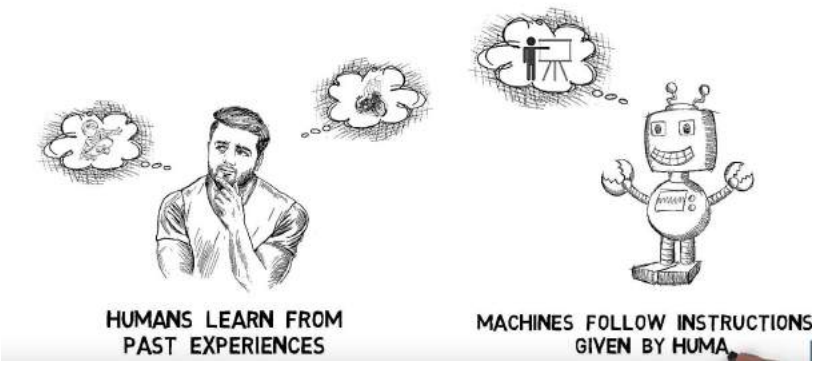
KHOA CNTT & TRUYỀN THÔNG
BM KHOA HỌC MÁY TÍNH

CÁC PHƯƠNG PHÁP HỌC TỪ DỮ LIỆU
Machine Learning

Giáo viên giảng dạy:
TS. TRẦN NGUYỄN MINH THƯ
tnmthu@cit.ctu.edu.vn

1

Máy học là gì?



2

Máy học là gì?

3

3

Máy học là gì?

Chương trình truyền thống

Nguyên lý máy học

4

4

Máy học là gì?

Máy học là chương trình máy tính cho phép **học tự động** từ **dữ liệu** để nhận dạng các mẫu phức tạp, tạo ra hành vi ứng xử thông minh với trường hợp mới đến [T. Mitchell, 1997]

T. Mitchell: machine Learning: improving performance via **experience**

Học= Cải thiện tác vụ (task) nào đó bằng kinh nghiệm

5

5

Máy học là gì?

T. Mitchell:
– machine learning: improving performance via experience
– Formally, A computer program is said to learn from experience E with respect to some class of tasks T and performance measure P, it its performance at tasks in T as measured by P, improves with experience.

(Mitchell, 1997): một chương trình máy tính được gọi là học từ kinh nghiệm E với một vài lớp của vấn đề T và độ đo hiệu quả P, nếu hiệu năng của vấn đề trong T, đánh giá theo tiêu chí P, được cải thiện từ kinh nghiệm E

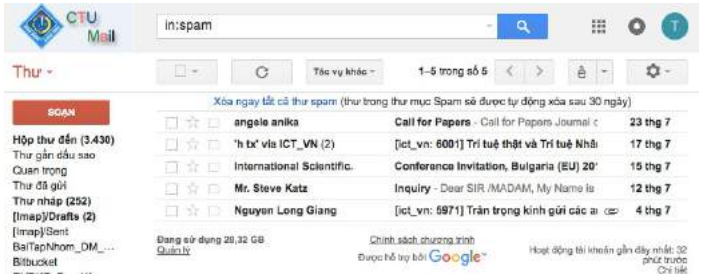
– Cải thiện tác vụ T,
– Với độ đo hiệu quả P
– Dựa trên kinh nghiệm E

6

6

Máy học là gì?

- Ví dụ: "Chương trình lọc email rác"



– Cải thiện tác vụ T (Task) ?
– Với độ đo hiệu quả P (Performance)?

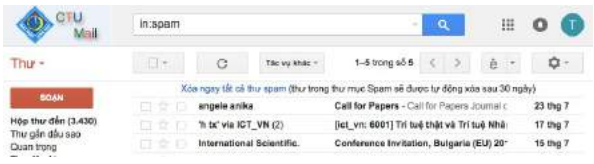
– Dựa trên kinh nghiệm E (Experience) ?

7

7

Máy học là gì?

- Ví dụ: "Chương trình lọc email rác"

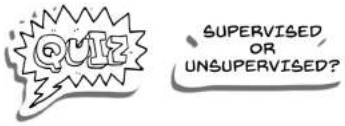


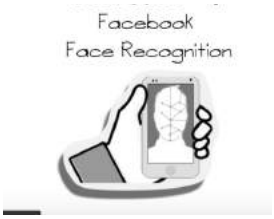
– Cải thiện tác vụ T (Task) ? => lọc được email rác hay không rác
– Với độ đo hiệu quả P (Performance)? => % email được gán nhãn/ xác định đúng là rác hoặc không rác.
– Dựa trên kinh nghiệm E (Experience) ? Kiểm tra trong danh sách email đã có email nào được gán nhãn là rác và email là không phải rác

8

8

Phân loại học máy dựa trên phương thức học





https://www.youtube.com/watch?v=ukzFI9rgwfU

9

9

Phân loại học máy

1. **Học không có giám sát – unsupervised learning**
2. Học có giám sát – supervised learning
3. Học bán giám sát- semi- supervised learning
4. Học củng cố / học tăng cường: reinforcement learning

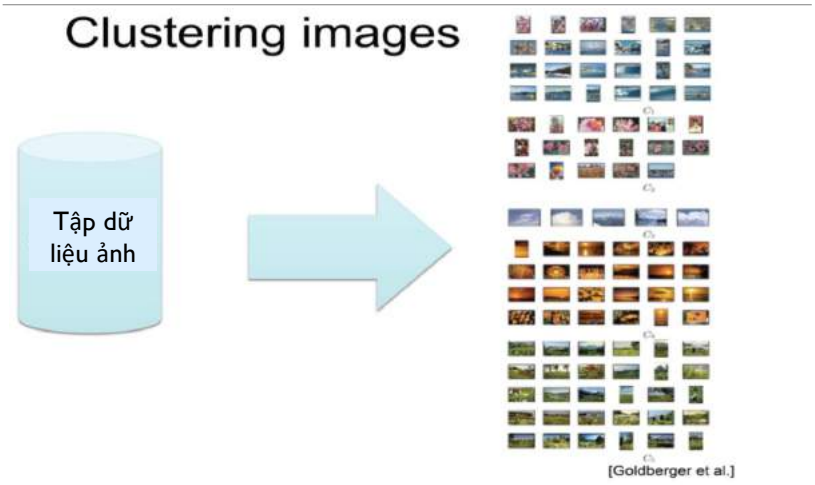
Phân loại học máy – **học không giám sát**

- Học không giám sát là thuật toán học thực hiện mô hình hoá một tập dữ liệu đầu vào**, không được gán nhãn** (lớp, giá trị cần dự báo)

  gom cụm, nhóm (clustering, unsupervised learning): xây dựng mô hình gom cụm dữ liệu tập học (không có nhãn) sao cho các dữ liệu cùng nhóm có các tính chất tương tự nhau và dữ liệu của 2 nhóm khác nhau sẽ có các tính chất khác nhau
- Xây dựng **mô hình H** từ ***tập dữ liệu*** $(X^1,X^2,..,X^m)$

sepal length (cm)	sepal width (cm)	petal length (cm)	petal width (cm)
5.8	2.8	5.1	2.4
6.3	3.4	5.6	2.4
7.2	3.2	6.0	1.8

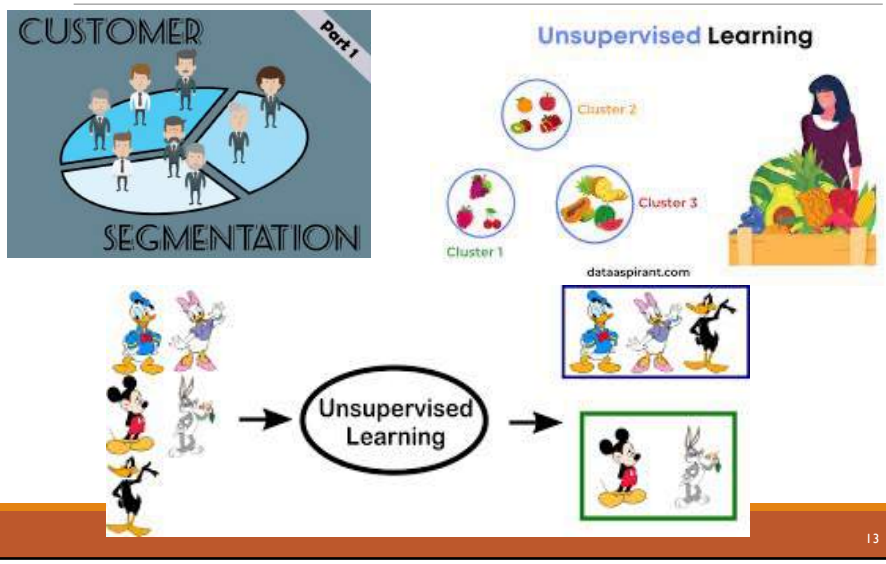
Phân loại học máy – **học không giám sát**



12

12

Phân loại học máy – **học không giám sát**

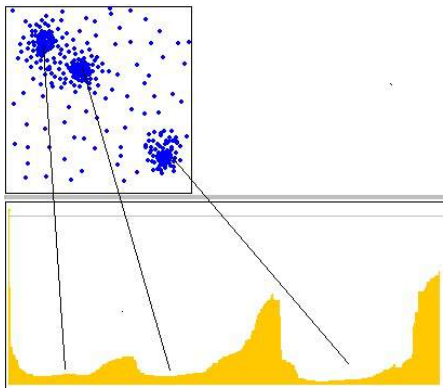


13

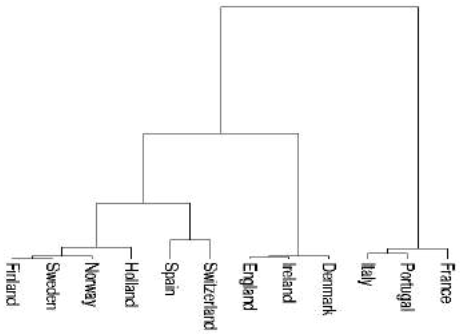
13

Phân loại học máy – **học không giám sát**

Kmeans



Hierachical Clustering



14

14

Phân loại học máy

1. Học không có giám sát – unsupervised learning
2. Học có giám sát – supervised learning
3. Học bán giám sát- semi- supervised learning
4. Học củng cố / học tăng cường: reinforcement learning

15

15

Phân loại học máy – **học có giám sát**

- Học có giám sát là thuật toán học tạo ra một hàm ánh xạ dữ liệu đầu vào tới kết quả đích mong muốn (nhãn, lớp, giá trị cần dự báo). Trong học có giám sát, tập dữ liệu dùng để huấn luyện phải **được gán nhãn, lớp hay giá trị cần dự báo**
- Xây dựng mô hình H được huấn luyện từ tập dữ liệu {($\mathbf{X}^1$, $\mathbf{y}^1$), ($\mathbf{X}^2$,y^2),..,($\mathbf{X}^n$,y^n)
  - **Bài toán hồi quy** (regression): y là giá trị liên tục
  - **Bài toán phân lớp**: y là giá trị **không** liên tục

16

16

Phân loại học máy – **học có giám sát**

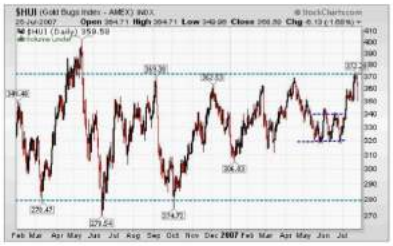
Classification Example: Weather Prediction

17

17

Phân loại học máy – **học có giám sát**

Regression example: Predicting Gold/Stock prices



Good ML can make you rich (but there is still some risk involved).

Given historical data on Gold prices, predict tomorrow's price!

18

18

Phân loại học máy – **học có giám sát**

Tập dữ liệu học/ huấn luyện $\{(x^1, y^1), (x^2,y^2),..,(x^m,y^m)\}$

Day	Outlook	Temperature	Humidity	Wind	PlayTennis
D1	Sunny	Hot	High	Weak	No
D2	Sunny	Hot	High	Strong	No
D3	Overcast	Hot	High	Weak	Yes
D4	Rain	Mild	High	Weak	Yes
D5	Rain	Cool	Normal	Weak	Yes
D6	Rain	Cool	Normal	Strong	No
D7	Overcast	Cool	Normal	Strong	Yes
D8	Sunny	Mild	High	Weak	No
D9	Sunny	Cool	Normal	Weak	Yes
D10	Rain	Mild	Normal	Weak	Yes
D11	Sunny	Mild	Normal	Strong	Yes
D12	Overcast	Mild	High	Strong	Yes
D13	Overcast	Hot	Normal	Weak	Yes
D14	Rain	Mild	High	Strong	No

19

19

Phân loại học máy – **học có giám sát**

- **phân lớp** (classification, supervised learning) : xây dựng mô hình phân loại dựa trên dữ liệu tập học đã có **nhãn** (lớp) **là kiểu liệt kê**

  VD: có sẵn tập dữ liệu thư điện tử, mỗi thư có 1 nhãn là thư rác hay thư bình thường, mục tiêu là xây dựng mô hình phân lớp tập dữ liệu thư điện tử thành thư rác hay thư bình thường để khi có một thư điện tử mới đến thì mô hình dự báo được thư này có phải là thư rác hay không

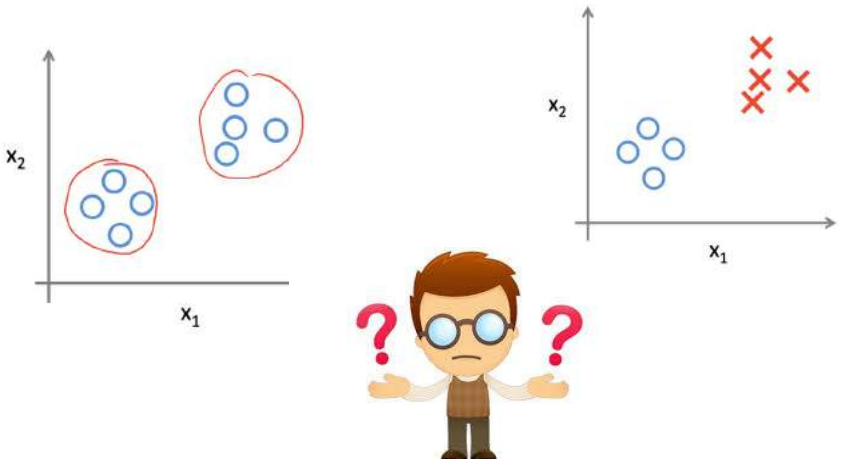
- **hồi quy (regression)** : xây dựng mô hình phân loại dựa trên dữ liệu tập học đã có nhãn (lớp) là **giá trị liên tục**.

VD. Xd mô hình dự báo mực nước sông Mekong từ các yếu tố thời tiết, mùa,…

20

20

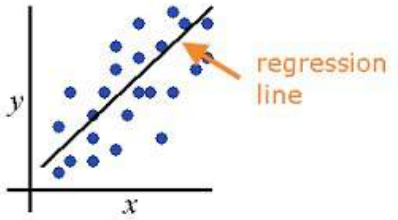
Bài toán phân lớp

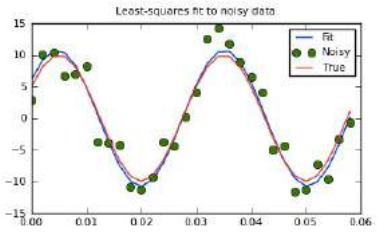


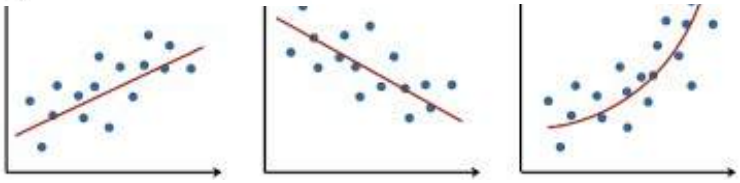
21

21

Bài toán hồi quy: Regression







22

22

Từ tập dữ liệu học/huấn luyện { $(x^1, y^1), (x^2,y^2),..,(x^m,y^m)$ }

Day	Outlook	Temperature	Humidity	Wind	PlayTennis
D1	Sunny	Hot	High	Weak	No
D2	Sunny	Hot	High	Strong	No
D3	Overcast	Hot	High	Weak	Yes
D4	Rain	Mild	High	Weak	Yes
D5	Rain	Cool	Normal	Weak	Yes
D6	Rain	Cool	Normal	Strong	No
D7	Overcast	Cool	Normal	Strong	Yes
D8	Sunny	Mild	High	Weak	No
D9	Sunny	Cool	Normal	Weak	Yes
D10	Rain	Mild	Normal	Weak	Yes
D11	Sunny	Mild	Normal	Strong	Yes
D12	Overcast	Mild	High	Strong	Yes
D13	Overcast	Hot	Normal	Weak	Yes
D14	Rain	Mild	High	Strong	No

[See: Tom M. Mitchell, *Machine Learning*, McGraw-Hill, 1997]

Chỉ ra thuộc tính? Nhãn/lớp của tập dữ liệu thời tiết trong bảng trên

23

Day	Outlook	Temp.	Humidity	Wind	Golf Players
1	Sunny	Hot	High	Weak	25
2	Sunny	Hot	High	Strong	30
3	Overcast	Hot	High	Weak	46
4	Rain	Mild	High	Weak	45
5	Rain	Cool	Normal	Weak	52
6	Rain	Cool	Normal	Strong	23
7	Overcast	Cool	Normal	Strong	43
8	Sunny	Mild	High	Weak	35
9	Sunny	Cool	Normal	Weak	38
10	Rain	Mild	Normal	Weak	46
11	Sunny	Mild	Normal	Strong	48
12	Overcast	Mild	High	Strong	52
13	Overcast	Hot	Normal	Weak	44
14	Rain	Mild	High	Strong	30

24

b. Dựa vào thông tin: số ca nhiễm, số người tử vong, số lượt di chuyển của người dân trong thành phố, dân số của thành phố, người ta cần xác định mức độ lây nhiễm của dịch Covid-19 theo 3 mức: nguy cơ thấp, nguy cơ, nguy cơ cao. Theo anh/chị, chúng ta nên sử dụng giải thuật gì (clustering/classification/regression) để xác định mức độ lây nhiễm của dịch Covid-19? Anh/chị hãy giải thích cho lựa chọn của mình?

25

25

Bài tập 1

Cho dataset như sau:

Example No.	Color	Type	Origin	Stolen?
1	Red	Sports	Domestic	Yes
2	Red	Sports	Domestic	No
3	Red	Sports	Domestic	Yes
4	Yellow	Sports	Domestic	No
5	Yellow	Sports	Imported	Yes
6	Yellow	SUV	Imported	No
7	Yellow	SUV	Imported	Yes
8	Yellow	SUV	Domestic	No
9	Red	SUV	Imported	No
10	Red	Sports	Imported	Yes

Cần dự đoán tình trạng xe dựa trên tập dữ liệu đã cho. Bài toán trên thuộc dạng nào?

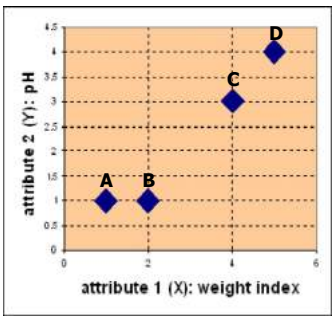
26

Bài tập 3

Cho 4 loại thuốc mỗi loại có 2 thuộc tính pH và Weight

Yêu cầu nhóm những loại thuốc này thành 2 nhóm sử dụng khoảng cách Euclidea với 2 điểm khởi tạo là A và B

Medicine	Weight	pH-Index
A	1	1
B	2	1
C	4	3
D	5	4



27

Bài tập 4

Cho tập dữ liệu hình bên, cần dự đoán giá trị y cho phần tử có x1 =2, x2=4, x3=0.7. Bài toán trên thuộc dạng nào?

x1	x2	x3	y
3	4	0	1
4	6	0.5	2.5
3.5	5	1.5	3.5
2	7	2.5	4
6	3	3	5.5

28

INTERESTING MACHINE LEARNING MODEL WHICH IS USED BY

UBER

DIFFERENTIAL PRICING IN REAL TIME BASED ON:

- DEMAND
- NUMBER OF CARS AVAILABLE
- BAD WEATHER
- RUSH HOUR

https://www.youtube.com/watch?v=ukzFI9rgwfU

29

29

However, for many problems, labeled data can be rare or expensive.

Need to pay someone to do it, requires special testing,…

Unlabeled data is much cheaper.

Can we make use of cheap unlabeled data?

30

Phân loại học máy – **học có giám sát**

Học bán giám sát: Học bán giám sát đối với trường hợp dữ liệu thu thập được có một phần nhỏ đã được gán nhãn, phần lớn còn lại chưa được gán nhãn trong quá trình học.

Semi-supervised learning (or classification)

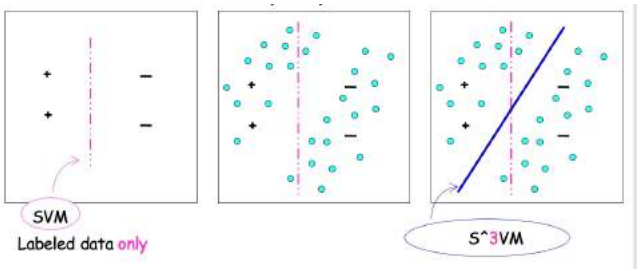
- LU learning: Learning with a small set of Labeled examples and a large set of Unlabeled examples
- PU learning: Learning with Positive and Unlabeled examples (no labeled negative examples).

31

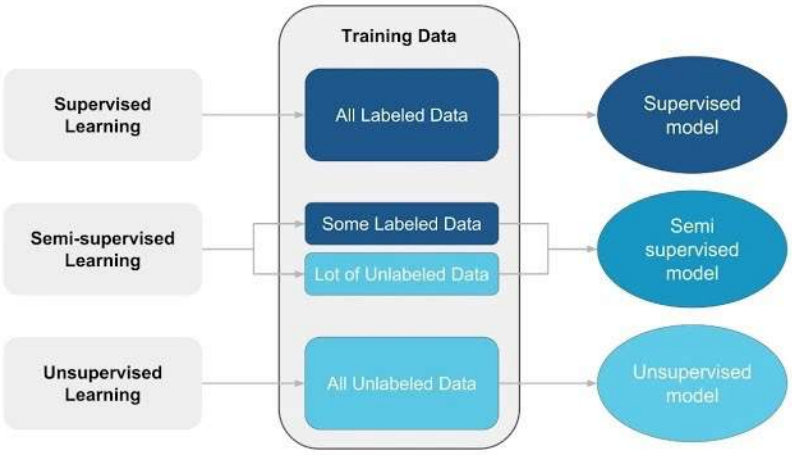
31

Phân loại học máy – **học có giám sát**

Học bán giám sát đối với trường hợp dữ liệu thu thập được có một phần nhỏ đã được gán nhãn, phần lớn còn lại chưa được gán nhãn trong quá trình học.



32



33

Phân loại học máy – **học tăng cường**

Học củng cố / học tăng cường: reinforcement learning:

- là một cách tiếp cận tập trung vào việc học để hoàn thành được mục tiêu bằng việc tương tác trực tiếp với môi trường.
- Đây là các bài toán giúp cho một hệ thống tự động xác định hành động dựa vào môi trường cụ thể để đạt được hiệu quả cao nhất.
- Bản chất của học tăng cường là trial-and-error, nghĩa là thử đi thử lại và rút ra kinh nghiệm sau mỗi lần thử như vậy. Đây là một nhánh học khá hấp dẫn trong máy học.

34

34

Phương pháp học tăng cường

- AlphaGo gần đây nổi tiếng với việc chơi cờ vây thắng cả con người.
- Cờ vây được xem là có độ phức tạp cực kỳ cao với tổng số nước đi là xấp xỉ 1076110761, so với cờ vua là 1012010120
- Vì vậy, thuật toán phải chọn ra 1 nước đi tối ưu trong số hàng nhiều tỉ tỉ lựa chọn, và tất nhiên, không thể áp dụng thuật toán tương tự như IBM Deep Blue (IBM Deep Blue đã thắng con người trong môn cờ vua 20 năm trước).
- AlphaGo bao gồm các thuật toán thuộc cả Supervised learning và Reinforcement learning.
- Supervised learning, dữ liệu từ các ván cờ do con người chơi với nhau được đưa vào để huấn luyện.
- sau khi *học* xong các ván cờ của con người, AlphaGo tự chơi với chính nó với hàng triệu ván chơi để tìm ra các nước đi mới tối ưu hơn. Thuật toán trong phần tự chơi này được xếp vào loại Reinforcement learning

35

Phương pháp học tăng cường

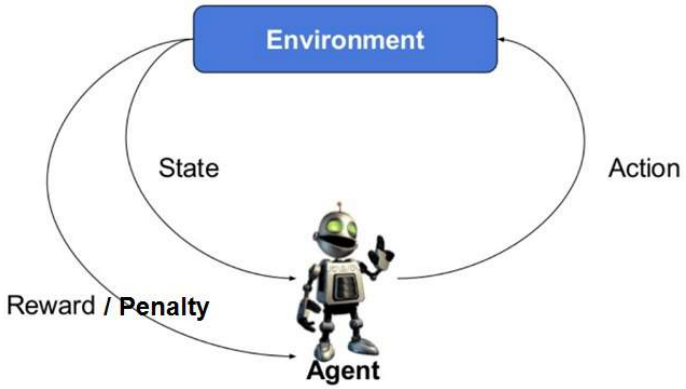
Huấn luyện cho máy tính chơi game Mario.

Ứng dụng học tăng cường sâu trong xây dựng hệ thống chơi trò chơi tự động

- tăng cường sâu (deep reinforcement learning) - Q-Learning kép (Hasselt et al., 2015) để "dạy" cho máy tính chơi trò chơi.
- Bài toán dạy/huấn luyện cho máy tính chơi trò chơi được mô hình hoá về bài toán dự báo điểm số của mỗi nước đi.
- Dữ liệu huấn luyện có thể được thu thập tự động khi người chơi thật chơi trò chơi kết hợp với việc sinh dữ liệu theo quy luật của trò chơi.
- Mô hình huấn luyện sử dụng 2 mạng nơ-ron: *model* và *target_model* để huấn luyện bộ dự báo. Quá trình huấn luyện cập nhật các trọn

36

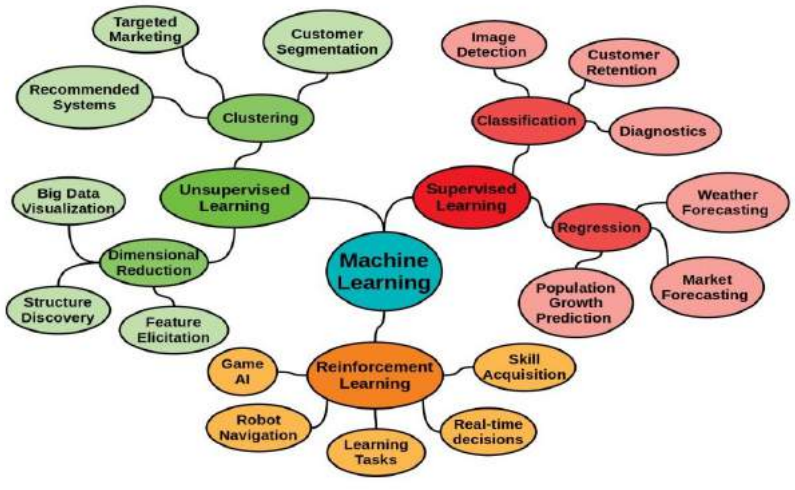
Typical RL scenario



37

Criteria	Supervised Learning	Unsupervised Learning	Reinforcement Learning
Definition	The machine learns by using labeled data	The machine is trained on unlabeled data without any guidance	An agent interacts with its environment by performing actions & learning from errors or rewards
Type of problems	Regression & classification	Association & clustering	Reward-based
Type of data	Labeled data	Unlabeled data	No predefined data
Training	External supervision	No supervision	No supervision
Approach	Maps the labeled inputs to the known outputs	Understands patterns & discovers the output	Follows the trial-and-error method

38



39

Bài tập 2

Một tập dữ liệu dùng chẩn đoán bệnh tim được cho như sau.
Bài toán trên thuộc dạng nào?

PATIENT ID	CHEST PAIN?	MALE?	SMOKES?	EXERCISES?	HEART ATTACK?
1.	yes	yes	no	yes	yes
2.	yes	yes	yes	no	yes
3.	no	no	yes	no	yes
4.	no	yes	no	yes	no
5.	yes	no	yes	yes	yes
6.	no	yes	yes	yes	no

40

Bài 5

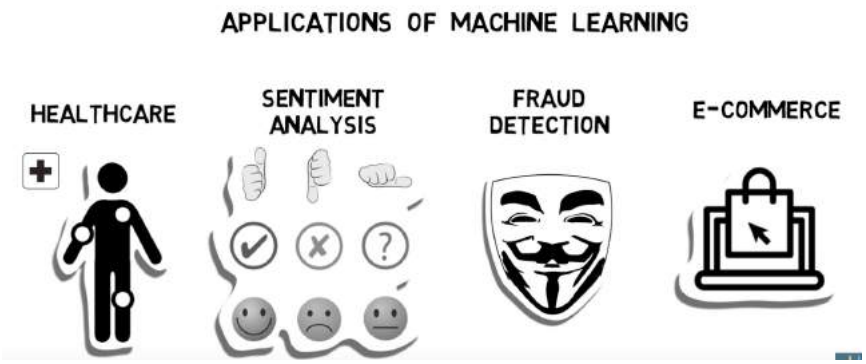
Một tập dữ liệu được cho trong bảng sau (từ A đến H) dùng để phân loại nấm ăn được hay không (giá trị Edible lần lượt là 1 hoặc 0) dựa vào các thuộc tính sau: NotHeavy, Smelly, Spotted, Smooth

Cần xây dựng mô hình dự đoán xem một loại nấm có ăn được hay ko?
Bài toán trên thuộc dạng nào?

Example	*NotHeavy*	*Smelly*	*Spotted*	*Smooth*	*Edible*
A	1	0	0	0	1
B	1	0	1	0	1
C	0	1	0	1	1
D	0	0	0	1	0
E	1	1	1	0	0
F	1	0	1	1	0
G	1	0	0	1	0
H	0	1	0	0	0
U	0	1	1	1	?
V	1	1	0	1	?
W	1	1	0	0	?

41

Ứng dụng của Nguyên lý Máy học



https://www.kdnuggets.com/2018/11/10-free-must-see-courses-machine-learning-data-science.html 42

42

Resources: Datasets

- UCI Repository: http://www.ics.uci.edu/~mlearn/MLRepository.html
- UCI KDD Archive: http://kdd.ics.uci.edu/summary.data.application.html
- Statlib: http://lib.stat.cmu.edu/
- Delve: http://www.cs.utoronto.ca/~delve/
- Kaggle: https://www.kaggle.com/datasets

43

43

Resources: Journals

Journal of Machine Learning Research www.jmlr.org

Machine Learning

IEEE Transactions on Neural Networks

IEEE Transactions on Pattern Analysis and Machine Intelligence

Annals of Statistics

Journal of the American Statistical Association

...

44

45