

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 THỦ
tnmthu@cit.ctu.edu.vn

1

Nguyên lý máy học là gì?



MACHINES FOLLOW INSTRUCTIONS
GIVEN BY HUMANS

2

Nguyên lý máy học là gì?



3

3

Nguyên lý máy học là gì?

Chương trình truyền thống



Nguyên lý máy học



4

4

Nguyên lý 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

Nguyên lý máy học là gì?

Customer purchase behavior:

<i>Customer103: (time=t0)</i>	<i>Customer103: (time=t1)</i>	...	<i>Customer103: (time=tn)</i>
Sex: M	Sex: M		Sex: M
Age: 53	Age: 53		Age: 53
Income: \$50k	Income: \$50k		Income: \$50k
Own House: Yes	Own House: Yes		Own House: Yes
MS Products: Word	MS Products: Word		MS Products: Word
Computer: 386 PC	Computer: Pentium		Computer: Pentium
Purchase Excel?: ?	Purchase Excel?: ?		Purchase Excel?: Yes

Customer retention:

<i>Customer103: (time=t0)</i>	<i>Customer103: (time=t1)</i>	...	<i>Customer103: (time=tn)</i>
Sex: M	Sex: M		Sex: M
Age: 53	Age: 53		Age: 53
Income: \$50k	Income: \$50k		Income: \$50k
Own House: Yes	Own House: Yes		Own House: Yes
Checking: \$5k	Checking: \$20k		Checking: \$0
Savings: \$15k	Savings: \$0		Savings: \$0
Current-customer?: yes	Current-customer?: yes		Current-customer?: No

6

6

Nguyên lý 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, if 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

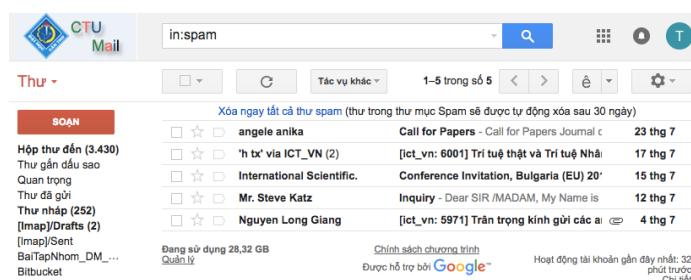
- Cải thiện tác vụ T,
- Với độ đo hiệu quả P
- Dựa trên kinh nghiệm E

7

7

Nguyên lý 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) ?

8

8

Nguyên lý máy học là gì?

- Ví dụ: “Chương trình lọc email rác”

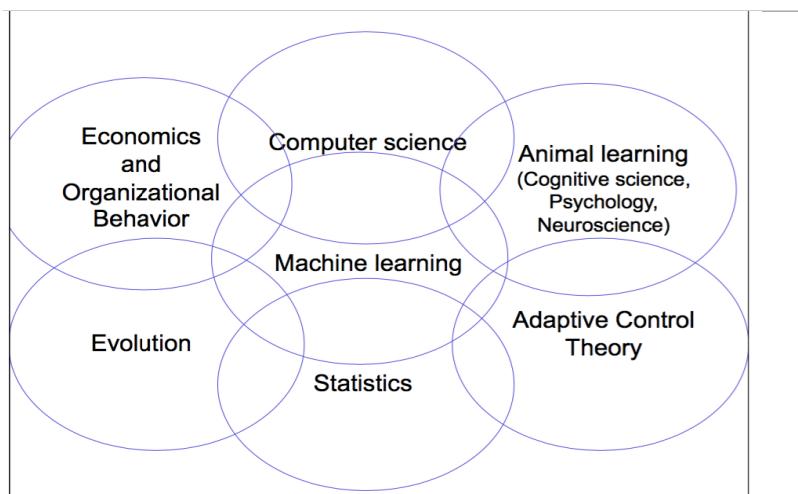


- Cải thiện tác vụ T (Task) ? => lọc được email rác hay không rác
- Với độ đo hiệu quả P (Performance)? => % email được gán nhãn/ xác định đúng là rác hoặc không rác.
- Dựa trên kinh nghiệm E (Experience) ? Kiểm tra trong danh sách email đã có email nào được gán nhãn là rác và email là không phải rác

9

9

Các kiến thức có liên quan



10

10

Phân loại học máy

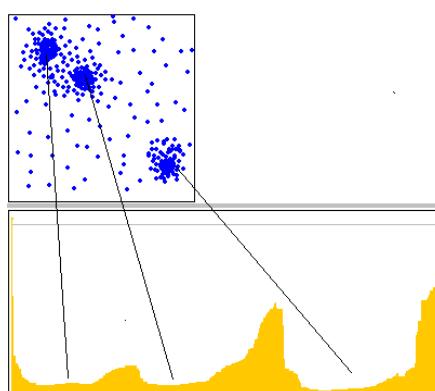
- 1. Học không có giám sát – unsupervised learning**
2. Học có giám sát – supervised learning
3. Học bán giám sát- semi- supervised learning
4. Học củng cố / học tăng cường: reinforcement learning

11

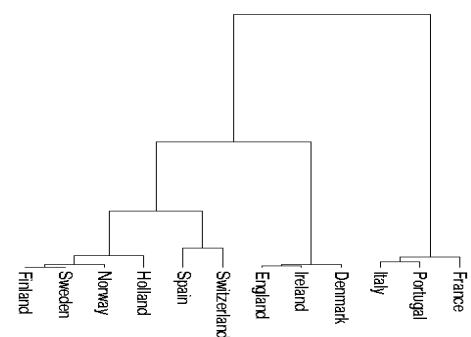
11

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

Kmeans



Hierarchical Clustering



12

12

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 classification): 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, \dots, X^m)
 - Hierarchical Clustering
 - Kmeans, ...

13

13

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**

14

14

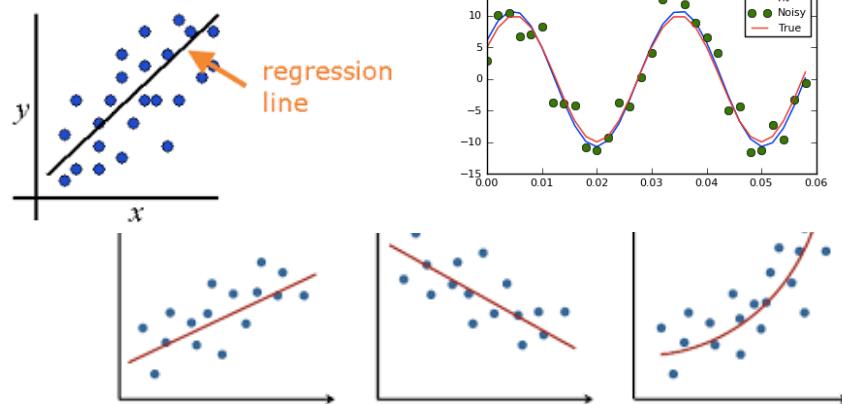
Phân loại học máy – **học có giám sát**

- 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 $\{(X^1, y^1), (X^2, y^2), \dots, (X^n, y^n)\}$
 - **Bài toán hồi quy (regression):** y là giá trị liên tục
 - **Bài toán phân lớp:** y là giá trị **không** liên tục

15

15

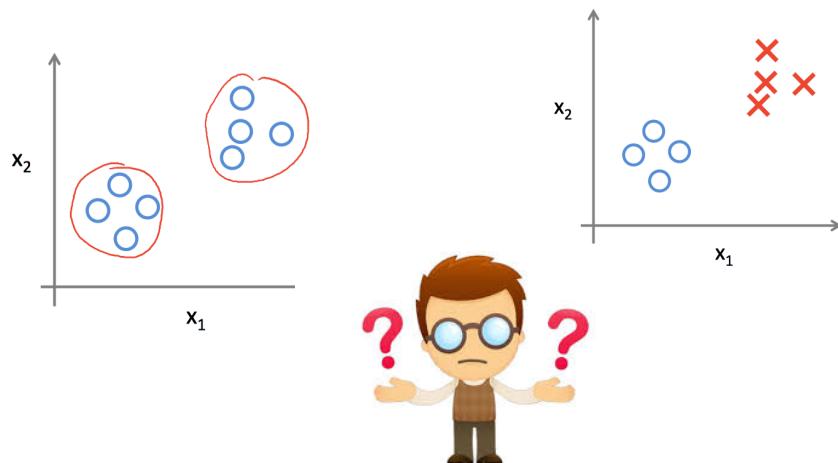
Bài toán hồi quy: Regression



16

16

Bài toán phân lớp



17

17

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

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

18

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

19

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?

20

20

Phân loại học máy

Học có giám sát

- **phân lớp** (classification, supervised classification) : 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,...

21

21

Phương pháp học tăng cường

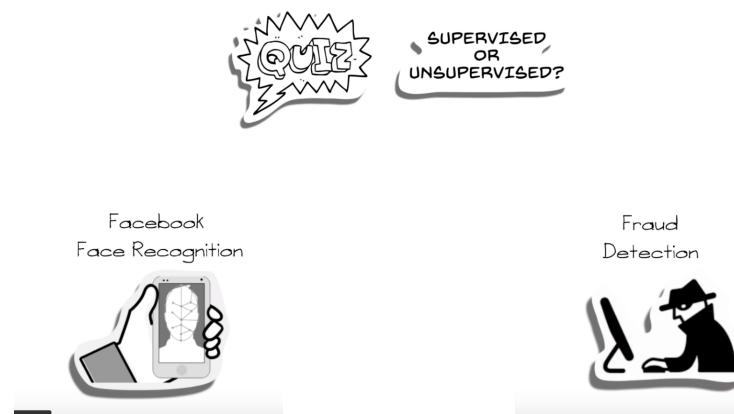
Học bán giám sát: kết hợp không giám sát và giám sát

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.²²

22

Phân loại học máy dựa trên phương thức học



<https://www.youtube.com/watch?v=ukzFl9rgwfU>

23

23

INTERESTING MACHINE LEARNING MODEL WHICH IS USED BY



DIFFERENTIAL PRICING IN REAL TIME BASED ON:

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

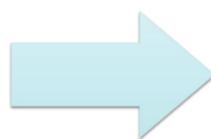
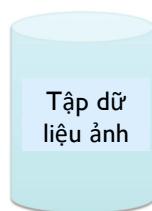
<https://www.youtube.com/watch?v=ukzFl9rgwfU>

24

24

Ứng dụng của Nguyên lý Máy học

Clustering images

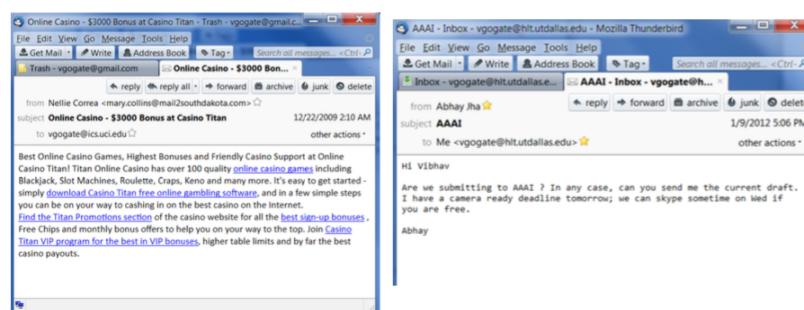


25

25

Ứng dụng của Nguyên lý Máy học

Classification Example: Spam Filtering



Classify as “Spam” or “Not Spam”

26

26

Ứng dụng của Nguyên lý Máy học

Classification Example: Weather Prediction

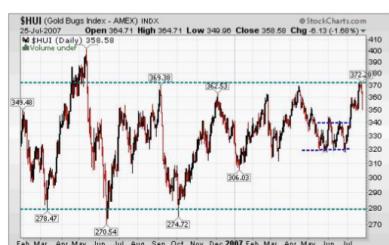


27

27

Ứng dụng của Nguyên lý Máy học

Regression example: Predicting Gold/Stock prices



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

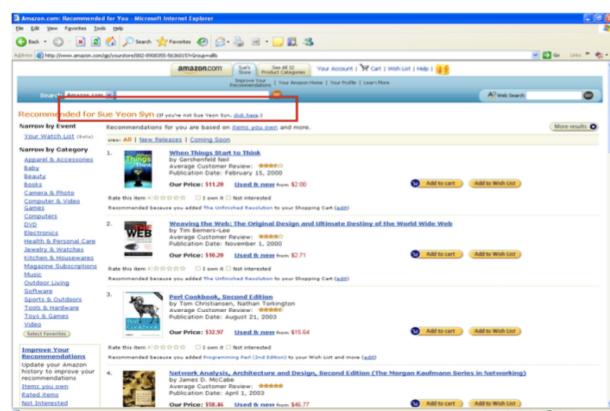
Given historical data on Gold prices, predict tomorrow's price!

28

28

Ứng dụng của Nguyên lý Máy học

Collaborative Filtering

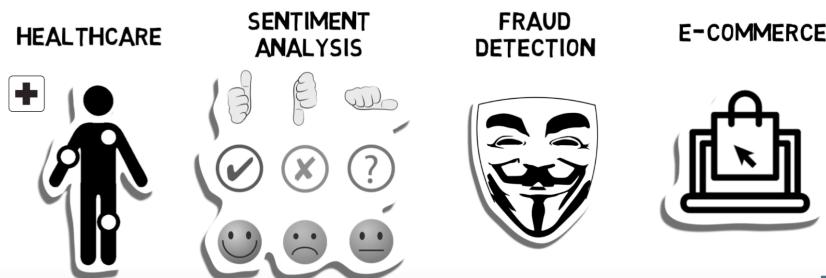


29

29

Ứng dụng của Nguyên lý Máy học

APPLICATIONS OF MACHINE LEARNING



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

30

30

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>

31

31

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

...

32

32



The
End

33

33