

KHOA CNTT & TRUYỀN THÔNG
BM KHOA HỌC MÁY TÍNH

NGUYÊN LÝ MÁY HỌC

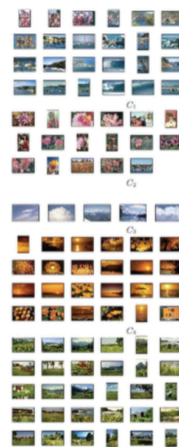
Machine Learning

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

1

Ứng dụng của Nguyên lý Máy học

Clustering images

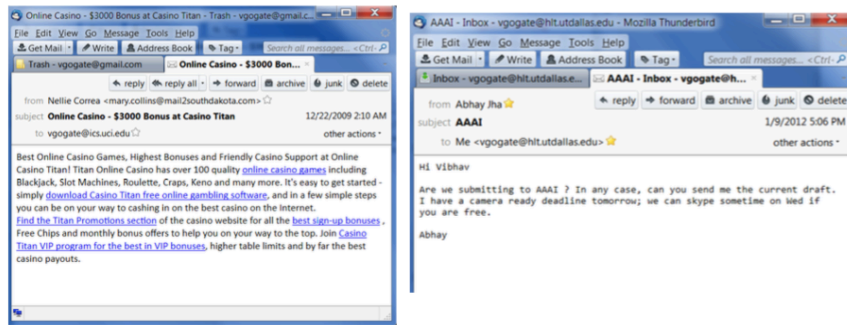


[Goldberger et al.]

2

Ứng dụng của Nguyên lý Máy học

Classification Example: Spam Filtering



Classify as “Spam” or “Not Spam”

3

Ứng dụng của Nguyên lý Máy học

Classification Example: Weather Prediction



4

Ứng dụng của Nguyên lý Máy học

Regression example: Predicting Gold/Stock prices



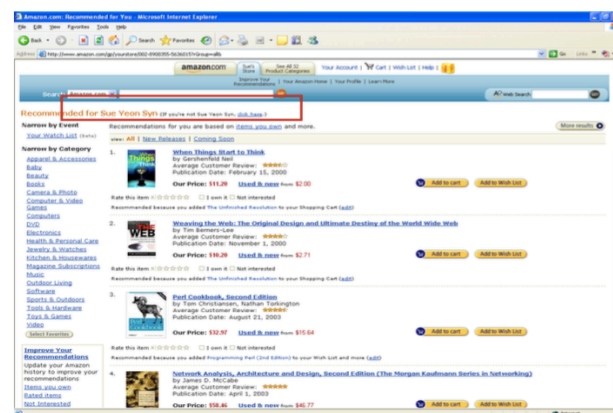
Good ML can make you rich (but there is still some risk involved).

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

5

Ứng dụng của Nguyên lý Máy học

Collaborative Filtering



6

Nguyên lý máy học là gì?

Chương trình truyền thống



Nguyên lý máy học



7

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:

– Well posed machine learning – Improving performance via experience

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

8

Nguyên lý máy học là gì?

T. Mitchell:

- Well posed 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.

(Mitchel, 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

9

Nguyên lý máy học là gì?

Ví dụ: **học chơi cờ**

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

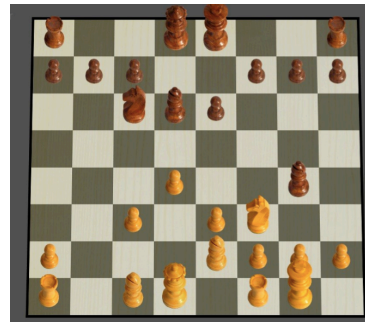


10

Nguyên lý máy học là gì?

Ví dụ: **học chơi cờ**

- Cải thiện tác vụ T ? => chơi cờ
- Với độ đo hiệu quả P ? => % ván thắng
- Dựa trên kinh nghiệm E ?
 - Kinh nghiệm gì?
 - Nên học gì?
 - Biểu diễn như thế nào?
 - Giải thuật nào để học?



11

Nguyên lý máy học là gì?

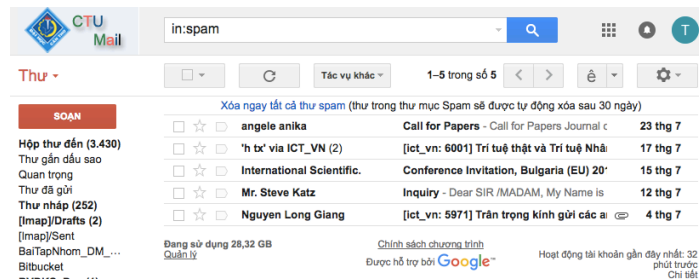
• Ví dụ: **học chơi cờ**

- T: chơi cờ
- P: % ván thắng
- E: kinh nghiệm gì? Nên học gì? Biểu diễn như thế nào? Giải thuật nào để học?
 - Kiểu dạy/ huấn luyện: Trực tiếp hay gián tiếp ?
 - Có thầy hay không có thầy ? Vấn đề: việc huấn luyện có phải là kinh nghiệm có thể biểu diễn được của mục tiêu hiệu quả ?
 - Lượng giá hàm mục tiêu:
 - Lựa chọn nước đi: Bàn cờ => Nước đi ??
 - Lượng giá bàn cờ ?

12

Nguyên lý máy học là gì?

- Ví dụ: “Chương trình lọc email rác”

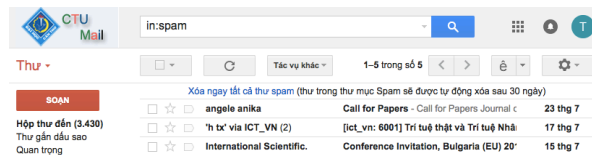


- Cải thiện tác vụ T (Task) ?
- Với độ đo hiệu quả P (Performance) ?
- Dựa trên kinh nghiệm E (Experience) ?

13

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

14

Các môn học khác có liên quan

Trí tuệ nhân tạo

Phương pháp Bayes

Lý thuyết độ phức tạp tính toán

Lý thuyết điều khiển

Lý thuyết thông tin

Triết học

Tâm lý học và thần kinh học (neurobiology)

Thống kê

...

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 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

17

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 $\{(x_1, y_1), (x_2, y_2), \dots, (x_m, y_m)\}$

- **Bài toán hồi quy** (regression): y là giá trị liên tục
- **Bài toán phân lớp**: y là giá trị **không** liên tục

18

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)

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,...