

HỌC MÁY

TS. Nguyễn Thị Kim Ngân

Email: ngannguyen@tlu.edu.vn

*Có tham khảo bài giảng của PGS.TS Nguyễn Hữu Quỳnh,
PGS.TS Nguyễn Thanh Tùng, khoa CNTT, TLU*



Thông tin môn học

- Là học phần cơ sở ngành bắt buộc cho các ngành CNTT, HTTT, CNPM
- Môn học tiên quyết: Toán rời rạc, Cấu trúc dữ liệu và giải thuật, Thống kê ứng dụng
- Kỹ năng lập trình: Python cơ bản



Mục tiêu môn học

- Cung cấp các kiến thức cơ bản về
 - Các mô hình học máy (không giám sát và có giám sát)
 - Một số bài toán cơ bản trong học máy: hồi quy, phân loại, phân cụm
 - Một số giải thuật học máy cơ bản: hồi quy tuyến tính, K-mean, Gradient, Học Perceptron, Decision tree, Hồi quy Logistic, SVM, Học kết hợp
 - Phương pháp đánh giá “độ tốt” của một hệ thống học
- Kỹ năng thực hành thuật toán học máy trên Python
 - Sinh viên cài đặt được một số thuật toán học máy cơ bản



Đánh giá

- Điểm quá trình: 50%
 - Bài tập: 20%
 - Kiểm tra trên lớp: 20%
 - Vắng <14 tiết: 10%
- Thi cuối kỳ (vấn đáp): 50%



Tài liệu tham khảo

- Vũ Hữu Tiệp, *Machine Learning cơ bản*, 2018. Link download <https://github.com/tiepvupsu/ebookMLCB>
 - Blog: [https:// machinelearningcoban.com](https://machinelearningcoban.com)
 - Facebook Page: [https:// www.facebook.com/machinelearningbasicvn/](https://www.facebook.com/machinelearningbasicvn/)
 - Facebook Group: [https:// www.facebook.com/ groups/machinelearningcoban/](https://www.facebook.com/groups/machinelearningcoban/)
 - Interactive Learning: <https://fundaml.com>



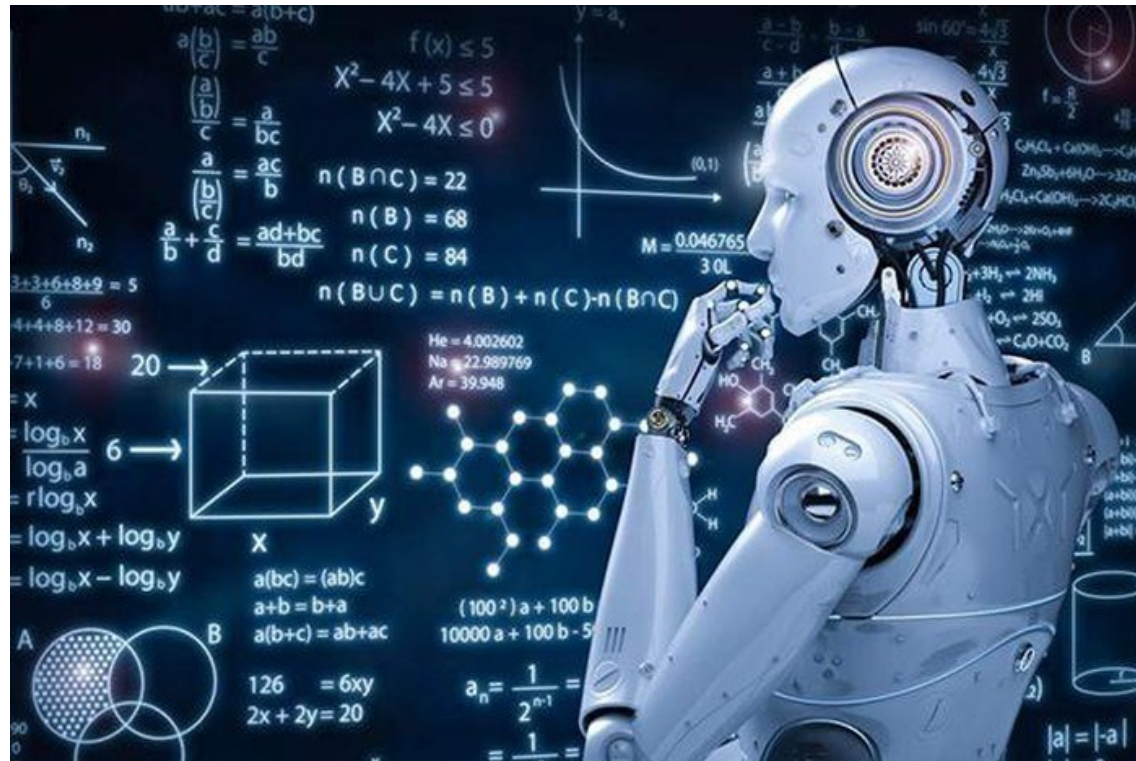
Ngôn ngữ lập trình python

Scikit-learn ([http:// scikit-learn.org/](http://scikit-learn.org/)): thư viện chứa rất nhiều các thuật toán machine learning cơ bản, dễ sử dụng

Numpy ([http:// www.numpy.org/](http://www.numpy.org/)): thư viện giúp xử lý các phép toán liên quan đến các mảng nhiều chiều, với các hàm gần gũi với đại số tuyến tính

Giới thiệu máy học

Machine Learning nổi lên như một bằng chứng của cuộc cách mạng công nghiệp lần thứ tư





Xe tự hành

Một **hệ thống tự hành** đúng nghĩa phải đáp ứng được hai yếu tố (**Sridhar Lakshmanan**, Đại học Michigan-Dearborn)

- Xử lý khối lượng lớn dữ liệu tương tự như một chiếc máy tính
- Thông minh như não người để thích ứng với môi trường mới lần cũ”

Demo



Xe tự hành

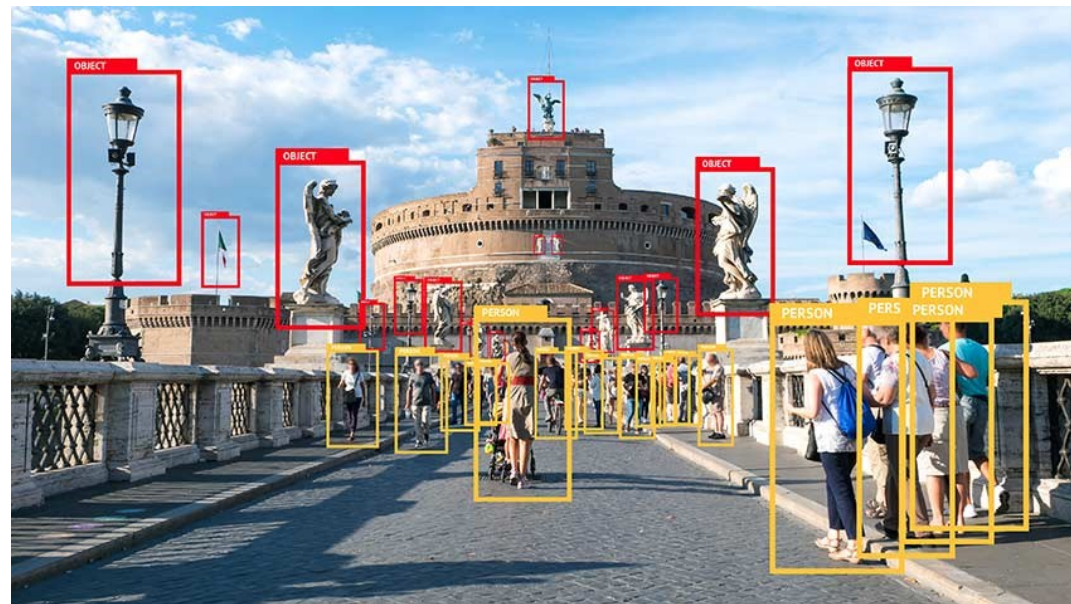
Một chiếc ô tô chỉ có thể tự lái được nếu đáp ứng được các tiêu chuẩn:

- Có hình dạng giống những chiếc ô tô thông thường
- Được trang bị hệ thống nhận diện các biến động trên đường:
 - GPS xác định điểm đầu và điểm cuối của hành trình dựa trên Google Maps
 - Hệ thống công nghệ hỗ trợ khác:
 - Camera: nhìn thấy tình trạng xung quanh xe
 - Radar: nhìn được đường phía trước (khoảng cách 100m)
 - Laser: quét các hiện tượng xảy ra xung quanh liên tục và gửi đến hệ thống máy tính
- Có hệ thống chuyển các thông tin từ GPS và các hệ thống hỗ trợ thành hành động thực tế trên đường

Hệ thống tự tag ảnh của Facebook

Dự án DeepFace của Facebook

- Nhận diện khuôn mặt và xác định đối tượng cụ thể trong ảnh
- Cung cấp Thẻ Alt (Thẻ thay thế) cho hình ảnh đã được tải lên trên Facebook





Trợ lý ảo

- Hỗ trợ tìm kiếm thông tin hữu ích, khi được yêu cầu qua văn bản hoặc giọng nói
- Lịch trình của tôi vào ngày mai là gì? Hoặc các chuyến bay có sẵn sắp tới cho chuyến công tác của tôi?

=> Trợ lý cá nhân của bạn tìm kiếm thông tin hoặc nhớ lại các truy vấn liên quan của bạn để thu thập thông tin

Hệ thống gợi ý sản phẩm của Amazon

amazon [Join Prime](#) [Thai-Nghe's Amazon.com](#) [Today's Deals](#) [Gift Cards](#) [Help](#)

Shop by Department ▾ Search

[Your Amazon.com](#) [Your Browsing History](#) **Recommended For You** [Amazon Betterizer](#) [Improve Your Recommendations](#) [Your Profile](#) [Le](#)


[Your Amazon.com](#) **Recommended for You**
(If you're not Thai-Nghe Nguyen, click here.)

Just For Today
[Browse Recommended](#)

Recommendations
[Amazon Instant Video](#)
[Amazon MP3 Store](#)
[Appliances](#)
[Appstore for Android](#)
[Arts, Crafts & Sewing](#)
[Automotive](#)
[Baby](#)
[Beauty](#)
[Books](#)

These recommendations are based on [items you own](#) and more.

view: **All** | [New Releases](#) | [Coming Soon](#)

- 

ArmorSuit MilitaryShield - Samsung Galaxy S3 Screen Protector, U.S. Cellular
by ArmorSuit (May 18, 2012)
Average Customer Review: ★★★★★ (344)
In Stock

List Price: \$42.95
Price: \$9.95
6 used & new from \$9.17

☐ I own it ☐ Not interested ☒ ★★★★★ Rate this item

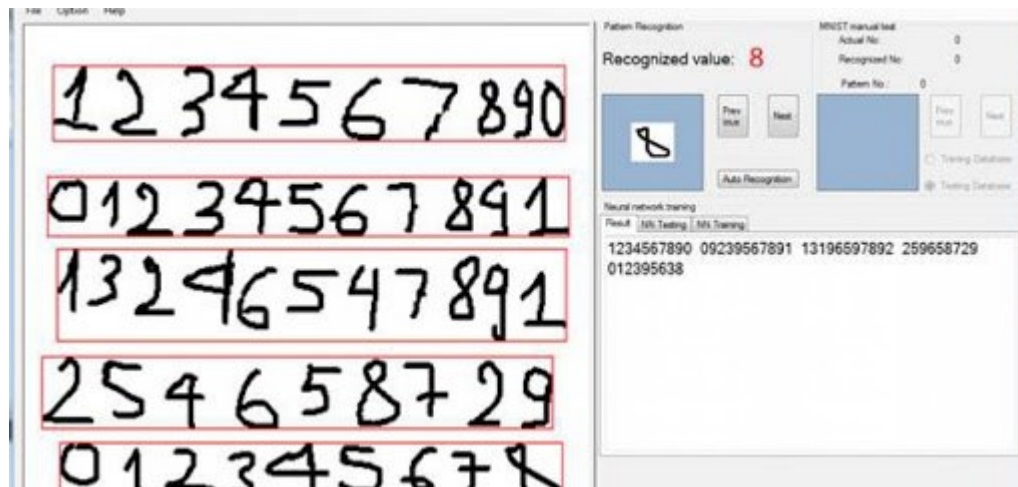
Recommended because you liked **Samsung Galaxy S III 4G Android Phone, Blue 16GB** (Fix

AlphaGo

Máy chơi cờ vây AlphaGo của Google DeepMind,



Nhận dạng chữ viết tay





Học máy

Học máy (Machine Learning): nghiên cứu và xây dựng các kĩ thuật cho phép các hệ thống “học” tự động từ dữ liệu để giải quyết những vấn đề cụ thể

Một chương trình máy tính được gọi là **học** từ *kinh nghiệm* E để hoàn thành *nhiệm vụ* T , với hiệu quả được đo bằng *phép đánh giá* P , nếu hiệu quả của nó khi thực hiện nhiệm vụ T , khi được đánh giá bởi P , cải thiện theo kinh nghiệm E .



Phân nhóm các thuật toán học máy

Phân nhóm dựa trên phương thức học:

- Supervised learning,

- Unsupervised learning,

- Semi-supervised learning

- Reinforcement learning

Phân nhóm dựa trên chức năng của các thuật toán:

- Regression Algorithms

- Classification Algorithms

- Clustering Algorithms

- Bayesian Algorithms



Supervised Learning (Học có giám sát)

Supervised learning là nhóm phổ biến nhất trong các thuật toán Machine Learning

Supervised learning là thuật toán dự đoán đầu ra (outcome) của một dữ liệu mới (new input) dựa trên các cặp (*data, label*) đã biết từ trước

Với tập ví dụ huấn luyện:

ID	Diện tích (m ²)	Số phòng ngủ	Giá bán (triệu VNĐ)
1	20	1	250.396
2	37	1	412.569
3	45	2	512.021
4	15	1	125.455
5	22	1	265.314
6	120	2	1.325.156
...

Cần trả lời:

Một căn phòng có: x_1 m², x_2 phòng ngủ sẽ có giá bao nhiêu?



Supervised Learning (Học có giám sát)

Một tập hợp biến đầu vào $X=\{x_1, x_2, \dots, x_N\}$ và một tập hợp nhãn tương ứng $Y=\{y_1, y_2, \dots, y_N\}$

Các cặp dữ liệu biết trước $(x_i, y_i) \in X \times Y$ được gọi là tập *training data*

Từ tập training data, chúng ta cần tạo ra một hàm số ánh xạ mỗi phần tử từ tập X sang một phần tử (xấp xỉ) tương ứng của tập Y :

$$y_i \approx f(x_i), \quad \forall i=1, 2, \dots, N$$

Mục đích là xấp xỉ hàm số f thật tốt để khi có một dữ liệu x mới, chúng ta có thể tính được nhãn $y=f(x)$



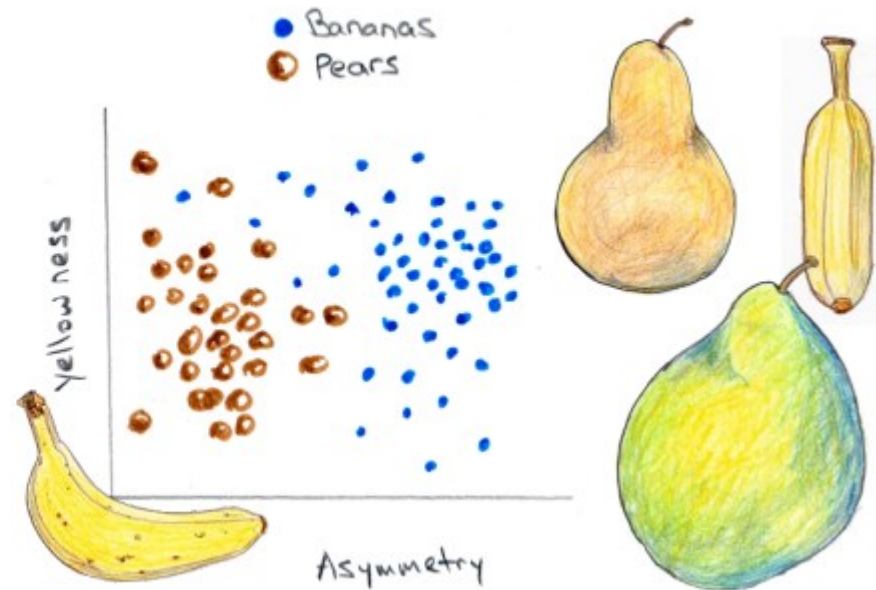
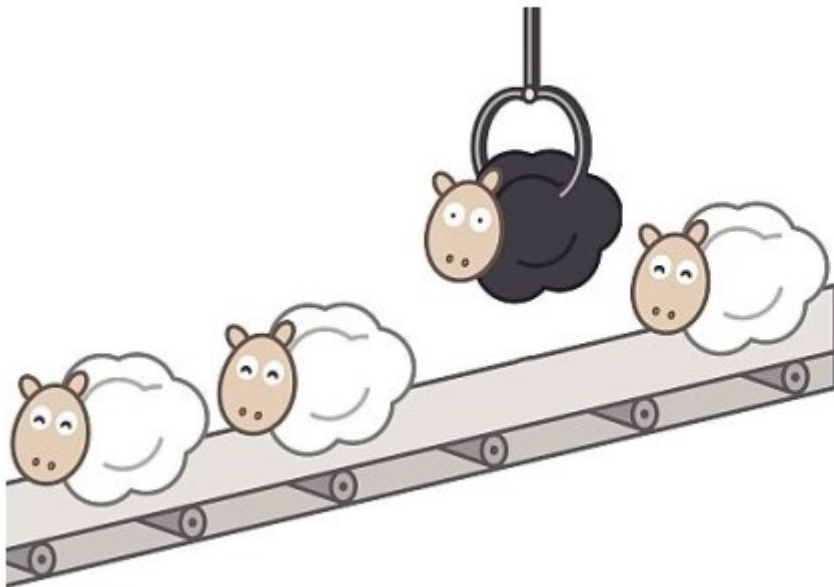
Supervised Learning (Học có giám sát)

Classification (Phân lớp): Một bài toán được gọi là *classification* nếu các *label* của *input data* được chia thành một số hữu hạn nhóm

Rec. ID	Age	Income	Student	Credit_Rating	Buy_Computer
1	Young	High	No	Fair	No
2	Young	High	No	Excellent	No
3	Medium	High	No	Fair	Yes
4	Old	Medium	No	Fair	Yes
5	Old	Low	Yes	Fair	Yes
6	Old	Low	Yes	Excellent	No
7	Medium	Low	Yes	Excellent	Yes
8	Young	Medium	No	Fair	No
9	Young	Low	Yes	Fair	Yes
10	Old	Medium	Yes	Fair	Yes
11	Young	Medium	Yes	Excellent	Yes
12	Medium	Medium	No	Excellent	Yes
13	Medium	High	Yes	Fair	Yes
14	Old	Medium	No	Excellent	No

Một sv trẻ với mức thu nhập trung bình, mức đánh giá tín dụng bình thường sẽ được phân vào lớp Yes hay No?

Supervised Learning (Học có giám sát)





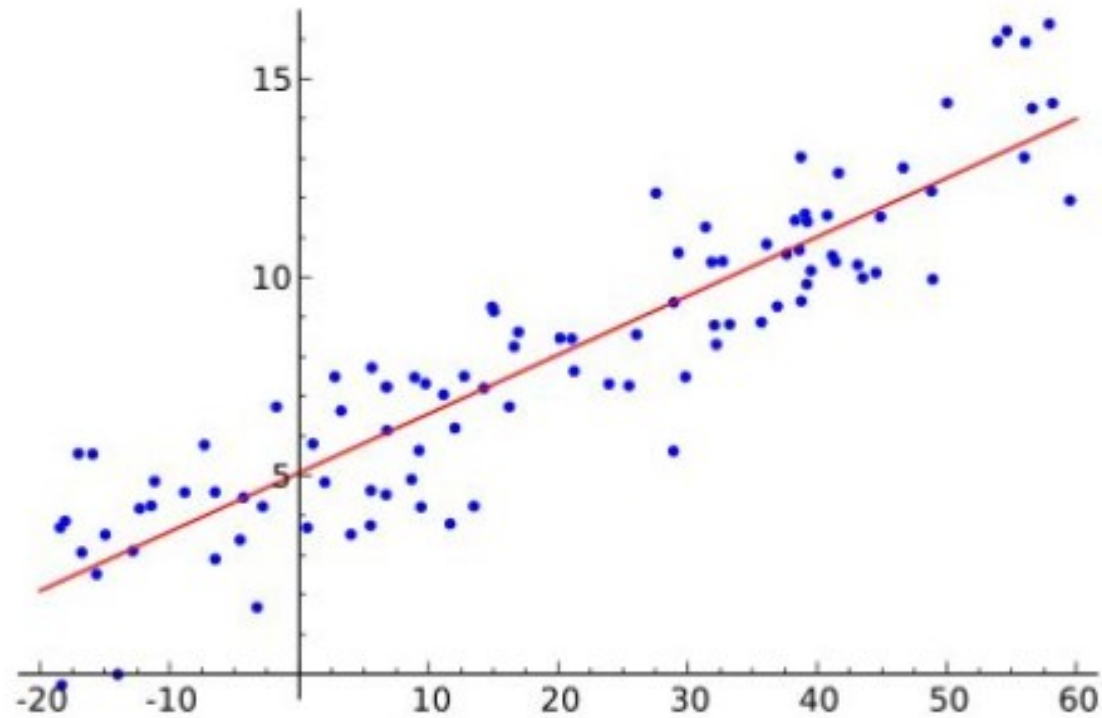
Supervised Learning (Học có giám sát)

Regression (Hồi quy): Nếu *label* không được chia thành các nhóm mà là một giá trị thực cụ thể.

Diện tích	Số phòng ngủ	Cách Hồ Gươm	Giá tiền
70	1	5 km	800 triệu
90	2	5 km	1.2 tỷ
120	3	15 km	1.1 tỷ

Hỏi: Một căn phòng có: x_1 m²; x_2 phòng ngủ và cách Hồ Gươm x_3 km, sẽ có giá bao nhiêu?

Supervised Learning (Học có giám sát)





Unsupervised Learning (Học không giám sát)

Unsupervised learning là khi chúng ta chỉ có dữ liệu vào X mà không biết *nhãn* Y
Thuật toán unsupervised learning sẽ dựa vào cấu trúc của dữ liệu để thực hiện một công việc nào đó, ví dụ như:

Phân nhóm (clustering)

Giảm số chiều của dữ liệu (dimension reduction)

Tên thuốc	Đặc trưng 1	Đặc trưng 2
A	1	1
B	2	1
C	4	3
D	5	4

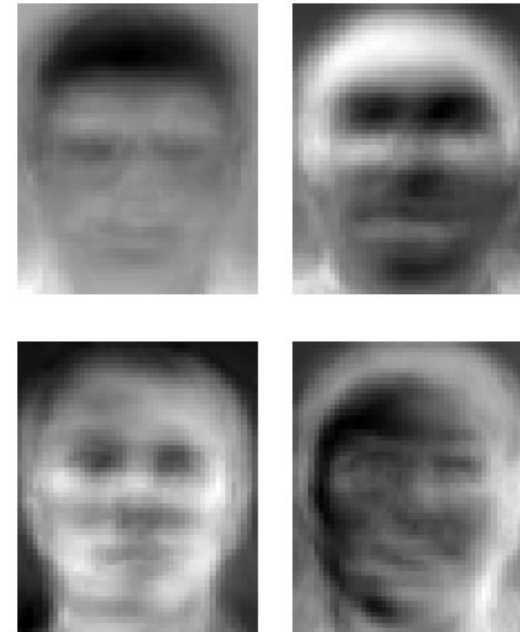


Unsupervised Learning (Học không giám sát)

Cho một tập các tài liệu văn bản, cần xác định tập tài liệu thuộc những chủ đề như thể thao, chính trị, ca nhạc,...

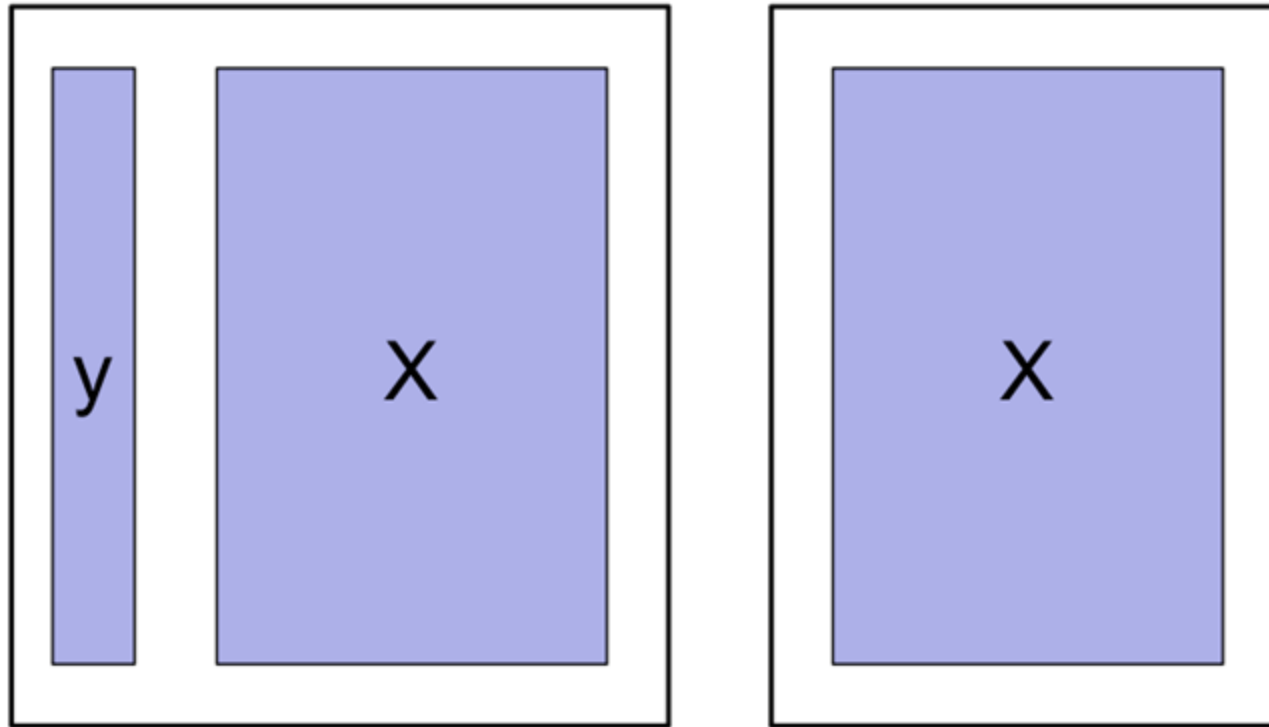
Unsupervised Learning (Học không giám sát)

Cho các ảnh khuôn mặt có số chiều cao, tìm một biểu diễn đơn giản/thu gọn của các ảnh này để đưa vào bộ phân lớp nhận dạng khuôn mặt





Học có giám sát so với không giám sát





Một số ký hiệu toán học

Các chữ cái in nghiêng biểu thị các số vô hướng

$$x_1, N, y, k$$

Các chữ cái thường in đậm biểu thị các véc tơ

$$\mathbf{y}, \mathbf{x}_1$$

Các chữ cái hoa in đậm biểu thị ma trận

$$\mathbf{X}, \mathbf{Y}, \mathbf{W}$$



Một số ký hiệu toán học

$\mathbf{x} = [x_1, x_2, \dots, x_n]$: véc tơ hàng

$\mathbf{x} = [x_1; x_2; \dots; x_n]$: véc tơ cột

$\mathbf{X} = [\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_n]$: ma trận với \mathbf{x}_j là các véc tơ cột

$\mathbf{X} = [\mathbf{x}_1; \mathbf{x}_2; \dots; \mathbf{x}_m]$: ma trận với \mathbf{x}_j là các véc tơ hàng

x_{ij} : phần tử hàng i , cột j

\mathbb{R}^n : tập hợp các véc tơ cột có n phần tử

$\mathbb{R}^{m \times n}$: tập hợp các ma trận có m hàng và n cột

\mathbf{w}_i : véc tơ cột thứ i của ma trận \mathbf{W}



Chuyển vị của ma trận

Cho $\mathbf{A} \in \mathbb{R}^{m \times n}$, ta nói $\mathbf{B} \in \mathbb{R}^{n \times m}$ là chuyển vị của \mathbf{A} nếu $b_{ij} = a_{ji}$, $\forall 1 \leq i \leq n, 1 \leq j \leq m$
Chuyển vị của một véc tơ \mathbf{x} là \mathbf{x}^T

$$\mathbf{x} = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix} \Rightarrow \mathbf{x}^T = [x_1 \ x_2 \ \dots \ x_n]$$

Chuyển vị của ma trận \mathbf{A} ký hiệu là \mathbf{A}^T

$$\mathbf{A} = \begin{bmatrix} a_{11} & a_{12} & \dots & a_{1n} \\ a_{21} & a_{22} & \dots & a_{2n} \\ \dots & \dots & \ddots & \dots \\ a_{m1} & a_{m2} & \dots & a_{mn} \end{bmatrix} \Rightarrow \mathbf{A}^T = \begin{bmatrix} a_{11} & a_{21} & \dots & a_{m1} \\ a_{12} & a_{22} & \dots & a_{m2} \\ \dots & \dots & \ddots & \dots \\ a_{1n} & a_{2n} & \dots & a_{mn} \end{bmatrix}$$

Nếu $\mathbf{A} \in \mathbb{R}^{m \times n}$ thì $\mathbf{A}^T \in \mathbb{R}^{n \times m}$

Nếu $\mathbf{A}^T = \mathbf{A}$ ta nói \mathbf{A} là ma trận đối xứng



Ma trận đơn vị

Đường chéo chính của một ma trận là tập hợp các điểm có chỉ số hàng và cột bằng nhau

Một ma trận bậc n là một ma trận đặc biệt trong $\mathbb{R}^{n \times n}$ với các phần tử trên đường chéo chính bằng 1, các phần tử còn lại bằng 0

$$\mathbf{I}_3 = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}, \quad \mathbf{I}_4 = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}$$

Nếu $A \in \mathbb{R}^{m \times n}$, $B \in \mathbb{R}^{n \times m}$ và I là ma trận đơn vị bậc n : $AI = A$, $IB = B$



Ma trận nghịch đảo

- Cho ma trận vuông $A \in \mathbb{R}^{n \times n}$, nếu tồn tại ma trận vuông $B \in \mathbb{R}^{n \times n}$ sao cho $AB = I_n$ thì
 - A là khả nghịch
 - B là ma trận nghịch đảo của A
- Nếu không tồn tại ma trận B thỏa mãn điều kiện trên, ta nói A là không khả nghịch
- Nếu A là khả nghịch, ma trận nghịch đảo của nó được ký hiệu là A^{-1} thì

$$\mathbf{A}^{-1}\mathbf{A} = \mathbf{A}\mathbf{A}^{-1} = \mathbf{I}$$



Chuẩn 2 của véc tơ

Độ dài của một véc tơ $\mathbf{x} \in \mathbb{R}^n$ chính là một chuẩn (norm) 2

$$\|\mathbf{x}\|_2 = \sqrt{x_1^2 + x_2^2 + \cdots + x_n^2}$$

$$\|\mathbf{x}\|_2^2 = \mathbf{x}^T \mathbf{x}.$$



Một số ký hiệu toán học

Ký hiệu	Ý nghĩa
x, y, N, k	in nghiêng, thường hoặc hoa, là các số vô hướng
\mathbf{x}, \mathbf{y}	in đậm, chữ thường, là các vector
\mathbf{X}, \mathbf{Y}	in đậm, chữ hoa, là các ma trận
\mathbb{R}	tập hợp các số thực
\in	phần tử thuộc tập hợp
\exists	tồn tại
\forall	mọi
x_i	phần tử thứ i (tính từ 1) của vector \mathbf{x}
a_{ij}	phần tử hàng thứ i , cột thứ j của ma trận \mathbf{A}
\mathbb{N}	tập hợp các số tự nhiên



Một số ký hiệu toán học

Ký hiệu	Ý nghĩa
\mathbf{A}^T	chuyển vị của ma trận \mathbf{A}
\mathbf{A}^{-1}	nghịch đảo của ma trận vuông \mathbf{A} , nếu tồn tại
\mathbf{A}^\dagger	giả nghịch đảo của ma trận không nhất thiết vuông \mathbf{A}
\mathbf{A}^{-T}	nghịch đảo rồi chuyển vị của ma trận \mathbf{A}
$\ \mathbf{x}\ _p$	norm p của vector \mathbf{x}



Dự đoán và suy diễn

- Dự đoán (prediction): Dự đoán biến đích Y khi cho tập dữ liệu đầu vào X , ta cần sử dụng hàm \hat{f} (là ước lượng thống kê của hàm f)
- Suy diễn (inference): Tìm hiểu mối quan hệ giữa Y và các biến độc lập X_i



Các mô hình học máy

- Các mô hình có tham số (parametric)
 - Đặt các giả thiết cho dạng (form) của hàm f
 - Sử dụng dữ liệu huấn luyện để xấp xỉ mô hình (ước lượng các tham số)
- Ưu điểm
 - Dễ tìm các tham số của f
- Nhược điểm:
 - Dạng của hàm f có thể thiếu chính xác (mô hình ước lượng)