CSE: **Faculty of Computer Science and Engineering**
Thuyloi University

Nhập môn học máy

Trình bày: PGS.TS Nguyễn Hữu Quỳnh

Email: quynhnh@tlu.edu.vn

Bài giảng được dựa trên giáo trình machine learning cơ bản và có tham khảo bài giảng của PGS.TS Nguyễn Thanh Tùng, Khoa CNTT, TLU

Thông tin môn học

- Các ngành học: CNTT
- Điều kiện: Toán rời rạc, Cấu trúc dữ liệu và giải thuật, thống kê ứng dụng
- Có kỹ năng lập trình cơ bản Python

Mục tiêu môn học

- Học máy là học phần cơ sở ngành bắt buộc cho các ngành CNTT, HTTT và CNPM và là kiến thức cần thiết để học các học phần nâng cao liên quan đến kỹ thuật học.
- Học phần này trang bị cho sinh viên các kiến thức cơ bản về:
  - các mô hình (không giám sát và có giám sát);
  - bài toán phân loại, phân cụm, và bài toán hồi quy;
  - các giải thuật học máy cơ bản như hồi quy tuyến tính, K-mean, Gradient, Học Perceptron, Decision tree, Hồi quy Logistic, SVM, Học kết hợp
  - Phương pháp đánh giá một hệ thống phân lớp.
  - Khi kết thúc học phần, sinh viên cài đặt được một số thuật toán học máy cơ bản.
- Kỹ năng thực hành thuật toán học máy trên Python

Tài liệu tham khảo

Machinelearningcoban.com

Bài tập

- Sinh viên phải hoàn thành 2 đầu điểm gồm:
  - điểm bài tập (trung bình của 2 bài),
  - điểm thi giữa kỳ
- Số tiết vắng không quá 30% tổng số tiết của học phần
- Nộp bài tập theo thời khóa biểu của môn học

Ngôn ngữ lập trình python

Python: www.python.org

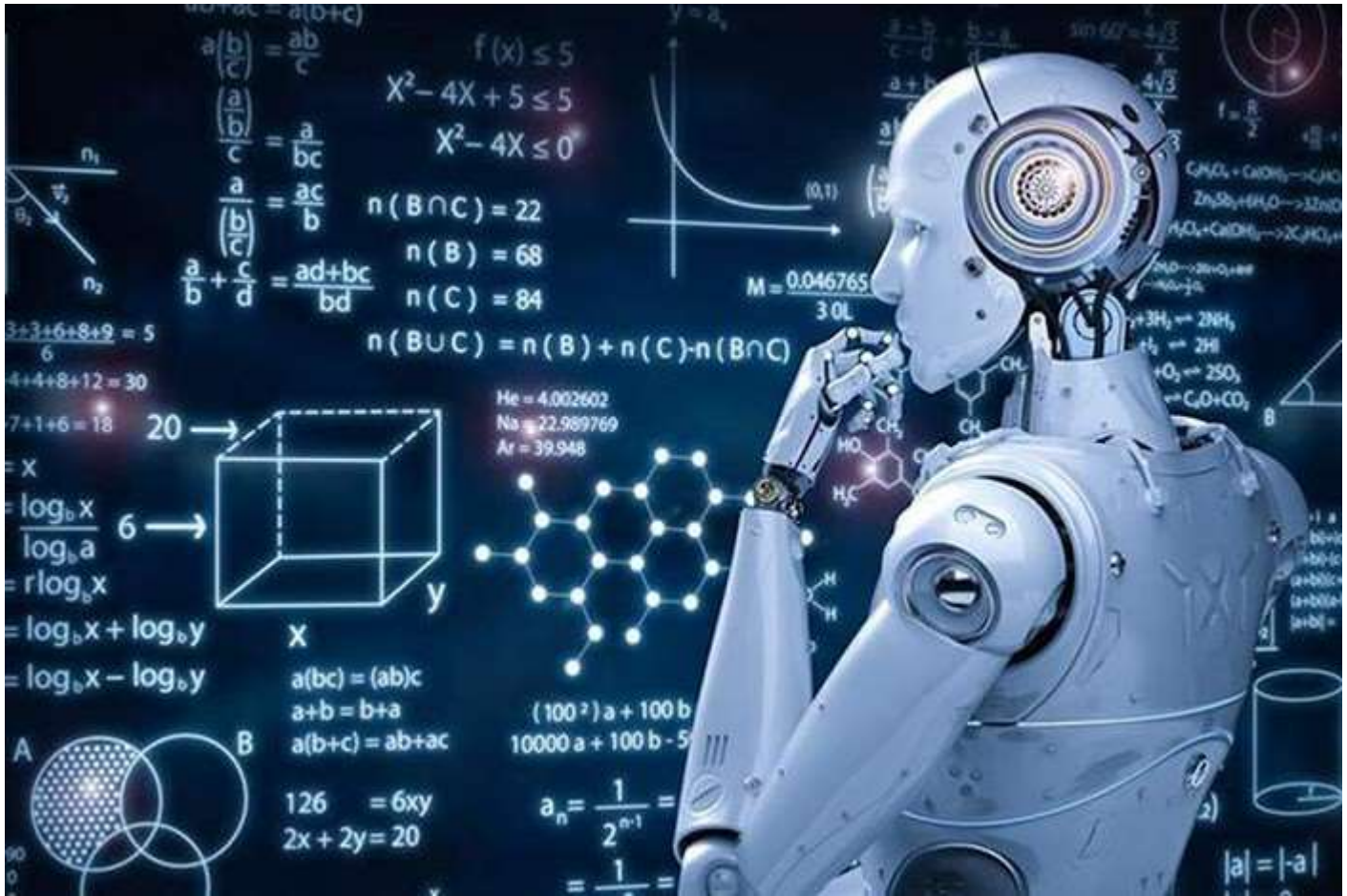
- scikit--learn: http://scikit--learn.org/

Hỏi & Đáp

- Đặt các câu hỏi liên quan đến môn học trên Piazza
- Website: https://piazza.com/class/k63hdmr4h043hf

Giới thiệu máy học

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

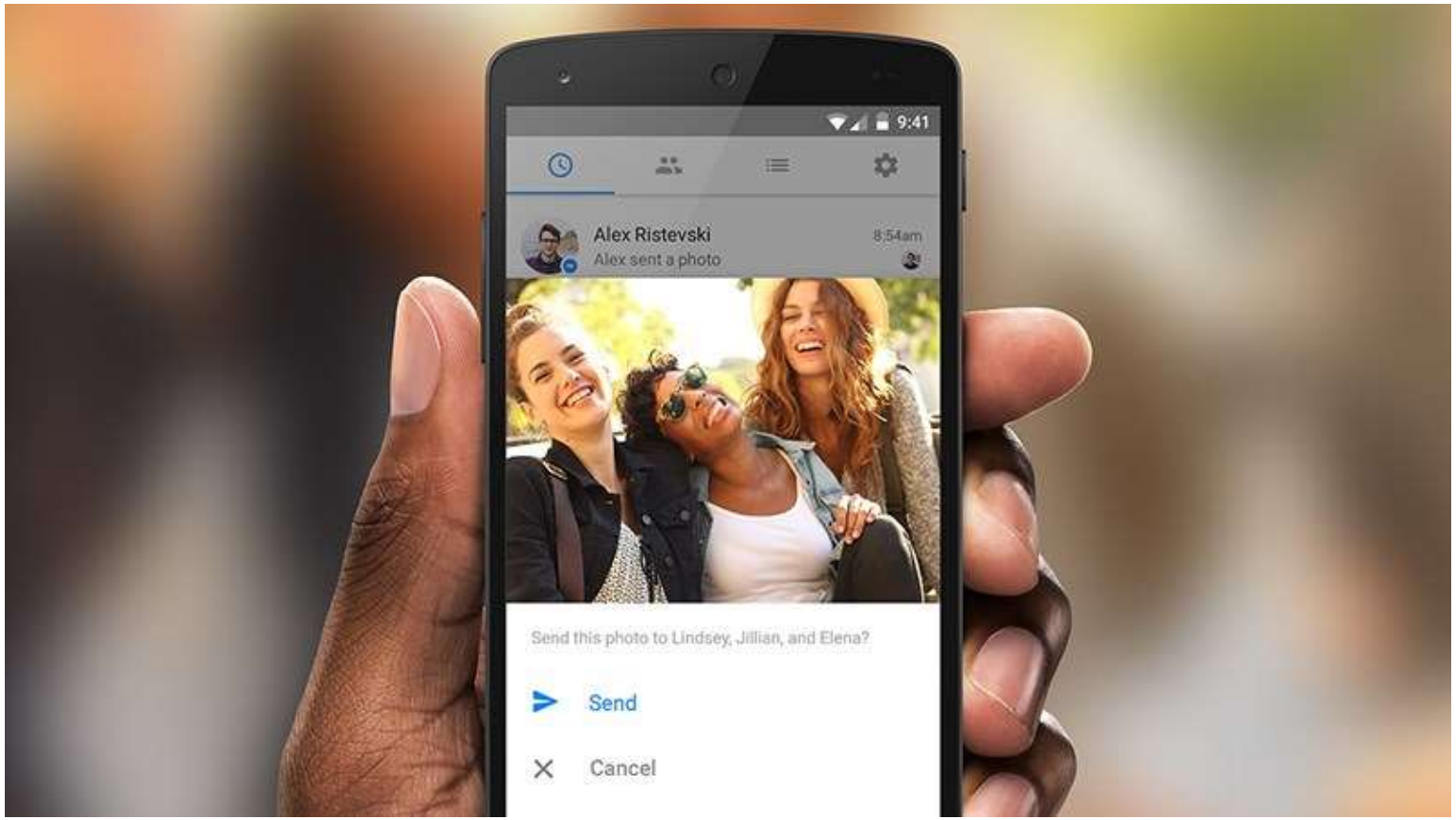


Giới thiệu máy học

- Machine Learning nổi lên như một bằng chứng của cuộc cách mạng công nghiệp lần thứ tư
- Một số ví dụ:
  - Xe tự hành của Google,

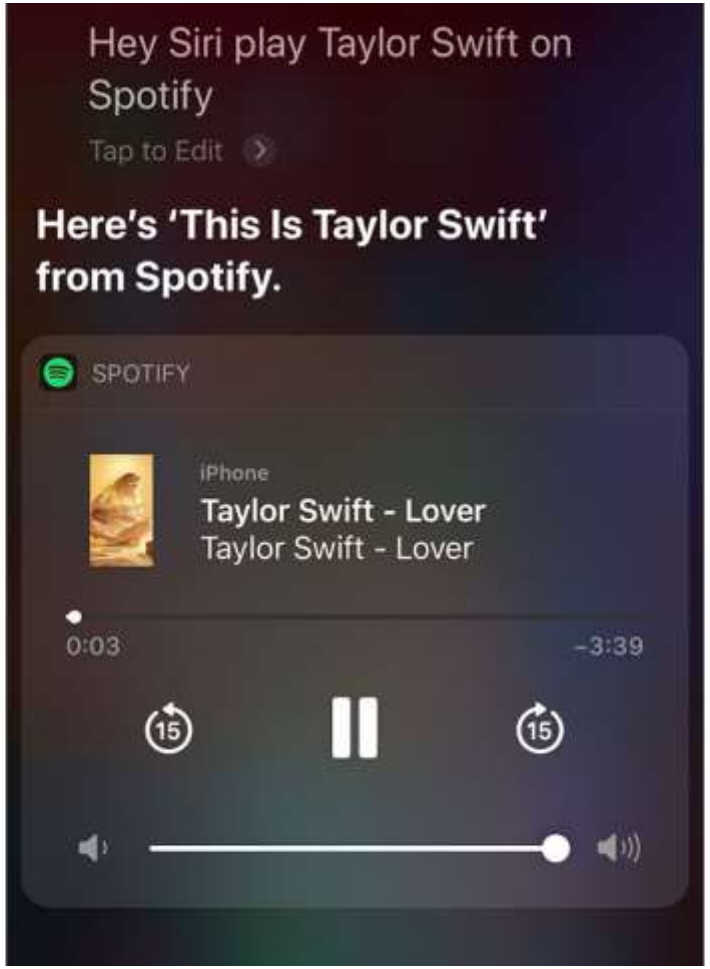
Giới thiệu máy học

- Hệ thống tự tag khuôn mặt trong ảnh của Facebook,



Giới thiệu máy học

- Trợ lý ảo Siri của Apple,



Giới thiệu máy học

- Hệ thống gợi ý sản phẩm của Amazon,

Giới thiệu máy học

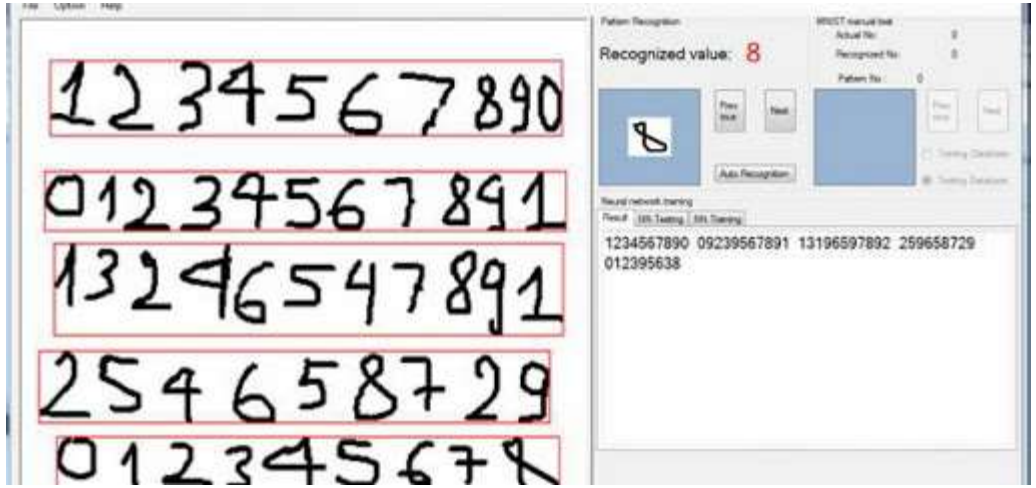
- Hệ thống gợi ý phim của Netflix,

Giới thiệu máy học

- Máy chơi cờ vây AlphaGo của Google DeepMind,

Giới thiệu máy học

- Nhận dạng chữ viết tay,



Giới thiệu máy học

- Machine Learning có khả năng tự học hỏi dựa trên dữ liệu đưa vào mà không cần phải được lập trình cụ thể

Phân nhóm các thuật toán học máy

- Phân nhóm dựa trên phương thức học:
  - Supervised learning,
  - Unsupervised learning,
  - Semi-supervised lerning
  - 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 m^2 , x_2 phòng ngủ sẽ có giá bao nhiêu?

Supervised Learning (Học có giám sát)

- Một tập hợp biến đầu vào $X=\{x_1,x_2,\ldots,x_N\}$ và một tập hợp nhãn tương ứng $Y=\{y_1,y_2,\ldots,y_N\}$
- Các cặp dữ liệu biết trước $(x_i,y_i)\in X\times Y$ được gọi là tập *training data*
- Từ tập traing 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:

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

- 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)

- 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 m^2 , x_2 phòng ngủ sẽ có giá bao nhiêu?

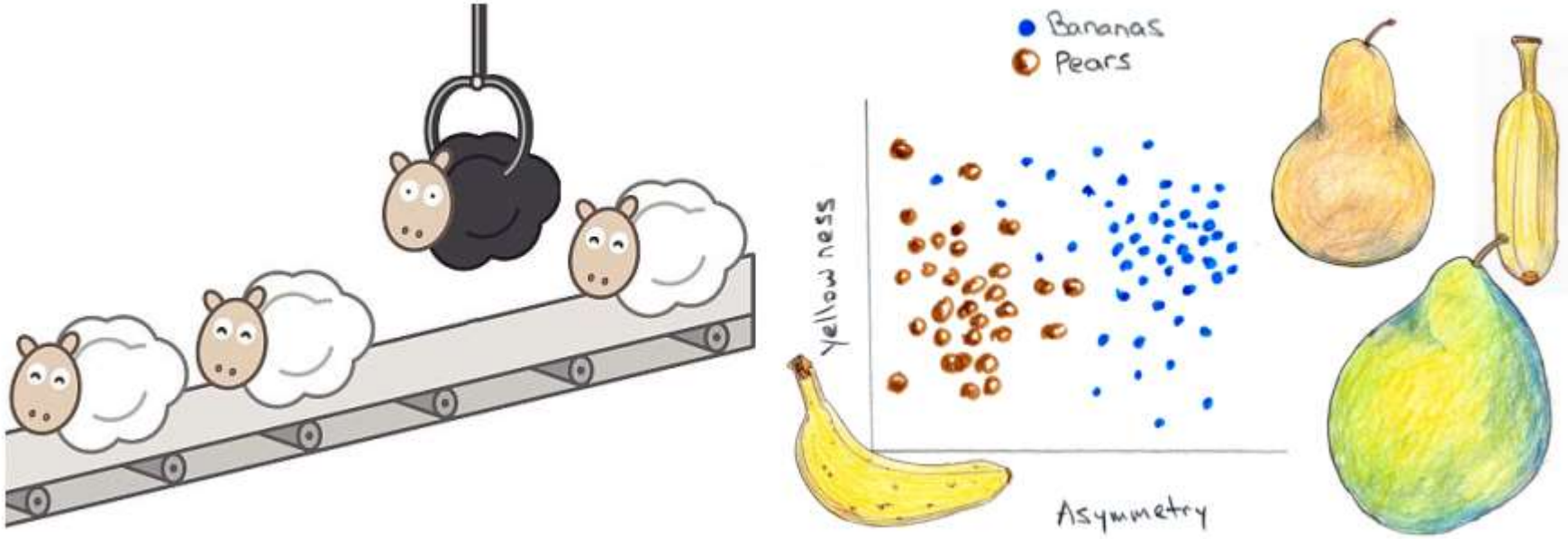
Supervised Learning (Học có giám sát)

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)



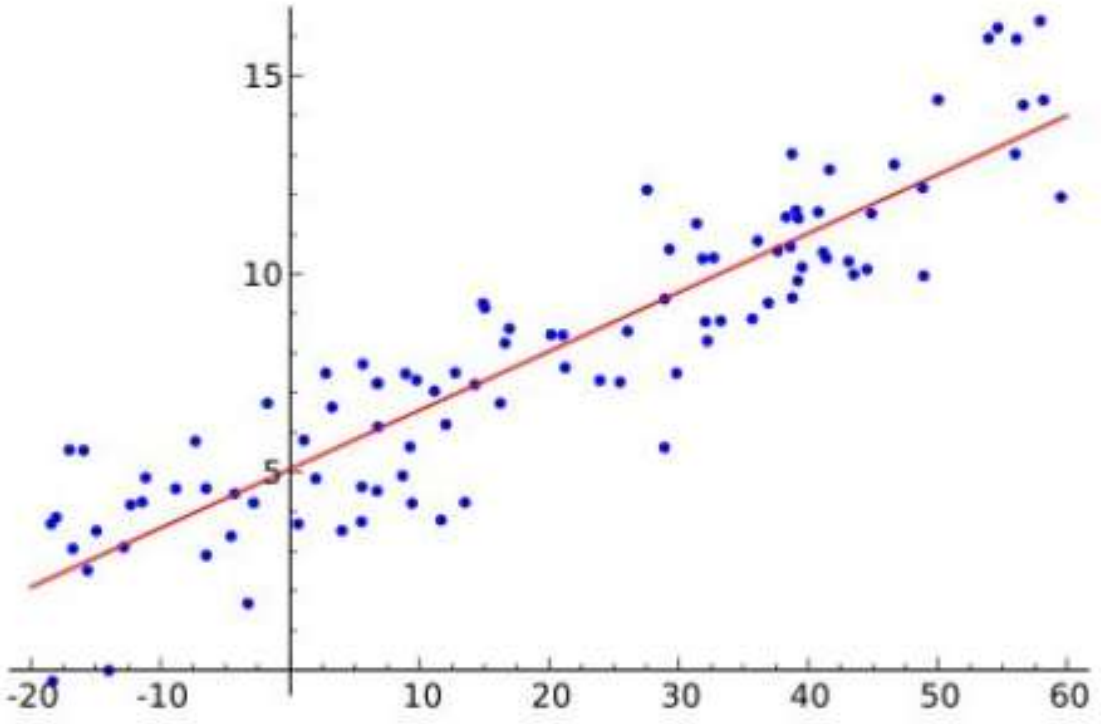
Supervised Learning (Học có giám sát)

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 m^2 ; x_2 phòng ngủ và cách Hồ Gươm x_3 km, sẽ có giá bao nhiêu?

Supervised Learning (Học có giám sát)



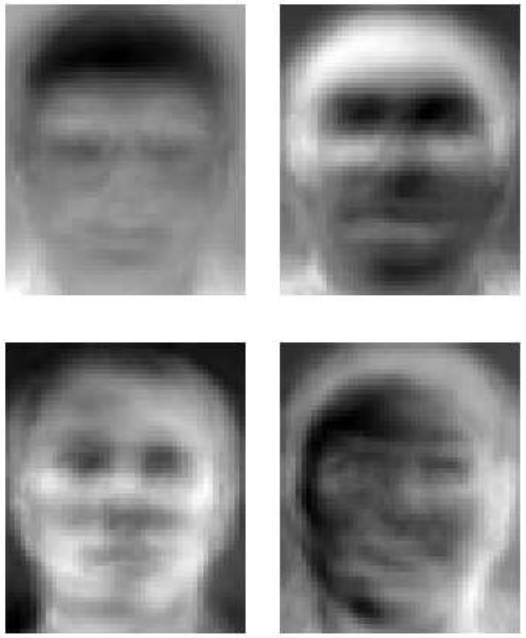
Unsupervised Learning (Học không giám sát)

- 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

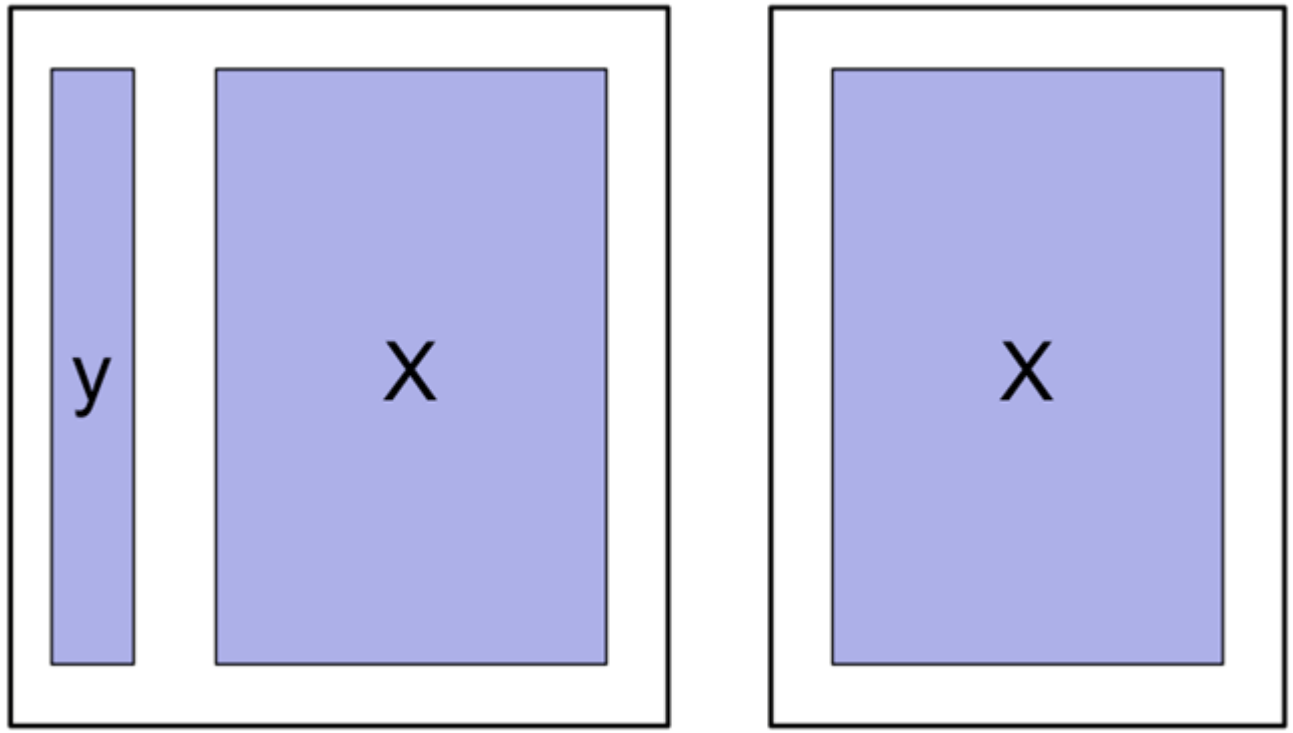
Unsupervised Learning (Học không giám sát)

- Ví dụ ứng dụng:
  - Cho một tập các tài liệu văn bản, cần xác định tập các tài liệu có chung chủ đề như thể thao, chính trị, ca nhạc,..
  - 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



(AT&T Laboratorie
Cambridge)

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

- $\mathbf{x} = [x_1, x_2, \ldots, x_n]$ là véc tơ hang
- $\mathbf{x} = [x_1; x_2; \ldots; x_n]$ là véc tơ cột
- $\mathbf{X} = [\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n]$ là ma trận với x_j là các véc tơ cột
- $\mathbf{X} = [\mathbf{x}_1; \mathbf{x}_2; \ldots; \mathbf{x}_m]$ là ma trận với x_j là các véc tơ hang
- x_{ij} là 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 hang và n cột
- $\mathbf{w}_i$ là véc tơ cột thứ i của ma trận **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 x là x^T

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

- Chuyển vị của ma trận A ký hiệu là 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 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 $\mathrm{A} \in \mathbb{R}^{m\times n}$, $\mathrm{B} \in \mathbb{R}^{n\times m}$ và I là ma trận đơn vị bậc n: $\mathrm{AI} = \mathrm{A}$, $\mathrm{IB} = \mathrm{B}$

Ma trận nghịch đảo

- Cho ma trận vuông $\mathrm{A} \in \mathbb{R}^{n \times n}$, nếu tồn tại ma trận vuông $\mathrm{B} \in \mathbb{R}^{n \times n}$ sao cho $\mathrm{AB} = \mathrm{I}_n$ 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}, ta cũng có:

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

Chuẩn 2 của véc tơ

- Độ dài của một véc tơ $x \in \mathbb{R}^n$ 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): để dự đoán biến đích Y khi cho tập dữ liệu đầu vào X, ta cần sử dụng hàm $\hat{f}$ (là ước lượng thống kế của hàm f)

- Suy diễn (inference): Tìm hiểu mỗi 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)