ĐẠI HỌC QUỐC GIA THÀNH PHỐ HỒ CHÍ MINH
TRƯỜNG ĐẠI HỌC KHOA HỌC TỰ NHIÊN

Cơ sở trí tuệ nhân tạo

Học từ dữ liệu

Nguyễn Ngọc Đức

2021

Nội dung

Học từ dữ liệu

Phê duyệt tín dụng

- Giả sử một ngân hàng nhận được hàng ngàn yêu cầu mở thẻ tín dụng mỗi ngày, và ngân hàng này muốn tự động hóa quá trình phê duyệt.
- Thông tin ứng viên

Tuổi	23
Giới tính	Nam
Lương	30000 $
Số năm làm việc	1 năm
Nợ hiện tại	15000 $

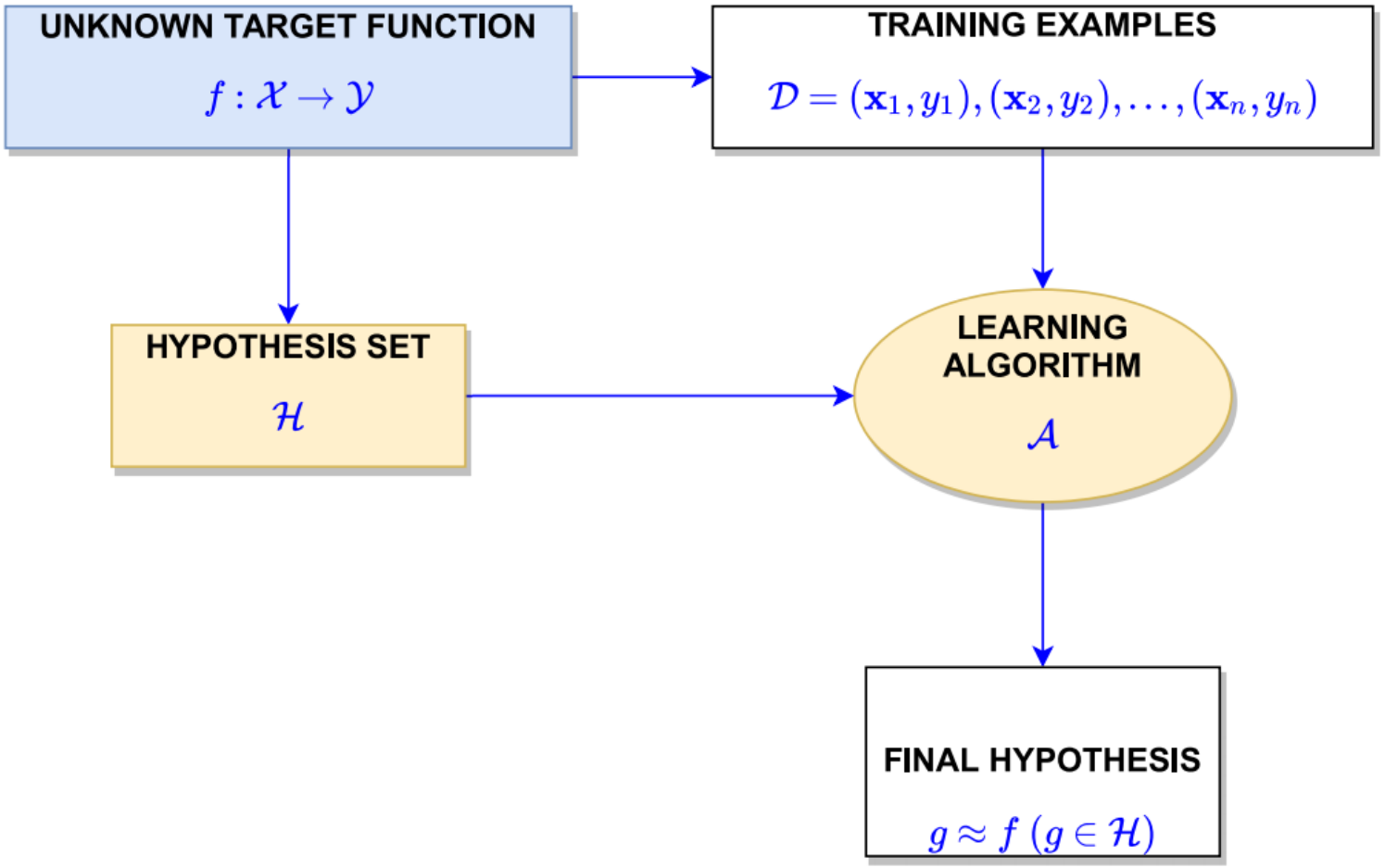
- Phê duyệt?

Mô hình hóa bài toán

Bài toán phê duyệt tín dụng

- Input: $\mathbf{x}$ thông tin ứng viên
- Output: y (khách hàng tốt / xấu)
- Data: $(\mathbf{x_1}, y), (\mathbf{x_2}, y), ..., (\mathbf{x_n}, y)$ (các kết quả trước)
- Hàm mục tiêu: $f : \mathcal{X} \to \mathcal{Y}$ (Hàm phê duyệt lý tưởng)
- Hàm phê duyệt xấp sỉ tốt nhất: $g : \mathcal{X} \to \mathcal{Y}$

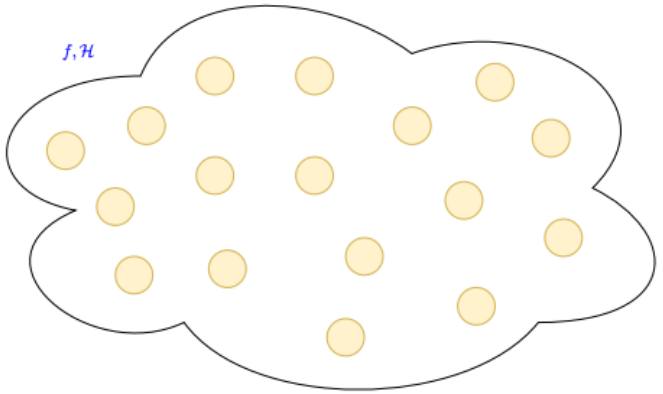
Quy trình học



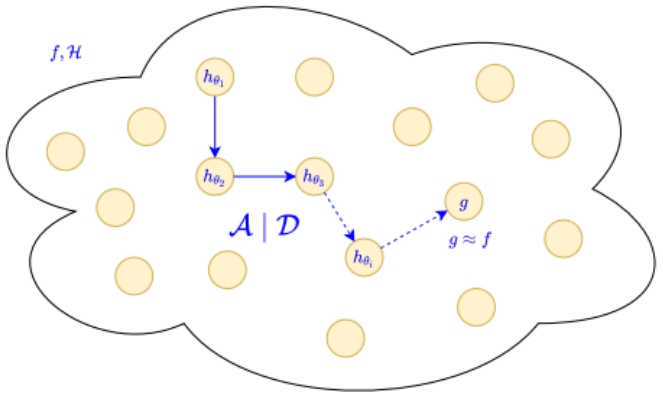
Mô hình học

Một **mô hình học** bao gồm 2 thành phần:

- Tập giả định $\mathcal{H}$ xây dựng từ bài toán

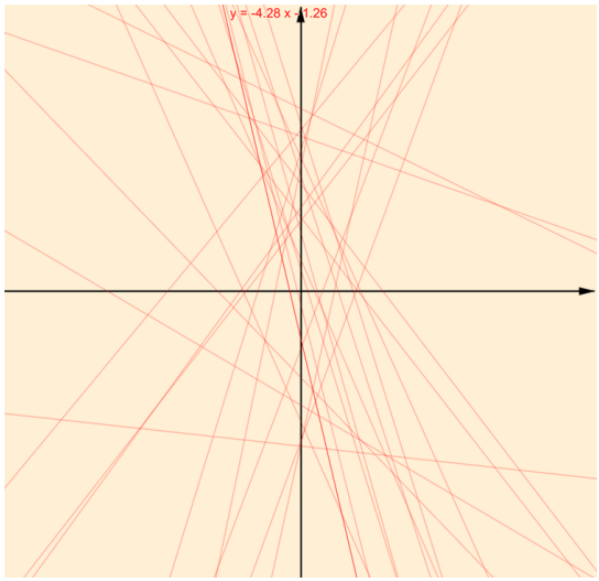


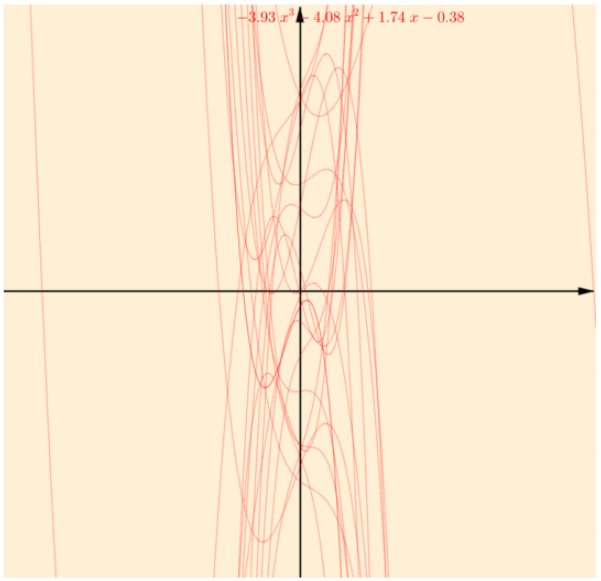
- Thuật toán học $\mathcal{A}$ là thuật toán tìm kiếm $g \in \mathcal{H}$ sao cho: $g \approx f$



Tập giả định

- Mỗi thành phần của tập giả định có thể được định nghĩa bằng các tham số (θ hoặc w).
- Ta có thể có **vô số** hàm giả định. Ví dụ, các tập giả định của phương trình $y = \theta_0 + \theta_1 x$ và $y = \theta_0 + \theta_1 x + \theta_2 x^2 + \theta_3 x^3$



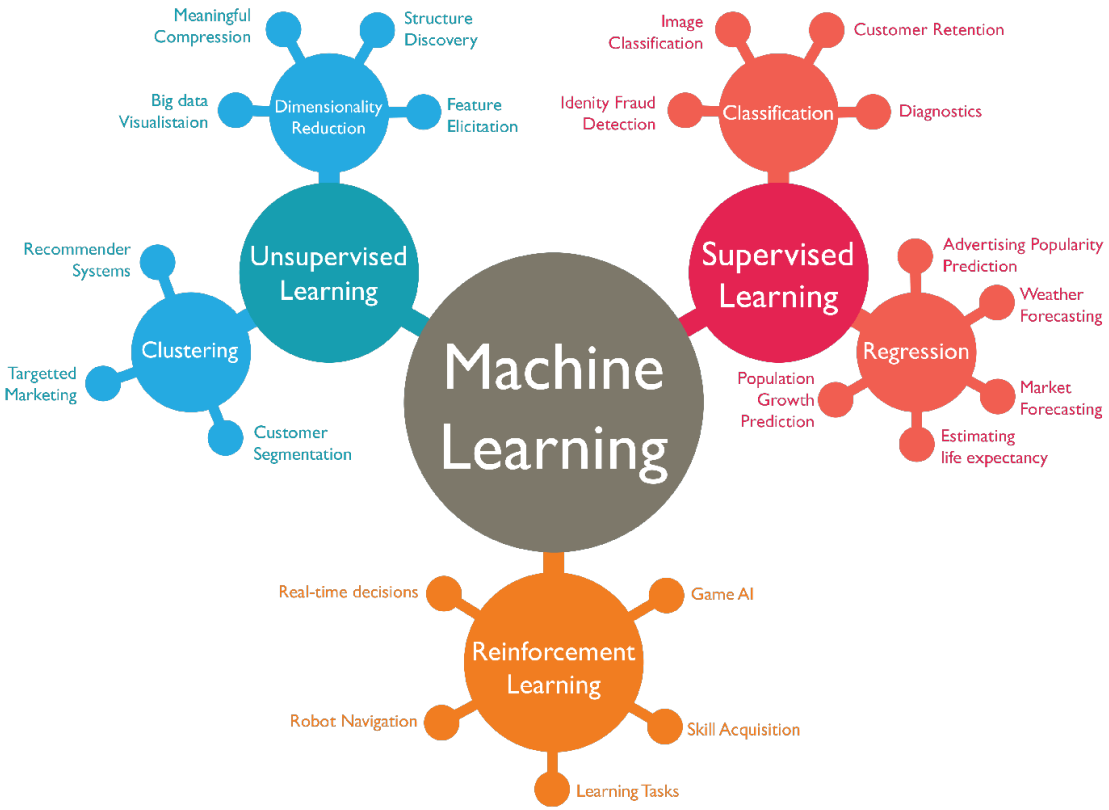


Các phương thức học

Các phương thức học

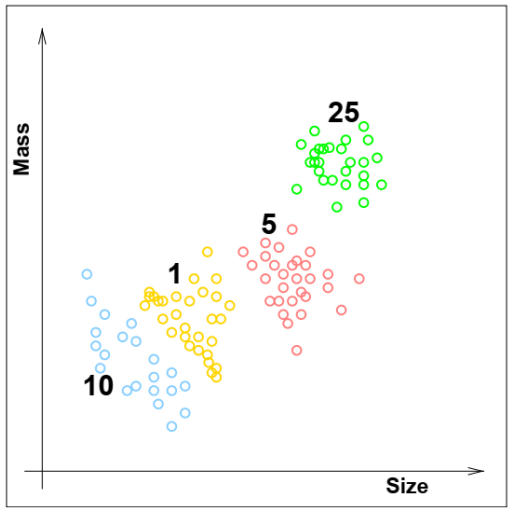
Các thuật toán học máy thường được chia làm 3 nhóm:

- Học có giám sát
- Học không giám sát
- Học tăng cường



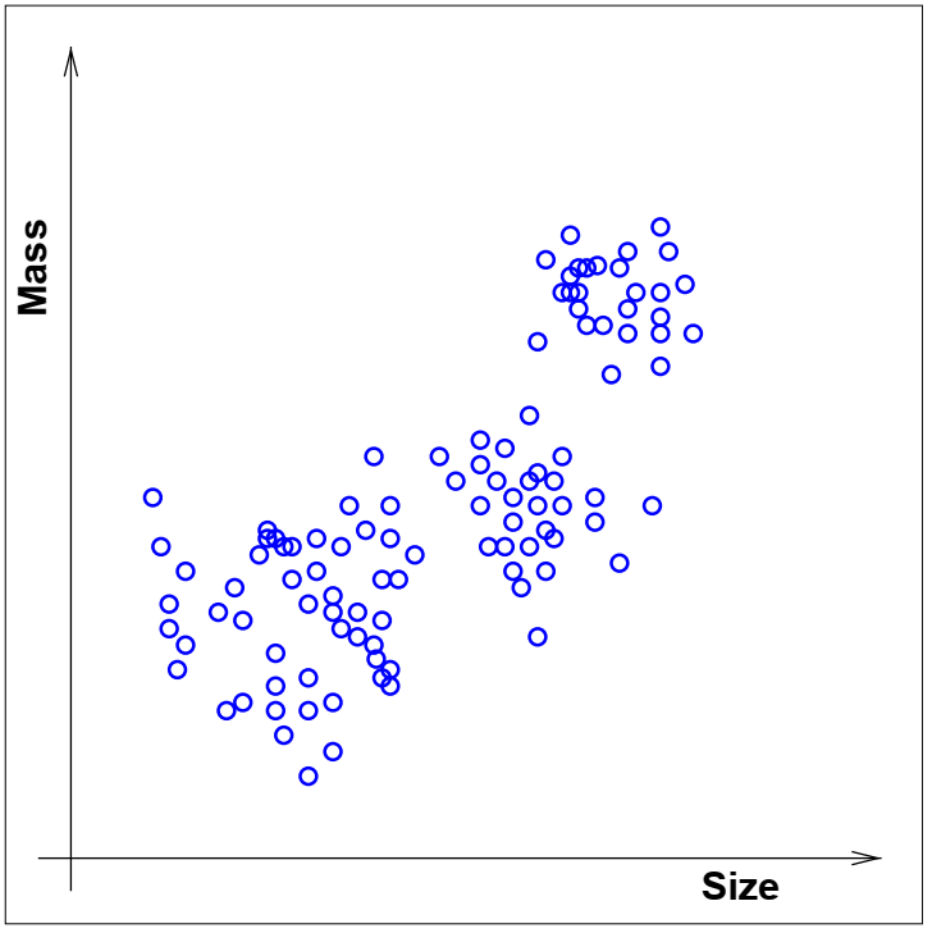
Học có giám sát

- Ta có tập dữ liệu $\mathcal{D}$ gồm: **(input, output)**
  - Nếu **output** là các giá trị hữu hạn, bài toán học được gọi là bài toán **phân lớp**.
  - Nếu **output** là các giá trị liên tục, bài toán học được gọi là bài toán **hồi quy**.



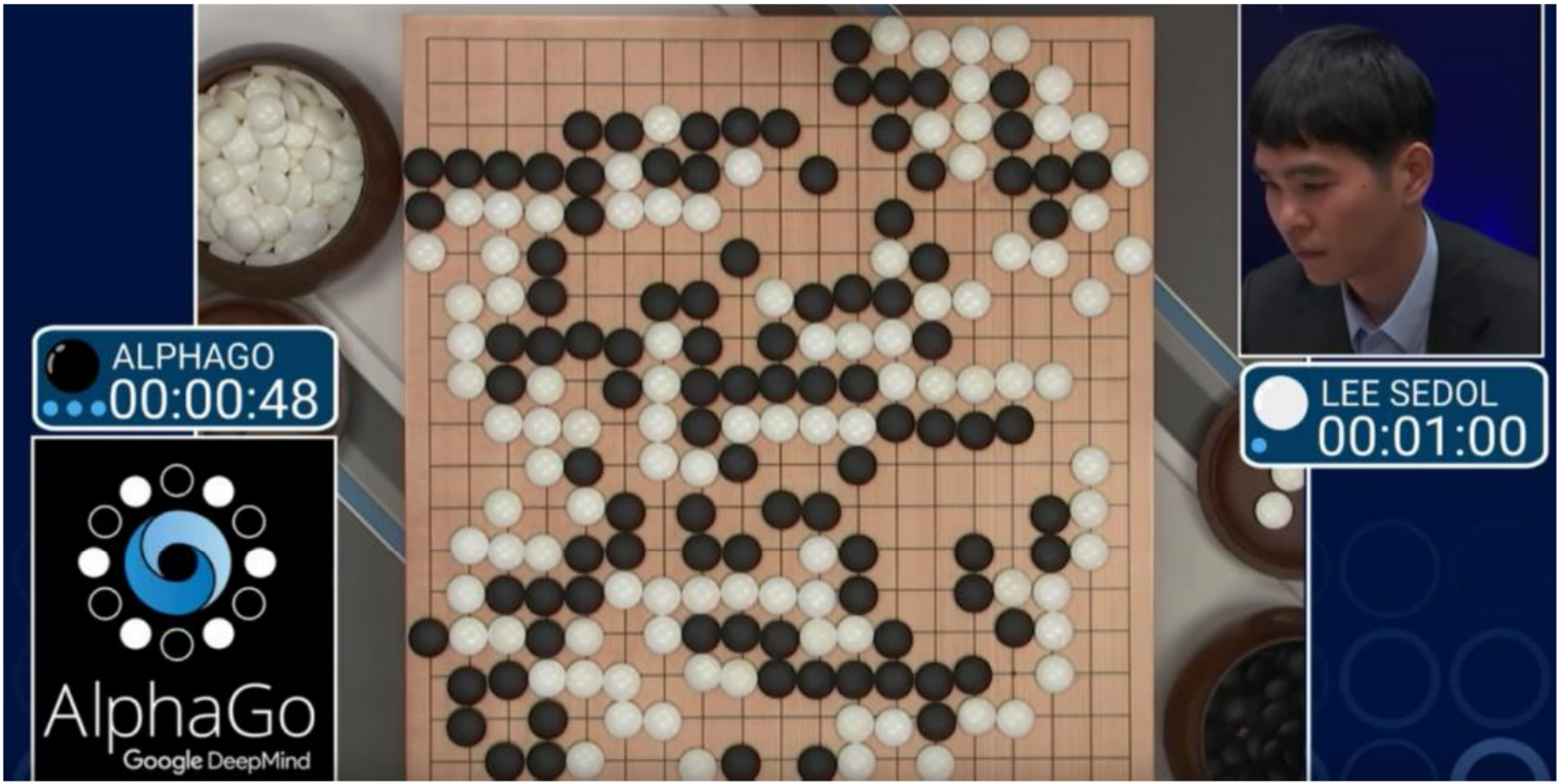
Học không giám sát

Thay vì **(input, output)**, ta có **(input, ?)**



Học tăng cường

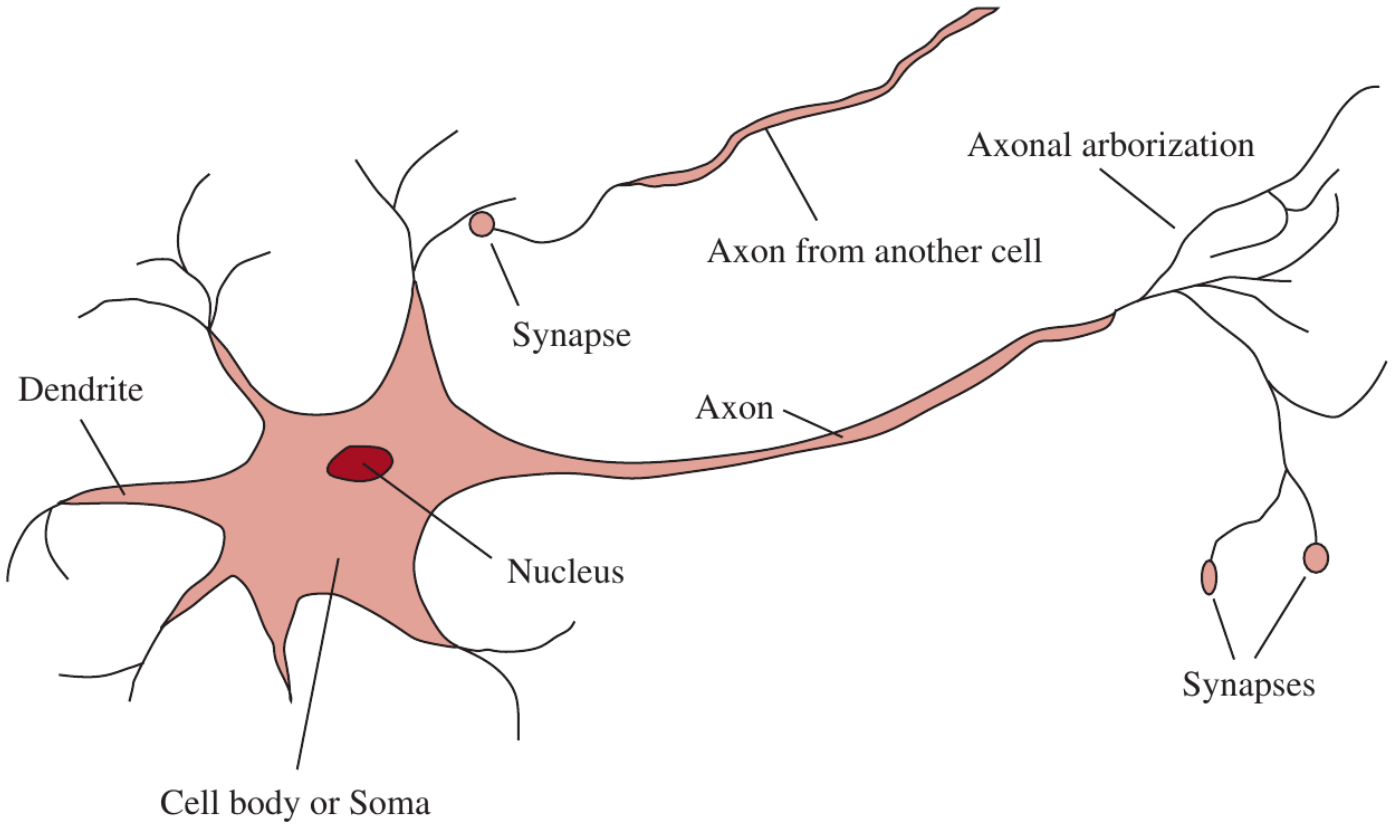
Thay vì **(input, output)**, ta có **(input, một số output, điểm)**



Mạng neuron nhân tạo

Mạng neuron sinh học

- Neuron: tế bào thần kinh, đơn vị xử lý thông tin cơ sở của bộ não.
- Thành phần: Soma, Synapse, Dedrites, Axon



Mạng neuron sinh học

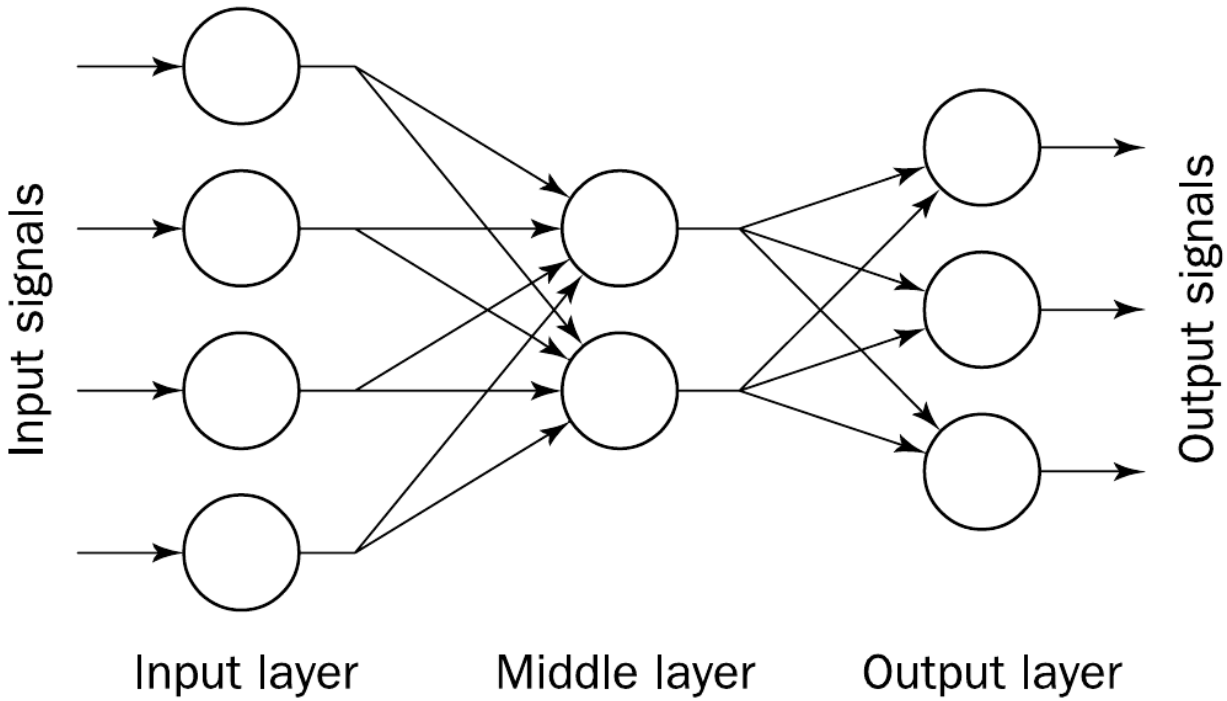
- Bộ não người có gần 10 tỉ neuron và 60.000 tỉ kết nối
- Bộ não con người xử lý thông tin song song, xuyên suốt mạng neuron
- Mạng neuron có thể tạo mới hoặc thay đổi trọng số của các kết nối

Mạng neuron sinh học

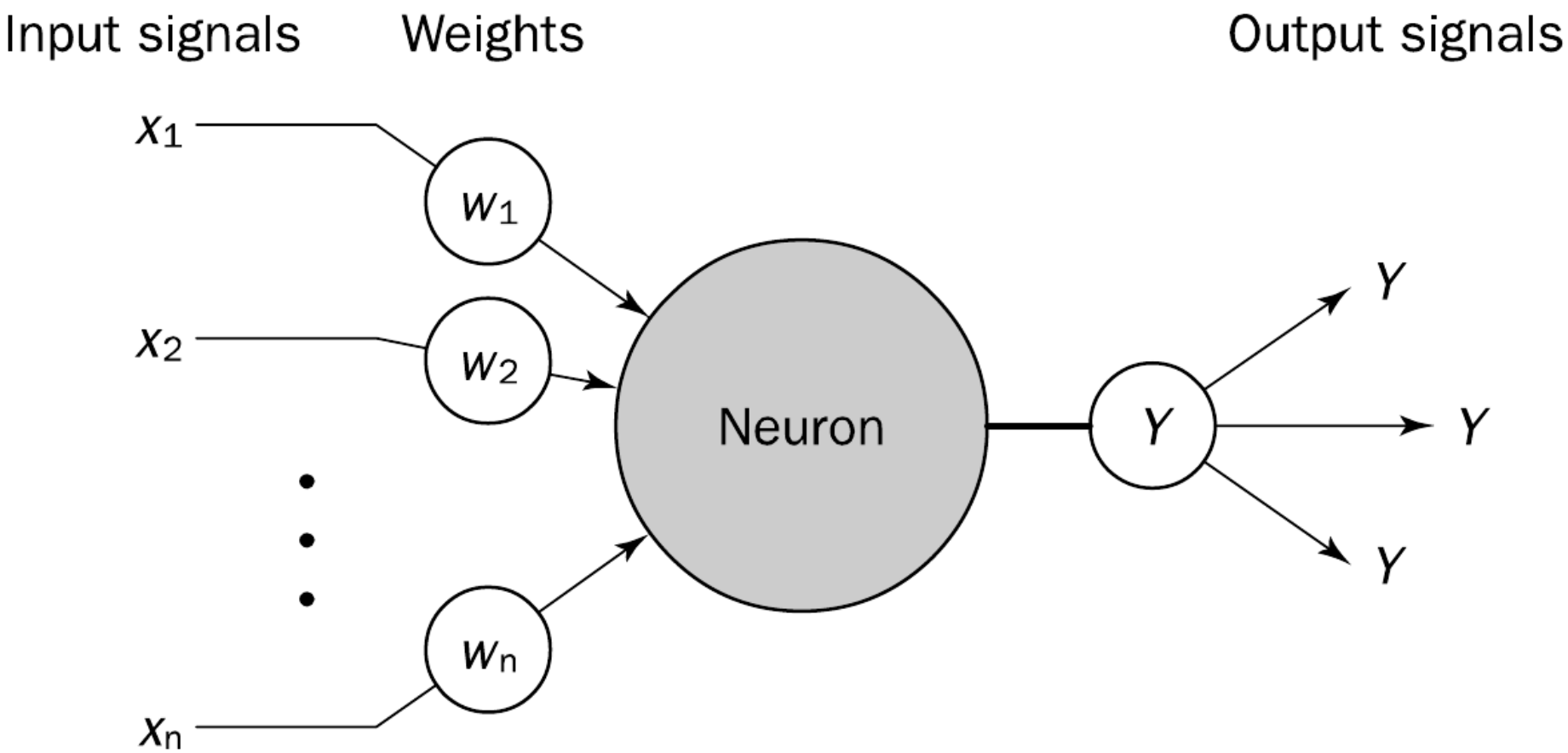
- Sự lan truyền tín hiệu trong mạng neuron thông qua các phản ứng điện hóa phức tạp
  - Các chất hóa học tiết ra từ synapses làm thay đổi điện năng trong soma.
  - Khi điện năng vượt một ngưỡng thì gây ra một xung điện gửi xuống axon.
  - Lan truyền đến các neuron khác.

Mạng neuron nhân tạo

- Mô phỏng mạng neuron sinh học.
- Giải quyết các bài toán dựa vào việc học và kinh nghiệm.
- Quá trình học là quá trình điều chỉnh trọng số.



Mạng neuron nhân tạo



Mạng neuron nhân tạo

- Nhược điểm
  - Phải huấn luyện
  - Kiến trúc mạng neuron khác kiến trúc vi xử lý hiện nay nên cần phải giả lập.
  - Mạng neuron lớn cần nhiều thời gian xử lý

Một số mô hình học đơn giản

Perceptron

- Tập giả định $\mathcal{H}$:

$$y \approx \hat{y} = h_{\mathbf{w}}(\mathbf{x}) = \sum_{i=1}^{d} w_i x_i + w_0$$

- Ta có thuộc tính nhân tạo $x_0 = 1$:

$$h_{\mathbf{w}}(\mathbf{x}) = \sum_{i=0}^{d} w_i x_i$$

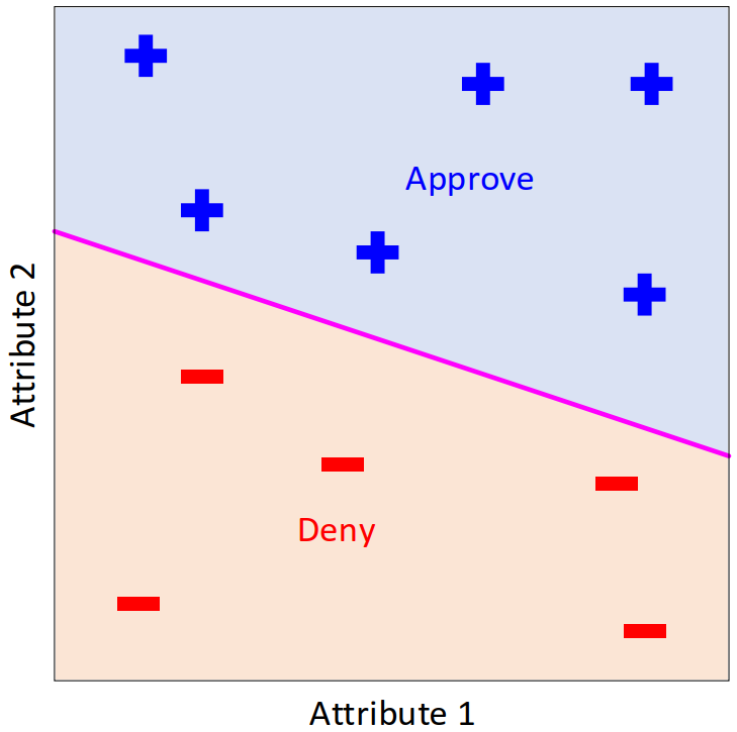
- Tập giả định dưới dạng vector:

$$h_{\mathbf{w}}(\mathbf{x}) = sign\left(\mathbf{w}^{\mathsf{T}}\mathbf{x}\right)$$

Perceptron

Quay lại với bài toán phê duyệt tín dụng:

- **Ranh giới phê duyệt:** đường thẳng
- **Phân vùng phê duyệt:** chấp thuật và không chấp thuận.



Perceptron

Thuật toán học

- Tập dữ liệu huấn luyện $\mathcal{D}_{\text{train}}$ ký hiệu $(\mathbf{X}, \mathbf{y})$ chứa N mẫu dữ liệu $(\mathbf{x}_1, y_1), (\mathbf{x}_2, y_2), ...(\mathbf{x}_n, y_n)$
- Lựa một điểm phân lớp sai $(\mathbf{x}_i, y_i)$:

$$sign\left(\mathbf{w}^{\mathsf{T}}\mathbf{x}_i\right) \neq y_i$$

- Cập nhật vector trọng số:

$$\mathbf{w} \leftarrow \mathbf{w} + y_i\mathbf{x}_i$$

Perceptron

Thuật toán học

- Tập dữ liệu huấn luyện $\mathcal{D}_{\text{train}}$ ký hiệu $(\mathbf{X}, \mathbf{y})$ chứa N mẫu dữ liệu $(\mathbf{x}_1, y_1), (\mathbf{x}_2, y_2), ...(\mathbf{x}_n, y_n)$
- Lựa một điểm phân lớp sai $(\mathbf{x}_i, y_i)$:

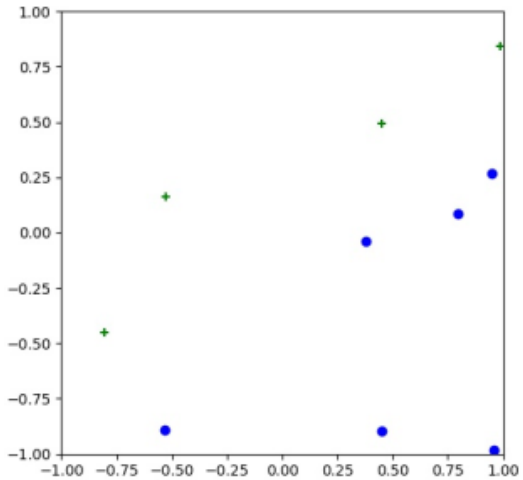
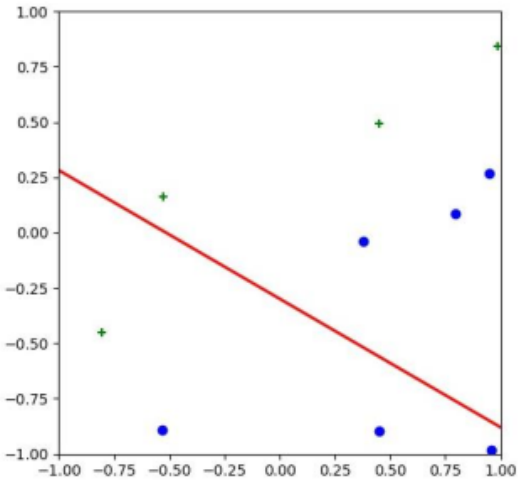
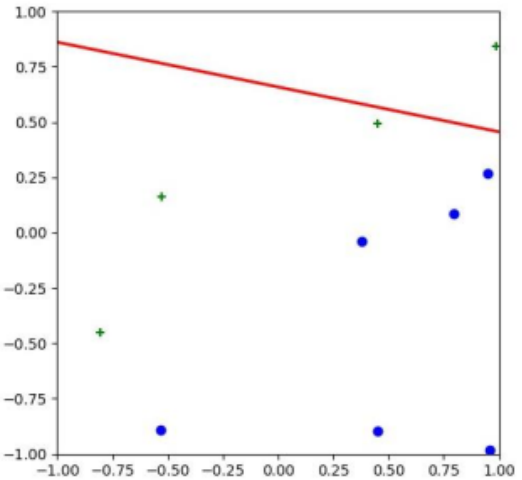
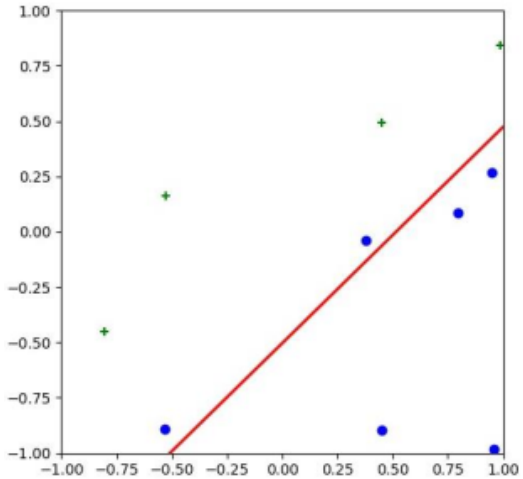
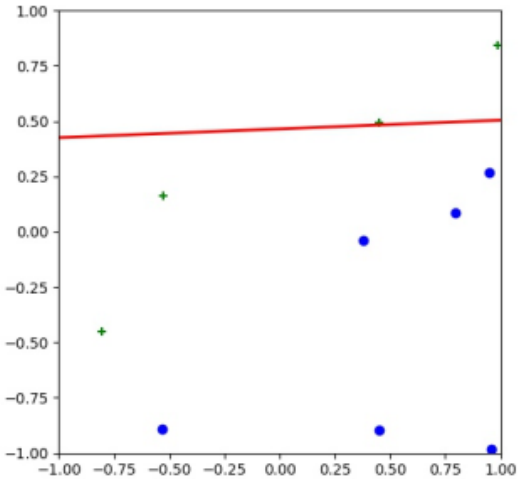
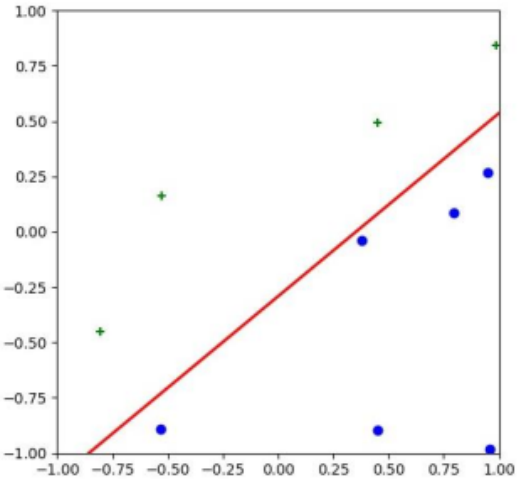
$$sign\left(\mathbf{w}^{\mathsf{T}}\mathbf{x}_i\right) \neq y_i$$

- Cập nhật vector trọng số:

$$\mathbf{w} \leftarrow \mathbf{w} + y_i\mathbf{x}_i$$

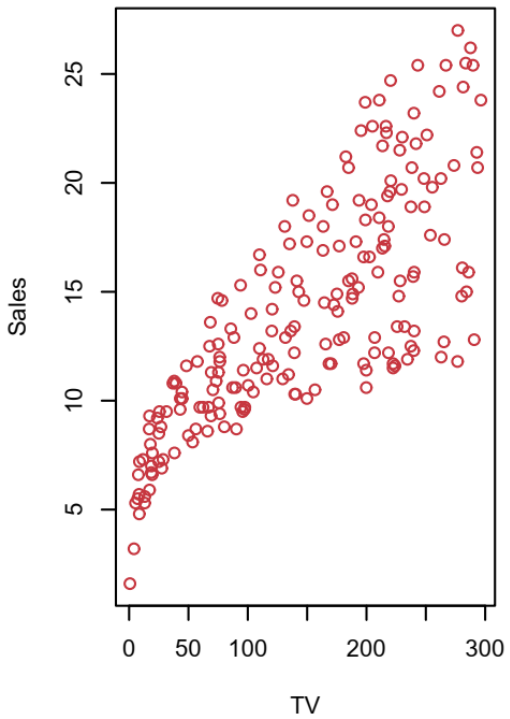
Perceptron là thuật toán học?

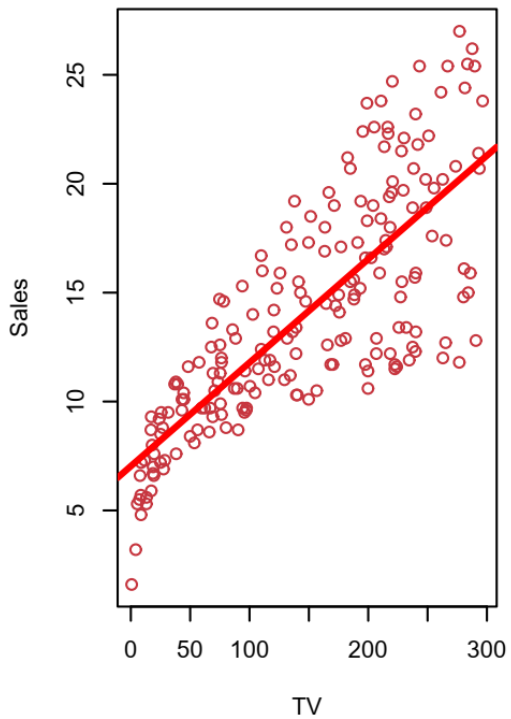
Perceptron là thuật toán học?

Bài toán

Xét một tập dữ liệu $\mathcal{D}$ chứa doanh số bán hàng của 200 cửa hàng cùng với ngân sách quảng cáo của mỗi cửa hàng trên TV. Tìm mối liên hệ giữa doanh số bán hàng và ngân sách quảng cáo trên TV.





Hồi quy tuyến tính

- Yêu cầu ở đây là xây dựng một hệ thống **dự đoán** giá trị $y \in \mathbb{R}$ từ input $x \in \mathbb{R}^{D+1}$
- Tập giả định $\mathcal{H}$:

$$y \approx \hat{y} = h_{\mathbf{w}}(\mathbf{x}) = \mathbf{w}^{\mathsf{T}}\mathbf{x}$$

với $\hat{y}$ là giá trị mô hình dự đoán và $\mathbf{w} \in \mathbb{R}^{D+1}$ là vector tham số của mô hình.

Hồi quy tuyến tính

- **Đánh giá mô hình**:

  Trung bình bình phương độ lỗi (MSE) của một mô hình trên tập dữ liệu huấn luyện $\mathcal{D}_{\text{train}}$ ký hiệu $(\mathbf{X}, \mathbf{y})$ chứa N mẫu dữ liệu $(\mathbf{x}_1, y_1), (\mathbf{x}_2, y_2), ...(\mathbf{x}_n, y_n)$:

$$MSE_{\text{train}} = \frac{1}{N} \|\hat{\mathbf{y}} - \mathbf{y}\|_2^2$$

Hồi quy tuyến tính

$$\underbrace{\mathbf{X} = \begin{bmatrix} \mathbf{x}_1^T \\ \mathbf{x}_2^T \\ \vdots \\ \mathbf{x}_n^T \end{bmatrix}}_{\text{ma trận đầu vào}}, \underbrace{\mathbf{y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}}_{\text{vector mục tiêu}}, \underbrace{\hat{\mathbf{y}} = \begin{bmatrix} \hat{y}_1 \\ \hat{y}_2 \\ \vdots \\ \hat{y}_n \end{bmatrix}}_{\text{vector đầu ra}}$$

- **Mục tiêu học**: tìm ra vector tham số $\mathbf{w}$ sao cho

$$\mathbf{w} = argmin_{\mathbf{w}}(MSE_{\text{train}})$$

Hồi quy tuyến tính

Lời giải

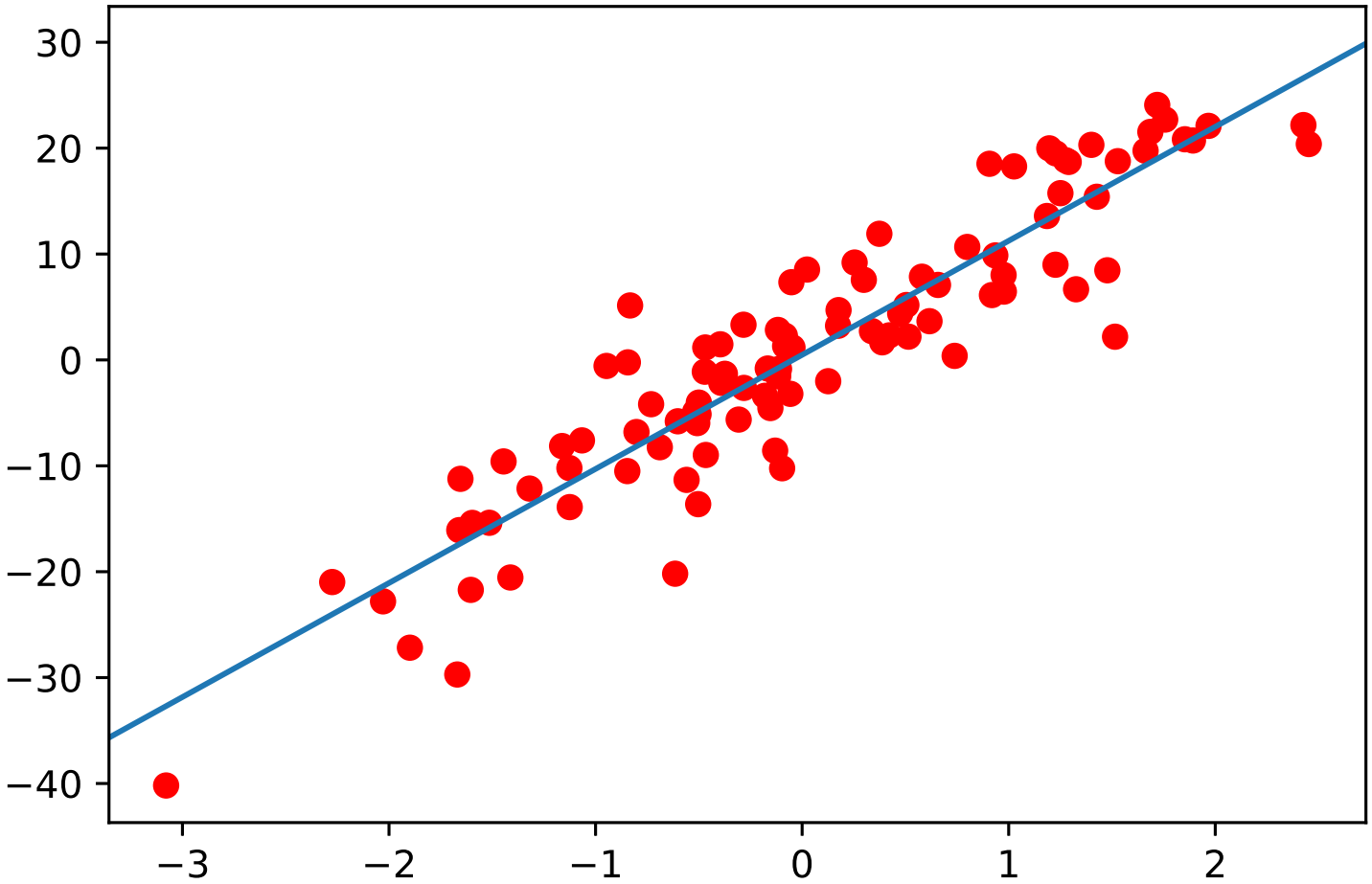
- Tính gradient của MSE_{train}:

$$\nabla_{\mathbf{w}}(MSE_{\text{train}}) = \nabla_{\mathbf{w}}\left(\mathbf{w}^{\mathsf{T}}\mathbf{X}^{\mathsf{T}}\mathbf{X}\mathbf{w} - \mathbf{w}^{\mathsf{T}}\mathbf{X}^{\mathsf{T}}\mathbf{y} - \mathbf{y}^{\mathsf{T}}\mathbf{X}\mathbf{w} + \mathbf{y}^{\mathsf{T}}\mathbf{y}\right)$$
$$= 2\mathbf{X}^{\mathsf{T}}\mathbf{X}\mathbf{w} - 2\mathbf{X}^{\mathsf{T}}\mathbf{y}$$

- MSE_{train} đạt được giá trị cực tiểu khi:

$$\nabla_{\mathbf{w}}(MSE_{\text{train}}) = 0$$
$$\mathbf{w} = \left(\mathbf{X}^{\mathsf{T}}\mathbf{X}\right)^{\dagger}\mathbf{X}^{\mathsf{T}}\mathbf{y}$$

Demo



Hồi quy logistic

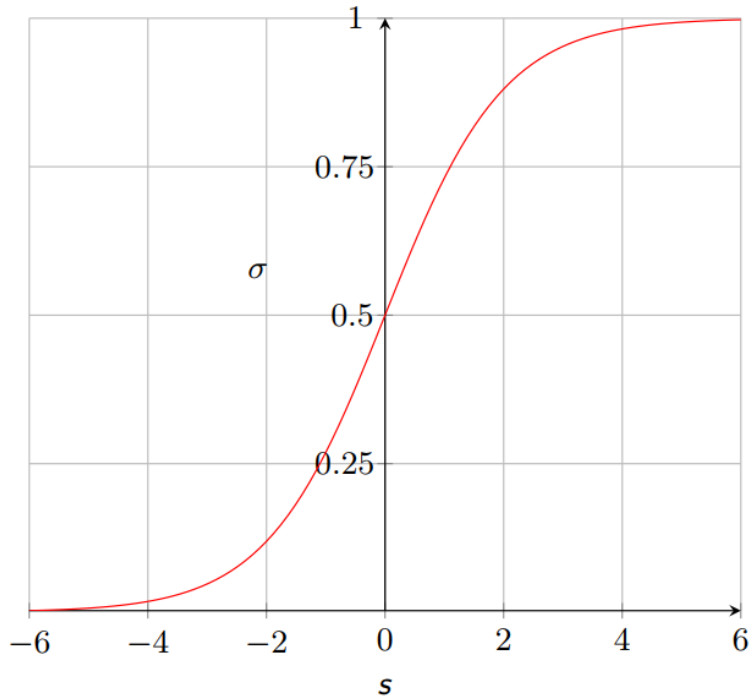
Hàm logistic

$$\sigma(s) = \frac{1}{1 + e^{-s}}$$

Đặc điểm

$$\sigma(-s) = 1 - \sigma(s)$$

$$\sigma'(s) = \sigma(s)(1 - \sigma(s))$$



Hồi quy logisitic

Phát biểu bài toán

- Hàm mục tiêu f là một phân phối xác suất

$$f : \mathbb{R}^D \to [0, 1]$$

- Tập giả định $h_{\mathbf{w}}(\mathbf{x}) = \sigma(\mathbf{w}^{\mathsf{T}}\mathbf{x})$ và phân phối xác suất có điều kiện:

$$P(y|\mathbf{x}, \mathbf{w}) = \begin{cases} h_{\mathbf{w}}(\mathbf{x}) & \text{nếu } y = 1 \\ 1 - h_{\mathbf{w}}(\mathbf{x}) & \text{nếu } y = 0 \end{cases}$$

Hồi quy logisitic

Đánh giá mô hình

- Likelihood của $\mathcal{D} = \{(\mathbf{x}_1, y_1), (\mathbf{x}_2, y_2), ...(\mathbf{x}_n, y_n)\}$ là:

$$\prod_{n=1}^{N} P(y_n|\mathbf{x}_n, \mathbf{w})$$

- Ước lượng cực đại likelihood:

$$\text{Maximize } \prod_{n=1}^{N} P\left(y_n|\mathbf{x}_n, \mathbf{w}\right)$$

$$\Leftrightarrow \text{Minimize } -log \prod_{n=1}^{N} P\left(y_n|\mathbf{x}_n, \mathbf{w}\right)$$

Hồi quy logisitic

Đánh giá mô hình

- Đánh giá lỗi:

$$E(h_{\mathbf{w}}) = -\sum_{n=1}^{N} (y_n log(h_{\mathbf{w}}(\mathbf{x}_n)) + (1 - y_n) log(1 - h_{\mathbf{w}}(\mathbf{x_n})))$$

- **Mục tiêu học:** tối thiểu hóa $E(h_{\mathbf{w}})$
- Vậy làm cách nào để tối thiểu hóa $E(h_{\mathbf{w}})$????

Hồi quy logisitic

Entropy: $J(w) = -(ylog(z) + (1-y)log(1-z))$

Chain rule: $\frac{\partial J(w)}{\partial w} = \frac{\partial J(w)}{\partial z}\frac{\partial z}{\partial h}\frac{\partial h}{\partial w}$

$$\frac{\partial J(w)}{\partial z} = -\left(\frac{y}{z} - \frac{1-y}{1-z}\right) = \frac{z-y}{z(1-z)}$$

$$\frac{\partial z}{\partial h} = z(1-z), \frac{\partial h}{\partial w} = X \rightarrow \frac{\partial J(w)}{\partial w} = X^T(z-y)$$

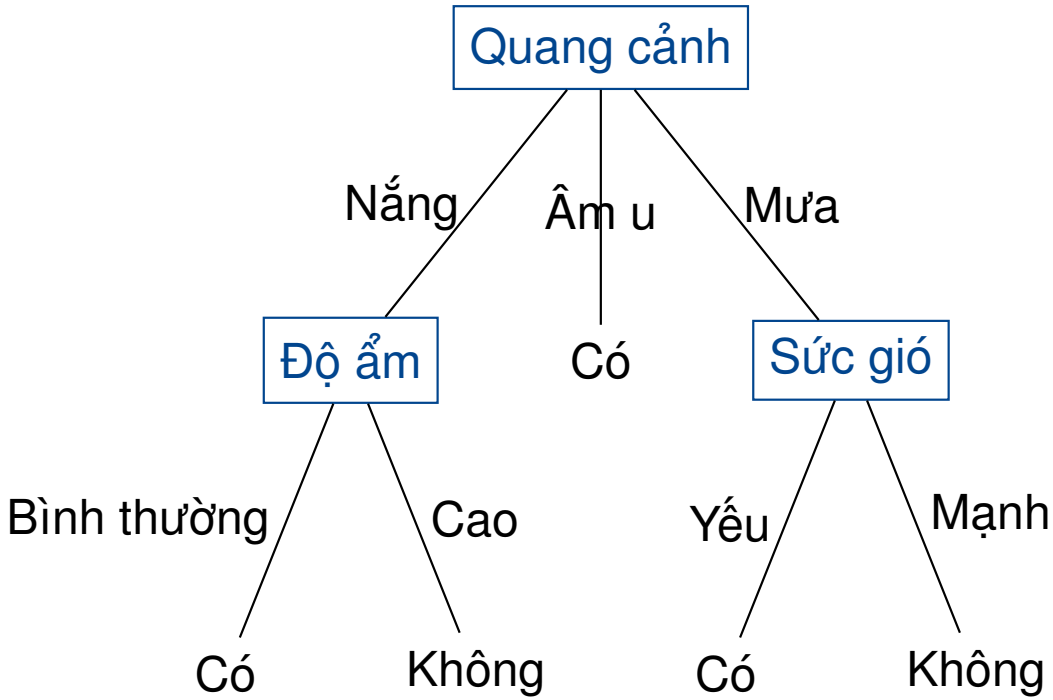
Cây quyết định

- Sử dụng cho bài toán phân lớp
- Giá trị thuộc tính là các giá trị rời rạc

Quang cảnh	Nhiệt độ	Độ ẩm	Sức gió	Đi chơi?
Nắng	Nóng	Cao	Yếu	Không
Nắng	Nóng	Cao	Mạnh	Không
Âm u	Nóng	Cao	Yếu	Có
Mưa	Ấm	Cao	Yếu	Có
Mưa	Mát	Bình thường	Yếu	Có
...	...	...	...	...

Biểu diễn cây quyết định

- Mỗi nút nội là một phép thử thuộc tính
- Mỗi nhánh tương ứng với một giá trị thuộc tính
- Mỗi nút lá là một giá trị phân lớp



Nguyên lý Occam's Razor

- Ta nên lựa chọn các thuộc tính theo thứ tự như thế nào?

Nguyên lý Occam's Razor

"The simplest model that fits the data is also the most plausible."

Tạm dịch: Mô hình đơn giản nhất có thể khớp với dữ liệu cũng là mô hình khả thi nhất.

Nguyên lý Occam's Razor

- Ta nên lựa chọn các thuộc tính theo thứ tự như thế nào?

Nguyên lý Occam's Razor

"The simplest model that fits the data is also the most plausible."

Tạm dịch: Mô hình đơn giản nhất có thể khớp với dữ liệu cũng là mô hình khả thi nhất.

- Cây càng nhỏ càng tốt $\rightarrow$ Thuộc tính có giá trị thông tin càng cao càng gần gốc.

Information Gain I

- Gọi S là tập các thể hiện dữ liệu với n lớp và p_i là xác suất thể hiện dữ liệu được gán nhãn i trong S
- **Entropy** ký hiệu $E(S)$ đo sự không thuần nhất dữ liệu trong S

$$E(S) = -\sum_{i=1}^{n} p_i \log_2 p_i$$

- Trung bình entropy của thuộc tính A

$$AE(S, A) = \sum_{v \in Values(A)} \frac{|S_v|}{|S|} E(S_v)$$

Information Gain II

- Information Gain là mức entropy kỳ vọng của S khi chia dữ liệu theo thuộc tính A

$$Gain(S, A) = E(S) - AE(S, A)$$

- Chọn thuộc tính có information gain cao nhất

Gini index

- Độ đo bất thuần nhất **Gini**:

$$G(S) = 1 - \sum_{i=1}^{C} p_i^2$$

- Gini index khi chia dữ liệu S theo thuộc tính A

$$G(S, A) = \sum_{v \in Values(A)} \frac{|S_v|}{|S|} G(S_v)$$

- Chọn thuộc tính có gini index thấp nhất

Xây dựng cây

Các bước xây dựng cây quyết định:

1. Chọn thuộc tính quyết định tốt nhất A
2. Phân chia dữ liệu dựa trên giá trị thuộc tính ($Values(A)$)
3. Phân loại các thể hiện dữ liệu của từng phân vùng
4. Nếu phân loại được toàn bộ dữ liệu, kết thúc xây dựng cây. Ngược lại quay lại bước 1 cho các phân vùng dữ liệu.

Tài liệu tham khảo I

- Goodfellow, I. and Bengio, Y. and Courville, A.

  Deep learning..

  MIT Press, 2016.

- Russell, S. and Norvig, P.

  Artificial intelligence: a modern approach.

  Pearson Education Limited, 2016.

- Michael, N.

  Artificial Intelligence: A Guide to Intelligent Systems.

  Pearson Education Limited, 2005.