

Nhận dạng

Tuần 3: Các kỹ thuật phân lớp mẫu
thông dụng

Lê Hoàng Thái



KHOA CÔNG NGHỆ THÔNG TIN
TRƯỜNG ĐẠI HỌC KHOA HỌC TỰ NHIÊN

Nhận dạng mẫu là gì?

“Gán một đối tượng cụ thể hoặc sự kiện về một bộ phân loại đã được xác định trước” -- Duda & Hart

- Mẫu (**pattern**) là một đối tượng, qui trình hoặc sự kiện đã được gắn liền với một cái tên cho trước.
- Lớp mẫu (**pattern class** (or category)) là một tập các mẫu có cùng chung thuộc tính và thường xuất phát từ cùng một nguồn.
- Trong nhận dạng(**recognition**) (hoặc phân lớp (**classification**)), các đối tượng cho trước sẽ được gán về những lớp được xác định trước.
- Một bộ phân lớp (**classifier**) là một máy dùng cho hoạt động phân loại.

Các ví dụ ứng dụng

• Nhận dạng ký tự quang (OCR)

- Chữ viết tay: sắp xếp các ký tự theo mã bưu điện, thiết bị đầu vào cho PDA.
- Văn bản in: máy đọc cho người khiếm thị, số hóa tài liệu văn bản.

• Sinh trắc học (Biometrics)

- Nhận dạng, xác thực và truy vấn khuôn mặt.
- Nhận dạng vân tay.
- Nhận dạng giọng nói.

• Các hệ thống chẩn đoán

- Chẩn đoán y khoa: X-Ray, EKG analysis.
- Máy chẩn đoán, phát hiện chất thải.

• Các ứng dụng quân sự

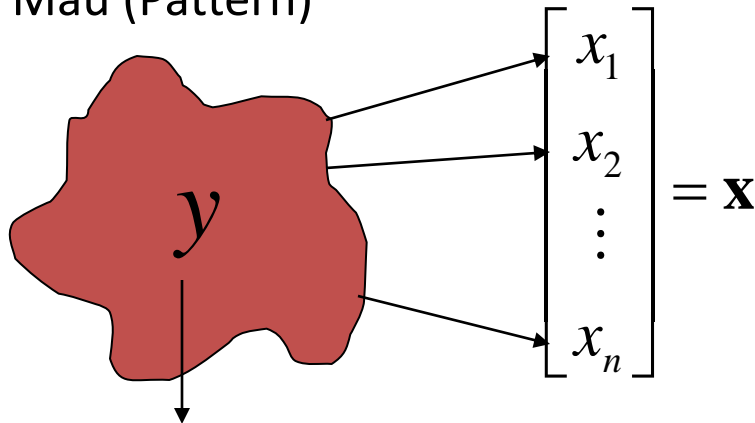
- Tự động xác định mục tiêu (ATR).
- Phân đoạn và phân tích ảnh (nhận dạng từ hình ảnh vệ tinh trên không).

Các phương pháp nhận dạng

- **Nhận dạng mẫu thống kê (Statistical PR):** dựa vào mô hình thống kê tập mẫu cơ bản và các lớp mẫu cho trước.
- **Nhận dạng mẫu theo cấu trúc (hoặc ngữ nghĩa) (Structural (or syntactic) PR):** Các lớp mẫu được biểu diễn bằng kỹ thuật của các cấu trúc chính thức như văn phạm (grammars), automata, strings, vv....
- **Mạng nơron nhân tạo (Neural networks):** Bộ phân lớp được biểu diễn là một mạng các tế bào mô hình hóa các nơron trong bộ não con người (cách tiếp cận nối kết).

Các khái niệm căn bản

Mẫu (Pattern)



Đặc trưng (Feature vector) $\mathbf{x} \in X$

- Một vector quan sát (được đo lường).
- \mathbf{x} là một điểm trong không gian X

Trạng thái ẩn (Hidden state) $y \in Y$

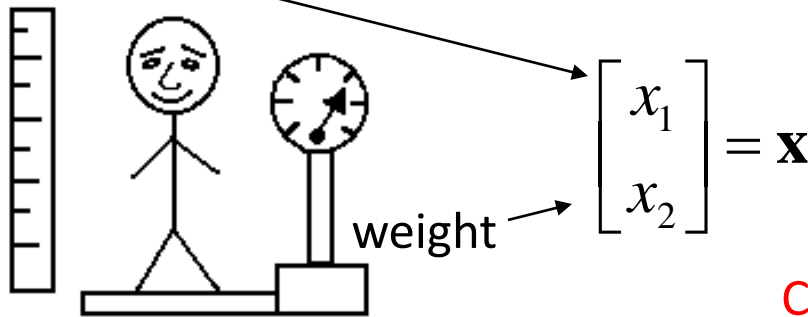
- Không đo lường trực tiếp được.
- Các mẫu với cùng trạng thái ẩn sẽ thuộc về cùng một lớp.

Nhiệm vụ (Task)

- Thiết kế một bộ phân lớp (decision rule) $q: X \rightarrow Y$
- Nó quyết định Trạng thái ẩn dựa trên quan sát.

Ví dụ

height



Yêu cầu (Task): Nhận dạng jockey-hoopster.

The set of hidden state is $Y = \{H, J\}$

The feature space is $X = \mathbb{R}^2$

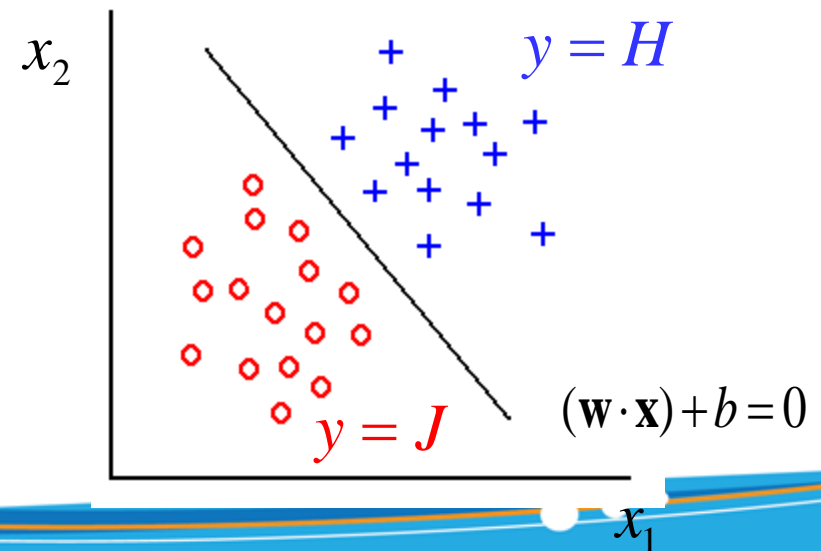
Các mẫu huấn luyện

(Training examples) $\{(\mathbf{x}_1, y_1), \dots, (\mathbf{x}_l, y_l)\}$

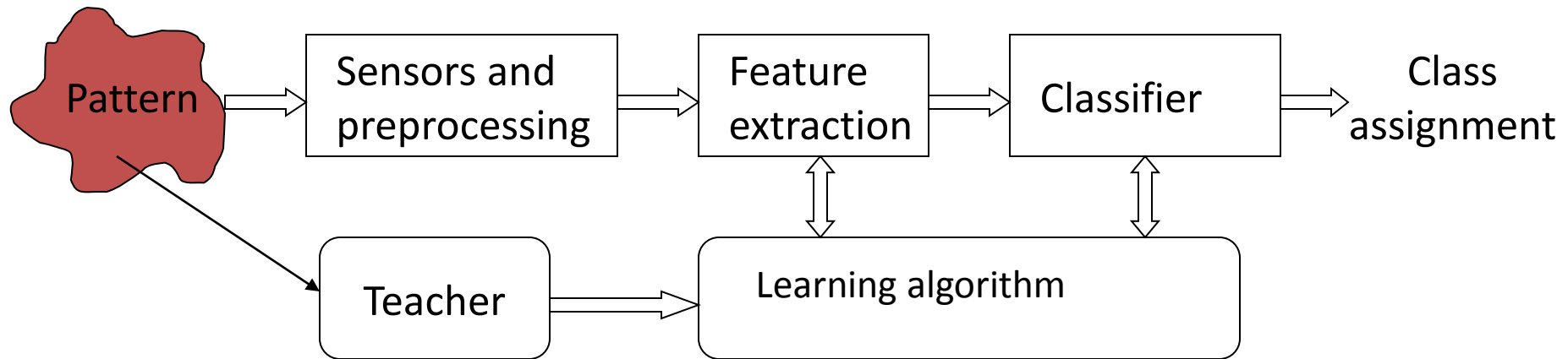
Bộ phân lớp tuyến tính

(Linear classifier):

$$q(\mathbf{x}) = \begin{cases} H & \text{if } (\mathbf{w} \cdot \mathbf{x}) + b \geq 0 \\ J & \text{if } (\mathbf{w} \cdot \mathbf{x}) + b < 0 \end{cases}$$



Các thành phần của hệ thống nhận dạng mẫu (PR)



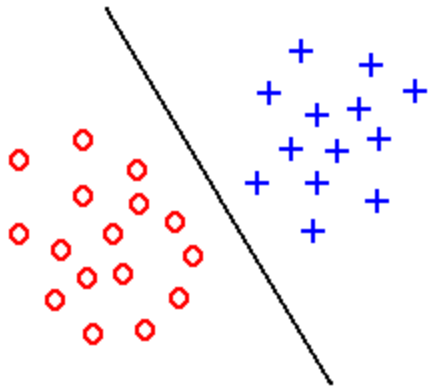
- **Sensors and preprocessing (Cảm biến và tiền xử lý).**
- **A feature extraction** tạo ra những đặc trưng tách lớp tốt cho việc phân lớp mẫu.
- **A classifier (Một bộ phân lớp).**
- **A teacher** cung cấp thông tin về Trạng thái ẩn (hidden state) -- Học có giám sát (supervised learning).
- **A learning algorithm** thiết lập PR từ tập mẫu huấn luyện.

Trích chọn đặc trưng

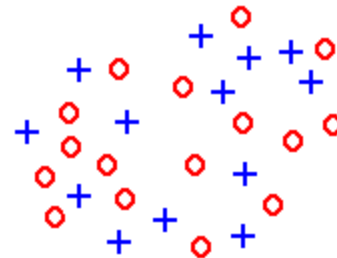
Nhiệm vụ (Task): trích chọn đặc trưng tốt cho phân lớp.

Các đặc trưng tốt:

- Các đối tượng thuộc cùng một lớp có các giá trị đặc trưng tương tự.
- Các đối tượng từ các lớp khác nhau có các giá trị đặc trưng khác.



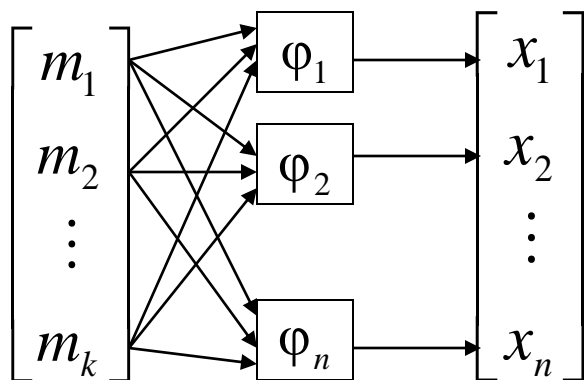
Các đặc trưng Tốt
“Good” features



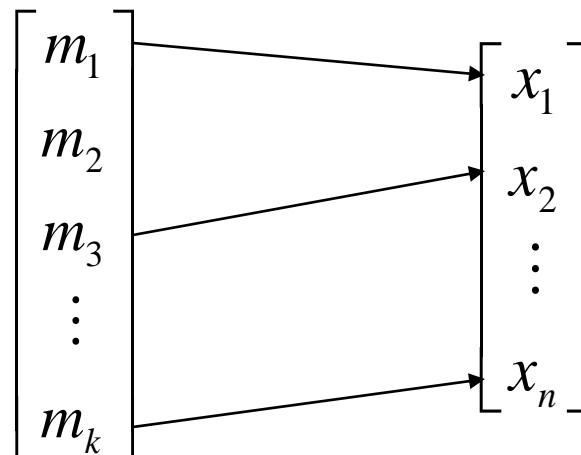
Các đặc trưng Tồi
“Bad” features

Các phương pháp trích chọn đặc trưng

Rút trích (Feature extraction)



Chọn lọc (Feature selection)



Vấn đề cần nhấn mạnh là xác định bộ tham số tối ưu của hàm rút trích Đ Trưng $\varphi(\theta)$.

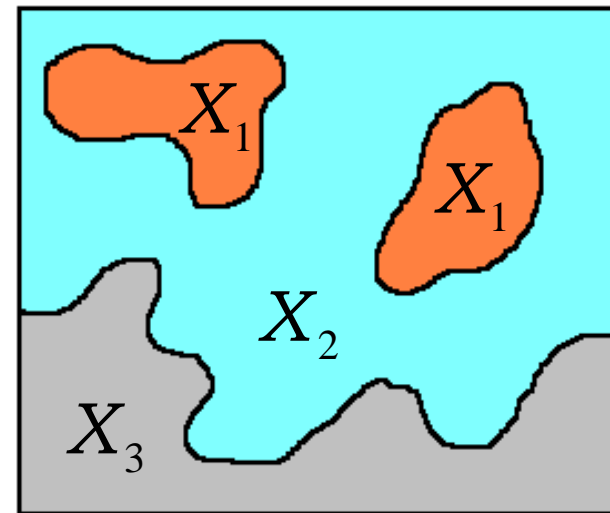
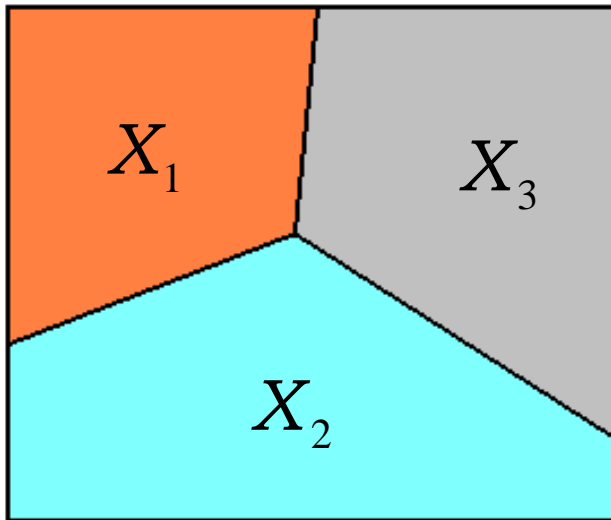
Các phương pháp học có giám sát: Hàm mục tiêu là tiêu chí tách lớp của các mẫu đã gán nhãn (discriminability). Ví dụ, Phân tích tách lớp tuyến tính (linear discriminat analysis (LDA)).

Các phương pháp học không giám sát: Biểu diễn số chiều dữ liệu thấp nhưng vẫn bảo đảm các đặc tính quan trọng của dữ liệu nhập. Ví dụ, Phân tích thành phần chính (principal component analysis (PCA)).

Bộ phân lớp

Một bộ phân lớp sẽ phân chia không gian đặc trưng X về các vùng lớp đã gán nhãn (**class-labeled regions**) sao cho

$$X = X_1 \cup X_2 \cup \dots \cup X_{|Y|} \quad \text{and} \quad X_1 \cap X_2 \cap \dots \cap X_{|Y|} = \{0\}$$



Việc phân lớp bao gồm việc xác định vùng, ở đó, vector đặc trưng \mathbf{x} thuộc về.

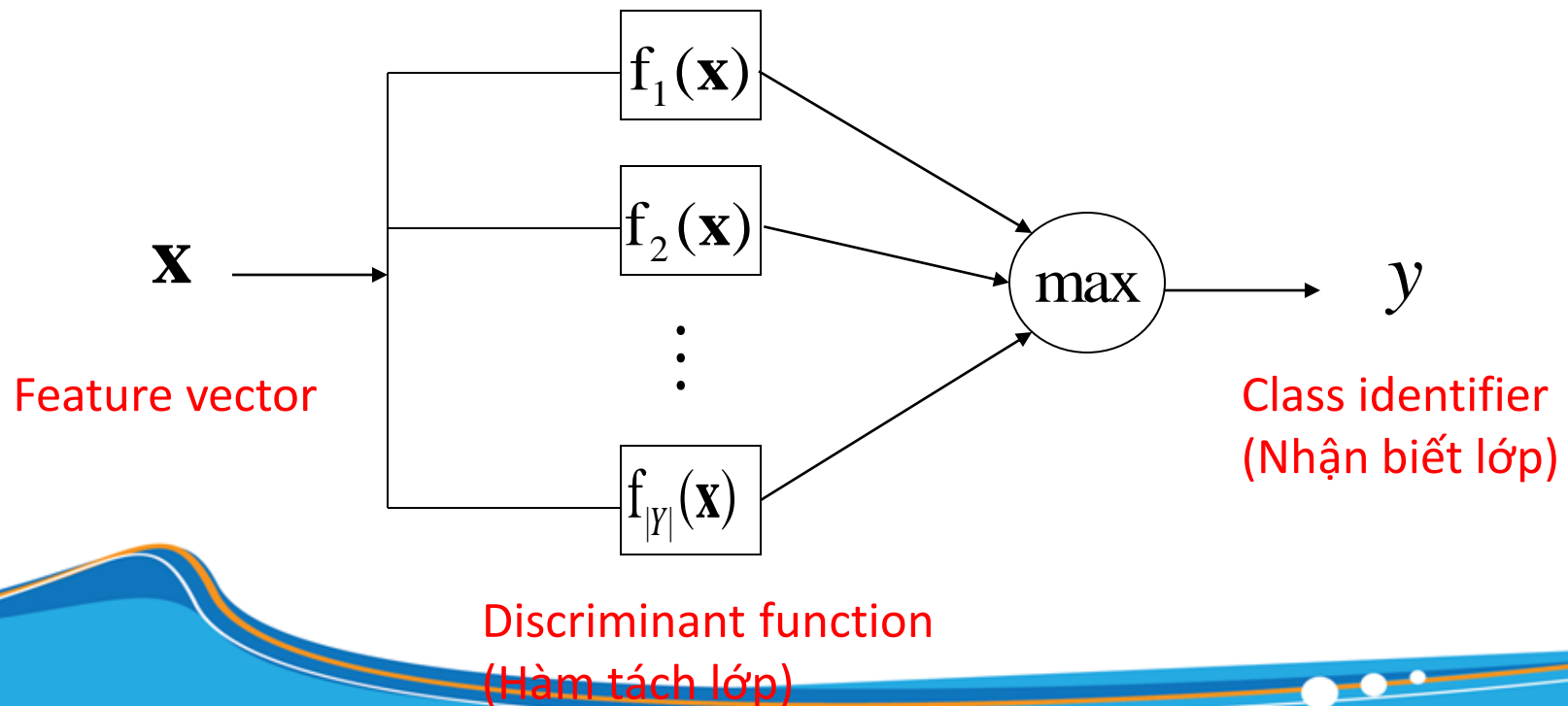
Khoanh vùng giữa các biên quyết định (**decision boundaries**) được gọi là các vùng quyết định.

Cách biểu diễn của bộ phân lớp

Một bộ phân lớp điển hình được biểu diễn như một tập các hàm tách lớp:

$$f_i(\mathbf{x}) : X \rightarrow \mathcal{R}, i = 1, \dots, |Y|$$

Một bộ phân lớp gán một feature vector \mathbf{x} đến lớp i nếu $f_i(\mathbf{x}) > f_j(\mathbf{x}) \quad \forall j \neq i$



Hình thành quyết định Bayesian (Bayesian decision making)

- Hình thành quyết định Bayesian là một phương pháp thống kê căn bản cho phép thiết kế một bộ phân lớp tối ưu nếu mô hình thống kê đầy đủ được biết (**statistical model is known**).

| | | | | |
|--------------------|---------------|-----|---------------------|--------------------------------|
| <u>Definition:</u> | Observations | X | A loss function | $W : Y \times D \rightarrow R$ |
| | Hidden states | Y | A decision rule | $q : X \rightarrow D$ |
| | Decisions | D | A joint probability | $p(\mathbf{x}, y)$ |

Task: thiết kế luật quyết định q để cực tiểu độ rủi ro Bayesian (to design decision rule q which minimizes Bayesian risk)

$$R(q) = \sum_{y \in Y} \sum_{x \in X} p(\mathbf{x}, y) W(q(\mathbf{x}), y)$$

Ví dụ cho công việc Bayesian

Task: Cực tiểu hóa lỗi phân lớp (minimization of classification error).


Một tập các quyết định D bằng với tập các trạng thái ẩn Y (hidden states Y).

0/1 - loss function used

$$W(q(\mathbf{x}), y) = \begin{cases} 0 & \text{if } q(\mathbf{x}) = y \\ 1 & \text{if } q(\mathbf{x}) \neq y \end{cases}$$

Độ rủi ro $R(q)$ (Bayesian risk $R(q)$) tương ứng với xác suất trượt lớp (misclassification).

Giải pháp của Bayesian là

$$q^* = \arg \min_q R(q) \Rightarrow y^* = \arg \max_y p(y | \mathbf{x}) = \arg \max_y \frac{p(\mathbf{x} | y) p(y)}{p(\mathbf{x})}$$


Giới hạn của phương pháp Bayesian

- Mô hình thống kê $p(\mathbf{x}, y)$ hầu như không được biết, bởi vậy, việc học phải được thực hiện để ước lượng $p(\mathbf{x}, y)$ từ tập mẫu huấn luyện $\{(\mathbf{x}_1, y_1), \dots, (\mathbf{x}_\ell, y_\ell)\}$ -- **plug-in Bayes**.
- Các phương pháp **Non-Bayesian** đưa ra những cách giải quyết sâu hơn:
 - Chỉ tồn tại một phần của mô hình thống kê:
 - $p(y)$ không biết hoặc không tồn tại.
 - $p(\mathbf{x}|y, \theta)$ chịu ảnh hưởng của một tác động phi ngẫu nhiên θ .
 - Hàm trượt (loss function) không xác định.
 - Ví dụ: Neyman-Pearson's task, Minimax task, vv.

Cách phương pháp tách lớp

Cho trước một lớp các luật phân lớp $q(\mathbf{x};\boldsymbol{\theta})$ được tham số hóa bởi $\boldsymbol{\theta} \in \Xi$ mục tiêu là tìm ra tham số “tối ưu (best)” $\boldsymbol{\theta}^*$ dựa trên tập mẫu huấn luyện $\{(\mathbf{x}_1, y_1), \dots, (\mathbf{x}_\ell, y_\ell)\}$ – học có giám sát (**supervised learning**).

Mục tiêu của quá trình học (**task of learning**): Nhận dạng luật phân lớp nào sẽ được sử dụng.

Thực hiện việc học được định nghĩa bởi việc chọn một nguyên tắc qui nạp (**inductive principle**).



Nguyên tắc giảm thiểu rủi ro thực nghiệm

Rủi ro dự kiến (expected risk $R(q)$) được xấp xỉ bởi rủi ro thực nghiệm (**empirical risk**)

$$R_{\text{emp}}(q(x; \theta)) = \frac{1}{\ell} \sum_{i=1}^{\ell} W(q(\mathbf{x}_i; \theta), y_i)$$

Được tính từ một tập dữ liệu gán nhãn cho trước $\{(\mathbf{x}_1, y_1), \dots, (\mathbf{x}_\ell, y_\ell)\}$.

Việc học dựa trên nguyên tắc cực tiểu thực nghiệm (**empirical minimization principle**) được xác định:

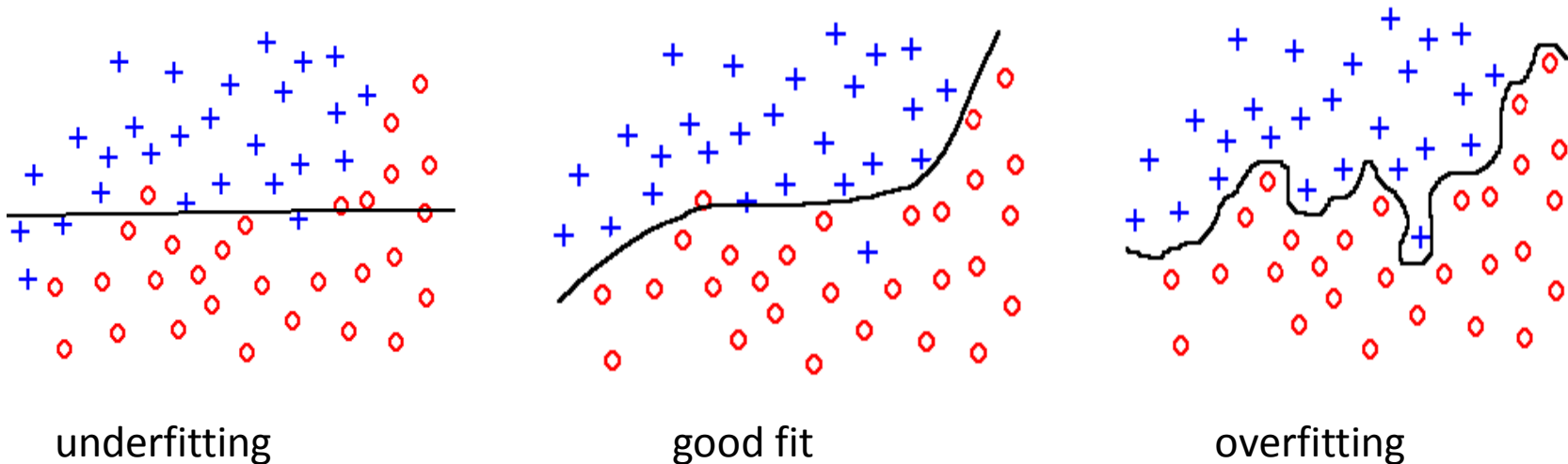
$$\theta^* = \arg \min_{\theta} R_{\text{emp}}(q(\mathbf{x}; \theta))$$

Ví dụ của các thuật toán: **Perceptron**, **Back-propagation**, vv.



Quá khớp và không khớp (Overfitting and underfitting)

Bài toán: Làm thế nào tìm được bộ phân lớp $q(\mathbf{x};\theta)$ để sử dụng.



Vấn đề khái quát hóa: Độ rủi ro thực nghiệm (empirical risk R_{emp}) nhỏ không đảm bảo độ rủi ro dự kiến (expected risk R) cũng là nhỏ.

Nguyên tắc cực tiểu độ rủi ro cấu trúc

Lý thuyết học thống kê -- Vapnik & Chervonenkis.

Cận trên của độ rủi ro dự kiến (expected risk) của một luật phân lớp $q \in Q$

$$R(q) \leq R_{emp}(q) + R_{str}\left(\frac{1}{\ell}, h, \log \frac{1}{\sigma}\right)$$

Trong đó, ℓ là số lượng các mẫu huấn luyện, h là chiều VC (VC-dimension) của lớp các hàm Q và $1-\sigma$ là độ tin cậy của cận trên.

Nguyên tắc SRM : Từ các lớp hàm lồng nhau cho trước Q_1, Q_2, \dots, Q_m , sao cho

$$h_1 \leq h_2 \leq \dots \leq h_m$$

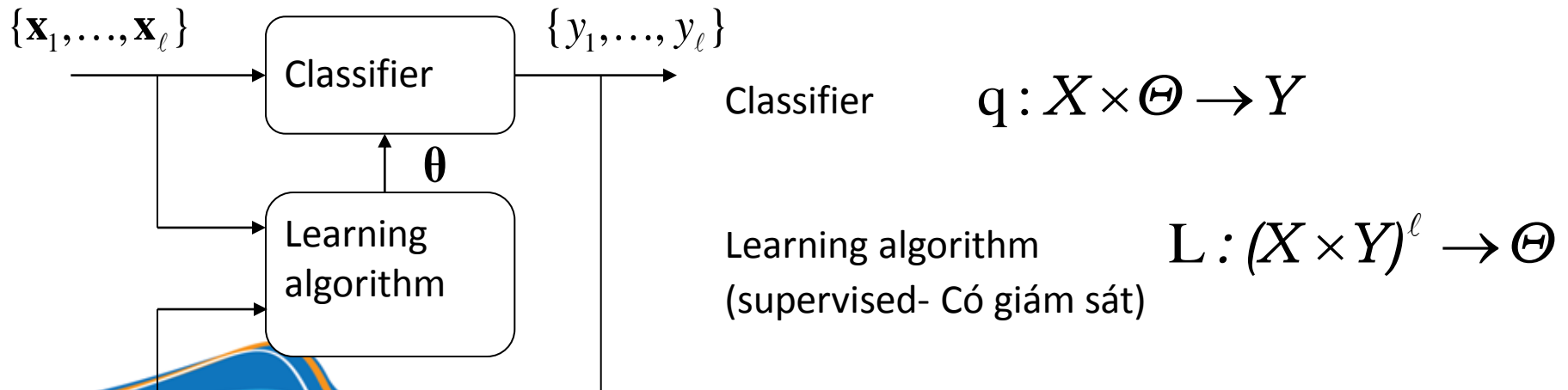
Chọn một luật q^* cực tiểu cận trên của độ rủi ro dự kiến (expected risk).

Học không giám sát

Input: Tập mẫu huấn luyện $\{\mathbf{x}_1, \dots, \mathbf{x}_\ell\}$ không có thông tin về trạng thái ẩn.

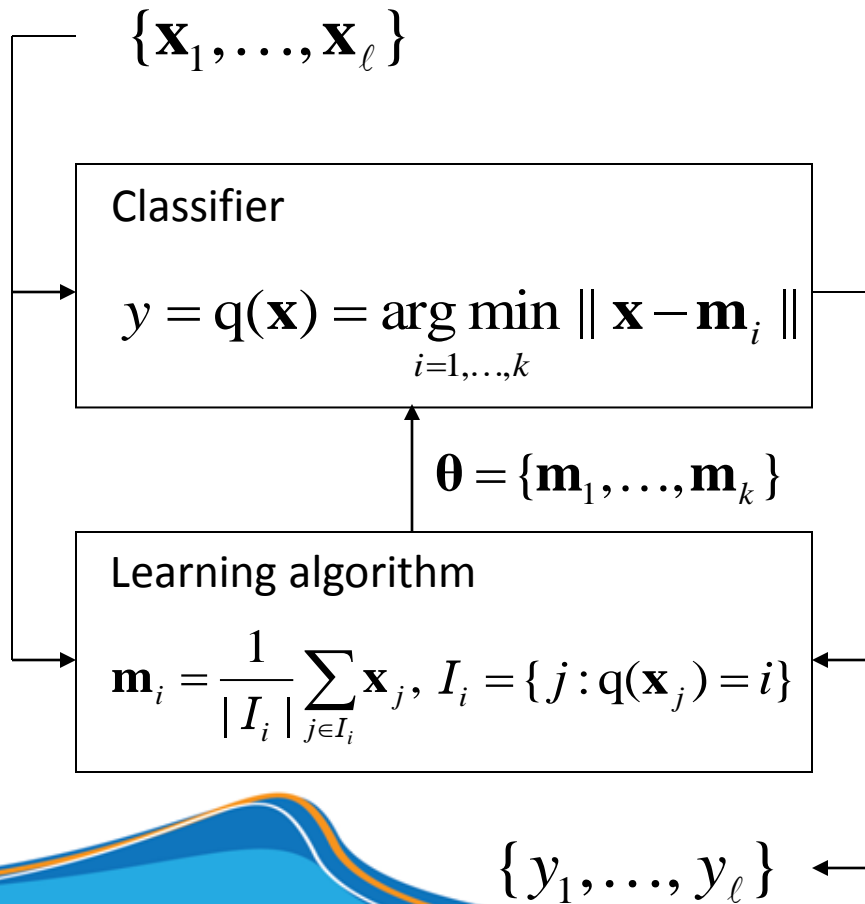
Clustering: Mục tiêu là tìm ra những nhóm dữ liệu có cùng chung tính chất.

Một lớp các thuật toán học không giám sát (**unsupervised learning algorithms**):



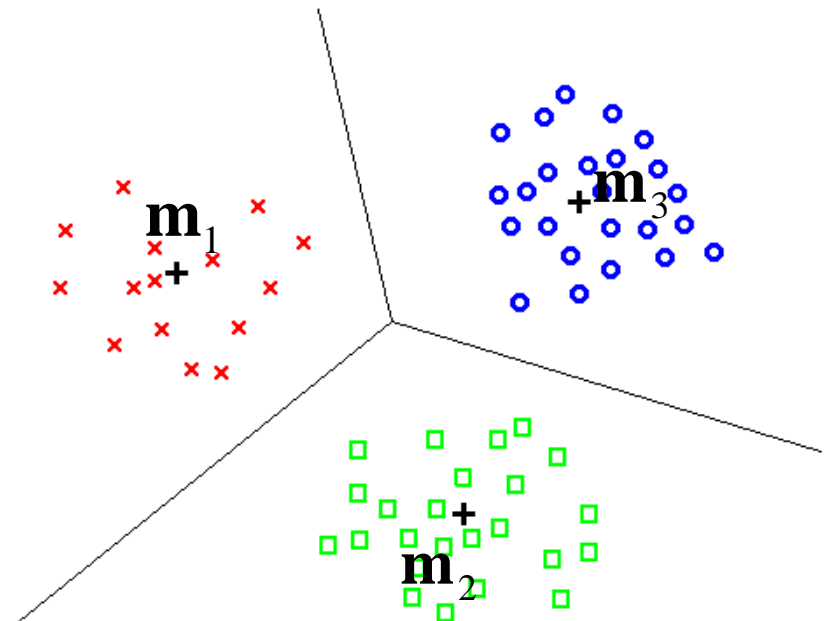
Ví dụ của thuật toán học không giám sát

k-Means clustering(gom cụm k-Means):



Mục tiêu là cực tiểu:

$$\sum_{i=1}^{\ell} \|\mathbf{x}_i - \mathbf{m}_{q(\mathbf{x}_i)}\|^2$$



Tham khảo (References)

Sách (Books)

Duda, Hart: Pattern Classification and Scene Analysis. J. Wiley & Sons, New York, 1982. (2nd edition 2000).

Fukunaga: Introduction to Statistical Pattern Recognition. Academic Press, 1990.

Bishop: Neural Networks for Pattern Recognition. Claredon Press, Oxford, 1997.

Schlesinger, Hlaváč: Ten lectures on statistical and structural pattern recognition. Kluwer Academic Publisher, 2002.

Tạp chí (Journals)

Journal of Pattern Recognition Society.

IEEE transactions on Neural Networks.

Pattern Recognition and Machine Learning.

