HỌC PHẦN: HỌC MÁY 2 (MACHINE LEARNING 2)

Số tín chỉ: 4

Số tiết lý thuyết: 30 tiết

Số tiết thực hành: 30 tiết

Số tiết tự học: 120 tiết

Tài liệu học tập:

[1] David Forsyth (2019). *Applied Machine Learning*. Swithzerland: Springer.

[2] José Unpingco (2019). *Python for Probability, Statistics, and Machine Learning*. Swithzerland: Springer.

Tài liệu khác

[3] Vũ Hữu Tiệp (2018). *Machine Learning Cơ Bản*. Thành phố Hồ Chí Minh: Nhà xuất bản Khoa học và Kỹ thuật.

GV: Quách Hải Thọ

Email: qhaitho@hueuni.edu.vn - Phone: 091.3439186

ĐÁNH GIÁ HỌC PHẦN

- Kiến thức và kỹ năng xây dựng mô hình phân lớp
- Kiến thức và kỹ năng xây dựng mạng nơ-ron nhân tạo
- Kiến thức và kỹ năng xây dựng mô hình phân lớp cây quyết định, SVM và các phương pháp xây dựng mô hình phân lớp dạng kết hợp
- Kiến thức và kỹ năng xây dựng phân cụm trong học máy
- Kiến thức và kỹ năng khai phá luật kết hợp
- Thực hành triển khai mô hình học máy
- Kiến thức và kỹ năng xây dựng mô hình học máy trong dự án cộng tác
- Kết quả đồ án
- Báo cáo đồ án

KẾ HOẠCH GIẢNG DẠY

Chương 1: Mô hình phân lớp và ứng dụng

Chương 2: Mạng nơ-ron nhân tạo

Chương 3: Mô hình cây quyết định, SVM & phương pháp kết hợp

Chương 4: Mô hình phân cụm

Chương 5: Khai phá luật kết hợp

Chương 6: Đồ án

- 6.1. Đồ án 1: Xây dựng mô hình mạng nơ-ron nhân tạo để nhận dạng ảnh
- 6.2. Đồ án 2: Phân cụm văn bản

KẾ HOẠCH THỰC HÀNH

Bài thực hành: Mô hình phân lớp và ứng dụng

Bài thực hành: Mạng nơ-ron nhân tạo

Bài thực hành: Mô hình cây quyết định, SVM & phương pháp kết hợp

Bài thực hành: Mô hình phân cụm

Bài thực hành: Khai phá luật kết hợp

Bài thực hành: Đồ án

Chương 1: Mô hình phân lớp và ứng dụng

1.1. Phân lớp nhị phân và Phân lớp đa nhãn

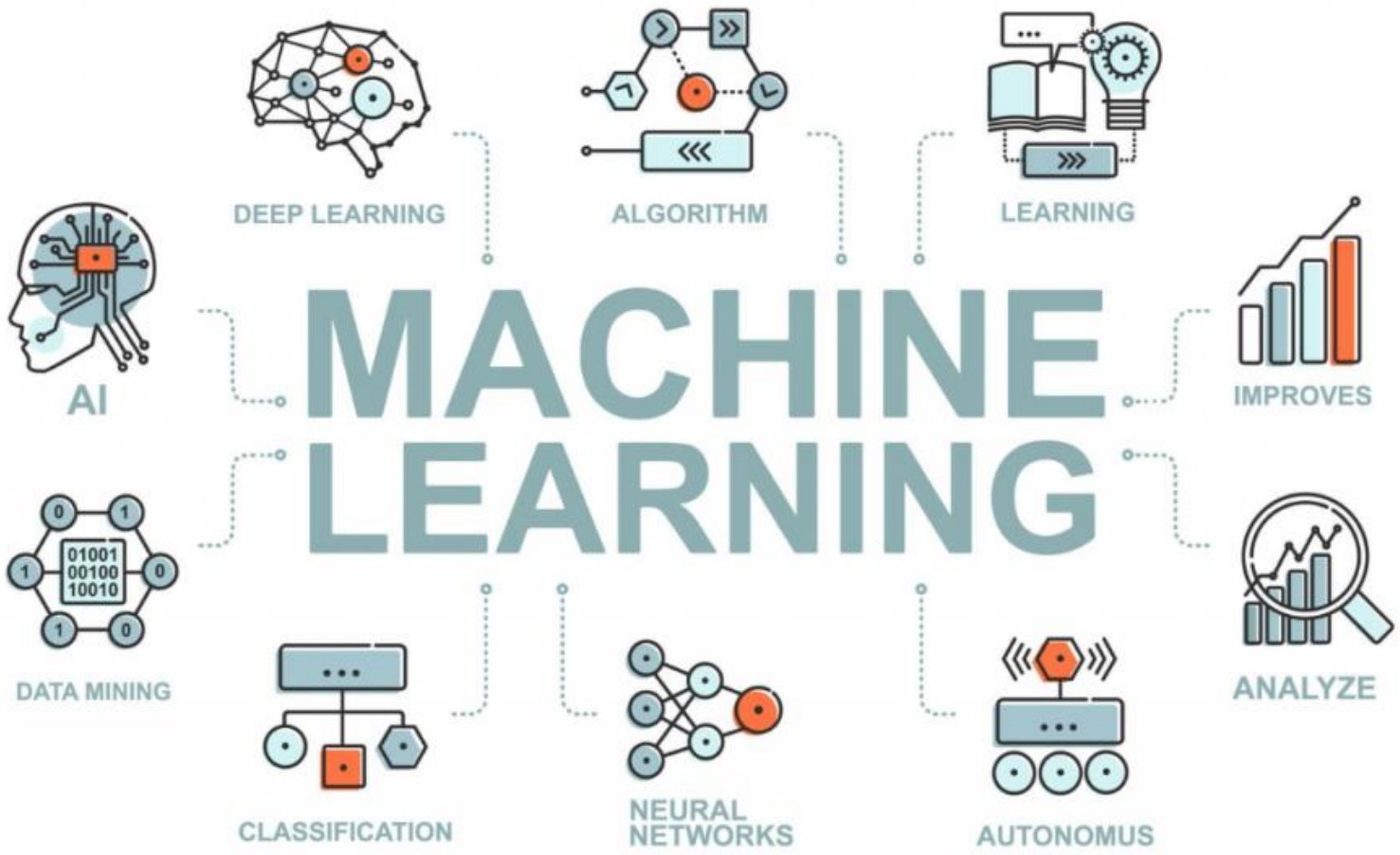
1.2. Hiện tượng overfitting

1.3. Phương pháp xác nhận chéo (cross validation)

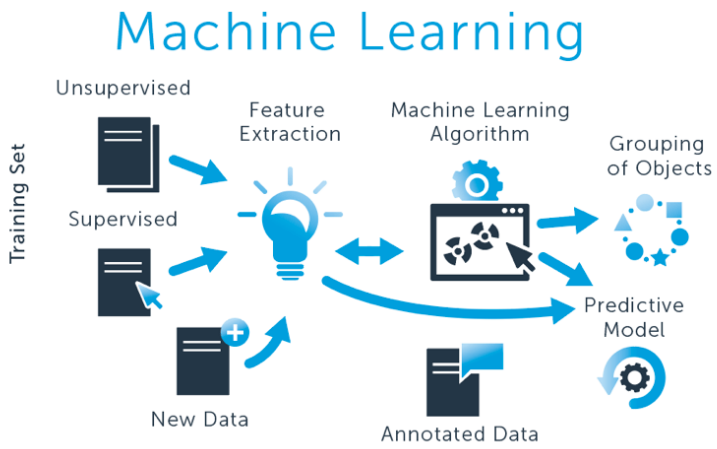
1.4. Đánh giá hiệu năng của mô hình phân lớp

1.5. Ứng dụng của mô hình phân lớp

Khái niệm machine learning



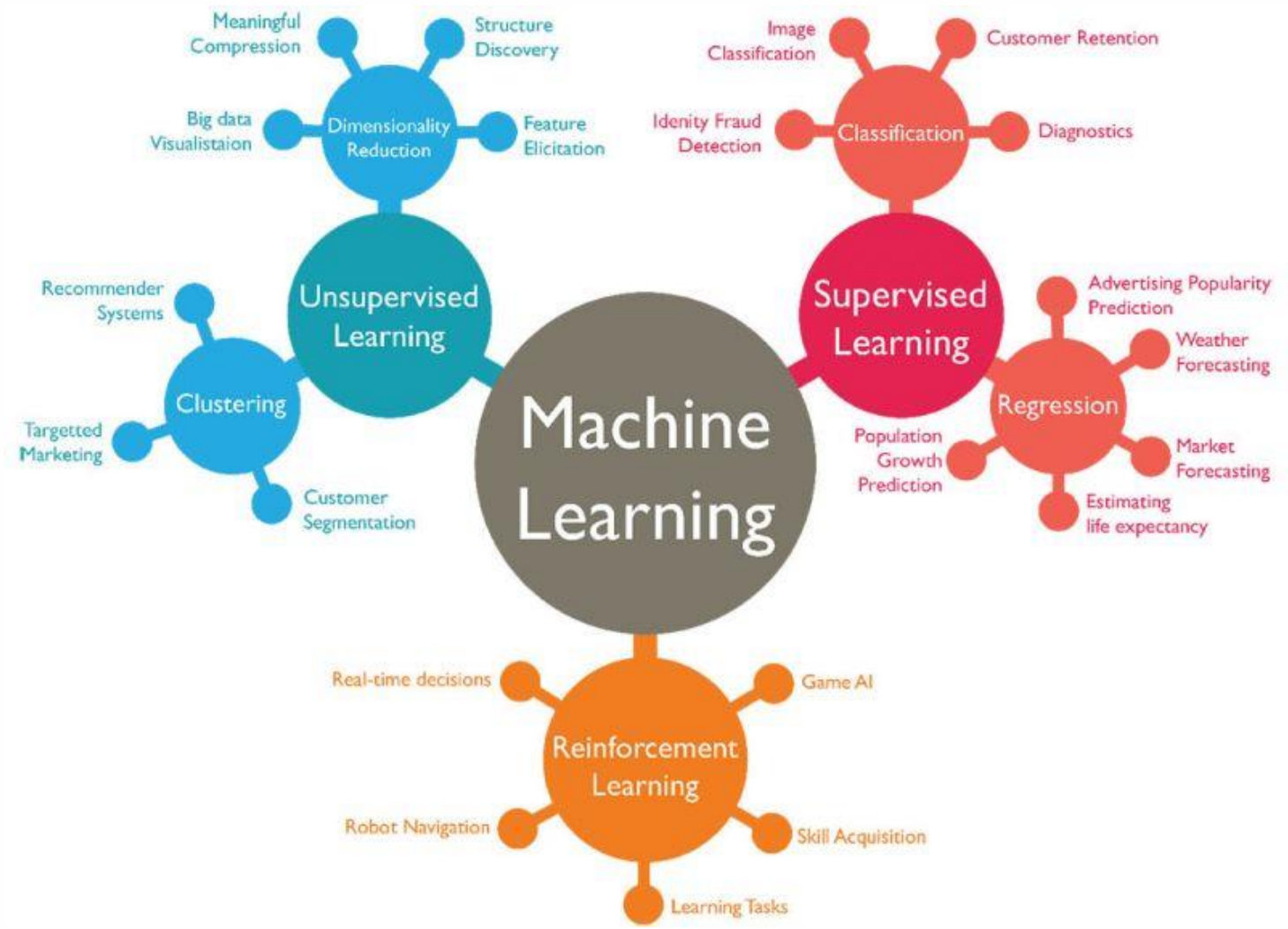
Cách thức hoạt động của machine learning



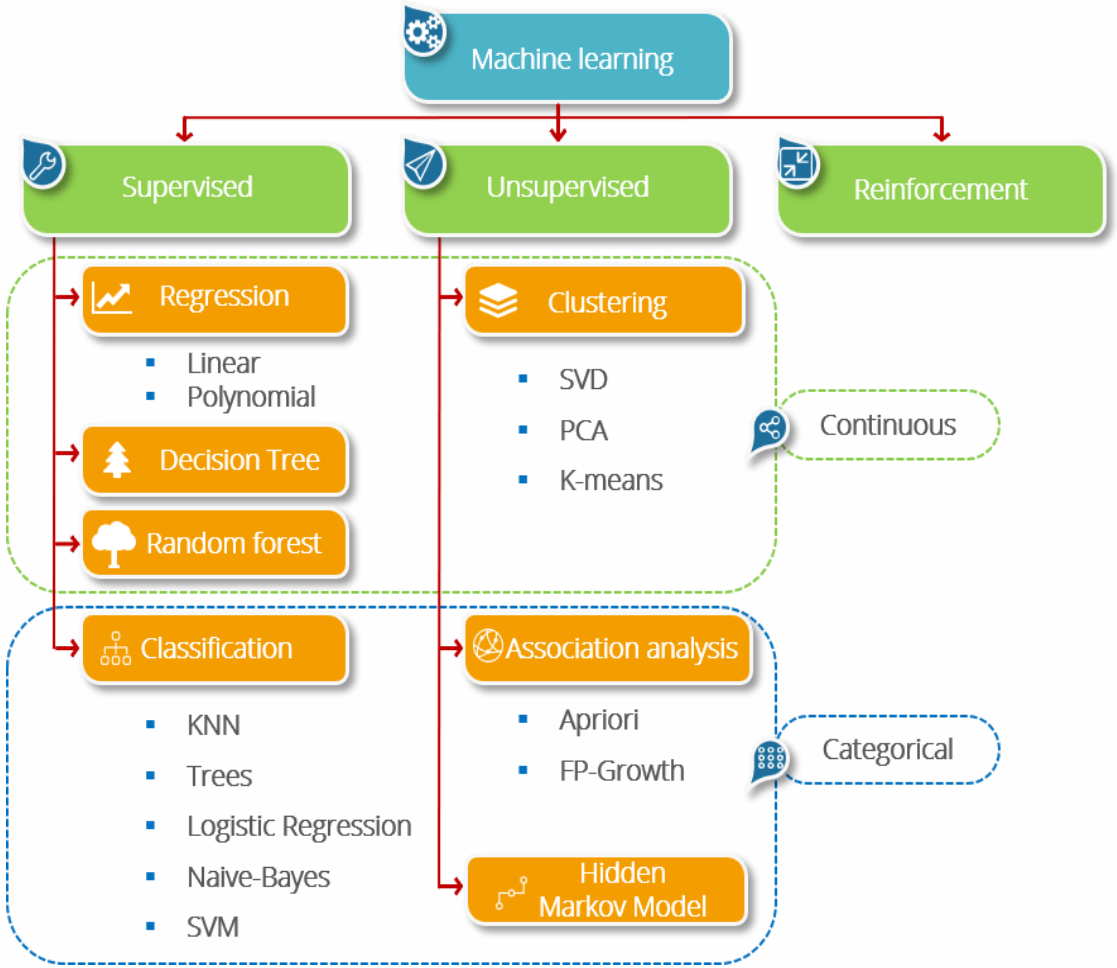
Một thuật toán Máy Học được chia thành ba phần chính:

1. **Một Quy trình Quyết định** (Decision Process)
2. **Một Hàm So lỗi** (Error Function)
3. **Một Quy trình Tối ưu hóa Mô hình** (Model Optimization Procress)

Các phương pháp học trong machine learning



Các phương pháp học trong machine learning



* Phương pháp học có giám sát (Supervised Learning): Đây là kỹ thuật học máy sử dụng cho các bài toán có sự phân lớp (Classification).

→ Máy vector hỗ trợ (Support Vector Machine – SVM); Cây quyết định (Decision Tree – DT); sử dụng mạng nơron (Neural Network – Net); dựa trên vector trọng tâm (Centroid– based vector); và tuyến tính bình phương nhỏ nhất (Linear Least Square Fit – LLSF).

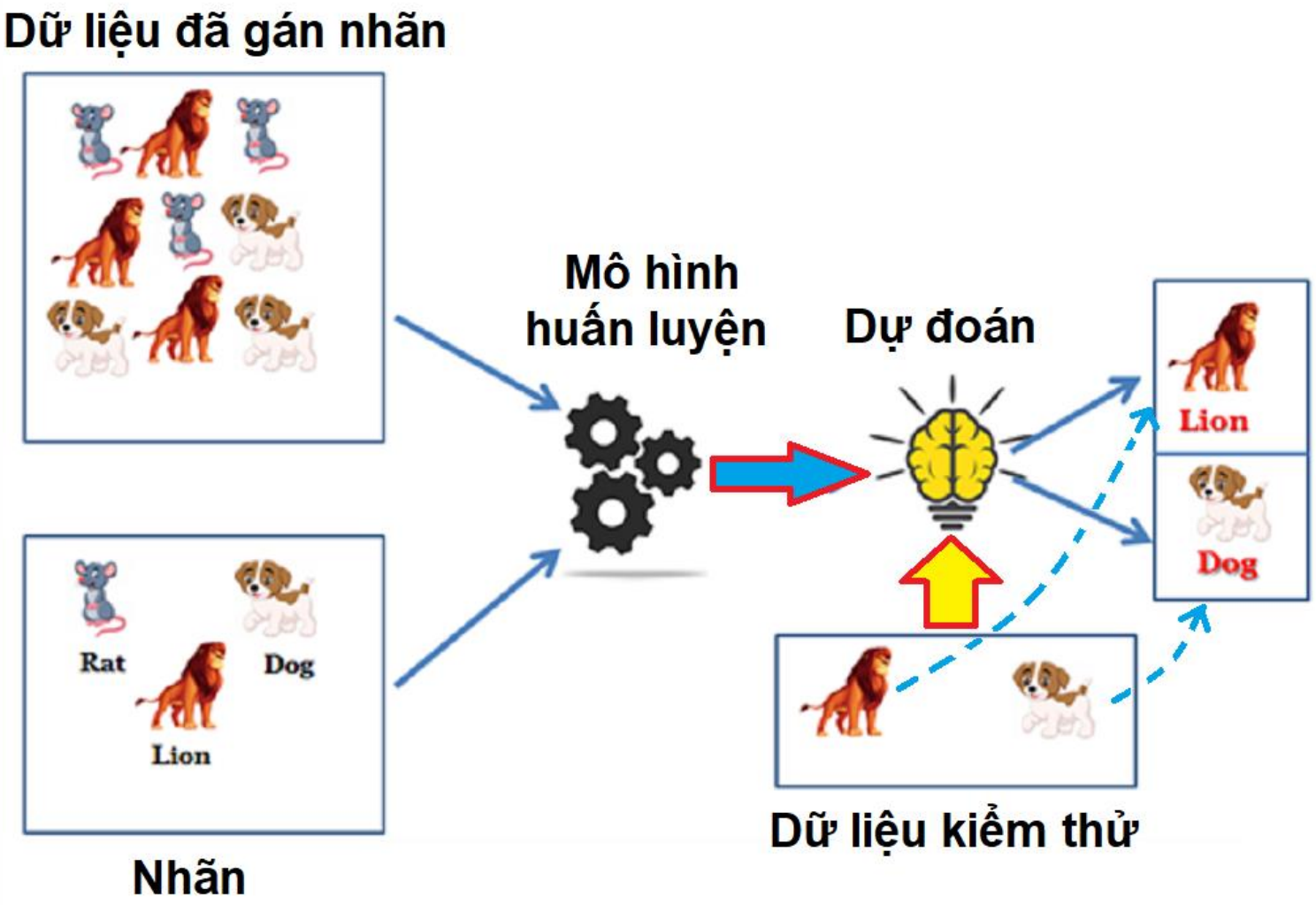
* Phương pháp học không giám sát (Unsupervised Learning): Đây là kỹ thuật học máy dùng cho các bài toán phân cụm, gom cụm (Clustering)

→ K-means, HAC (Hierarchical Agglomerative Clustering), SOM (Self-Organizing Map), DBSCAN, FCM,…

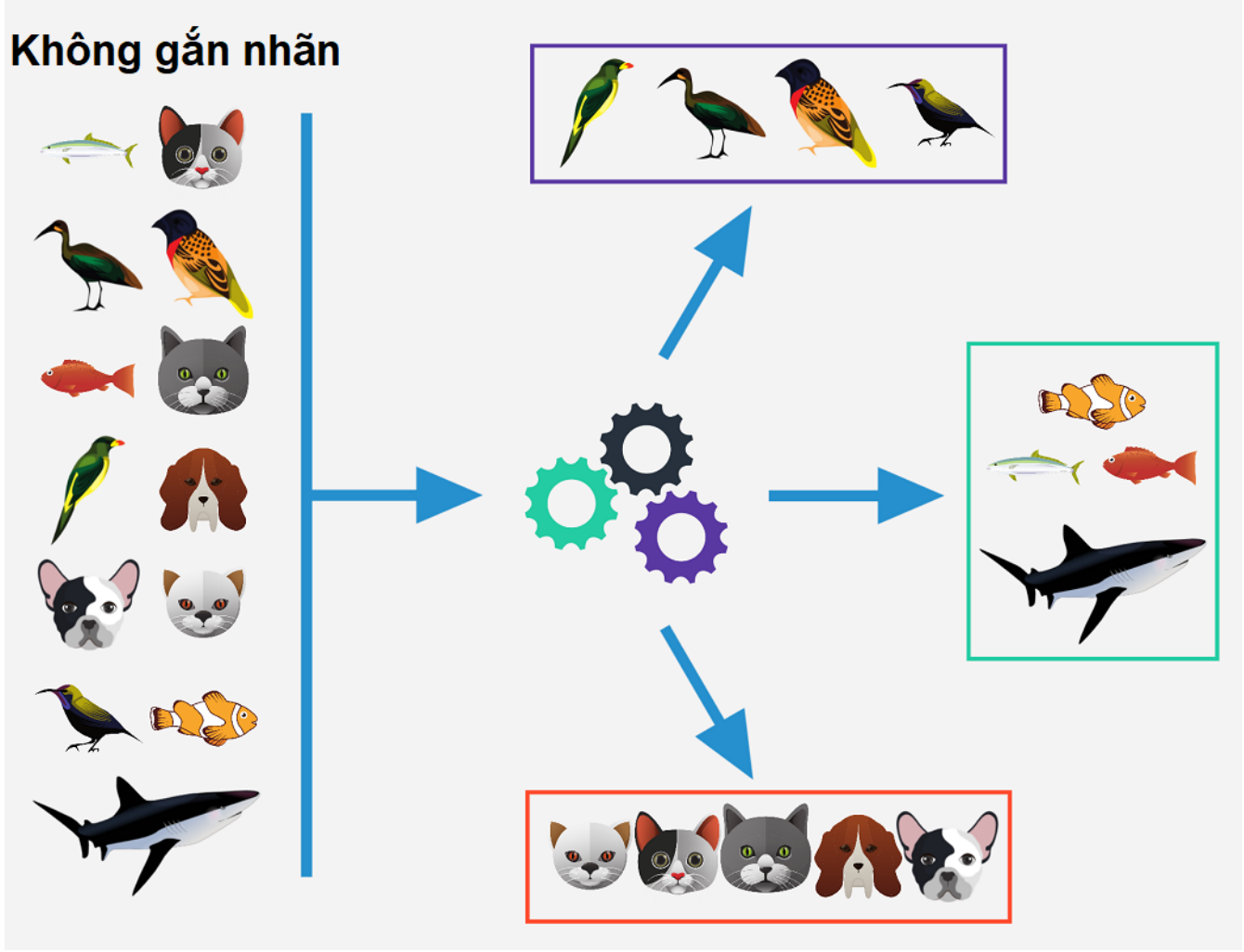
* Phương pháp học bán giám sát (Semi-Supervised Learning): Đây là một lớp của kỹ thuật học máy, sử dụng cả 2 dạng dữ liệu đã gán nhãn và chưa gán nhãn để hướng dẫn – điển hình như một lượng nhỏ dữ liệu có gán nhãn cùng với lượng lớn dữ liệu chưa gán nhãn.

→ Thuật toán Cực đại kỳ vọng (EM – Expectation Maximization), SVM truyền dẫn (TSVM – Transductive Support Vector Machine), Self-training, Co-training và các phương pháp dựa trên đồ thị (graph-based).

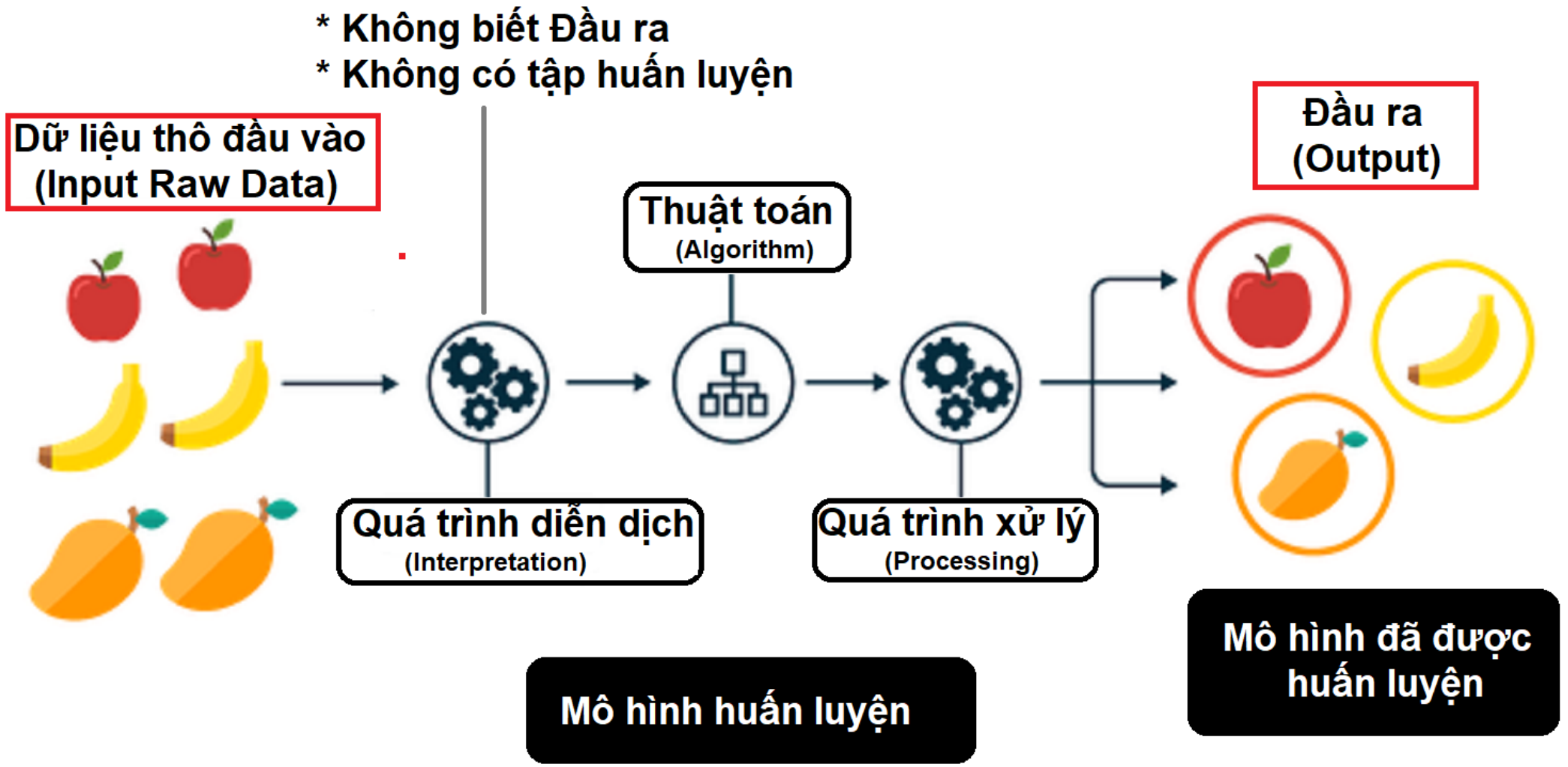
Học có giám sát (Supervised learning)



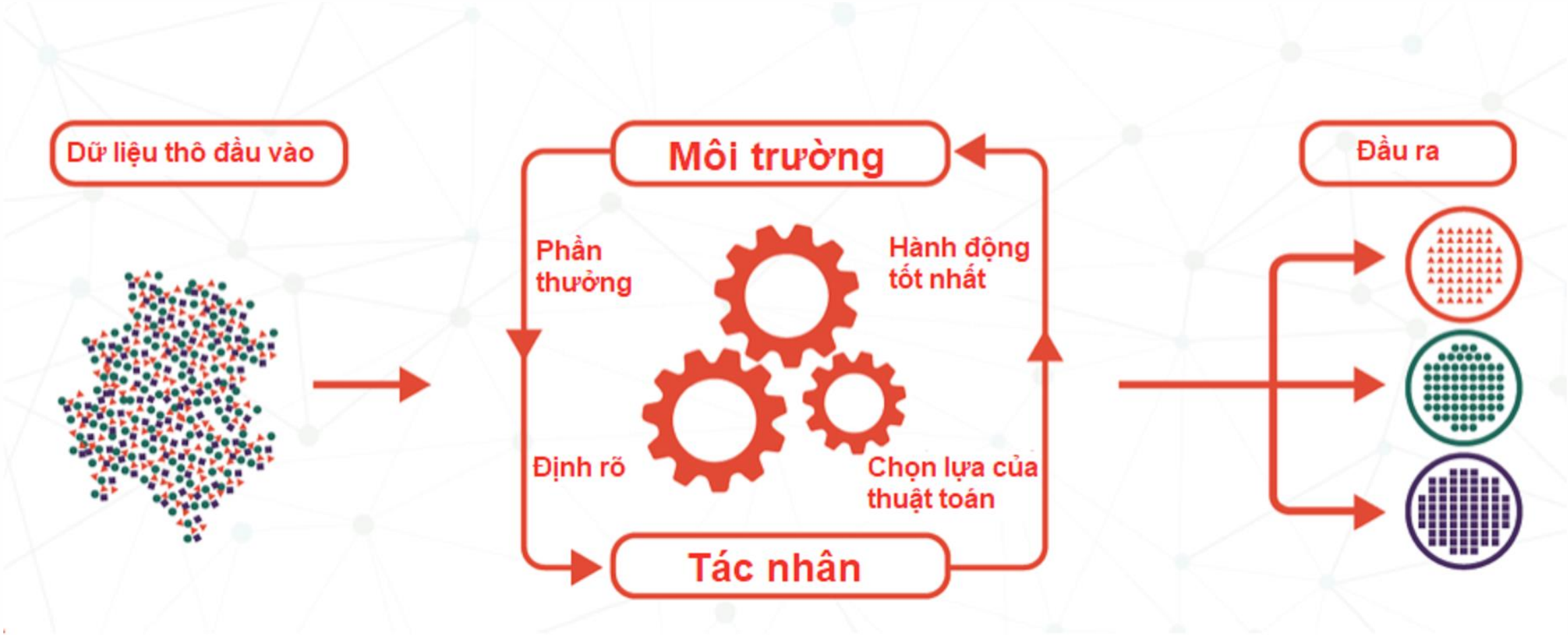
Học không giám sát (Unsupervised learning)



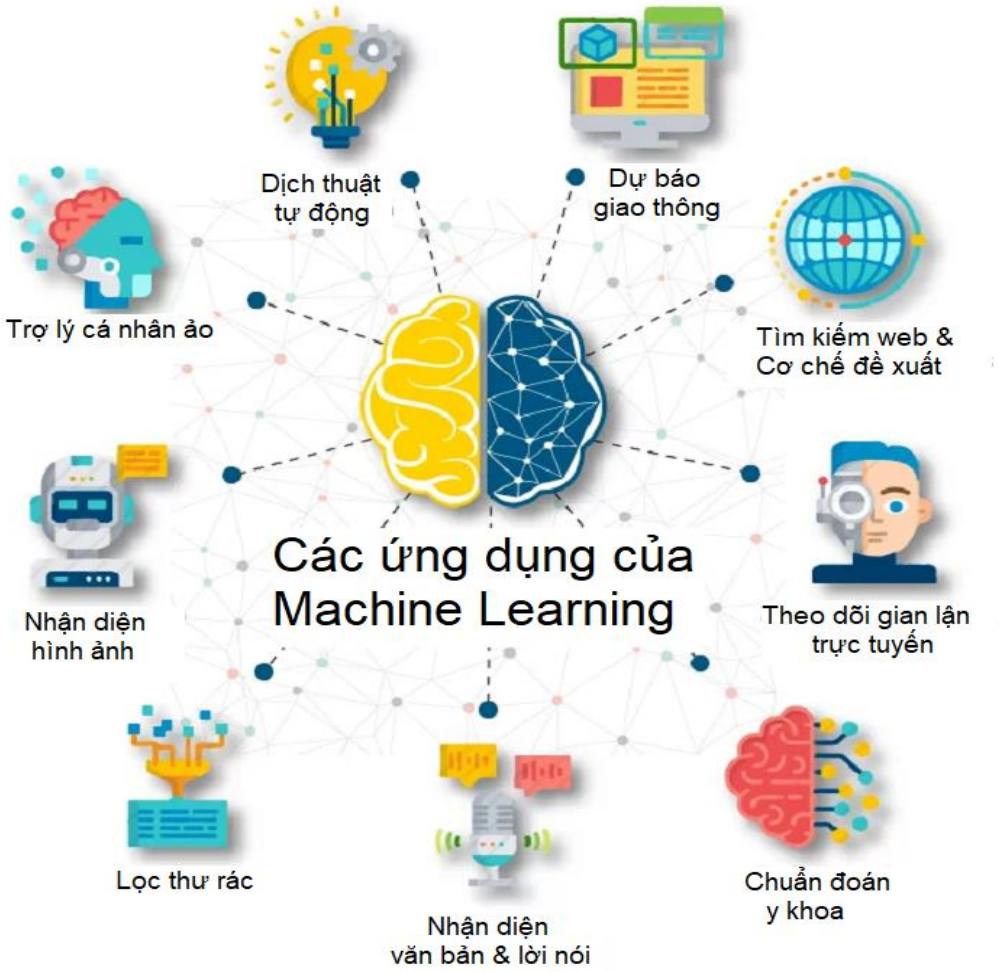
Học bán giám sát (Semi-supervised learning)



Học tăng cường (Reinforcement learning)



Các ứng dụng machine learning trong thực tiễn



* Ứng dụng trong dịch vụ tài chính
* Ứng dụng trong Chính phủ
* Ứng dụng trong chăm sóc sức khỏe
* Ứng dụng trong bán lẻ
* Ứng dụng trong vận tải

Phân lớp là gì ?

- **Mục đích:** Để dự đoán những nhãn phân lớp cho các bộ dữ liệu/mẫu mới
- **Đầu vào:** một tập các mẫu dữ liệu huấn luyện, với một nhãn phân lớp cho mỗi mẫu dữ liệu.
- **Đầu ra:** mô hình (bộ phân lớp) dựa trên tập huấn luyện và những nhãn phân lớp

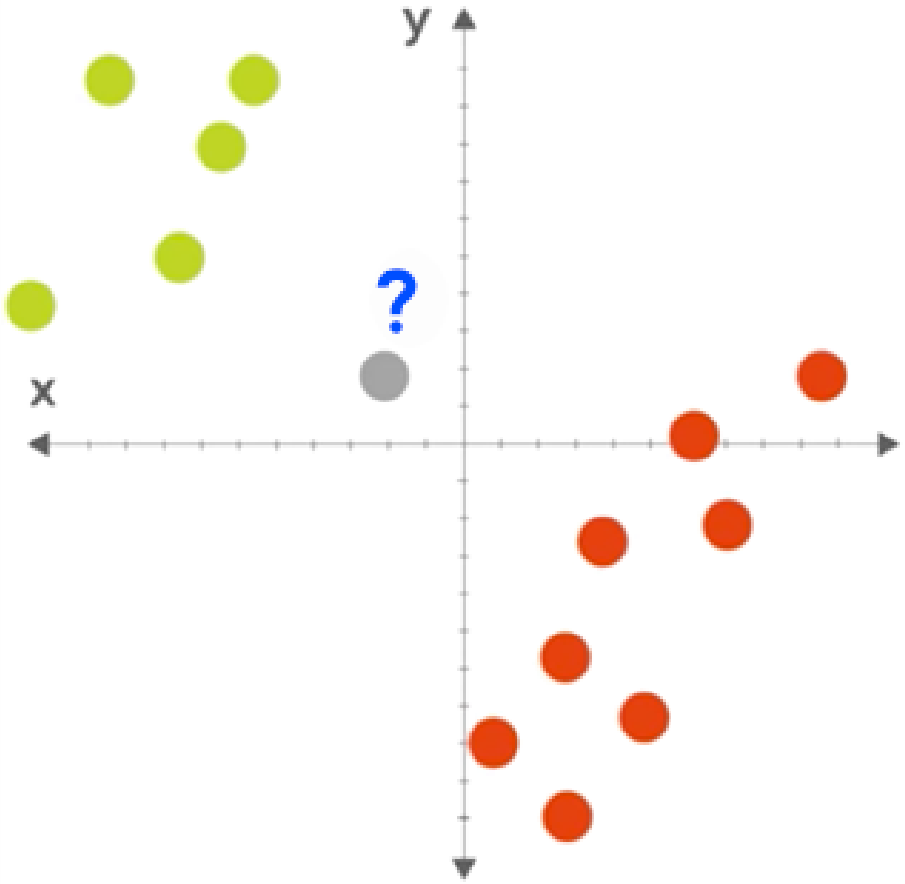
Bài toán phân lớp

Bài toán phân lớp là quá trình phân lớp 1 đối tượng dữ liệu vào 1 hay nhiều lớp đã cho trước nhờ 1 mô hình phân lớp (model).

+ Mô hình này được xây dựng dựa trên 1 tập dữ liệu được xây dựng trước đó có gán nhãn (hay còn gọi là tập huấn luyện).

+ Quá trình phân lớp là quá trình gán nhãn cho đối tượng dữ liệu.

→ Như vậy, nhiệm vụ của bài toán phân lớp là cần tìm 1 **mô hình phần lớp** để khi có dữ liệu mới thì có thể xác định được dữ liệu đó thuộc vào phân lớp nào.



Chấm xám sẽ thuộc lớp xanh hay đỏ?

Bài toán phân lớp

- Đầu vào
  - Tập dữ liệu D = $\{d_i\}$
  - Tập các lớp $C_1, C_2, \ldots, C_k$ mỗi dữ liệu d thuộc một lớp C_i
  - Tập mẫu $D_{exam} = D_1+D_2+ \ldots + D_k$ với $D_i=\{d \in D_{exam}$: d thuộc $C_i\}$
  - Tập mẫu D_{exam} đại diện cho tập D
- Đầu ra
  - Mô hình phân lớp: ánh xạ từ D sang C
- Sử dụng mô hình
  - $d \in D \setminus D_{exam}$: xác định lớp của đối tượng d

Quá trình phân lớp dữ liệu

Để xây dựng được mô hình phân lớp và đánh giá được mô hình phải trải qua các quá trình sau.

1. Chuẩn bị tập dữ liệu huấn luyện (dataset) và rút trích đặc trưng (feature extraction)
2. Xây dựng mô hình phân lớp (classifier model)
3. Kiểm tra dữ liệu với mô hình (making predictions)
4. Đánh giá mô hình phân lớp và chọn ra mô hình tốt nhất

Phân lớp: Quá trình hai pha

Pha 1: Xây dựng mô hình từ tập huấn luyện

Pha 2: Sử dụng mô hình – kiểm tra tính đúng đắn của mô hình và dùng nó để phân lớp dữ liệu mới

PHA 1

- **Mỗi bộ/mẫu dữ liệu** được phân vào một lớp được xác định trước
- Lớp của một bộ/mẫu dữ liệu được xác định bởi **thuộc tính gán nhãn lớp**
- Tập các bộ/mẫu dữ liệu huấn luyện – **tập huấn luyện** – được dùng để **xây dựng mô hình**
- Mô hình được biểu diễn bởi các **luật phân lớp, các cây quyết định** hoặc **các công thức toán học**

PHA 2

- **Phân lớp cho những đối tượng mới hoặc chưa được phân lớp**
- **Đánh giá độ chính xác của mô hình**
  + Lớp biết trước của một mẫu/bộ dữ liệu đem kiểm tra được so sánh với kết quả thu được từ mô hình
  + Tỷ lệ chính xác = phần trăm các mẫu/bộ dữ liệu được phân lớp đúng bởi mô hình trong số các lần kiểm tra

Phân lớp: Quá trình hai pha

- Xây dựng mô hình: Tìm mô tả cho tập lớp đã có
  - Cho trước tập lớp C = {C_1, C_2, …, C_k}
  - Cho ánh xạ (chưa biết) từ miền D sang tập lớp C
  - Có tập mẫu $D_{exam}=D_1+D_2+ \ldots + D_k$ với $D_i=\{d \in D_{exam}: d \in C_i\}$; D_{exam} được gọi là tập mẫu.
  - Xây dựng ánh xạ (mô hình) phân lớp trên: Dạy bộ phân lớp.
  - Mô hình: Luật phân lớp, cây quyết định, công thức toán học…

→ Pha 1: Huấn luyện bộ phân lớp

  - Tách D_{exam} thành D_{train} (2/3) + D_{test} (1/3). D_{train} và D_{test} "tính đại diện" cho miền ứng dụng
  - D_{train} : xây dựng mô hình phân lớp (xác định tham số mô hình)
  - D_{test} : đánh giá mô hình phân lớp (các độ đo hiệu quả)
  - Chọn mô hình có chất lượng nhất

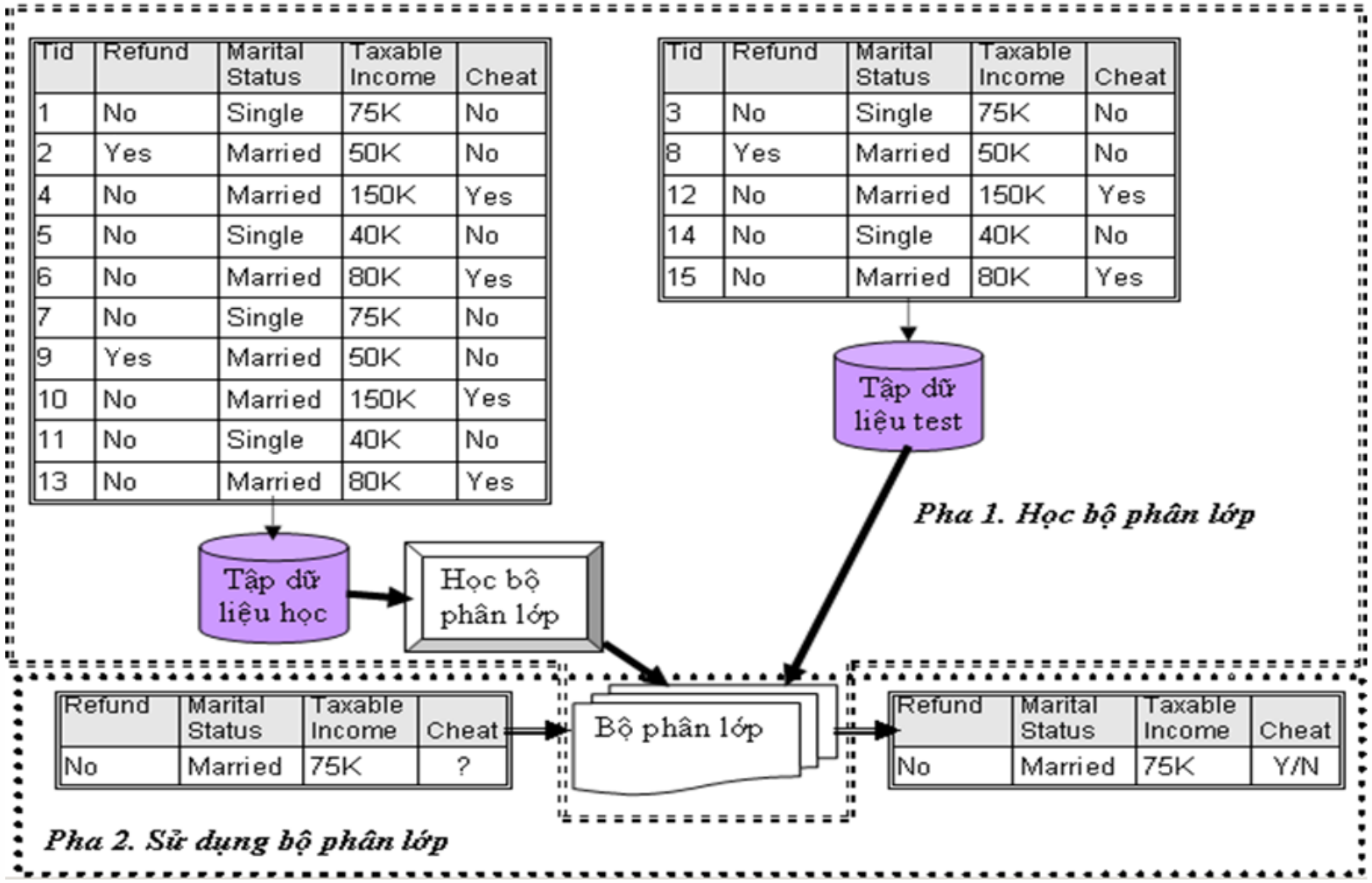
→ Pha 2: Sử dụng bộ phân lớp

  - $d \in D \setminus D_{exam}$: xác định lớp của d.

Ví dụ phân lớp: Bài toán cho vay

Tid	Refund	Marital Status	Taxable Income	Cheat
1	No	Single	75K	No
2	Yes	Married	50K	No
3	No	Single	75K	No
4	No	Married	150K	Yes
5	No	Single	40K	No
6	No	Married	80K	Yes
7	No	Single	75K	No
8	Yes	Married	50K	No
9	Yes	Married	50K	No
10	No	Married	150K	Yes
11	No	Single	40K	No
12	No	Married	150K	Yes
13	No	Married	80K	Yes
14	No	Single	40K	No
15	No	Married	80K	Yes

Phân lớp: Quá trình hai pha



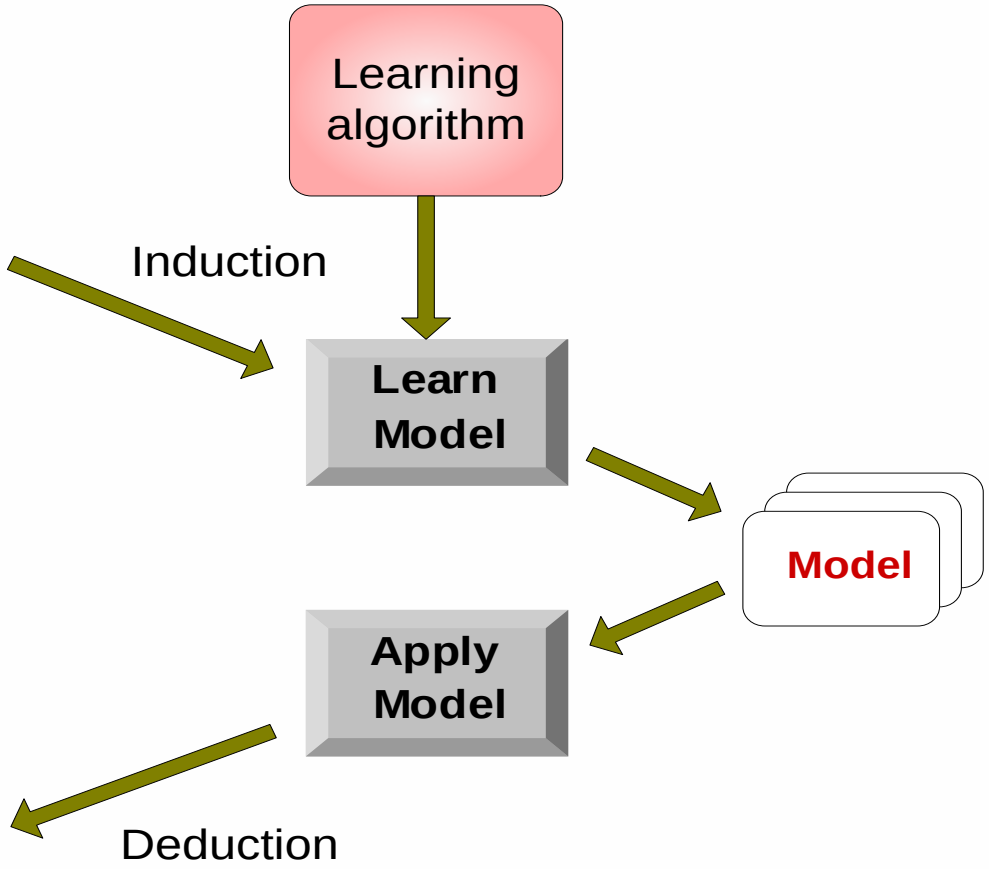
Phân lớp: Quá trình hai pha

Tid	Attrib1	Attrib2	Attrib3	Class
1	Yes	Large	125K	No
2	No	Medium	100K	No
3	No	Small	70K	No
4	Yes	Medium	120K	No
5	No	Large	95K	Yes
6	No	Medium	60K	No
7	Yes	Large	220K	No
8	No	Small	85K	Yes
9	No	Medium	75K	No
10	No	Small	90K	Yes

Training Set

Tid	Attrib1	Attrib2	Attrib3	Class
11	No	Small	55K	?
12	Yes	Medium	80K	?
13	Yes	Large	110K	?
14	No	Small	95K	?
15	No	Large	67K	?

Test Set



Các loại phân lớp

→ Phân lớp nhị phân/đa lớp

Nhị phân: hai lớp (|C| = 2)

Đa lớp: số lượng lớp > 2 (|C| > 2)

- Bài toán phân lớp nhị phân là bài toán gán nhãn dữ liệu cho đối tượng vào 1 trong 2 lớp khác nhau dựa vào việc dữ liệu đó có hay không có các đặc trưng (feature) của bộ phân lớp.
- Bài toán phân lớp đa lớp là quá trình phân lớp dữ liệu với số lượng lớp lớn hơn 2.

→ Phân lớp đơn nhãn/đa nhãn/phân cấp

Đơn nhãn: Một đối tượng chỉ thuộc duy nhất một lớp

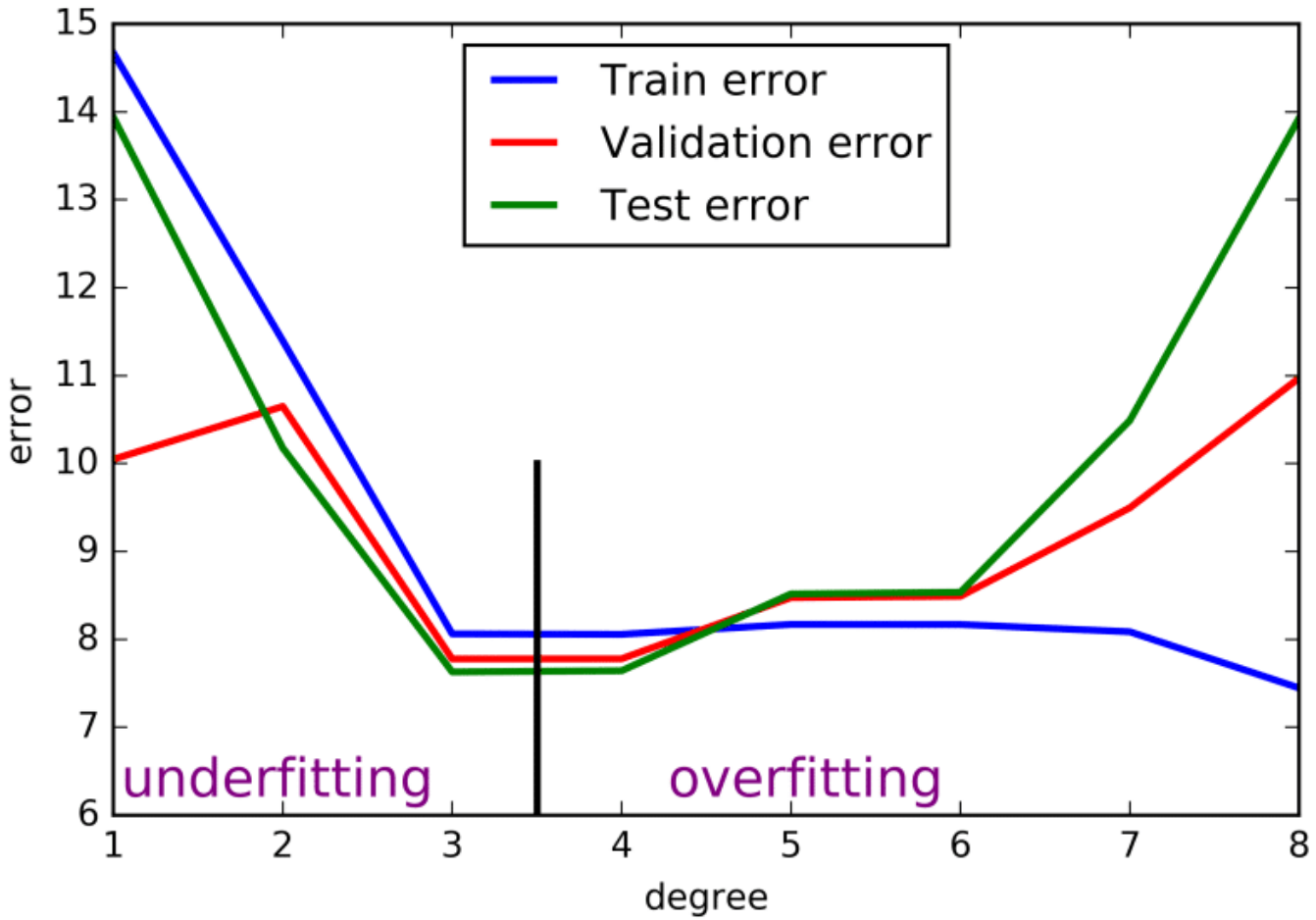
Đa nhãn: Một đối tượng thuộc một hoặc nhiều lớp

Phân cấp: Lớp này là con của lớp kia

Các loại phân lớp

- **Multi-class classification** (phân loại nhiều lớp): mỗi quan sát x chỉ nhận 1 nhãn trong tập nhãn lớp $\{c_1, c_2, \ldots, c_L\}$
  - Lọc Spam: y thuộc {*spam, normal*}
  - Đánh giá nguy cơ tín dụng: y thuộc {*high, normal*}
  - Phán đoán tấn công mạng: ?
- **Multi-label classification** (phân loại đa nhãn): mỗi đầu ra là một tập nhỏ các lớp; mỗi quan sát x có thể có nhiều nhãn
  - Image tagging: y = {*birds, nest, tree*}
  - sentiment analysis

Hiện tượng overfitting



→ Hiện tượng Overfitting thường xảy ra trong các mô hình phi tham số hoặc phi tuyến, những mô hình có sự linh hoạt cao trong xây dựng hàm mục tiêu.

→ Cả hai hiện tượng Overfitting và Underfitting đều khiến mô hình xây dựng có độ chính xác kém.

Làm thế nào để tránh Overfitting?

* Sử dụng kỹ thuật lấy lại mẫu để ước lượng độ chính xác của mô hình
* Sử dụng tập Validation test

Ví dụ, bài toán cây quyết định là một thuật toán học máy phi tham số. Đây là thuật toán thường xảy ra hiện tượng Overfitting. Ta có thể tránh hiện tượng này bằng phương pháp cắt tỉa cây (pruning).

Validation set (tập tối ưu)

Validation:

Training error + test error đều nhỏ → Mô hình tốt

- Khi xây dựng một mô hình chỉ dựa trên training set, làm thế nào để biết được chất lượng của nó trên test set ?

→ Phương pháp: Trích từ training set ra một tập con nhỏ và thực hiện việc đánh giá mô hình trên tập con nhỏ này.

- **Validation set (tập tối ưu):** Tập con nhỏ được trích ra từ training set

→ **Traning set** mới là phần còn lại của training set ban đầu

- 03 đại lượng cần được quan tâm:
    + Training set trên training set mới
    + Validation error được định nghĩa tương tự trên validation set
    + Test error trên test set

Validation set (tập tối ưu)

Tập tối ưu (Validation set):

- Các mẫu trong tập kiểm thử không thể được sử dụng (theo bất kỳ cách nào!) trong quá trình huấn luyện hệ thống

■ Trong một số bài toán, quá trình huấn luyện hệ thống bao gồm 2 giai đoạn

- Giai đoạn thứ 1: Huấn luyện hệ thống (= Học hàm mục tiêu)
- Giai đoạn thứ 2: Tối ưu giá trị các tham số của hệ thống

■ Tập kiểm thử không thể được sử dụng cho mục đích tối ưu (điều chỉnh) tham số!

■ Chia tập toàn bộ các mẫu *D* thành 3 tập con không giao nhau: **tập huấn luyện, tập tối ưu, và tập kiểm thử**

■ Tập tối ưu (validation set) được sử dụng để tối ưu giá trị các tham số được sử dụng

→ Đối với một tham số, giá trị tối ưu là giá trị giúp sinh ra ***hiệu năng cực đại*** đối với tập tối ưu

Đánh giá hiệu năng hệ thống học máy

+ Làm thế nào để thu được một đánh giá đáng tin cậy về hiệu năng của hệ thống?
  - Chiến lược đánh giá
  - Lựa chọn tham số tốt
+ Làm thế nào để lựa chọn tốt các tham số cho một phương pháp học máy?
+ Làm thế nào để so sánh hiệu quả của hai phương pháp học máy, với độ tin cậy cao?

Đánh giá hiệu năng hệ thống học máy

+ **Đánh giá lý thuyết (theoretical evaluation):** nghiên cứu các khía cạnh lý thuyết của một hệ thống mà có thể chứng minh được.
    - Tốc độ học, thời gian học,
    - Bao nhiêu mẫu học là đủ?
    - Độ chính xác trung bình của hệ thống,
    - Khả năng chống nhiễu,…
+ **Đánh giá thực nghiệm (experimental evaluation):** quan sát hệ thống làm việc trong thực tế, sử dụng một hoặc nhiều tập dữ liệu và các tiêu chí đánh giá. Tổng hợp đánh giá từ các quan sát đó.

→ Chúng ta sẽ nghiên cứu cách đánh giá thực nghiệm.

Đánh giá hiệu năng hệ thống học máy

+ **Bài toán đánh giá (model assessment):** cần đánh giá hiệu năng của phương pháp (model) học máy A, chỉ dựa trên bộ dữ liệu đã quan sát D.
+ Việc đánh giá hiệu năng của hệ thống
  - Thực hiện một cách tự động, sử dụng một tập dữ liệu.
  - Không cần sự tham gia (can thiệp) của người dùng.
+ **Chiến lược đánh giá (evaluation strategies)** → Làm sao có được một đánh giá đáng tin cậy về hiệu năng của hệ thống?
+ **Các tiêu chí đánh giá (evaluation metrics)** → Làm sao để đo hiệu năng của hệ thống?

Các phương pháp đánh giá hệ thống học máy

+ **Hold-out** (chia đôi)
+ **Stratified sampling** (lấy mẫu phân tầng)
+ **Repeated hold-out** (chia đôi nhiều lần)
+ **Cross-validation** (đánh giá chéo)
    - k-fold
    - Leave-one-out
+ **Bootstrap sampling**

Phương pháp Hold-out (chia đôi)

Toàn bộ tập mẫu *D* được chia thành 2 tập con không giao nhau

- Tập huấn luyện D_{train} – để huấn luyện hệ thống
- Tập kiểm thử D_{test} – để đánh giá hiệu năng của hệ thống đã học

→ $D = D_{train}\ U\ D_{test}$, và thường là $|D_{train}| >> |D_{test}|$

+ Các yêu cầu:
  - Bất kỳ mẫu nào thuộc vào tập kiểm thử D_{test} đều không được sử dụng trong quá trình huấn luyện hệ thống
  - Bất kỳ mẫu nào được sử dụng trong giai đoạn huấn luyện hệ thống (i.e., thuộc vào D_{train}) đều không được sử dụng trong giai đoạn đánh giá hệ thống
  - Các mẫu kiểm thử trong D_{test} cho phép một đánh giá không thiên vị đối với hiệu năng của hệ thống

+ Các lựa chọn thường gặp: $|D_{train}|=(2/3).|D|$, $|D_{test}|=(1/3).|D|$
+ Phù hợp khi ta có tập mẫu *D* có kích thước lớn

Phương pháp Stratified sampling (lấy mẫu phân tầng)

+ Đối với các tập mẫu có kích thước nhỏ hoặc không cân xứng (unbalanced datasets), các mẫu trong tập huấn luyện và thử nghiệm có thể không phải là đại diện
  - Ví dụ: Có (rất) ít mẫu đối với một số lớp
+ Mục tiêu: Phân bố lớp (class distribution) trong tập huấn luyện và tập kiểm thử phải xấp xỉ như trong tập toàn bộ các mẫu (*D*)
+ Lấy mẫu phân tầng (Stratified sampling)
  - Là một phương pháp để cân xứng (về phân bố lớp)
  - Đảm bảo tỷ lệ phân bố lớp (tỷ lệ các mẫu giữa các lớp) trong tập huấn luyện và tập kiểm thử là xấp xỉ nhau
+ Phương pháp lấy mẫu phân tầng không áp dụng được cho bài toán hồi quy (vì giá trị đầu ra của hệ thống là một giá trị số thực, không phải là một nhãn lớp)

Phương pháp Repeated hold-out (chia đôi nhiều lần)

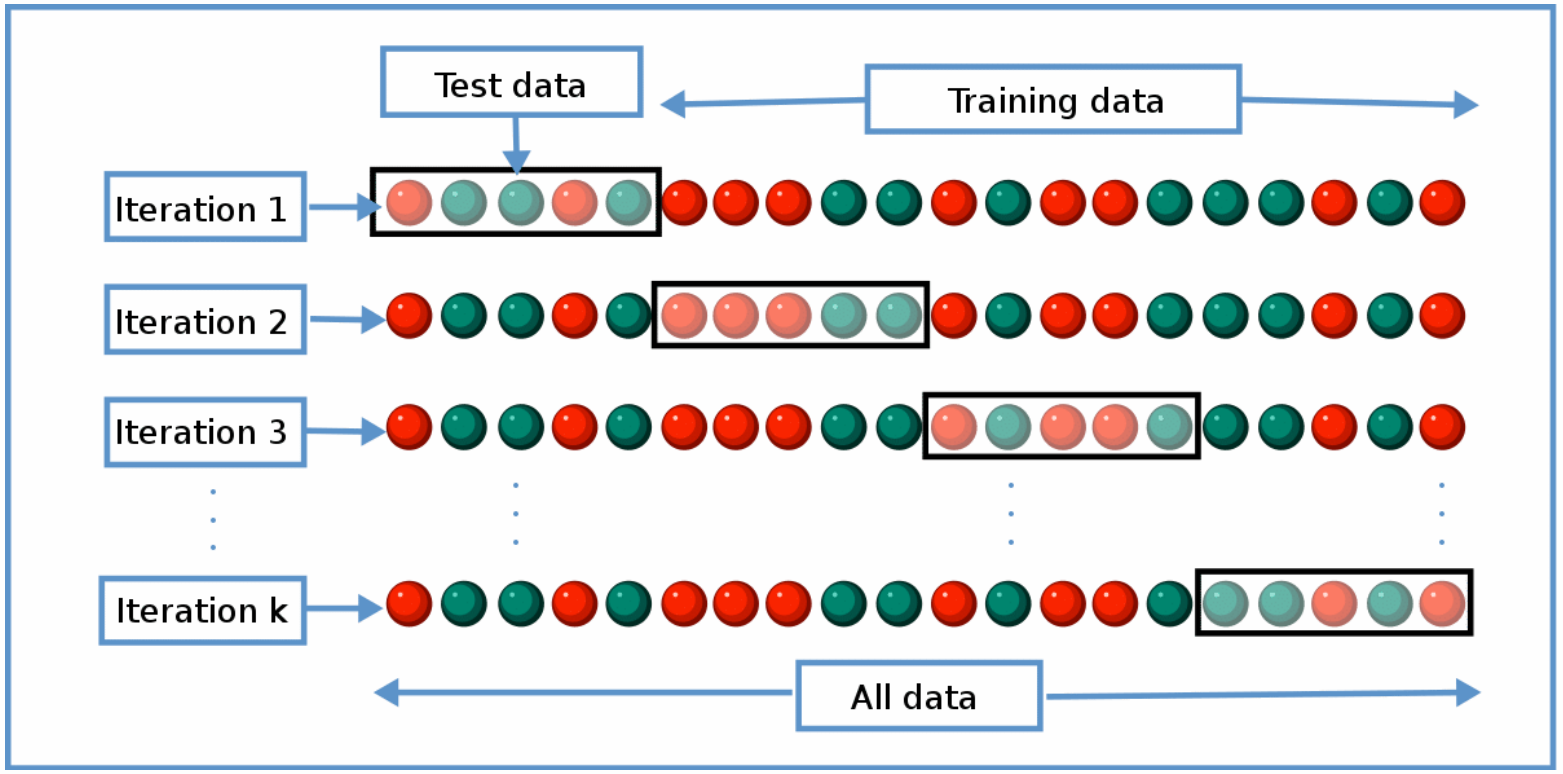
+ Áp dụng phương pháp đánh giá Hold-out nhiều lần, để sinh ra (sử dụng) các tập huấn luyện và thử nghiệm khác nhau
  - Trong mỗi bước lặp, một tỷ lệ nhất định của tập D được lựa chọn ngẫu nhiên để tạo nên tập huấn luyện (có thể sử dụng kết hợp với phương pháp lấy mẫu phân tầng – stratified sampling)
  - Các giá trị lỗi (hoặc các giá trị đối với các tiêu chí đánh giá khác) ghi nhận được trong các bước lặp này được lấy trung bình cộng (averaged) để xác định giá trị lỗi tổng thể
+ **Phương pháp này vẫn không hoàn hảo**
  - Mỗi bước lặp sử dụng một tập kiểm thử khác nhau
  - Có một số mẫu trùng lặp (được sử dụng lại nhiều lần) trong các tập kiểm thử này

Phương pháp Cross-validation (đánh giá chéo)

+ Để tránh việc trùng lặp giữa các tập kiểm thử (một số mẫu cùng xuất hiện trong các tập kiểm thử khác nhau)
+ k-fold cross-validation
  - Tập toàn bộ các mẫu D được chia thành k tập con không giao nhau (gọi là “fold”) có kích thước xấp xỉ nhau
  - Mỗi lần (trong số k lần) lặp, một tập con được sử dụng làm tập kiểm thử, và (k-1) tập con còn lại được dùng làm tập huấn luyện
  - k giá trị lỗi (mỗi giá trị tương ứng với một fold) được tính trung bình cộng để thu được giá trị lỗi tổng thể
+ Các lựa chọn thông thường của k: 10, hoặc 5
+ Thông thường, mỗi tập con (fold) được lấy mẫu phân tầng (xấp xỉ phân bố lớp) trước khi áp dụng quá trình đánh giá Cross-validation
+ **Phù hợp khi ta có tập mẫu D vừa và nhỏ**

Phương pháp Cross-validation (đánh giá chéo)

Cross Validation – đánh giá chéo là một quy trình lấy mẫu lại được sử dụng để đánh giá các mô hình học máy trên một mẫu dữ liệu hạn chế.



Phương pháp Leave-one-out cross-validation

+ Một trường hợp (kiểu) của phương pháp Cross-validation
  - Số lượng các nhóm (folds) bằng kích thước của tập dữ liệu (k=|D|)
  - Mỗi nhóm (fold) chỉ bao gồm một mẫu
+ Khai thác tối đa (triệt để) tập mẫu ban đầu
+ Không hề có bước lấy mẫu ngẫu nhiên (no random subsampling)
+ Áp dụng lấy mẫu phân tầng (stratification) không phù hợp

  → Vì ở mỗi bước lặp, tập thử nghiệm chỉ gồm có một mẫu
+ Chi phí tính toán (rất) cao
+ Phù hợp khi ta có một tập mẫu D (rất) nhỏ

Phương pháp Bootstrap sampling

Phương pháp Bootstrap sampling sử dụng việc **lấy mẫu có lặp lại** (sampling with replacement) để tạo nên tập huấn luyện

- Giả sử tập toàn bộ *D* bao gồm *n* mẫu
- Lấy mẫu có lặp lại *n* lần đối với tập *D*, để tạo nên tập huấn luyện D_{train} gồm *n* mẫu
  - ➢ Từ tập *D*, lấy ra ngẫu nhiên một mẫu *x* (nhưng **không loại bỏ** *x* khỏi tập *D*)
  - ➢ Đưa mẫu x vào trong tập huấn luyện: $D_{train} = D_{train} \cup x$
  - ➢ Lặp lại 2 bước trên n lần
- Sử dụng tập D_{train} để huấn luyện hệ thống
- Sử dụng tất cả các mẫu thuộc *D* **nhưng không thuộc** D_{train} để tạo nên tập thử nghiệm: $D_{test} = \{z \in D;\ z \notin D_{train}\}$

+ Phù hợp khi ta có một tập dữ liệu D có kích thước (rất) nhỏ

Đánh giá hiệu năng của mô hình phân lớp

→ Các phương pháp đánh giá hiệu quả

Câu hỏi: Làm thế nào để đánh giá được hiệu quả của một mô hình?

→ Độ đo để đánh giá hiệu quả

Câu hỏi: Làm thế nào để có được ước tính đáng tin cậy?

→ Phương pháp so sánh mô hình

Câu hỏi: Làm thế nào để so sánh hiệu quả tương đối giữa các mô hình có tính cạnh tranh?

Các tiêu chí đánh giá

Tính chính xác (Accuracy): Mức độ dự đoán (phân lớp) chính xác của hệ thống (đã được huấn luyện) đối với các mẫu kiểm chứng (test instances).

Tính hiệu quả (Efficiency): Chi phí về thời gian và tài nguyên cần thiết cho việc huấn luyện và kiểm thử hệ thống.

Khả năng xử lý nhiễu (Robustness): Khả năng xử lý (chịu được) của hệ thống đối với các mẫu nhiễu (lỗi) hoặc thiếu giá trị.

Khả năng mở rộng (Scalability): Hiệu năng của hệ thống (tốc độ học/ phân loại) thay đổi như thế nào đối với kích thước của tập dữ liệu.

Khả năng diễn giải (Interpretability): Mức độ dễ hiểu (đối với người sử dụng) của các kết quả và hoạt động của hệ thống.

Mức độ phức tạp (Complexity): Mức độ phức tạp của mô hình hệ thống (hàm mục tiêu) học được

Các tiêu chí đánh giá

Tính chính xác (Accuracy)

Độ chính xác là tỉ lệ giữa số điểm dữ liệu được dự đoán đúng và tổng số điểm dữ liệu.

$$Tính\ chính\ xác\ (Accuracy) = \frac{Số\ phán\ đoán\ chính\ xác}{Tổng\ số\ phán\ đoán}$$

Tuy nhiên, một mô hình có độ chính xác cao chưa hẳn đã tốt. Accuracy lộ rõ hạn chế khi được sử dụng trên bộ dữ liệu không cân bằng (imbalanced dataset)

Đối với bài toán hồi quy (dự đoán): Giá trị đầu ra của hệ thống là một giá trị số

$$\text{Error} = \frac{1}{|data\_test|} \sum_{x \in data\_test} Error(x)$$

$$Error(x) = |d(x) - o(x)|$$

o(x): Giá trị đầu ra (dự đoán) bởi hệ thống đối với mẫu *x*

d(x): Giá trị đầu ra thực sự (đúng) đối với mẫu x

Đánh giá phân lớp nhị phân

Ma trận nhầm lẫn (Confusion matrix)

→ Theo dữ liệu test: Giá trị thực: P (dương) / N (âm) và Giá trị qua phân lớp: T (đúng)/F (sai)

→ Sử dụng các ký hiệu: TP (true positives), TN (true negatives), FP (false positives), FN (false negatives)

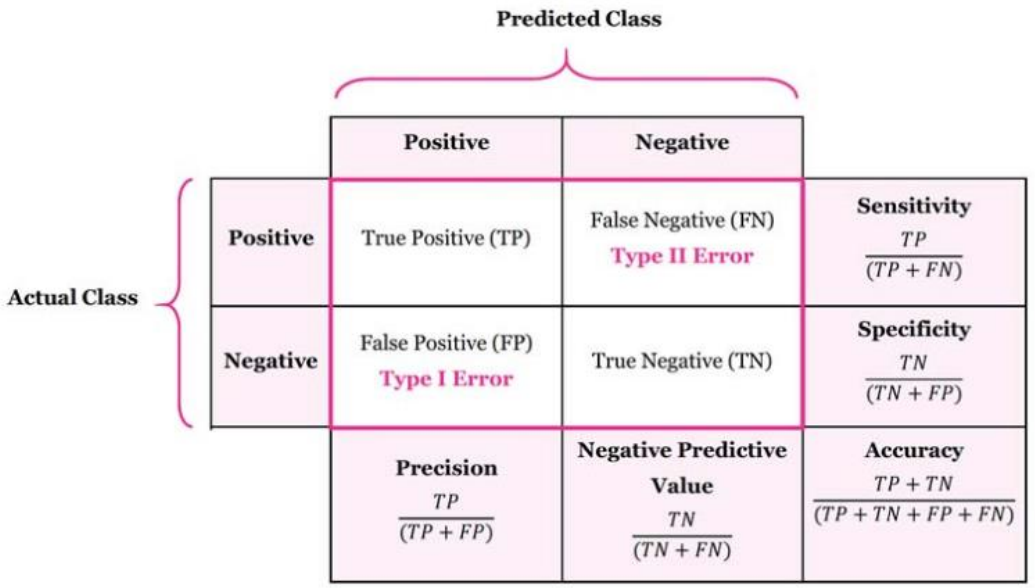
Khi thực hiện bài toán phân loại, có 4 trường hợp của dự đoán có thể xảy ra:

TP: đối tượng ở lớp Positive, mô hình phân đối tượng vào lớp Positive (dự đoán đúng)

TN: đối tượng ở lớp Negative, mô hình phân đối tượng vào lớp Negative (dự đoán đúng)

FP: đối tượng ở lớp Negative, mô hình phân đối tượng vào lớp Positive (dự đoán sai) – Type I Error

FN: đối tượng ở lớp Positive, mô hình phân đối tượng vào lớp Negative (dự đoán sai) – Type II Error



Đánh giá phân lớp nhị phân

→ Độ hồi tưởng ρ, độ chính xác π, các độ đo F_1 và F_β

$$\rho = \frac{TP}{TP + FP} \qquad \pi = \frac{TP}{TP + FN} \qquad f_\beta = \frac{(\beta^2 + 1)\pi\rho}{\beta^2\pi + \rho} \qquad f_1 = \frac{2\pi\rho}{\pi + \rho}$$

Độ hồi tưởng (Recall) đối với lớp C: Tổng số các mẫu thuộc lớp C_i được phân loại chính xác chia cho tổng số các mẫu **thuộc** lớp C

Độ chính xác (Precision) đối với lớp C: Tổng số các mẫu thuộc lớp C được phân loại chính xác chia cho tổng số các mẫu **được phân loại** vào lớp C

Tiêu chí đánh giá F1 là sự kết hợp của 2 tiêu chí đánh giá Độ chính xác (Precision) và độ hồi tưởng (Recall), F1 là một trung bình điều hòa (harmonic mean) của các tiêu chí Precision và Recall

→ F1 có xu hướng lấy giá trị gần với giá trị nào nhỏ hơn giữa 2 giá trị Precision và Recall

→ F1 có giá trị lớn nếu cả 2 giá trị Precision và Recall đều lớn

Đánh giá phân lớp nhị phân

→ Phương án khác đánh giá mô hình nhị phân theo độ chính xác (accuracy) và hệ số lỗi (Error rate)

		Lớp dự báo	
		Lớp = 1	Lớp = 0
Lớp thực sự	Lớp = 1	f_{11}	f_{10}
	Lớp = 0	f_{01}	f_{00}

$$\text{Accuracy} = \frac{\text{Number of correct predictions}}{\text{Total number of predictions}} = \frac{f_{11} + f_{00}}{f_{11} + f_{10} + f_{01} + f_{00}}$$

$$\text{Error rate} = \frac{\text{Number of wrong predictions}}{\text{Total number of predictions}} = \frac{f_{10} + f_{01}}{f_{11} + f_{10} + f_{01} + f_{00}}$$

Đánh giá phân lớp đa lớp

→ Bài toán ban đầu: C gồm có k lớp

Đối với mỗi lớp C_i , cho thực hiện thuật toán với các dữ liệu thuộc D_{test} nhận được các đại lượng TP_i, TF_i, FP_i, FN_i (như bảng dưới đây)

Lớp C_i		Giá trị thực	
		Thuộc lớp C_i	Không thuộc lớp C_i
Giá trị qua bộ phân lớp đa lớp	Thuộc lớp C_i	TP_i	FN_i
	Không thuộc lớp C_i	FP_i	TN_i

Đánh giá phân lớp đa lớp

→ Tương tự bộ phân lớp hai lớp (nhị phân)

- Độ chính xác Pr_i của lớp C_i là tỷ lệ số mẫu dương được thuật toán phân lớp cho giá trị đúng trên tổng số mẫu được thuật toán phân lớp vào lớp C_i :

$$Pr_i = \frac{TP_i}{TP_i + FN_i}$$

- Độ hồi tưởng Re_i của lớp C_i là tỷ lệ số mẫu dương được thuật toán phân lớp cho giá trị đúng trên tổng số mẫu dương thực sự thuộc lớp C_i:

$$\mathrm{Re}_i = \frac{TP_i}{TP_i + FP_i}$$

Đánh giá phân lớp đa lớp

→ Làm thế nào để tính toán được giá trị độ Chính xác và độ Hồi tưởng:

- Các giá trị ρ_i và π_i : độ hồi phục và độ chính xác đối với lớp C_i.
- Đánh giá theo các độ đo:

Trung bình vi mô (Micro averaging): ρ^{μ} và π^{μ}

$$\rho^{\mu} = \frac{\sum_{c=1}^{K} TP_c}{\sum_{c=1}^{K} (TP_c + FP_c)} \qquad \pi^{\mu} = \frac{\sum_{c=1}^{K} TP_c}{\sum_{c=1}^{K} (TP_c + TN_c)}$$

Trung bình vĩ mô (Macro averaging): ρ^{M} và π^{M}

$$\rho^{M} = \frac{1}{K} \sum_{c=1}^{K} \rho_c \qquad \pi^{M} = \frac{1}{K} \sum_{c=1}^{K} \pi_c$$

Ứng dụng của mô hình phân lớp

- Tín dụng
- Tiếp thị
- Chẩn đoán y khoa
- Phân tích hiệu quả điều trị
- Phân tích thị trường và chứng khoán
- Phát hiện gian lận
- Quản lý rủi ro và phân tích doanh nghiệp
- Phân tích giá trị trọn đời của khách hàng

Chương 1: Mô hình phân lớp và ứng dụng

1.1. Phân lớp nhị phân và Phân lớp đa nhãn

1.2. Hiện tượng overfitting

1.3. Phương pháp xác nhận chéo (cross validation)

1.4. Đánh giá hiệu năng của mô hình phân lớp

1.5. Ứng dụng của mô hình phân lớp