2102470 Học máy

Bài giảng: Giới thiệu

Chương 1: Giới thiệu

VTPham@IUH, 2024

Ôn lại bài học trước

- Bạn có nhớ ? % ?

Nội dung chính

- 1.1 Giới thiệu
- 1.2 Các giải thuật trong học máy
- 1.3 Xây dựng hệ thống học máy
- 1.4 Phương pháp trích chọn đặc trưng của dữ liệu
- 1.5 Tập dữ liệu chuẩn (Dataset)

1.1 Giới thiệu

1.1.1 Học máy là gì

"[Machine Learning is the] field of study that gives computers the ability to learn without being explicitly programmed."

Arthur Samuel, 1959

"A computer program is said to learn from experience *E* with respect to some task *T* and some performance measure *P*, if its performance on *T*, as measured by *P*, improves with experience *E*."

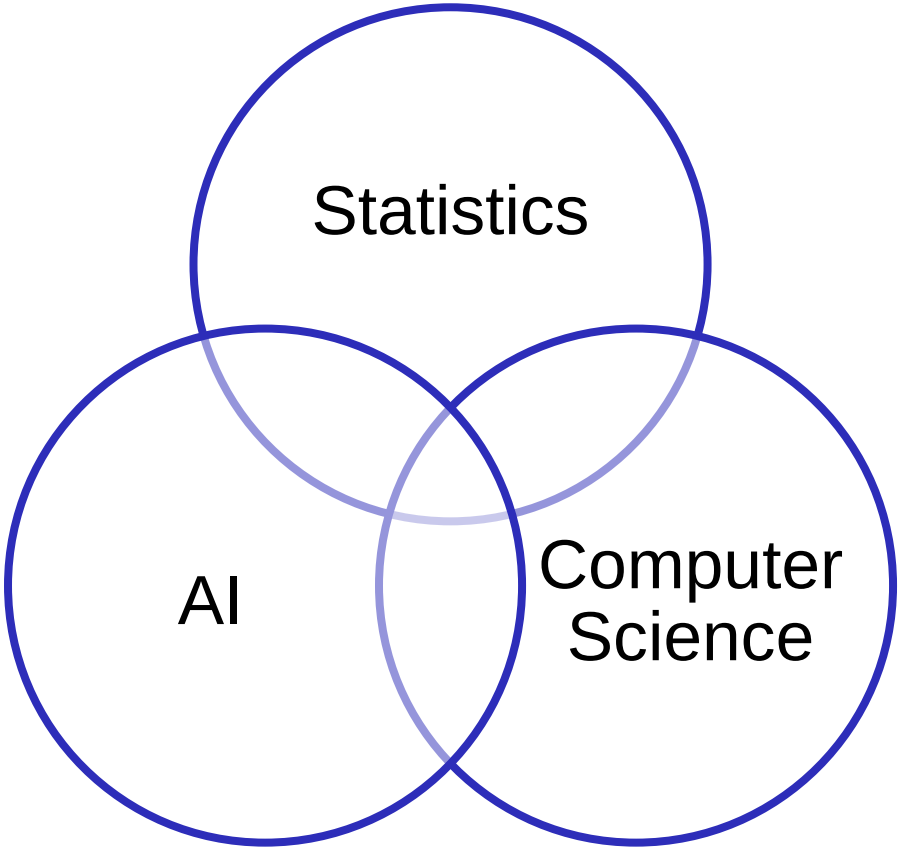
Tom Mitchell, 1997

- ML

Machine learning

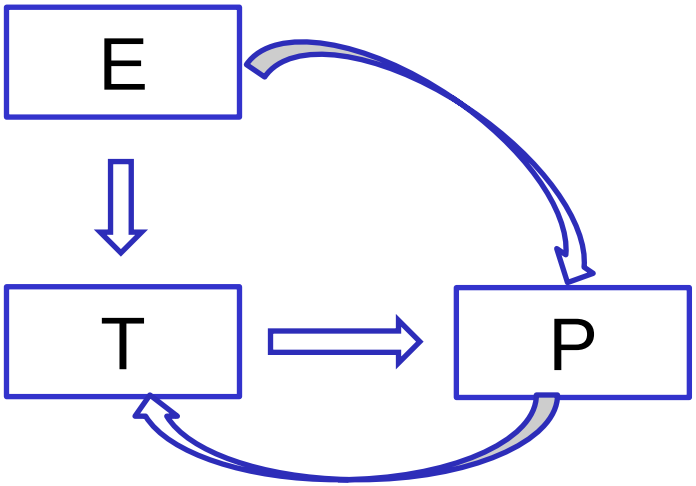
Statistical learning

Predictive analysis



P, E, T

- PET
  - P: Performance measure
  - E: Experience
  - T: Task

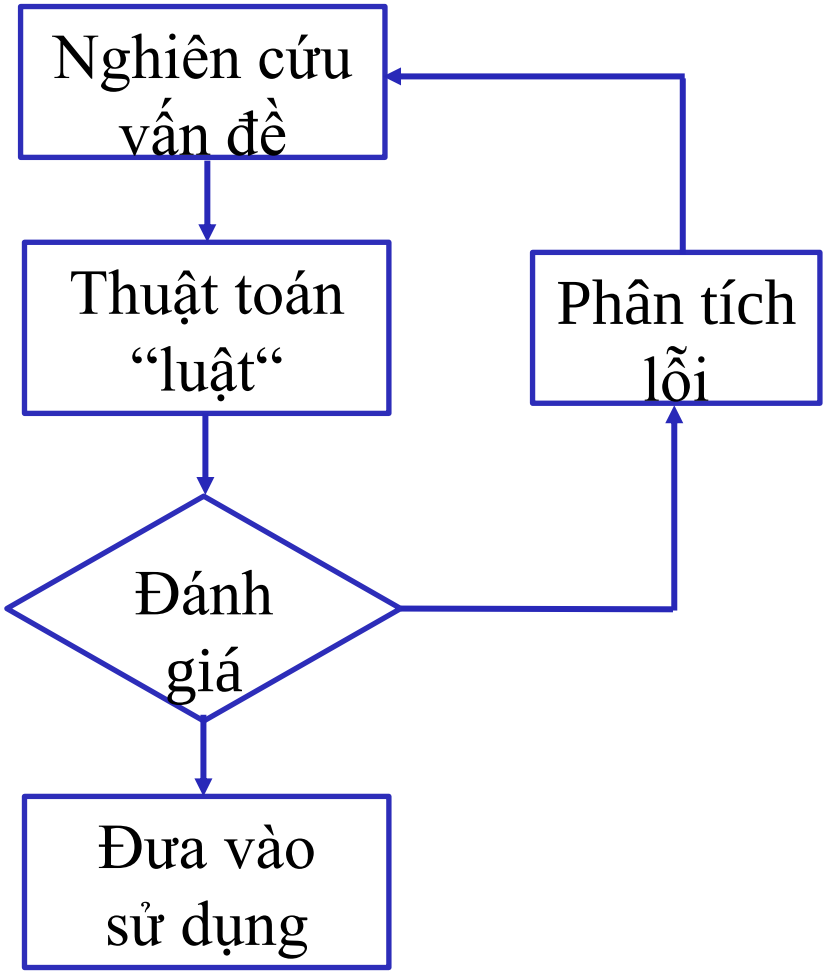


Ví dụ

- Bộ lọc thư rác
  - P: độ chính xác
  - E: dữ liệu đào tạo
  - T: 0: nonspam >< 1: spam

Ví dụ

- Bộ lọc thư rác
  - Cách tiếp cận truyền thống

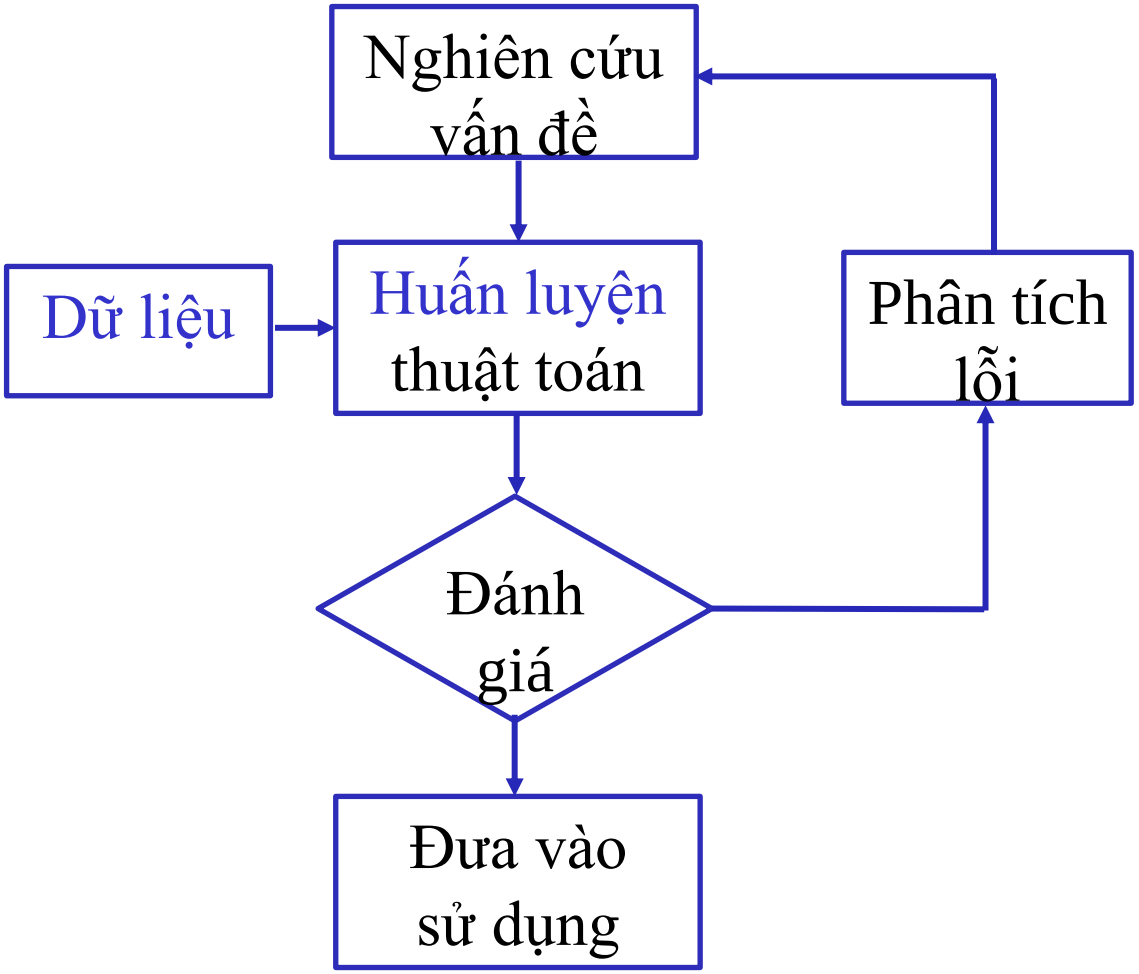


Thuật toán:
- danh sách dài các luật
- khó bảo trì

Khi vấn đề quá phức tạp hoặc không có giải thuật sẵn có ?

Ví dụ

- Bộ lọc thư rác
  - Cách tiếp cận sử dụng ML

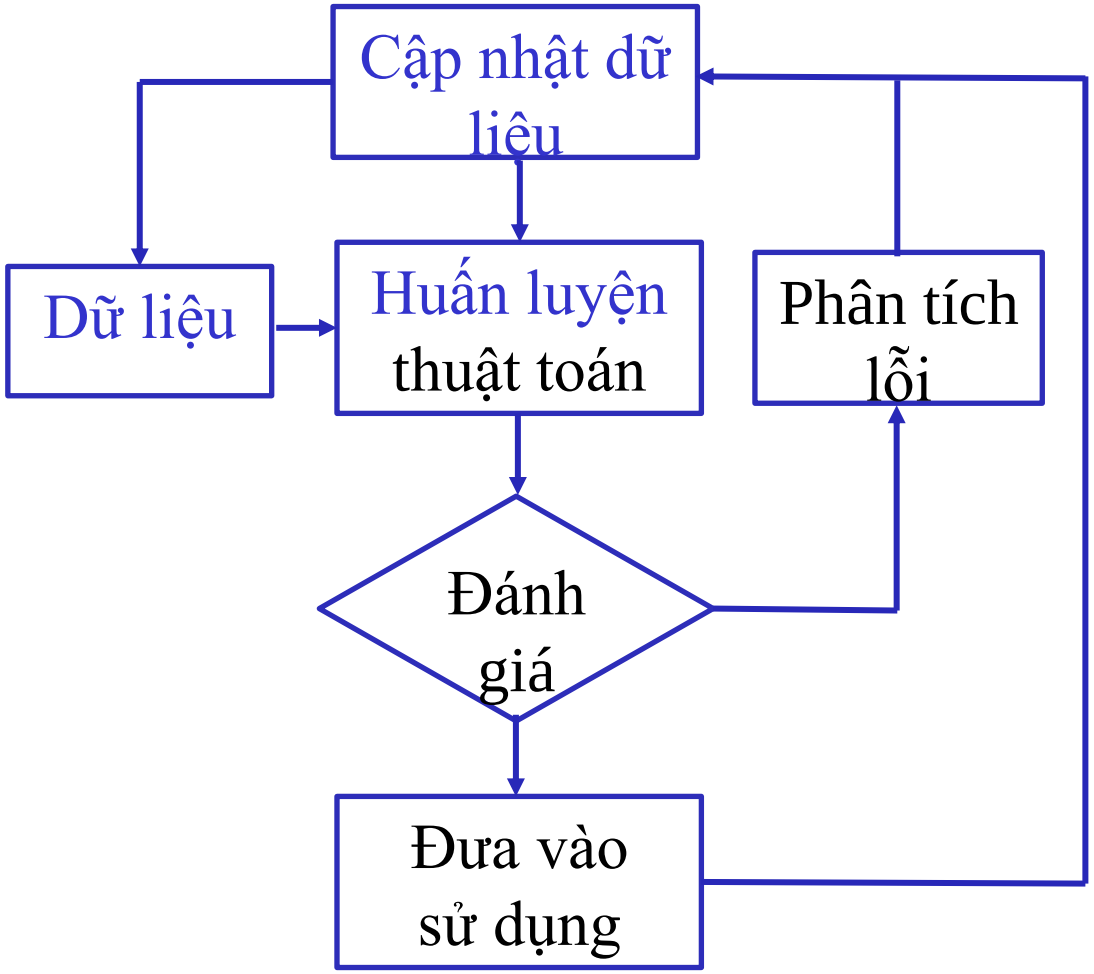


Dữ liệu

Huấn luyện, đào tạo

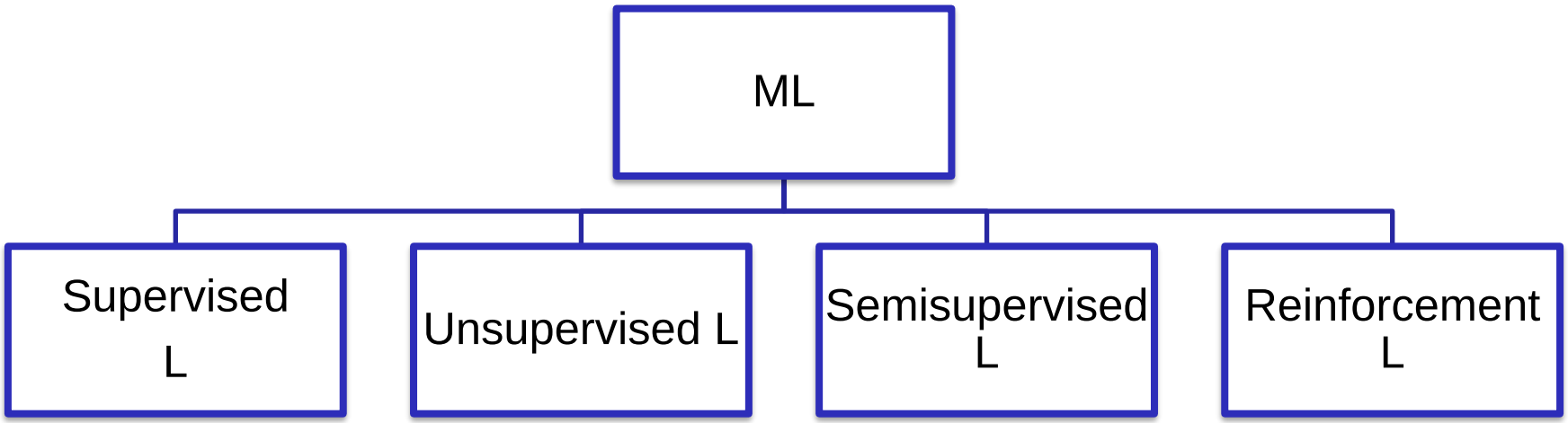
Ví dụ

- Bộ lọc thư rác
  - Cách tiếp cận sử dụng ML



1.1.2 Các loại học máy

- Có nhiều cách phân loại khác nhau
  - Khi xem xét đến số lượng và loại giám sát trong quá trình huấn luyện

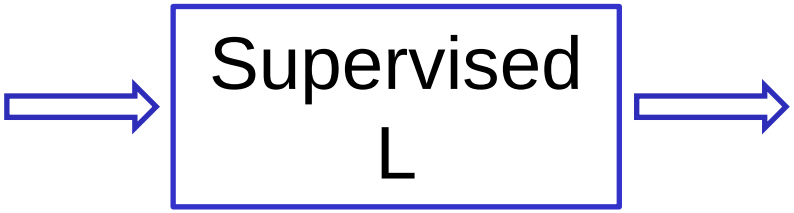


Học có giám sát

- Bộ dữ liệu huấn luyện được đánh nhãn

instance + label

mẫu *i* + nhãn *i*



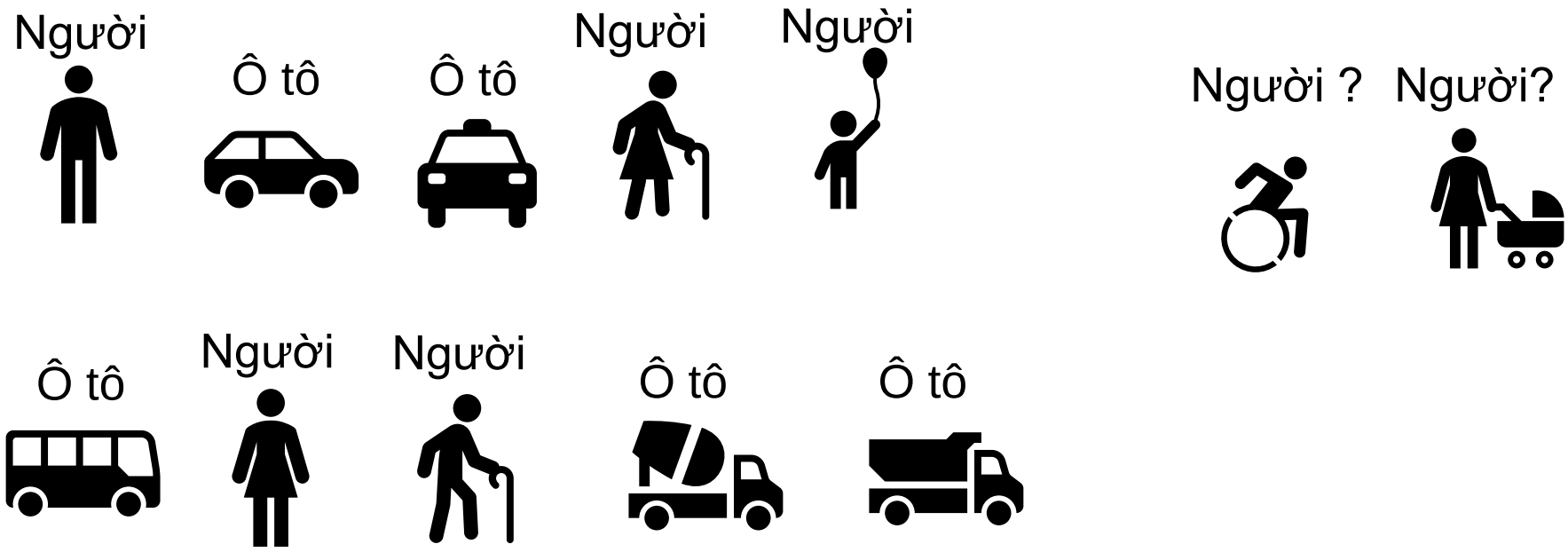
mẫu *x* — nhãn *x* — Hồi quy: Liên tục

mẫu *y* — nhãn *y* — Phân lớp: Rời rạc

Học có giám sát

- Bộ dữ liệu huấn luyện được đánh/gán nhãn

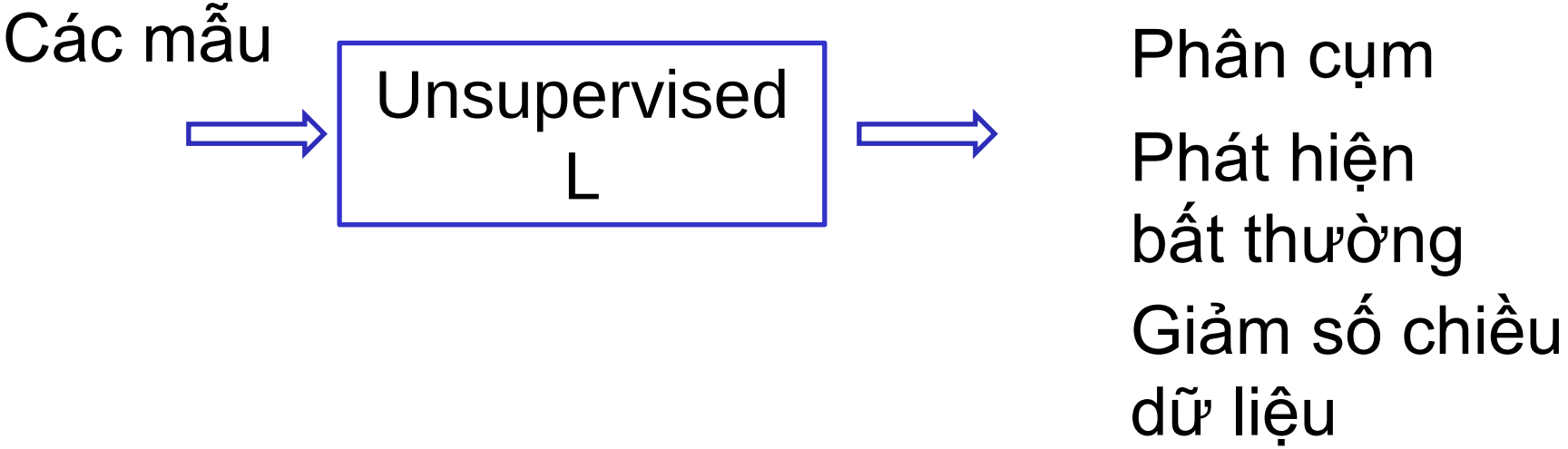
Phân lớp



Học không giám sát

- Bộ dữ liệu (huấn luyện) không đánh nhãn
  - Phân cụm, phát hiện bất thường, giảm số chiều, …

instance + ~~label~~



Học bán giám sát

- Bộ dữ liệu huấn luyện được đánh nhãn một phần

  instance + label
  instance

Học bán giám sát

- Dữ liệu được đánh nhãn một phần

- Do việc đánh nhãn dữ liệu tốn thời gian và chi phí
- Thường chỉ có 1 số lượng nhỏ dữ liệu được đánh nhãn.
- Phần lớn dữ liệu không được đánh nhãn

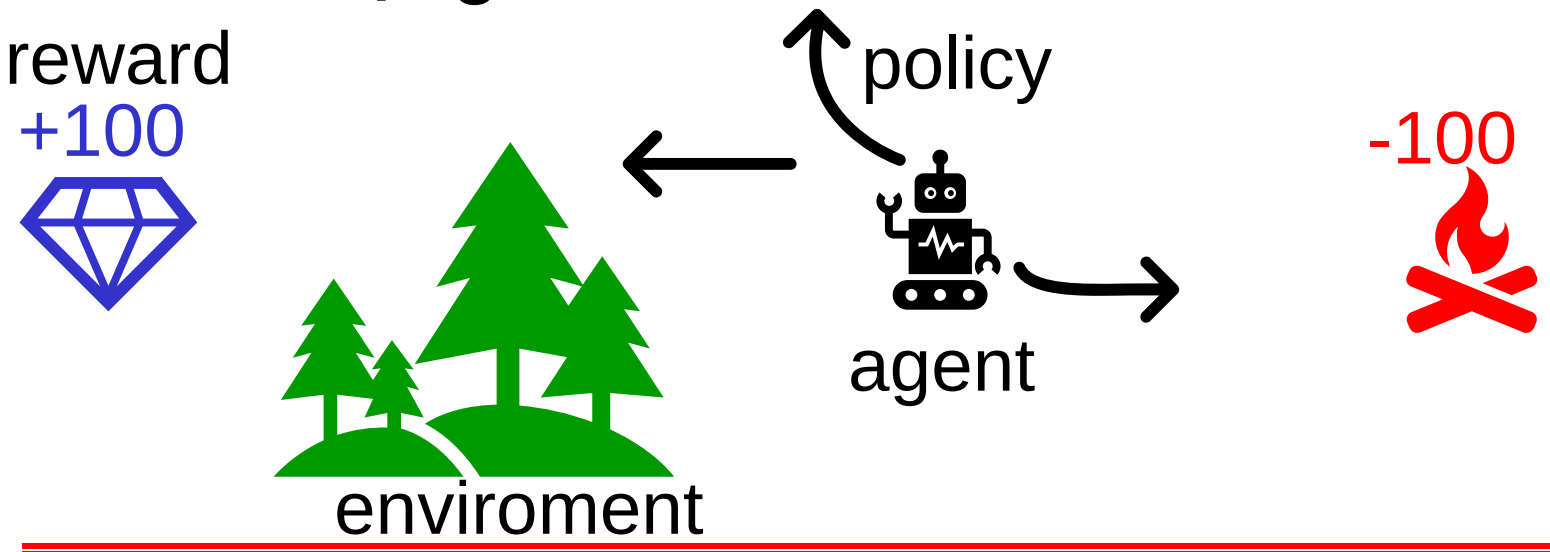
Học tăng cường

- Tự học để tìm ra chiến lược tốt nhất, nhằm đạt được nhiều điểm thưởng nhất

- Agent: quan sát môi trường, lựa chọn và hành động, để nhận được phần thưởng (hoặc bị phạt)
- Policy: định ra hành động mà agent nên lựa chọn khi nó ở trong một tình huống xác định

Học tăng cường

- Tự học để tìm ra chiến lược tốt nhất
  - Quan sát môi trường
  - Lựa chọn, thực hiện hành động
    - Nhận thưởng/phạt
- Trong một tình huống nhất định: biết nên hành động như thế nào



1.2 Các giải thuật trong học máy

- Học có giám sát
  - Linear regression
  - K-nearest neighbor
  - Support vector machine (SVM)
  - Decision tree
  - Logistic regression
  - Random forests

Các giải thuật trong học máy

- Học không có giám sát
  - Phân cụm
    - K-Means
    - DBSCAN
    - Hierarchical Cluster Analysis (HCA)

Các giải thuật trong học máy

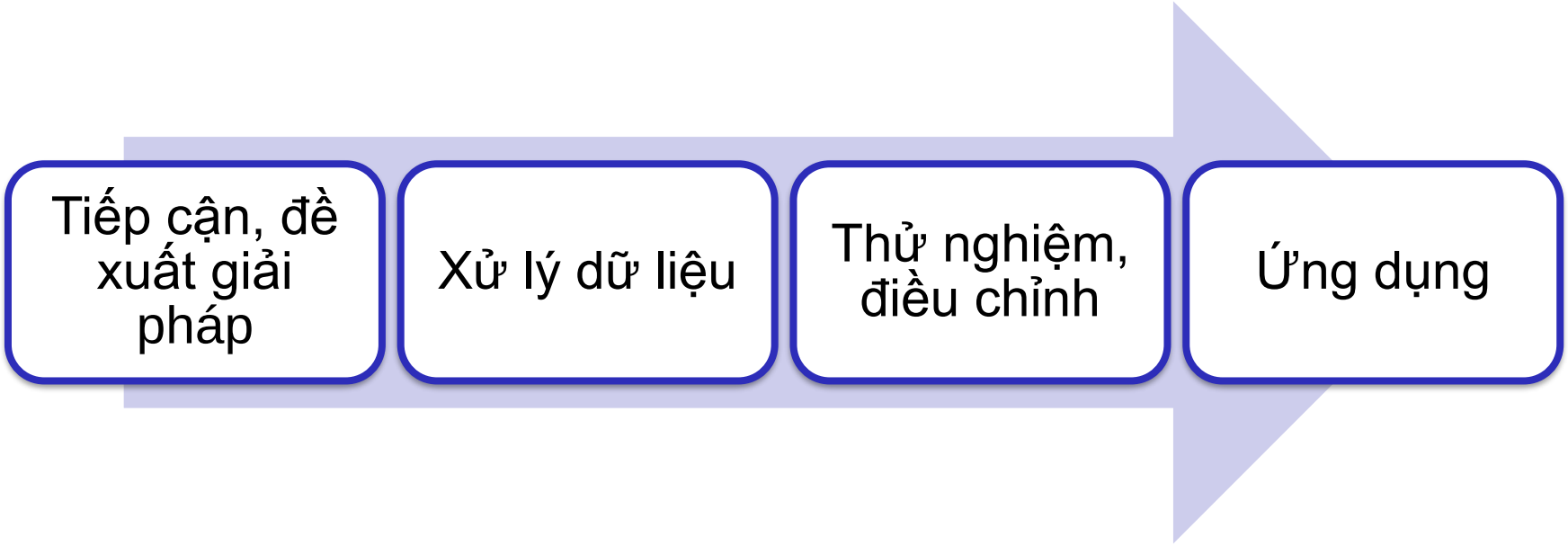
- Học tăng cường
  - Q-Learning
  - Deep Q-Learning

1.3 Xây dựng hệ thống học máy

1.3.1 Các bước xây dựng bài toán học máy

- Các bước chính
  - B1: Nhận định vấn đề và nắm được bức tranh chung
  - B2: Thu thập dữ liệu
  - B3: Khám phá và trực quan hóa dữ liệu
  - B4: Chuẩn bị dữ liệu cho các thuật toán ML
  - B5: Lựa chọn mô hình và huấn luyện
  - B6: Tinh chỉnh mô hình
  - B7: Trình bày giải pháp
  - B8: Phát hành, giám sát và duy trì hệ thống

Các bước xây dựng hệ thống học máy



Bước 1

Bước 2

Bước 3

Bước 4

Bước 5

Bước 5

Bước 6

Bước 7

Bước 8

Các bước

- B1: Nhận định vấn đề và nắm được bức tranh chung
  - Định nghĩa mục tiêu từ góc nhìn kinh doanh
  - Giải pháp sẽ được dùng như thế nào
  - Hiện nay đã có giải pháp nào chưa
  - Nhận định vấn đề cần giải quyết thế nào
  - Đo lường hiệu quả như thế nào
  - Những hiệu quả tối thiểu nào cần để đạt được mục đích kinh doanh
  - Có tồn tại những vấn đề tương tự không. Có thể sử dụng lại kinh nghiệm hay công cụ sẵn có không
  - Chuyên gia có sẵn không
  - Giải quyết vấn đề này một cách thủ công như thế nào
  - Liệt kê, kiểm chứng những giả thuyết
  - …

Các bước

- B2: Thu thập dữ liệu
  - Liệt kê dữ liệu cần và cần bao nhiêu
  - Tìm kiếm và ghi lại nơi nào có thể thu thập dữ liệu
  - Kiểm tra xem không gian lưu trữ như thế nào
  - Kiểm tra tính pháp lý, xin cấp phép nếu cần
  - Nhận cấp phép truy cập
  - Tạo không gian làm việc (với đủ không gian lưu trữ)
  - Thu thập dữ liệu
  - Chuyển dữ liệu vào định dạng có thể dễ dàng xử lý
  - Đảm bảo các thông tin nhạy cảm đã được xóa hoặc bảo vệ
  - Kiểm tra kích thước và kiểu dữ liệu
  - Chuẩn bị bộ dữ liệu kiểm tra, để riêng ra và không qua tâm đến nó
  - …

Các bước

- B3: Khám phá và trực quan hóa dữ liệu
  - Tạo ra một bản sao của dữ liệu để dùng trong việc khám phá dữ liệu
  - Tạo một Jupyter notebook để lưu trữ bản ghi của việc khám phá dữ liệu
  - Nghiên cứu từng thuộc tính và các đặc điểm của nó
  - Đối với học có giám sát, xác định thuộc tính đích
  - Trực quan hóa dữ liệu
  - Nghiên cứu sự tương quan giữa các thuộc tính
  - Nghiên cứu xem có thể giải quyết vấn đề một cách thủ công như thế nào
  - Xác định những phép biến đổi hứa hẹn có thể áp dụng
  - Xác định dữ liệu bổ sung có thể hữu ích
  - Ghi lại những điều đã học được
  - …

Các bước

- B4: Chuẩn bị dữ liệu cho các thuật toán ML
  - Làm sạch dữ liệu
  - Thực hiện lựa chọn đặc trưng
  - Thực hiện “feature engineering“ ở những nơi thích hợp
    - Rời rạc hóa các đặc trưng liên tục
    - Phân tích đặc trưng
    - Thêm vào những biến đổi hứa hẹn của đặc trưng
  - Thực hiện tỷ lệ hóa đặc trưng
  - …

Các bước

- B5: Lựa chọn mô hình và huấn luyện
  - Huấn luyện nhiều mô hình từ các nhóm khác nhau sử dụng các hệ số chuẩn
  - Đo lường và so sánh hiệu quả của chúng
  - Phân tích những biến có ý nghĩa nhất đối với từng thuật toán
  - Phân tích các kiểu lỗi mà mô hình tạo ra
  - Lập danh sách rút gọn từ 3 đến 5 mô hình hứa hẹn nhất, chú ý ưu tiên các mô hình tạo ra các kiểu lỗi khác nhau
  - …

Các bước

- B6: Tinh chỉnh mô hình
  - Tinh chỉnh các siêu hệ số sử dụng xác nhận chéo
  - Thử các phương pháp kết hợp.
    - Kết hợp các mô hình tốt nhất thường tạo ra hiệu quả cao hơn khi dùng riêng từng mô hình
  - Một khi đã tự tin về mô hình cuối cùng, đo lường hiệu quả của mô hình đó trên tập dữ liệu kiểm tra để ước lượng lỗi tổng quát hóa
  - …

Các bước

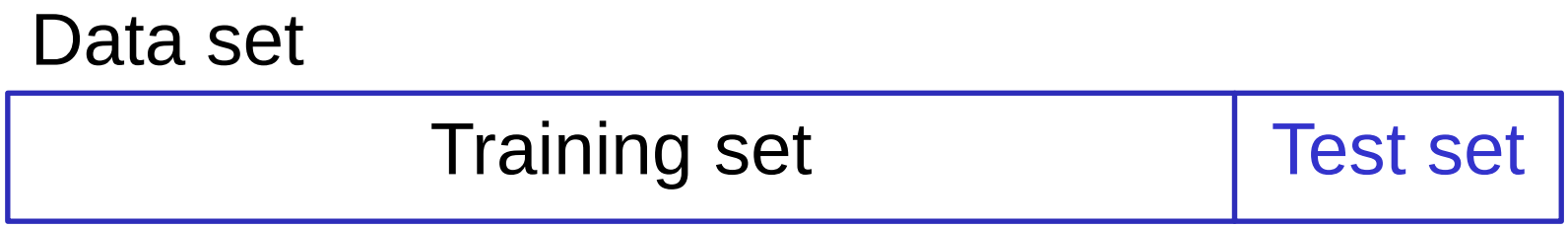
- B7: Trình bày giải pháp
  - Ghi lại những gì đã làm
  - Chuẩn bị bài trình bày một cách tốt nhất
  - Chắc chắn đã nhấn mạnh vào bức tranh chung (đã hình thành từ bước đầu tiên)
  - Giải thích tại sao giải pháp đạt được mục tiêu kinh doanh
  - Trình bày những điểm thú vị, đáng chú ý trong quá trình thực hiện
  - Đảm bảo những phát hiện trọng tâm được truyền tải thông qua đồ họa đẹp và câu phát biểu dễ nhớ
  - …

Các bước

- B8: Phát hành, giám sát và duy trì hệ thống
  - Đưa giải pháp đã sẵn sàng vào sản phẩm
  - Viết chương trình giám sát để kiểm tra hiệu quả hoạt động thực của hệ thống tại từng khoảng thời gian và các cảnh báo khi hệ thống bị lỗi
  - Huấn luyện lại mô hình dựa trên dữ liệu được làm mới
  - …

1.3.2 Đánh giá chất lượng mô hình

- Đánh giá mô hình: sử dụng bộ/tập dữ liệu kiểm tra



Dùng cho huấn luyện

⇩

Lỗi huấn luyện
Training error

Dùng cho
đánh giá/kiểm tra

⇩

Lỗi tổng quát
Generalization error
(test error)

Tỷ lệ lỗi trên các mẫu chưa biết/mới!!!

Overfitting

- Mô hình ML quá phù hợp/quá khớp với bộ dữ liệu huấn luyện
  - Thường gặp, Không mong muốn
  - Chú ý, sử dụng các kỹ thuật để tránh overfitting

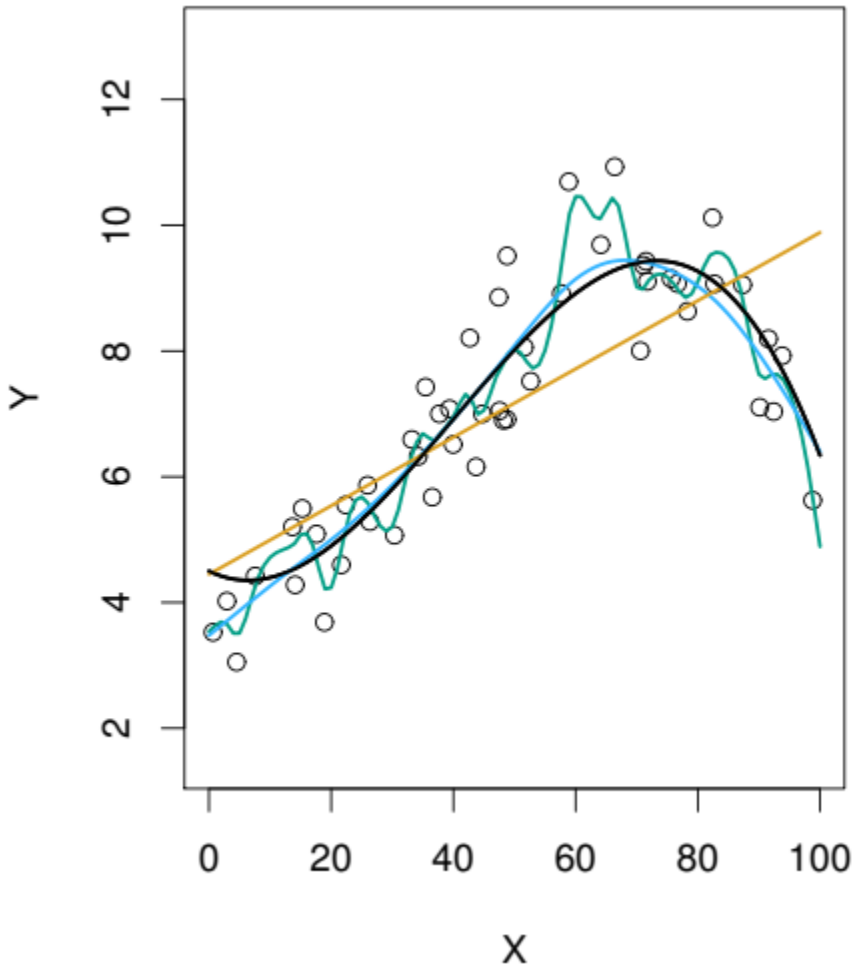
Data set

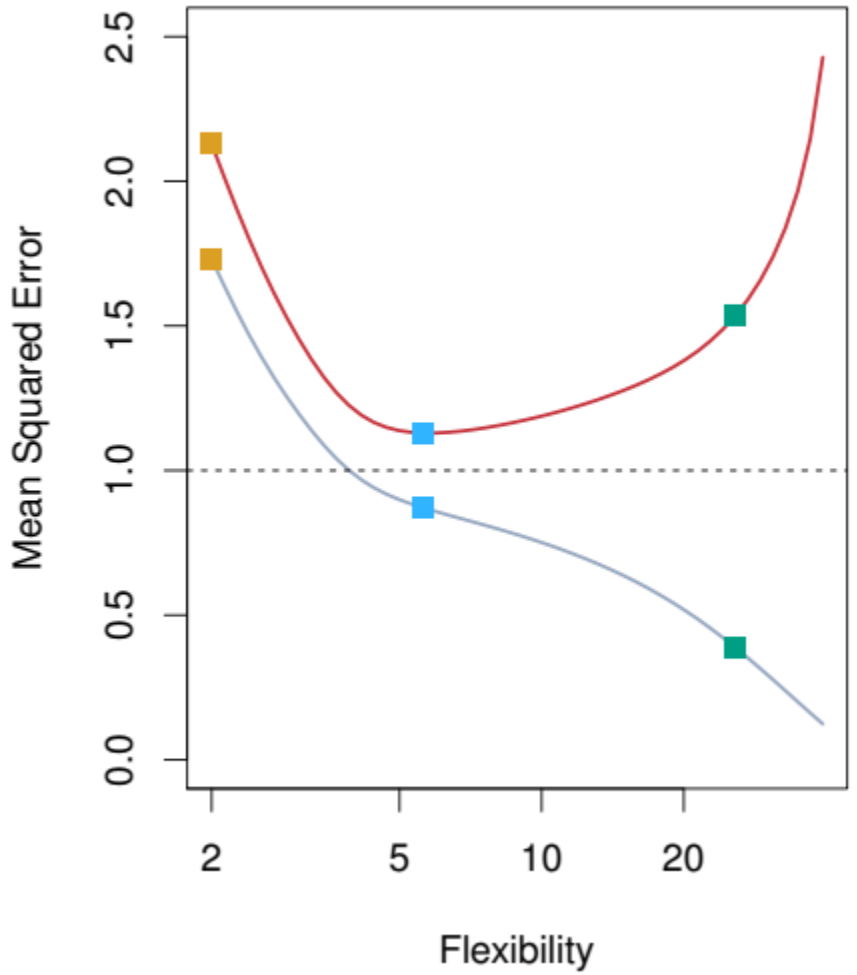
Lỗi huấn luyện: **thấp**

Lỗi kiểm tra: **cao**

Overfitting >< Underfitting

- Ví dụ:

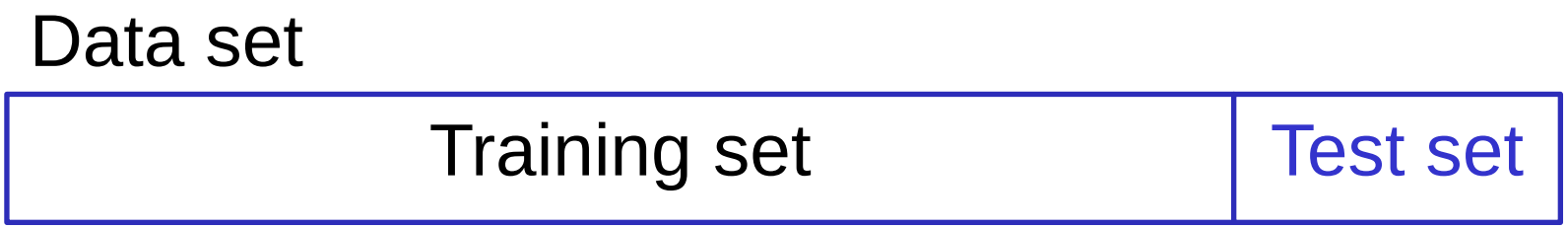




Textbook: The Elements of Statistical Learning Data Mining, Inference, and Prediction

Uderfitting

- Mô hình ML không phù hợp với bộ dữ liệu huấn luyện
  - Thường gặp, Không mong muốn
  - Khi chọn mô hình quá đơn giản



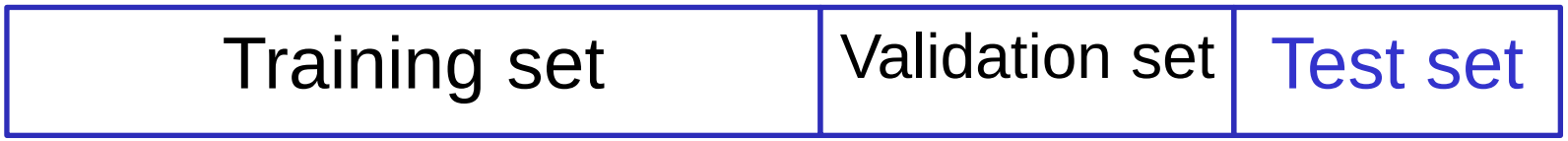
Lỗi huấn luyện: **cao**

Lỗi kiểm tra: **cao**

Bộ dữ liệu xác thực

- validatation set/development set/ dev set

Data set



Dùng cho huấn luyện

Lỗi huấn luyện
Training error

Dùng cho
xác thực

Lỗi xác thực
Validation error

Dùng cho
đánh giá/kiểm tra

Lỗi tổng quát
Generalization error
(test error)

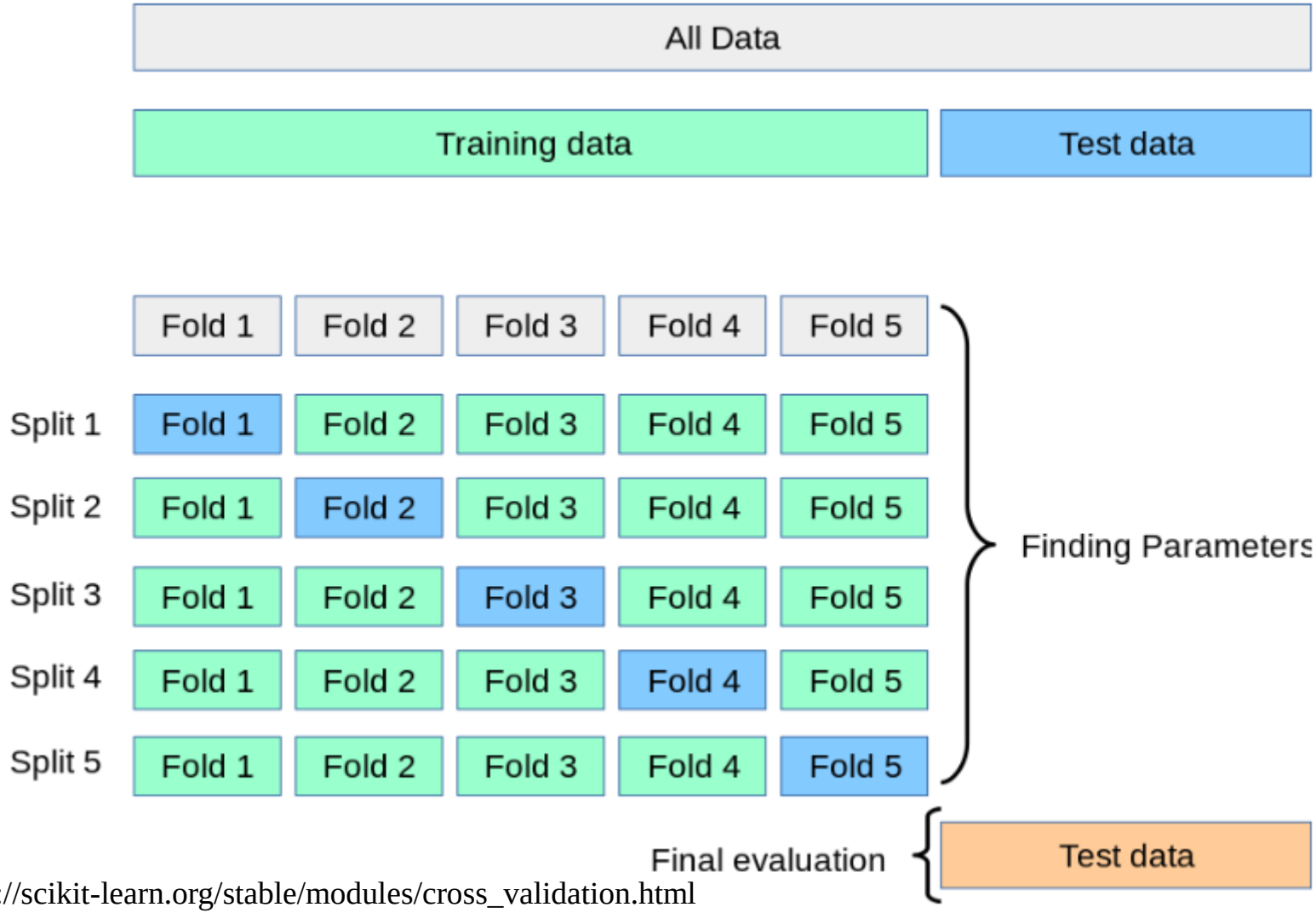
Tỷ lệ lỗi trên các mẫu chưa biết/mới!!!

Xác thực chéo

- Cross-validation
  - Thường chia tập dữ liệu ra *k* tập con (*k* = 5 hoặc *k* = 10) không giao nhau, có kích thước bằng nhau (*k*-fold cross-validation)
  - Nhược điểm: thời gian huấn luyện tỷ lệ với số lượng tập con *k*

Xác thực chéo

- Ví dụ: five-fold cross validation (scikit-learn)



https://scikit-learn.org/stable/modules/cross_validation.html

Regularization

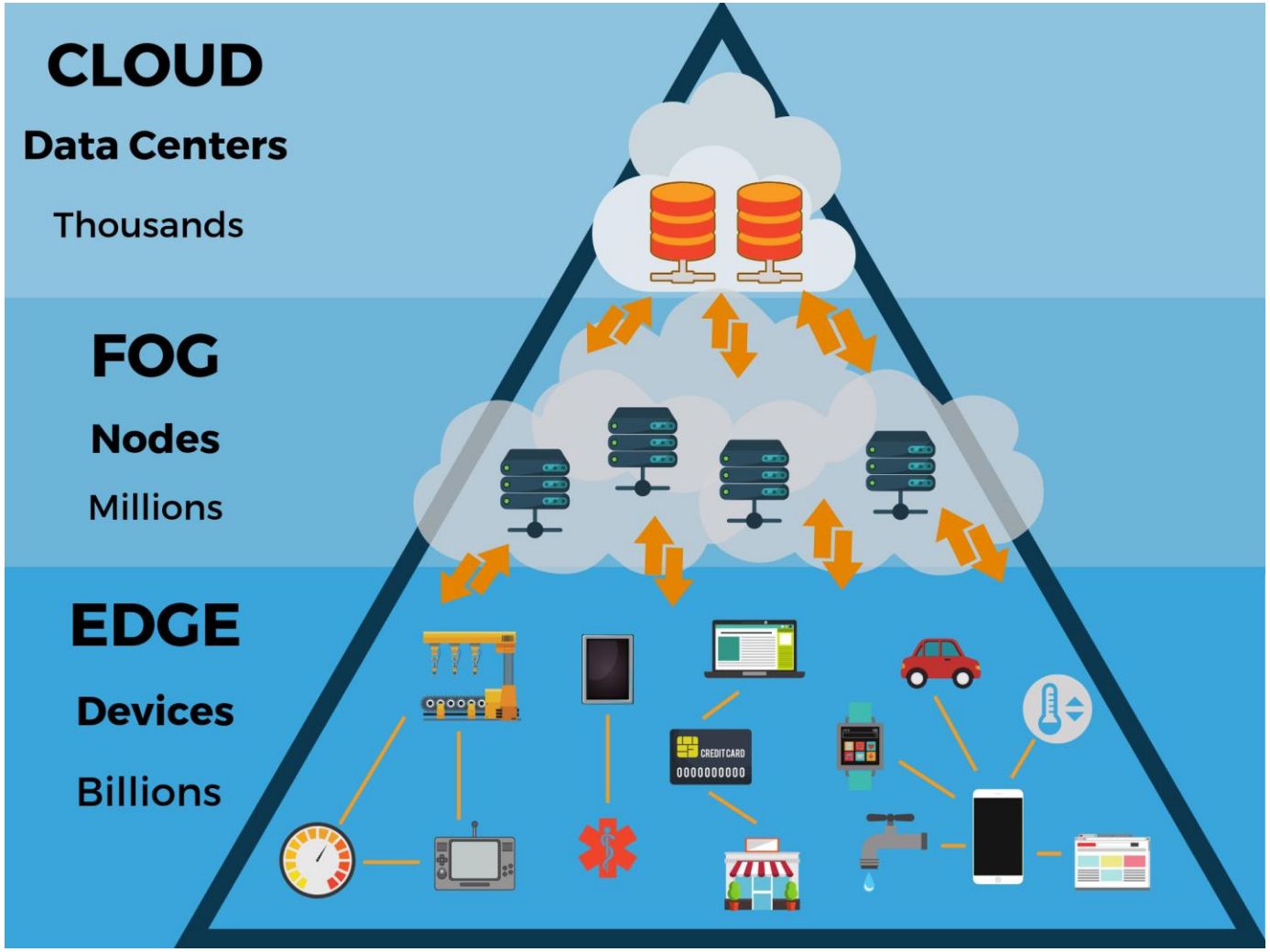
- Chấp nhận tăng training error
- Giúp giảm độ phức tạp của mô hình => tránh overfitting

Tinh chỉnh mô hình

- Grid search
- Evaluation metrics, scoring

1.3.3 Triển khai mô hình

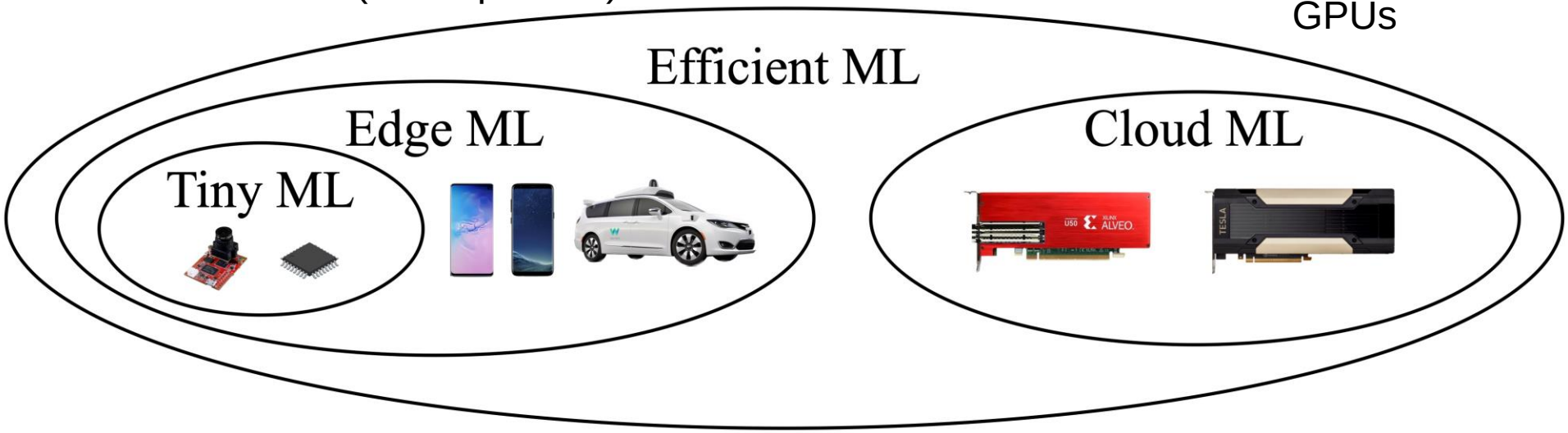
Cloud/fog/edge computing



https://www.thinkebiz.net/what-edge-computing/

Portable devices
(smart phones)

GPUs



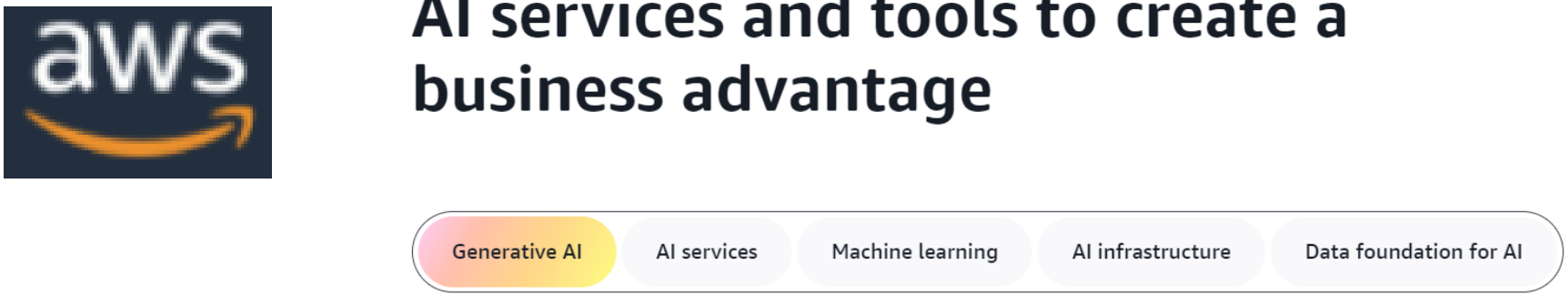
Ultra-low power
devices
(MCUs)

https://arxiv.org/html/2403.19076v1

- [TLHT2]

Amazon Web Services

- Cơ sở hạ tầng dưới dạng dịch vụ (IaaS) và nền tảng dưới dạng dịch vụ (PaaS)
  - Cung cấp các giải pháp có thể mở rộng cho điện toán, lưu trữ, cơ sở dữ liệu, phân tích, v.v.

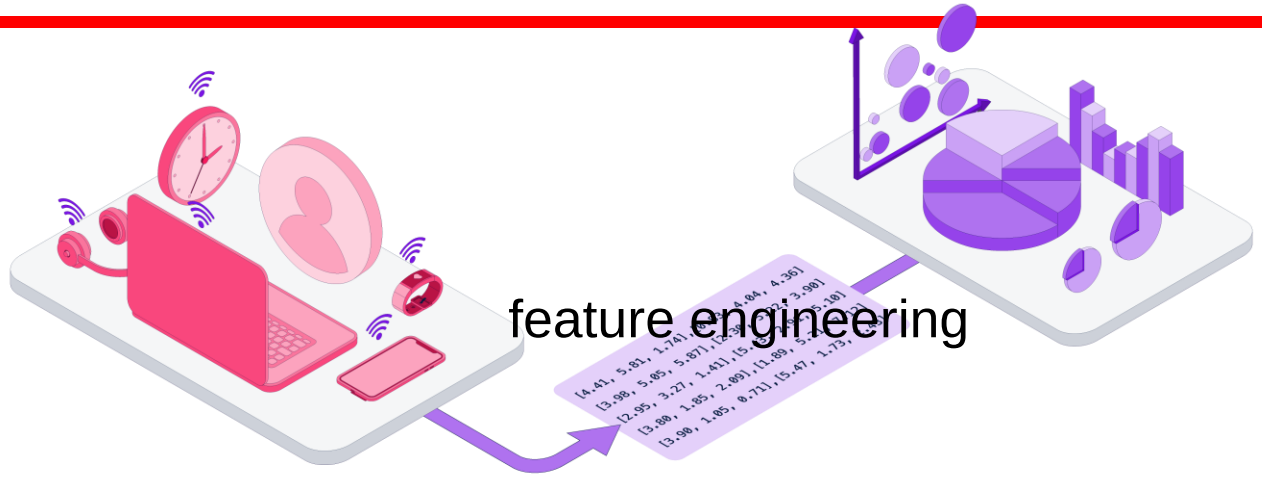


https://aws.amazon.com/

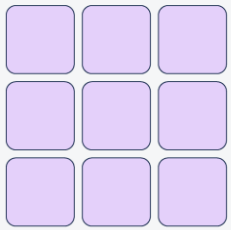
1.4 PP trích chọn đặc trưng của dữ liệu

- Liên quan đến việc chuyển đổi dữ liệu thô thành các biểu diễn có ý nghĩa mà các thuật toán ML có thể sử dụng một cách hiệu quả

Feature engineering



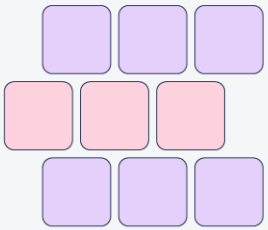
Structured



CLEAR SCHEMA, HIGHLY ORGANIZED

E.g., Name, address, geolocation, dates, contact, payment information, etc.

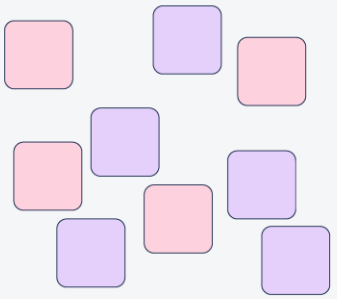
Semi-structured



LOOSE SCHEMA, TAGS, & SEMANTIC MARKERS

E.g., Email content by, from, to, inbox, sent, drafts; social media posts with hashtags, etc.

Unstructured



NO SCHEMA, QUALITATIVE

E.g., email body text, product descriptions, web text, speech, video, images, etc.

https://www.featureform.com/post/feature-engineering-guide

Đặc trưng số

- Dùng thống kê cơ bản
  - Giá trị trung bình, trung vị, phương sai, độ lệch, chuẩn …
- Giá trị tối đa, giá trị tối thiểu

Đặc trưng phân loại

- Categorical features (discrete features)
  - Thường không biểu diễn dưới dạng số
- One-hot-encoding

	age	workclass	education	gender	hours-per-week	occupation	income
0	39	State-gov	Bachelors	Male	40	Adm-clerical	<=50K
1	50	Self-emp-not-inc	Bachelors	Male	13	Exec-managerial	<=50K
2	38	Private	HS-grad	Male	40	Handlers-cleaners	<=50K
3	53	Private	11th	Male	40	Handlers-cleaners	<=50K
4	28	Private	Bachelors	Female	40	Prof-specialty	<=50K
5	37	Private	Masters	Female	40	Exec-managerial	<=50K
6	49	Private	9th	Female	16	Other-service	<=50K
7	52	Self-emp-not-inc	HS-grad	Male	45	Exec-managerial	>50K

[TLHT1]

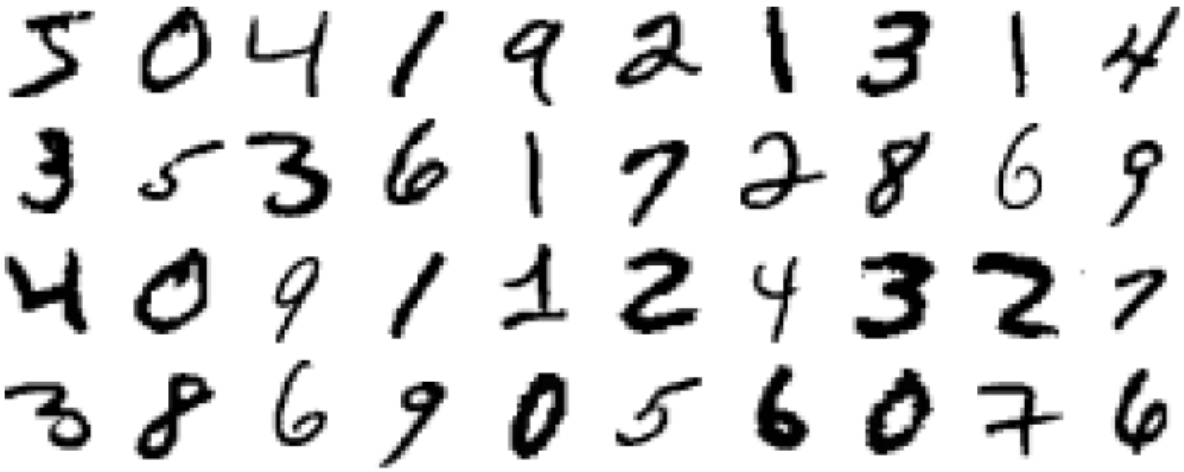
workclass	Government Employee	Private Employee	Self Employed	Self Employed Incorporated
Government Employee	1	0	0	0
Private Employee	0	1	0	0
Self Employed	0	0	1	0
Self Employed Incorporated	0	0	0	1

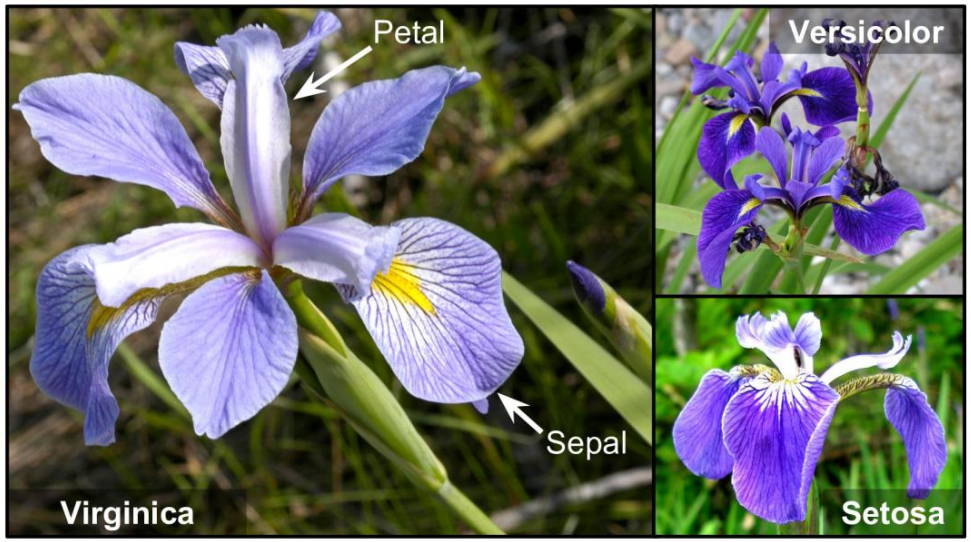
Văn bản

- Dữ liệu văn bản => biểu diễn số
  - BoW (Bag of Words)
    - Trình bày văn bản dưới dạng tập hợp số lượng từ hoặc tần số từ
  - TF-IDF (Term Frequency-Inverse Document Frequency)
    - Phản ánh tầm quan trọng của một từ trong văn bản so với toàn bộ văn bản
  - Word Embeddings:
    - Ví dụ: Word2Vec, GloVe, FastText, …
    - Nắm bắt ngữ nghĩa
  - …

Hình ảnh

- Biểu diễn như thế nào?





[TLHT2]

Xử lý tín hiệu

- Dữ liệu tín hiệu, chuỗi thời gian
  - Biến đổi Fourier
    - Chuyển đổi tín hiệu từ miền thời gian sang miền tần số
  - Biến đổi Wavelet
    - Phân tách tín hiệu thành các thành phần tần số khác nhau

Giảm số chiều của dữ liệu

- Giảm số lượng đặc trưng, vẫn giữ được thông tin cần thiết
  - PCA (Principal Component Analysis) – Phân tích thành phần chính
    - Chiếu dữ liệu vào không gian có chiều thấp hơn
  - LDA (Linear Discriminant Analysis) – Phân tích phân biệt tuyến tính
    - Tìm sự kết hợp tuyến tính của các đặc điểm phân tách các lớp
  - …
- Giảm: khối lượng tính toán, lưu trữ

Thảo luận

- Chủ đề: Feature engineering

1.5 Tập dữ liệu chuẩn (Dataset)

- Tập hợp dữ liệu được sử dụng để huấn luyện/đào tạo và đánh giá các mô hình học máy
- Điểm dữ liệu: mẫu và các thuộc tính (mô tả các khía cạnh khác nhau của dữ liệu)

Tập dữ liệu chuẩn

- Cấu trúc
  - Mẫu
    - Mỗi hàng đại diện cho một mẫu
  - Thuộc tính
    - Mỗi cột đại diện cho một thuộc tính
  - Nhãn
    - Trong học có giám sát
    - Cột biểu thị kết quả (dự đoán) hoặc lớp mà mô hình ML hướng tới

Dataset

- Chú ý

Sự liên quan	Chất lượng
Số lượng	Sự đa dạng

Dataset

- Sự liên quan
  - Tập dữ liệu phải phù hợp với vấn đề cần giải quyết
  - Chú ý tới các đặc trưng không liên quan hoặc dư thừa
  - Kỹ thuật trích chọn đặc trưng của dữ liệu đóng vai trò là rất quan trọng
    - để rút ra những hiểu biết có ý nghĩa
    - cải thiện độ chính xác của mô hình ML

Dataset

- Chất lượng
  - Cần dữ liệu chính xác, đầy đủ và mang tính đại diện
  - Các bước tiền xử lý dữ liệu
    - Lấy mẫu, làm sạch, chuẩn hóa và xử lý các giá trị bị thiếu …

Dataset

- Số lượng
  - Cần có đủ số lượng dữ liệu để huấn luyện mô hình ML
  - Nhiều dữ liệu hơn có thể giúp xây dựng mô hình ML hiệu quả
  - Chất lượng dữ liệu thường quan trọng hơn số lượng
    - Ví dụ: Một lượng lớn dữ liệu nhiễu hoặc không liên quan

Dataset

- Sự đa dạng
  - Bao gồm nhiều tình huống, trường hợp, biến thể khác nhau
    - Giúp tạo ra các mô hình có tính khái quát hơn
  - Giúp tránh những thành kiến và cải thiện khả năng ứng dụng của mô hình ML

Dataset

- Ví dụ
  - Dữ liệu số: Giá nhà đất, điểm thi, …
  - Dữ liệu phân loại: Các trường đại học, phản hồi khảo sát, danh mục sản phẩm, danh sách người dùng, …
  - Dữ liệu văn bản: Email, bài đăng trên mạng xã hội, đánh giá sản phẩm, …
  - Dữ liệu hình ảnh: hình ảnh y tế, hình ảnh vệ tinh, …
  - Dữ liệu chuỗi thời gian: thời tiết, chỉ số cảm biến, giao dịch tài chính, …
  - …

Dataset

- Kho dữ liệu mở phổ biến
  - UC Irvine Machine Learning Repository
  - Kaggle datasets
  - Amazon’s AWS datasets
  - OpenML platform
  - …

Tổng kết

- Hình dung được một bức tranh tổng quát về ML
- Phân biệt được các loại học máy
- Nắm được cách xây dựng hệ thống học máy
- Hình dung được các khó khăn, thách thức có thể gặp phải khi triển khai các ứng dụng thực tế

Hoạt động sau buổi học

- Ôn tập lại các kiến thức về toán học sử dụng trong ML
- Xem lại các câu lệnh cơ bản trong Python
- Làm BTVN

Chuẩn bị cho buổi học tiếp theo

- Đọc tài liệu tham khảo
  - Tóm tắt theo ý hiểu của mình về học có giám sát (supervised learning)
  - Tìm hiểu về 2 bài toán chính trong học có giám sát
    - Xấp xỉ hàm
    - Phân lớp

Tài liệu tham khảo

- Feature engineering
  https://www.featureform.com/post/feature-engineering-guide