MACHINE LEARNING

Tổng quan về học máy

TS. Trần Quang Quý

Các nội dung chính

Giới thiệu chung

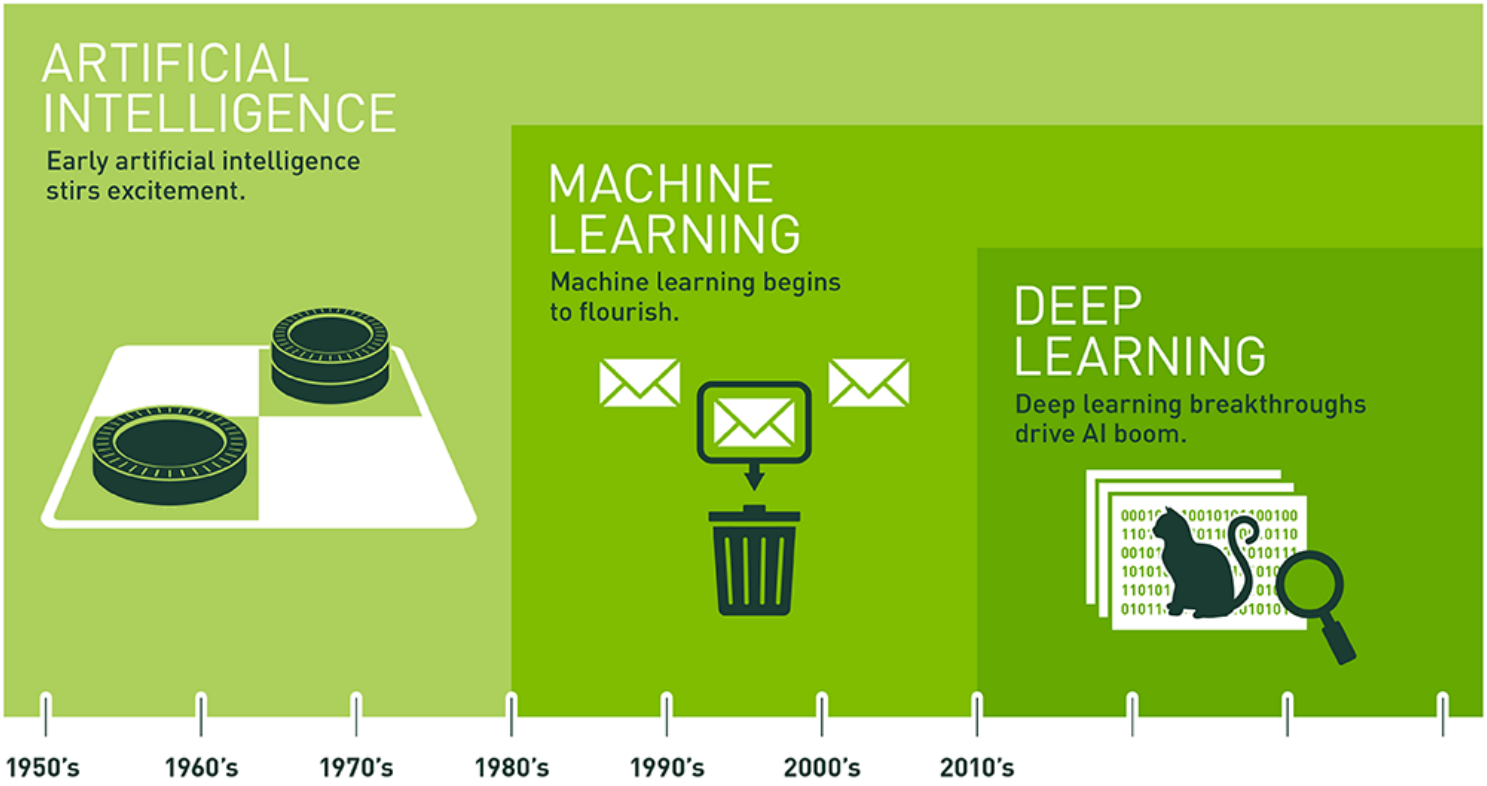
Trong nội dung này, giới thiệu tổng quan về học máy, các khái niệm và một số thuật ngữ chuyên ngành được dùng trong học máy. Các nội dung kiến thức được truyền tải như sau:

- Khái niệm về học máy, sử dụng học máy khi nào? Một số ứng dụng tiêu biểu của học máy.
- Phân loại học máy.
- Một số thách thức chính của học máy.
- Đánh giá và tinh chỉnh mô hình.
- Dự án học máy đầu tiên.

Khái niệm học máy

Những năm gần đây, **AI - Artificial Intelligence (Trí Tuệ Nhân Tạo)**, và cụ thể hơn là Machine Learning (Học Máy hoặc Máy Học) nổi lên như một bằng chứng của cuộc cách mạng công nghiệp lần thứ tư (1 - động cơ hơi nước, 2 - năng lượng điện, 3 - công nghệ thông tin). Trí Tuệ Nhân Tạo đang len lỏi vào mọi lĩnh vực trong đời sống mà có thể chúng ta không nhận ra. Xe tự hành của Google và Tesla, hệ thống tự tag khuôn mặt trong ảnh của Facebook, trợ lý ảo Siri của Apple, hệ thống gợi ý sản phẩm của Amazon, hệ thống gợi ý phim của Netflix, máy chơi cờ vây AlphaGo của Google DeepMind, ..., chỉ là một vài trong vô vàn những ứng dụng của AI/Machine Learning.

Mối quan hệ



Since an early flush of optimism in the 1950s, smaller subsets of artificial intelligence – first machine learning, then deep learning, a subset of machine learning – have created ever larger disruptions.

Figure 1: Mối quan hệ giữa AI, Machine learning và Deep learning

Khái niệm

- Học máy là một môn khoa học về cách lập trình máy tính để chúng có thể học từ dữ liệu.
- Học máy là lĩnh vực nghiên cứu nhằm giúp máy tính có khả năng học mà không cần lập trình một cách tường minh.
- Một chương trình máy tính được goi là "học tập" từ kinh nghiệm **E** để hoàn thành nhiệm vụ **T** với hiệu quả được đo bằng phép đánh giá **P**, nếu hiệu quả của nó khi thực hiện nhiệm vụ **T,** khi được đánh giá bởi **P,** cải thiện theo kinh nghiệm **E**.

Nói chung trong thực tế, một thuật toán học máy là một thuật toán có khả năng học từ dữ liệu.

Ví dụ về chương trình Học máy

Bộ lọc thư rác chính là một chương trình Học máy có khả năng học để phân loại đâu là thư rác từ các mẫu cho trước (thư rác được đánh dấu bởi người dùng) và các mẫu thư thường (không phải thư rác). Tập hợp các mẫu mà hệ thống dùng để học được gọi là *tập huấn luyện*. Mỗi mẫu dữ liệu huấn luyện được gọi là *mẫu huấn luyện*. Trong ví dụ này, tác vụ **T** là việc gán nhãn thư rác cho thư điện tử mới, kinh nghiệm **E** chính là dữ liệu huấn luyện, và ta cần định nghĩa thêm phép đo chất lượng **P**. Một lựa chọn khả thi cho P là tỷ lệ phân loại thư đúng, và phép đo chất lượng cụ thể này gọi là *độ chính xác*. Phép đo này thường được dùng trong các bài toán phân loại.

Tại sao cần Học máy?

Chúng ta cùng quan sát một bộ lọc thư rác bằng kỹ thuật lập trình truyền thống như hình dưới đây:

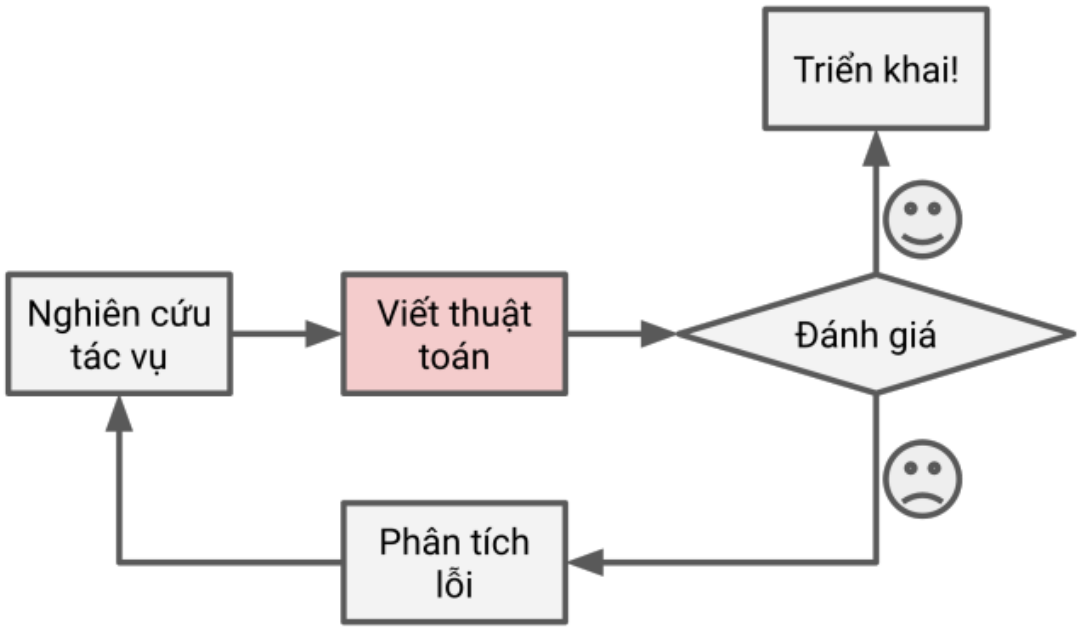


Figure 2: Tiếp cận theo lập trình truyền thống

Tiếp cận theo Học máy

Hình dưới đây miêu tả cho cách tiếp cận bài toán theo hướng Học máy, các kỹ thuật Học máy sẽ tự học và nhận diện các cụm từ nào là thư rác và có tần suất từ xuất hiện bất bình thường:

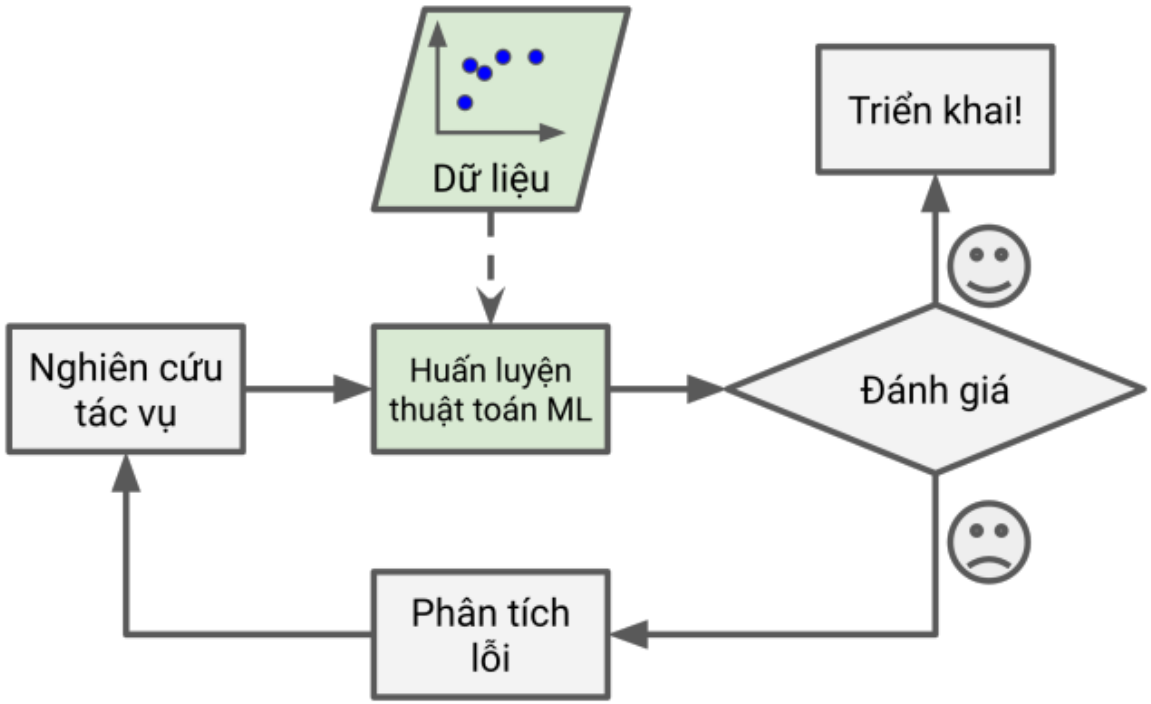


Figure 3: Tiếp cận theo hướng Học máy

Giúp con người học

Học máy có thể giúp con người học. Ta có thể kiểm tra các thuật toán học máy để biết những gì chúng đã học được. Việc áp dụng kỹ thuật học máy vào khai phá lượng dữ liệu lớn có thể giúp tìm ra các khuôn mẫu mà ta không thấy trực tiếp. Đây gọi là *khai phá dữ liệu (data mining).*

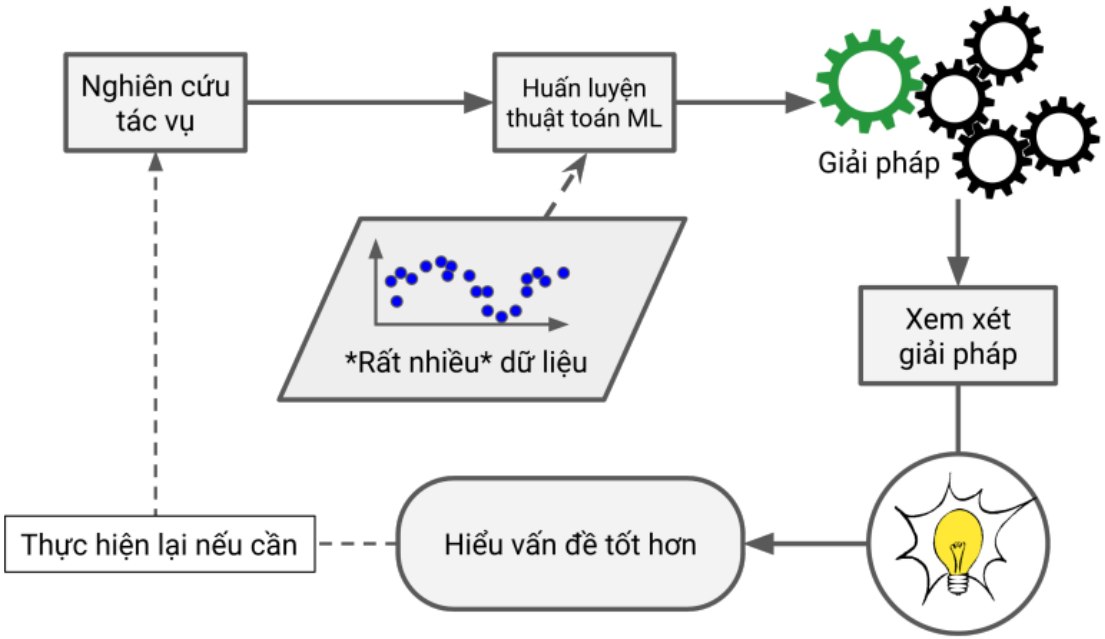


Figure 4: Học máy giúp con người học

Các ứng dụng tiêu biểu của học máy

1. Sử dụng mạng Nơ ron tích chập (CNN-Convolutional Neural Network) để phân tích hình ảnh và tự động phân loại sản phẩm trên dây chuyền sản xuất.
2. Phát hiện khối U trong ảnh quét não (bài toán phân vùng theo nhóm - semantic segmentation) sử dụng CNN.
3. Phân loại tin tức tự động (Là bài toán xử lý ngôn ngữ tự nhiên NLP - Natural Language Processing, sử dụng mạng nơ ron hồi tiếp RNN, CNN hoặc Transformer).
4. Tóm tắt tài liệu tự động
5. Tạo chatbot hoặc trợ lý cá nhân (Sử dụng NLP và hệ thống hỏi đáp).
6. Dự báo doanh thu của công ty, doanh nghiệp (sử dụng mô hình hồi quy - Hồi quy tuyến tính, hồi quy đa thức, SVM hồi quy, Rừng ngẫu nhiên Hồi quy hoặc ANN).

Các ứng dụng tiêu biểu của học máy (tiếp)

- Phát hiện gian lận thẻ tín dụng (bài toán phát hiện bất thường - anomaly dectection).
- Phân nhóm khách hàng dựa trên sản phẩm tiêu thụ và thiết kế chiến lược tiếp thị (bài toán phân cụm - clustering).
- Biểu diễn một tập dữ liệu phức tạp, nhiều chiều trong một biểu đồ một cách rõ ràng (bài toán trực quan dữ liệu, các kỹ thuật giảm chiều).
- Xây dựng hệ thống gợi ý khách hàng (ANN).
- Xây dựng bot thông minh biết chơi trò chơi (Thông qua học tăng cường - Reinforcement Learning - RL, ý tưởng là huấn luyện tác nhân để chọn các hành động sao cho phần thưởng được cực đại hóa theo thời gian, ví dụ: Con bot có thể được thưởng mỗi khi người chơi bị thua hoặc mất máu). AlphaGo là một chương trình rất nổi tiếng từng đánh bại nhà vô địch thế giới thông qua RL.

Các kiểu hệ thống Học máy

Theo phương thức học, các thuật toán Machine Learning thường được chia làm 4 nhóm:

1. Supervised learning.
2. Unsupervised learning.
3. Semi-supervised learning.
4. Reinforcement learning

Supervised learning - Học có giám sát

1. Supervised learning là thuật toán dự đoán đầu ra (outcome) của một dữ liệu mới (new input) dựa trên các cặp (*input, outcome*) đã biết từ trước. Cặp dữ liệu này còn được gọi là (*data, label*), tức (*dữ liệu, nhãn*). Supervised learning là nhóm phổ biến nhất trong các thuật toán Machine Learning. Bài toán Hồi quy và phân lớp là hai bài toán điển hình học máy có giám sát

2. Một cách toán học, Supervised learning là khi chúng ra có một tập hợp biến đầu vào $\mathcal{X} = \{\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_N\}$ và một tập hợp nhãn tương ứng $\mathcal{Y} = \{\mathbf{y}_1, \mathbf{y}_2, \dots, \mathbf{y}_N\}$, trong đó $\mathbf{x}_i, \mathbf{y}_i$ là các vector. Các cặp dữ liệu biết trước $(\mathbf{x}_i, \mathbf{y}_i) \in \mathcal{X} \times \mathcal{Y}$ được gọi là tập training data (dữ liệu huấn luyện). Từ tập training data này, chúng ta cần tạo ra một hàm số ánh xạ mỗi phần tử từ tập $\mathcal{X}$ sang một phần tử (xấp xỉ) tương ứng của tập $\mathcal{Y}$

$$\mathbf{y}_i \approx f(\mathbf{x}_i), \quad \forall i = 1, 2, \dots, N$$

Supervised learning - Học có giám sát

Trong học có giám sát, tập huấn luyện mà ta đưa vào thuật toán đã bao gồm cả kết quả mong muốn, kết quả đó được gọi là **nhãn**.

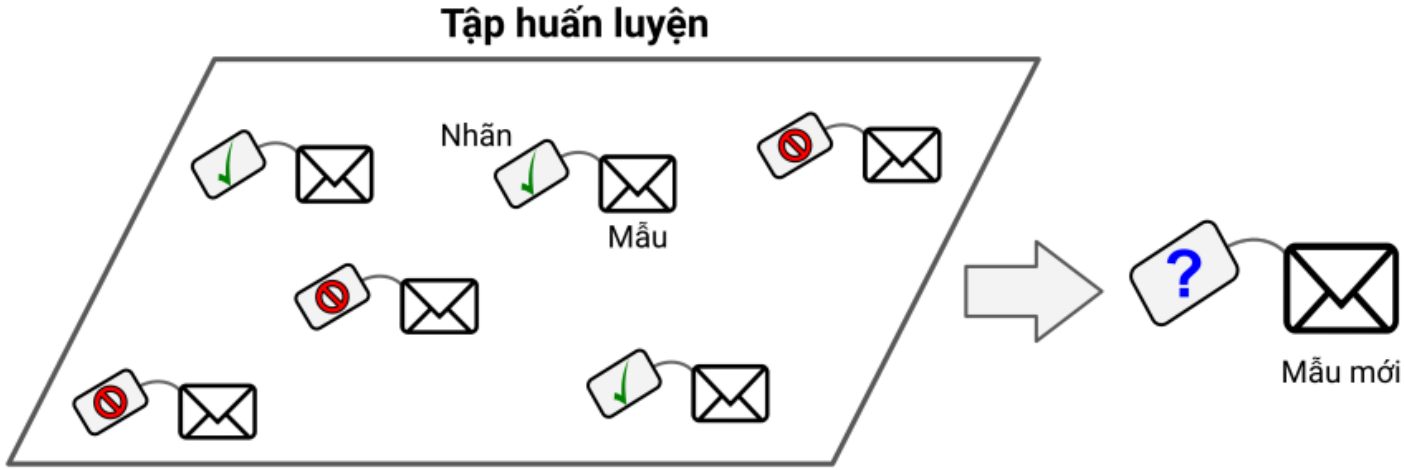


Figure 5: Bài toán phân loại - classification

Một số thuật toán học máy có giám sát

1. KNN - K điểm gần nhất
2. Linear Regression - Hồi quy tuyến tính
3. Logistic Regression - Hồi quy Logistic
4. SVM - Máy véc tơ hỗ trợ
5. Decision Tree - Cây quyết định
6. NN - mạng Nơ ron

Ví dụ về học có giám sát

1. **Ví dụ 1:** Thuật toán dò các khuôn mặt trong một bức ảnh đã được phát triển từ rất lâu. Thời gian đầu, facebook sử dụng thuật toán này để chỉ ra các khuôn mặt trong một bức ảnh và yêu cầu người dùng *tag friends* - tức gán nhãn cho mỗi khuôn mặt. Số lượng cặp dữ liệu (*khuôn mặt, tên người*) càng lớn, độ chính xác ở những lần tự động *tag* tiếp theo sẽ càng lớn.
2. **Ví dụ 2:** trong nhận dạng chữ viết tay, ta có ảnh của hàng nghìn ví dụ của mỗi chữ số được viết bởi nhiều người khác nhau. Chúng ta đưa các bức ảnh này vào trong một thuật toán và chỉ cho nó biết mỗi bức ảnh tương ứng với chữ số nào. Sau khi thuật toán tạo ra một mô hình, tức một hàm số mà đầu vào là một bức ảnh và đầu ra là một chữ số, khi nhận được một bức ảnh mới mà mô hình **chưa nhìn thấy bao giờ**, nó sẽ dự đoán bức ảnh đó chứa chữ số nào.

Ví dụ về học có giám sát

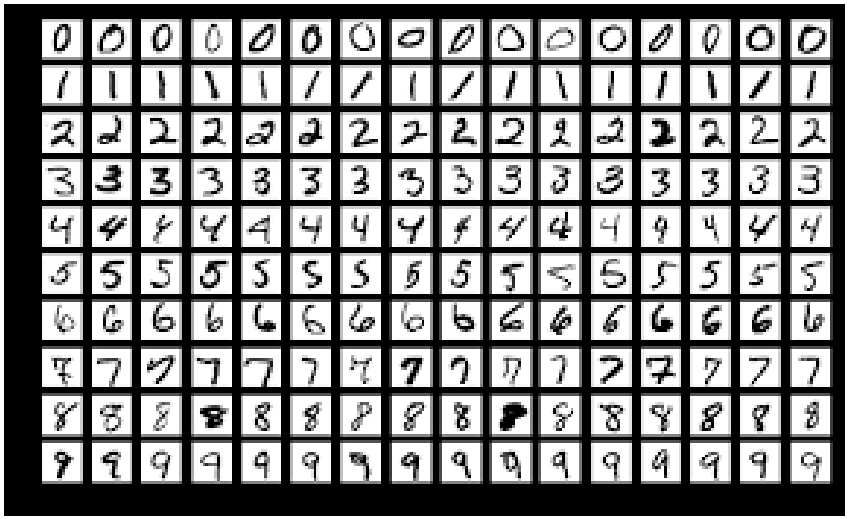


Figure 6: Tập dữ liệu MNIST

Cơ sở dữ liệu MNIST (tiếng Anh: **MNIST database**, viết tắt từ *Modified National Institute of Standards and Technology database*) là một cơ sở dữ liệu lớn chứa các chữ số viết tay thường được dùng trong việc huấn luyện các hệ thống xử lý hình ảnh khác nhau. Cơ sở dữ liệu này cũng được sử dụng rộng rãi để huấn luyện và kiểm thử trong lĩnh vực học máy

Unsupervised Learning (Học không giám sát)

1. Trong thuật toán này, chúng ta không biết được *outcome* hay *nhãn* mà chỉ có dữ liệu đầu vào. 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) hoặc giảm số chiều của dữ liệu (dimension reduction) để thuận tiện trong việc lưu trữ và tính toán.
2. Một cách toán học, Unsupervised learning là khi chúng ta chỉ có dữ liệu vào x mà không biết nhãn y tương ứng.
3. Những thuật toán loại này được gọi là Unsupervised learning vì không giống như Supervised learning, chúng ta không biết câu trả lời chính xác cho mỗi dữ liệu đầu vào. Giống như khi ta học, không có thầy cô giáo nào chỉ cho ta biết đó là chữ A hay chữ B. Cụm không giám sát được đặt tên theo nghĩa này.

Unsupervised Learning (Học không giám sát)

Trong học không giám sát. dữ liệu huấn luyện không được gán nhãn. Hệ thống cố gắng tự học mà không cần người quan sát (hướng dẫn). Một số thuật toán học máy không giám sát phổ biến bao gồm:

1. K-điểm trung bình (K-means)
2. DBSCAN
3. Phân cụm phân cấp (HCA - Hierarchical Cluster Analysis)
4. SVM một lớp (One - class SVM)
5. Rừng cô lập (Isolation Forest)
6. PCA - Phân tích thành phần chính (Principal Component Analysis)
7. Luật kết hợp - Association rules (Apriori)

Một số ví dụ học không giám sát

Phân cụm dữ liệu

Phương pháp này sử dụng khoảng cách giữa các điểm dữ liệu để tìm ra tâm cụm

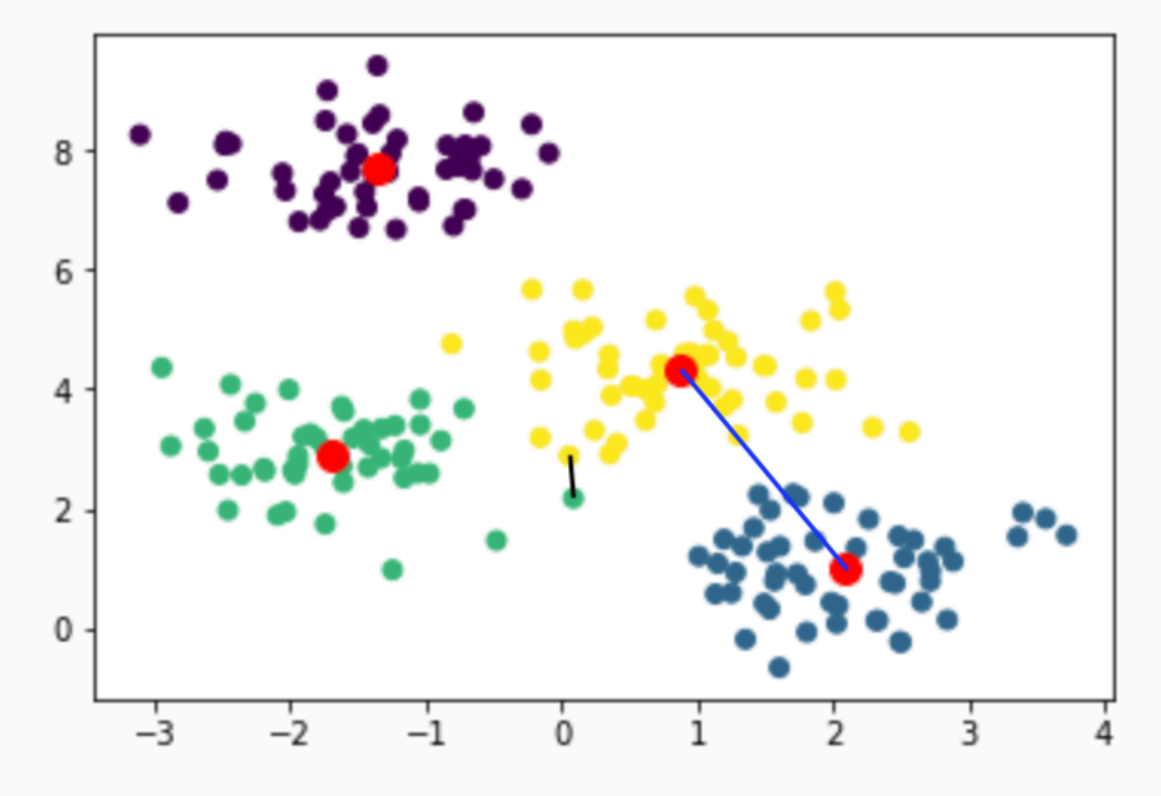


Figure 7: Clustering - Phân cụm K-means

Một số ví dụ học không giám sát

Khai phá luật kết hợp - Association rule

Phương pháp này sử dụng trong các bài toán liên quan đến phân tích giỏ hàng kinh doanh, thuật toán Apriori thường được sử dụng:

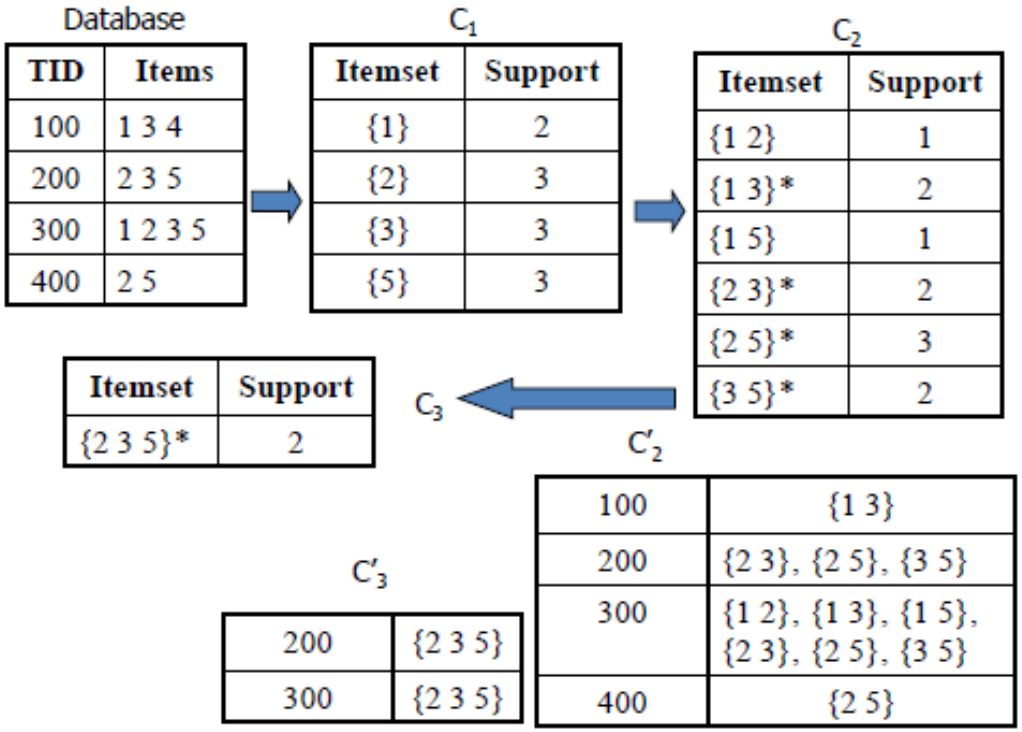


Figure 8: Apriori Algorithm

Semi-Supervised Learning (Học bán giám sát)

1. Các bài toán khi chúng ta có một lượng lớn dữ liệu $\mathcal{X}$ nhưng chỉ một phần trong chúng được gán nhãn được gọi là Semi-Supervised Learning. Những bài toán thuộc nhóm này nằm giữa hai nhóm được nêu bên trên.
2. Một ví dụ điển hình của nhóm này là chỉ có một phần ảnh hoặc văn bản được gán nhãn (ví dụ bức ảnh về người, động vật hoặc các văn bản khoa học, chính trị) và phần lớn các bức ảnh/văn bản khác chưa được gán nhãn được thu thập từ internet. Thực tế cho thấy rất nhiều các bài toán Machine Learning thuộc vào nhóm này vì việc thu thập dữ liệu có nhãn tốn rất nhiều thời gian và có chi phí cao. Rất nhiều loại dữ liệu thậm chí cần phải có chuyên gia mới gán nhãn được (ảnh y học chẳng hạn). Ngược lại, dữ liệu chưa có nhãn có thể được thu thập với chi phí thấp từ internet.

Ví dụ học bán giám sát

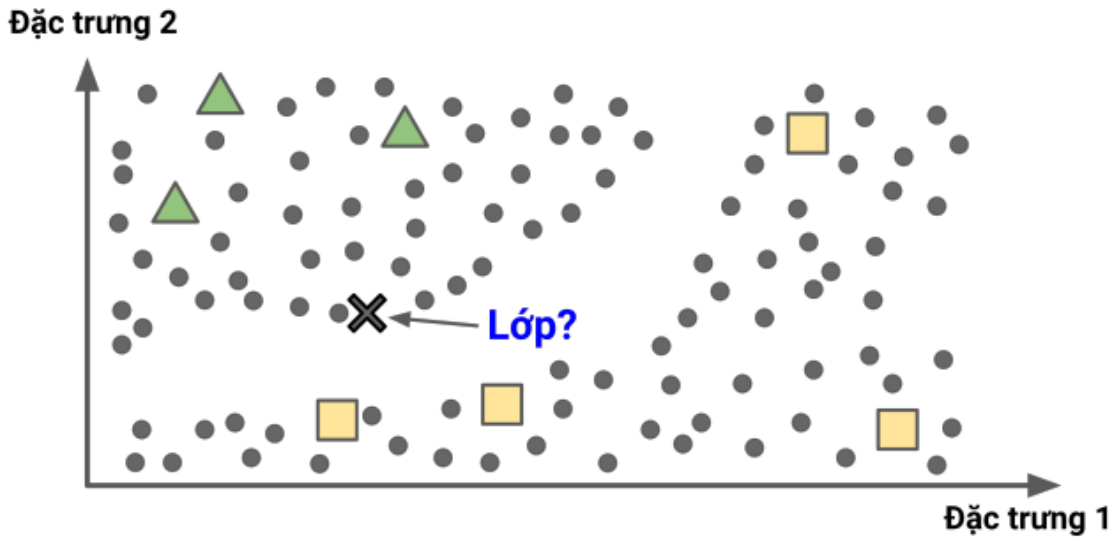


Figure 9: Học bán giám sát

Các dịch vụ lưu trữ của Google Photos là một tiêu biểu cho học bán giám sát. Bằng thuật toán phân cụm, dữ liệu ảnh tải lên sẽ được tự nhận dạng (ví dụ người A trong ảnh 1,5,11. Còn người B trong ảnh 2,5,7; Nhiệm vụ cuối cùng là chỉ ra những người này là ai bằng cách gán nhãn)

Reinforcement Learning (Học tăng cường)

1. **Reinforcement learning** là các bài toán giúp cho một hệ thống tự động xác định hành vi dựa trên hoàn cảnh để đạt được lợi ích cao nhất (maximizing the performance). Hiện tại, Reinforcement learning chủ yếu được áp dụng vào Lý Thuyết Trò Chơi (Game Theory), các thuật toán cần xác định nước đi tiếp theo để đạt được điểm số cao nhất.
2. **Huấn luyện cho máy tính chơi game Mario**. Đây là một chương trình thú vị dạy máy tính chơi game Mario. Đầu vào của thuật toán là sơ đồ của màn hình tại thời điểm hiện tại, nhiệm vụ của thuật toán là với đầu vào đó, tổ hợp phím nào nên được bấm. Việc huấn luyện này được dựa trên điểm số cho việc di chuyển được bao xa trong thời gian bao lâu trong game, càng xa và càng nhanh thì được điểm thưởng càng cao (điểm thưởng này không phải là điểm của trò chơi mà là điểm do chính người lập trình tạo ra). Thông qua huấn luyện, thuật toán sẽ tìm ra một cách tối ưu để tối đa số điểm trên, qua đó đạt được mục đích cuối cùng là cứu công chúa.

Reinforcement Learning (Học tăng cường)

Học tăng cường (RL) là phương pháp có cấu trúc rất khác. Hệ thống học trong RL được gọi là tác nhân (agent). Nó có thể quan sát môi trường xung quanh và chọn việc hành động, sau đó nhận về điểm thưởng (reward) hoặc lượng phạt (penalty) dưới dạng điểm thưởng âm, hệ thống sau đó cần tự học chiến lược tốt nhất để nhận về nhiều điểm thưởng nhất qua thời gian, và chiến lược này được gọi là chính sách (policy).

Reinforcement Learning (Học tăng cường)

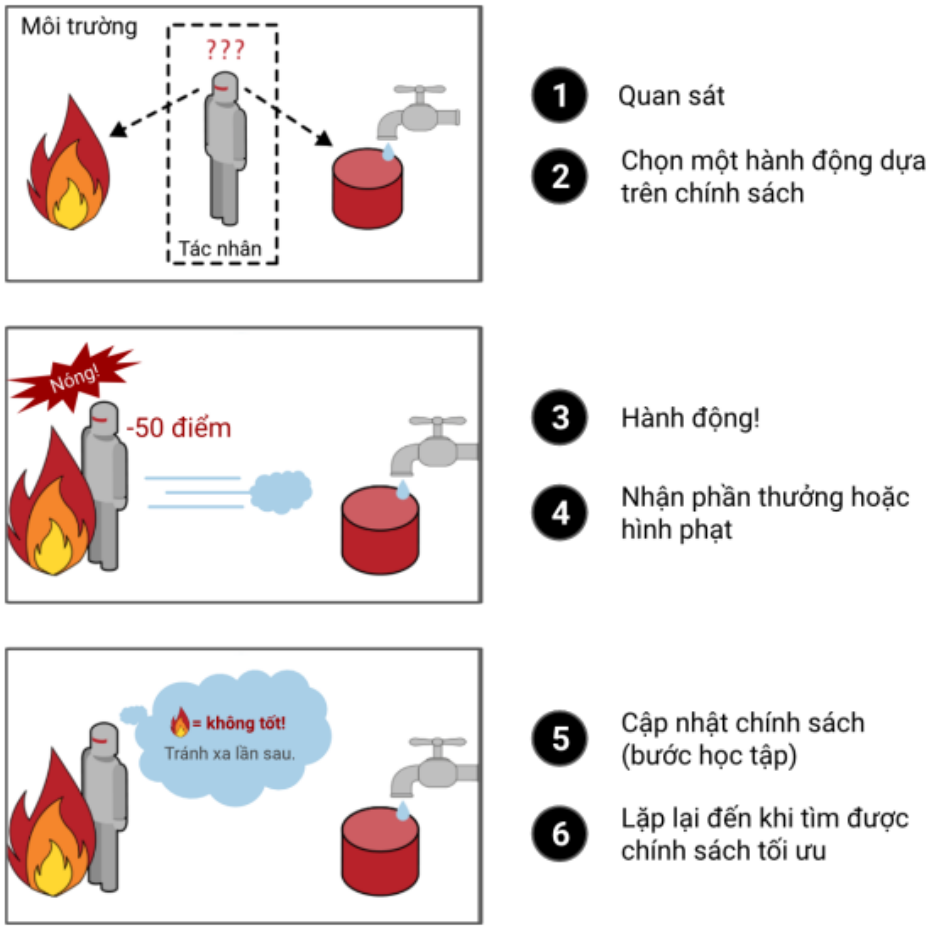


Figure 10: Học tăng cường

Huấn luyện game Mario

Xem tại link sau cách huấn luyện game Mario để đạt được điểm cao nhất sử dụng học tăng cường:

Bấm vào đây

Các bài toán cơ bản trong học máy

Trong học máy, tùy thuộc vào các dạng học máy mà chúng ta có những bài toán cơ bản trong học máy như sau:

1. Bài toán phân lớp - Classification
2. Bài toán hồi quy - Regression
3. Bài toán máy dịch
4. Bài toán phân cụm - Clustering
5. Bài toán hoàn thiện dữ liệu

Bài toán phân lớp

Khái niệm

Đây là dạng bài toán được nghiên cứu nhiều nhất trong học máy. Trong bài toán này, chương trình được yêu cầu xác định *lớp/nhãn* của một điểm dữ liệu trong số $\mathcal{C}$ nhãn khác nhau. Cặp (dữ liệu, nhãn) được ký hiệu là (x,y) với y nhận một trong $\mathcal{C}$ giá trị trong tập đích $\mathcal{Y}$. Việc xây dựng mô hình tương đương với việc đi tìm hàm số f ánh xạ một điểm dữ liệu x vào một phần tử $y \in \mathcal{Y} : y = f(x)$ trên tập dữ liệu huấn luyện đã biết sao cho sau đó mô hình được áp dụng trên các dữ liệu mới với sai số chấp nhận được. Khi đó ta xác định được nhãn của dữ liệu mới.

Bài toán hồi quy

Khái niệm

Nếu tập đích $\mathcal{Y}$ gồm các gía trị là các số thực thì bài toán được gọi là *hồi quy*. Trong bài toán này, ta cần xây dựng một hàm số $f : \mathbb{R}^d \to \mathbb{R}^m$, ánh xạ mỗi véc tơ dữ liệu với nhãn $y \in \mathcal{Y}$, đầu ra mong muốn dự đoán.
Ví dụ: Ước lượng giá của một căn nhà rộng $x\ m^2$, có y phòng ngủ và cách trung tâm thành phố $z\ km$. Ta mong muốn xây dựng hàm dự đoán dạng: $f : \mathbb{R}^3 \to \mathbb{R}$, với giá $= f(x, y, z)$
Bài toán hồi quy có thể mở rộng ra việc dự đoán nhiều đầu ra cùng một lúc, khi đó hàm cần tìm sẽ là $f : \mathbb{R}^d \to \mathbb{R}^m$. Một ví dụ là bài toán tạo ảnh độ phân giải cao từ một ảnh có độ phân giải thấp hơn. Khi đó, việc dự đoán giá trị các điểm ảnh trong đầu ra là một bài toán hồi quy nhiều đầu ra.

Bài toán máy dịch

Khái niệm

Trong bài toán **máy dịch**, chương trình máy tính được yêu cầu dịch một đoạn văn từ một ngôn ngữ sang một ngôn ngữ khác. Dữ liệu huấn luyện là các cặp văn bản song ngữ. Các văn bản này có thể chỉ gồm hai ngôn ngữ đang xét hoặc có thêm các ngôn ngữ trung gian.
Ví dụ: Google Translate là một ứng dụng (với nhiều kỹ thuật học sâu khác được áp dụng) học máy trong lĩnh vực máy dịch.

Bài toán phân cụm

Khái niệm

Bài toán phân cụm là bài toán chia tập dữ liệu $\mathcal{X}$ thành các cụm nhỏ dựa trên sự liên quan nào đó giữa các dữ liệu trong mỗi cụm. Trong bài toán này, dữ liệu huấn luyện không có nhãn, mô hình tự phân chia dữ liệu thành các cụm khác nhau dựa vào độ đo sự gần gũi của các đặc trưng dữ liệu.

Ví dụ: Phân cụm khách hàng dựa trên hành vi mua hàng. Dựa trên việc mua bán và theo dõi của người dùng trên một trang web thương mại điện tử, mô hình có thể phân người dùng vào các cụm theo sở thích mua hàng. Từ đó, mô hình có thể quảng cáo các mặt hàng mà người dùng có thể quan tâm trên cơ sở những nhóm người đã mua, đã quan tâm.

Bài toán hoàn thiện dữ liệu

Khái niệm

Bài toán hoàn thiện dữ liệu là bài toán dự đoán các trường dữ liệu còn thiếu trong dữ liệu đích. Nhiệm vụ của bài toán này là dựa trên mối tương quan giữa các điểm dữ liệu để dự đoán những giá trị còn thiếu. Các hệ thống khuyến nghị là một dạng điển hình của bài toán này. Các phần mềm tô màu ảnh cũ, khôi phục ảnh sử dụng học máy thuộc phạm vi của lĩnh vực hoàn thiện dữ liệu.

Câu hỏi ôn tập

Nội dung

Bài toán phân loại e-mail rác là bài toán phân loại điển hình trong học máy. Hãy cho biết:

- Nhiệm vụ T trong bài toán là gì?
- Phép đánh giá P trong bài toán học máy này là gì?
- Kinh nghiệm E trong bài toán này là gì?

Hàm mất mát và các tham số của mô hình

Hàm mất mát

Mỗi một mô hình học máy được mô tả bởi các bộ **tham số mô hình** . Công việc của mỗi thuật toán học máy là đi tìm các tham số mô hình tối ưu cho mỗi bài toán, nghĩa là đạt được tốt nhất theo phép đánh giá P. Ví dụ trong bài toán hồi quy, kết quả tốt là khi sự sai lệch giữa đầu ra dự đoán và đầu ra thực sự của dữ liệu là nhỏ.

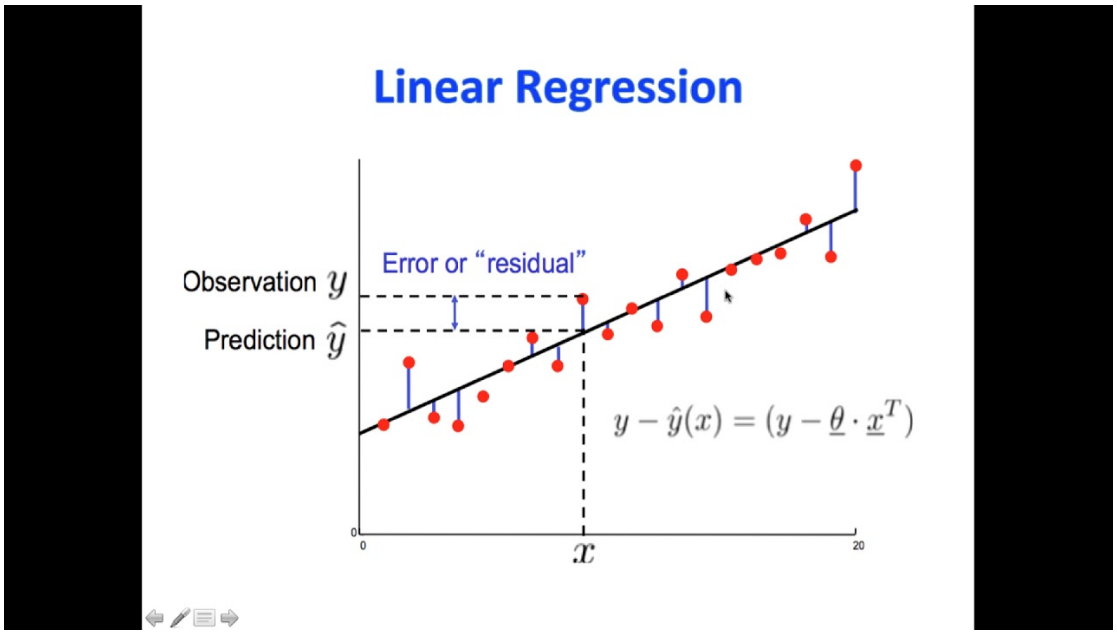


Figure 11: Loss function

Hàm mất mát

Quan hệ giữa một phép đánh giá và các tham số của mô hình được mô tả thông qua một hàm số được gọi là **hàm mất mát**. Như vậy việc xây dựng một mô hình học máy chính là việc đi giải một bài toán tối ưu. Quá trình đó được coi là quá trình *học* của *máy*.

Tập hợp các tham số mô hình thường được ký hiệu là θ, hàm mất mát của mô hình được ký hiệu là $\mathcal{L}(\theta)$ hoặc $\mathcal{J}(\theta)$. Bài toán đi tìm tham số của mô hình tương đương với bài toán tối thiểu hàm mất mát:

$$\theta^* = \arg\min_{\theta} \mathcal{L}(\theta)$$

Trong đó, ký hiệu $\arg\min_{\theta} \mathcal{L}(\theta)$ được hiểu là giá trị của θ để hàm số $\mathcal{L}(\theta)$ đạt giá trị nhỏ nhất.

Sai số huấn luyện và sai số kiểm tra

Khái niệm

Sai số huấn luyện chính là đại lượng đo sự khác biệt giữa giá trị dự đoán của mô hình và giá trị thực tế của dữ liệu, trong thực tế, đây chính là dữ liệu của hàm mất mát khi áp dụng lên dữ liệu huấn luyện, hàm mất mát cần có một thừa số gọi là N_{hl} chính là số lượng dữ liệu huấn luyện, để tính giá trị trung bình mất mát trên mỗi điểm dữ liệu. Với bài toán hồi quy, đại lượng này được đo bởi bình phương sai số:

$$sai\ s\ hun\ luyn = \frac{1}{N_{hl\ tp\ hun\ luyn}} \sum \|y - \hat{y}\|^2$$

Với $\|y - \hat{y}||^2$ là chuẩn Euclid của véc tơ.
Sai số kiểm tra hoàn toàn được tính giống như trong sai số huấn luyện, tuy nhiên khi này áp dụng vào trong tập dữ liệu kiểm tra chứ không phải trong tập dữ liệu huấn luyện.

Thách thức đối với học máy

Vì nhiệm vụ chính của học máy là chọn và huấn luyện một thuật toán trên một tập dữ liệu, nên hai vấn đề có thể xảy ra là thuật toán kém và dữ liệu xấu. Thách thức đầu tiên là dữ liệu xấu, bao gồm:

1. Dữ liệu xấu là một nguyên nhân quan trọng trong học máy, việc dữ liệu không đủ để huấn luyện sẽ dẫn đến hệ thống học máy kém, các thuật toán học máy cần rất nhiều dữ liệu để học và hoạt động hiệu quả. Với dữ liệu như giọng nói, hình ảnh cần tới hàng triệu mẫu.
2. Dữ liệu huấn luyện không mang tính đại diện.
3. Các đặc trưng không liên quan.

Thuật toán học máy kém

Ví dụ về hồi quy đa thức

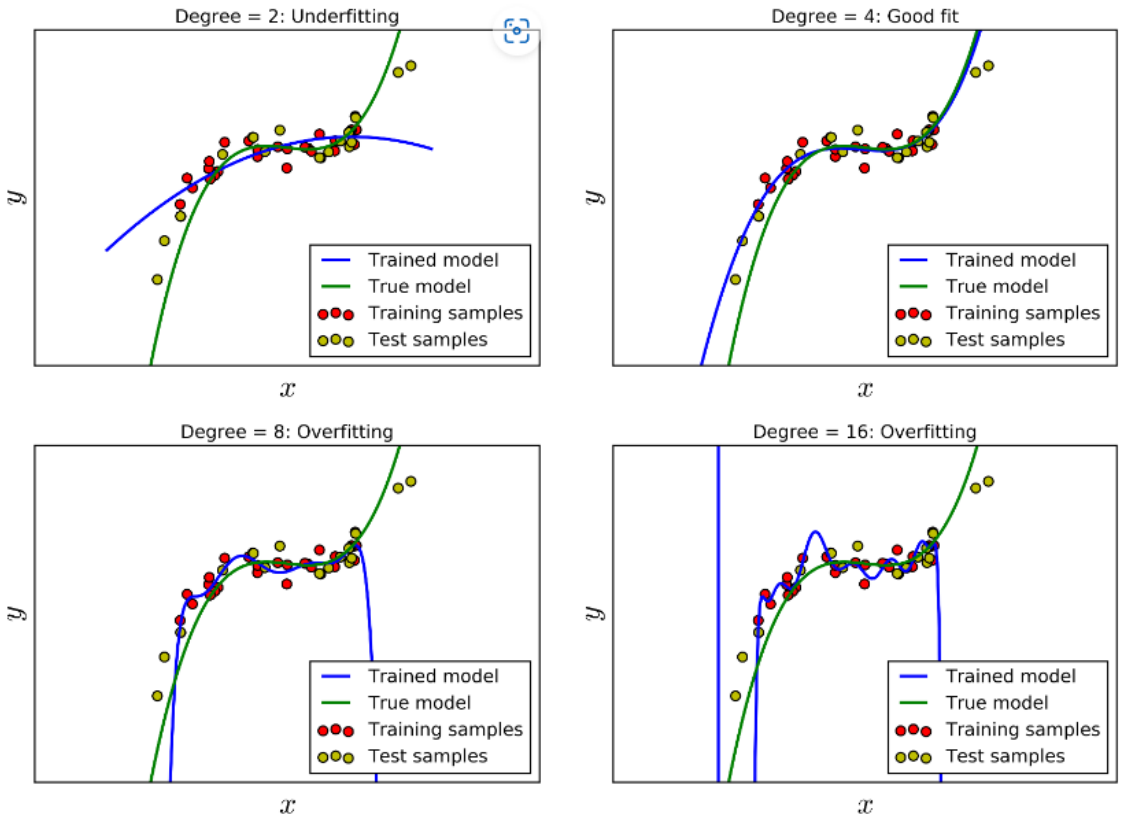


Figure 12: Hiện tượng overfitting

Overfitting

Khái niệm

- Overfitting là hiện tượng mô hình tìm được *quá khớp* với dữ liệu training. Việc *quá khớp* này có thể dẫn đến việc dự đoán nhầm nhiễu, và chất lượng mô hình không còn tốt trên dữ liệu test nữa. Dữ liệu test được giả sử là không được biết trước, và không được sử dụng để xây dựng các mô hình Machine Learning.
- Về cơ bản, overfitting xảy ra khi mô hình quá phức tạp để mô phỏng training data. Điều này đặc biệt xảy ra khi lượng dữ liệu training quá nhỏ trong khi độ phức tạp của mô hình quá cao
- Nếu train error thấp và test error cao thì sẽ dẫn đến hiện tượng overfitting (quá khớp).

Underfitting

Khái niệm

- Underfitting (chưa khớp) là hiện tượng khi mô hình xây dựng chưa có độ chính xác cao trong tập dữ liệu huấn luyện cũng như tổng quát hóa với tổng thể dữ liệu. Khi hiện tượng Underfitting xảy ra, mô hình đó sẽ không phải là tốt với bất kì bộ dữ liệu nào trong vấn đề đang nhắc tới.
- Hiện tượng Underfitting thường ít xảy ra trong bài toán hơn. Khi Underfitting xảy ra, ta có thể khắc phục bằng cách thay đổi thuật toán hoặc là bổ sung thêm dữ liệu đầu vào.

Good fitting

Khái niệm

- Good Fitting (vừa khớp) là nằm giữa Underfitting và Overfitting. Mô hình cho ra kết quả hợp lý với cả tập dữ liệu huấn luyện và các tập dữ liệu mới. Đây là mô hình lý tưởng mang được tính tổng quát và khớp được với nhiều dữ liệu mẫu và cả các dữ liệu mới.
- Good Fitting là mục tiêu của mỗi bài toán. Tuy nhiên, trên thực tế, vấn đề này rất khó thực hiện. Để tìm được điểm Good Fitting, ta phải theo dõi hiệu suất của thuật toán học máy theo thời gian khi thuật toán thực hiện việc học trên bộ dữ liệu huấn luyện. Ta có thể mô tả và thể hiện các thông số mô hình, độ chính xác của mô hình trên cả hai tập dữ liệu huấn luyện và đào tạo.

Good fitting

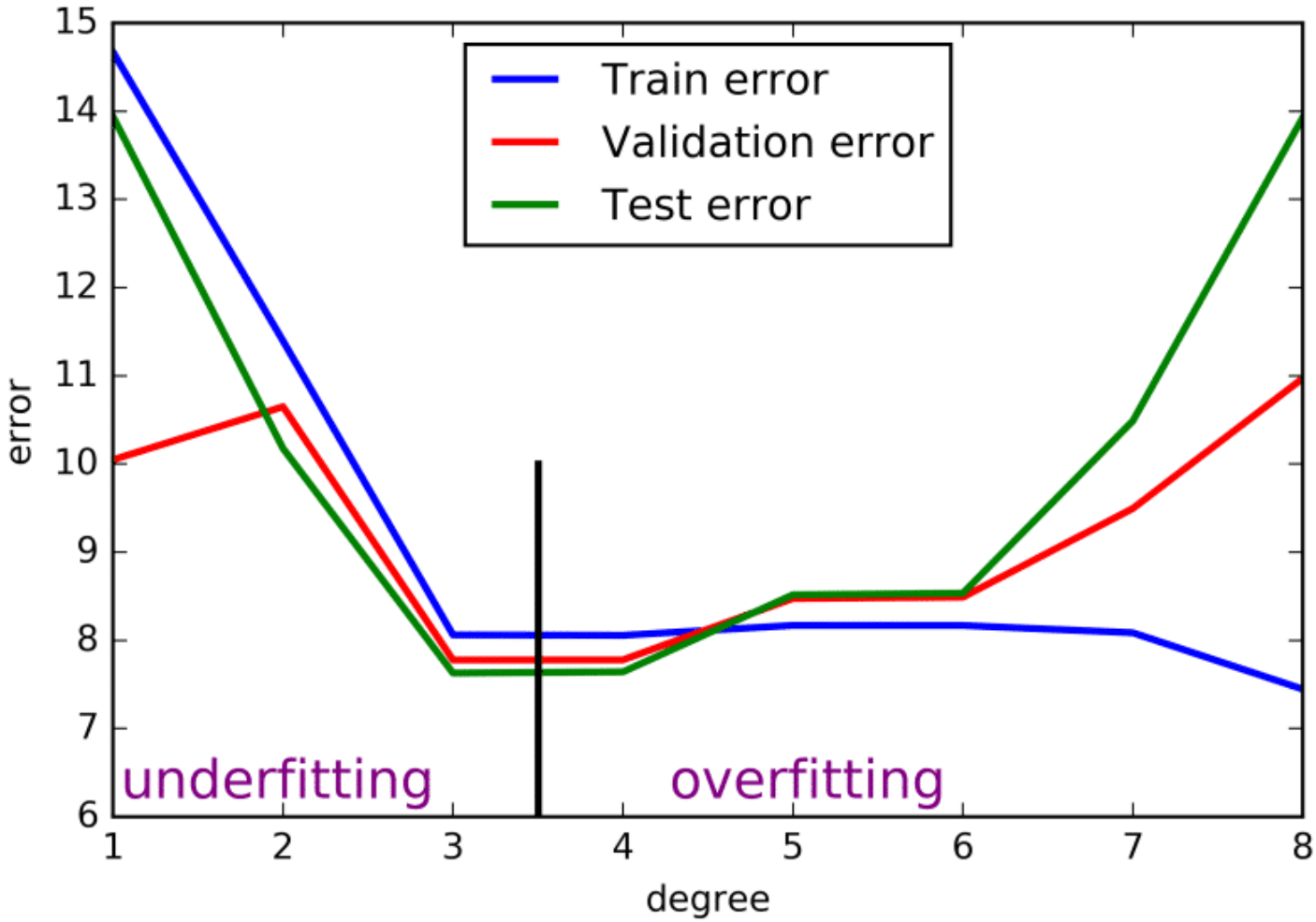


Figure 13: Good fitting

Tránh overfitting bằng k-cross validation

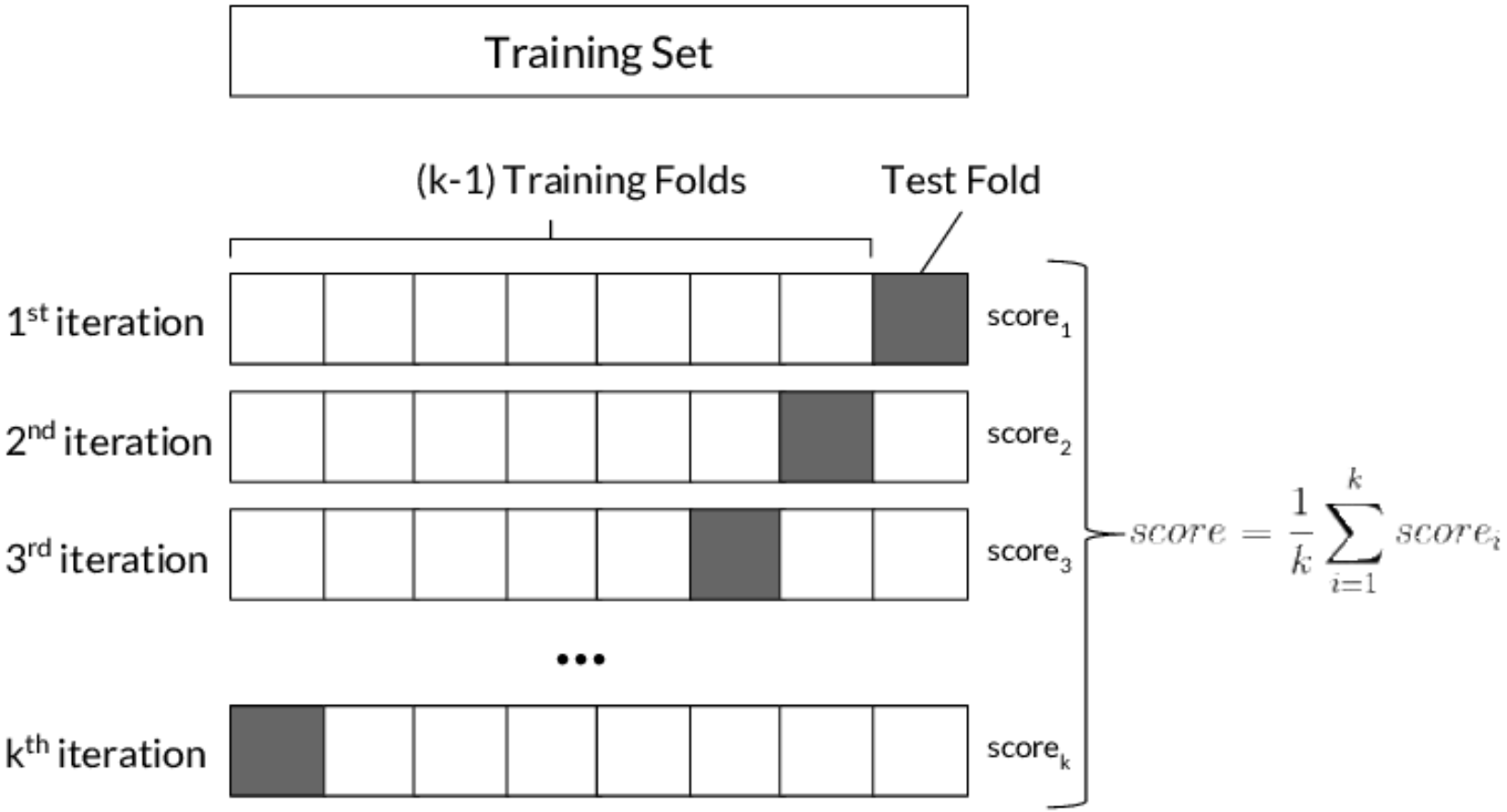


Figure 14: Phương pháp k-cross validation

K-cross validation

Khái niệm

Cross validation là một cải tiến của *validation* với lượng dữ liệu trong tập validation là nhỏ nhưng chất lượng mô hình được đánh giá trên nhiều tập *validation* khác nhau. Một cách thường đường sử dụng là chia tập training ra k tập con không có phần tử chung, có kích thước gần bằng nhau. Tại mỗi lần kiểm thử , được gọi là *run*, một trong số k tập con được lấy ra làm *validate set*. Mô hình sẽ được xây dựng dựa vào hợp của $k-1$. tập con còn lại. Mô hình cuối được xác định dựa trên trung bình của các *train error* và *validation error*. Cách làm này còn có tên gọi là **k-fold cross validation**.

Câu hỏi ôn tập

1. Nêu khái niệm học máy?
2. Hãy nêu một số ví dụ ứng dụng tiêu biểu trong thực tế sử dụng học máy?
3. Theo em, phương pháp tiếp cận theo lập trình truyền thống và tiếp cận theo phương pháp học máy có gì khác nhau, hãy trình bày chi tiết?
4. Dựa theo phương thức học, học máy chia làm mấy dạng, hãy kể tên và nêu chi tiết các bài toán trong các dạng đó?
5. Học tăng cường là gì? Hãy trình bày về học tăng cường?
6. Các bài toán cơ bản trong học máy là gì? Liệt kê chi tiết?
7. Tại sao cần sử dụng hàm mất mát?
8. Overfitting và underfitting là gì?
9. Hãy trình bày tổng quan về phương pháp k-cross validation?
10. Theo em mô hình như thế nào gọi là tốt (good fitting)?