Chủ đề 7

Giải quyết vấn đề với sự trợ giúp của máy tính

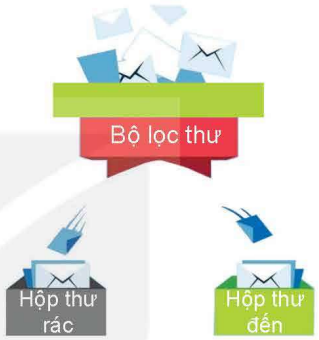
BÀI 25

LÀM QUEN VỚI HỌC MÁY

SAU BÀI HỌC NÀY EM SẼ:

- Giải thích được sơ lược về khái niệm Học máy.
- Nêu được vai trò của Học máy trong những công việc như lọc thư rác, chẩn đoán bệnh, phân tích thị trường, nhận dạng tiếng nói và chữ viết, dịch tự động,...

Khi truy cập tài khoản thư điện tử, ngoài các thư trong Hộp thư đến (Inbox) em có thể thấy nhiều thư được tự động phân loại vào Hộp thư rác (Spam). Hãy quan sát Hình 25.1 và cho biết việc phân loại này được thực hiện như thế nào?



Hình 25.1. Mô hình phân loại thư

1. TÌM HIỂU SƠ LƯỢC VỀ HỌC MÁY

Hoạt động 1 **Tìm hiểu bộ lọc thư điện tử**

Bộ lọc thư điện tử, thường là tập hợp các quy tắc, được thiết kế để phát hiện và đánh dấu các thư rác trước khi chúng được chuyển vào hộp thư của người dùng. Có quy tắc chỉ đơn giản là trong nội dung hoặc tiêu đề thư có các cụm từ đáng ngờ như "miễn phí", "giảm giá", "rẻ bất ngờ",... hay địa chỉ thư của người gửi hoặc địa chỉ của máy chủ gửi thư thuộc vào một "danh sách đen" xác định. Theo em, có thể xây dựng các bộ lọc thư này bằng cách nào?

Trong thực tế sử dụng thư điện tử, các loại thư rác mới xuất hiện ngày càng nhiều và đa dạng. Do vậy, việc xây dựng bộ lọc thư điện tử bằng cách thủ công nói chung tốn nhiều công sức và không hiệu quả. Thay vào đó, có thể sử dụng Học máy giúp máy tính tự xây dựng bộ lọc để phân loại thư điện tử.

Học máy là một lĩnh vực của AI tập trung vào việc phát triển các thuật toán và mô hình cho phép máy tính tự học và cải thiện từ dữ liệu để đưa ra dự đoán hoặc quyết định dựa trên dữ liệu mà không cần lập trình rõ ràng.

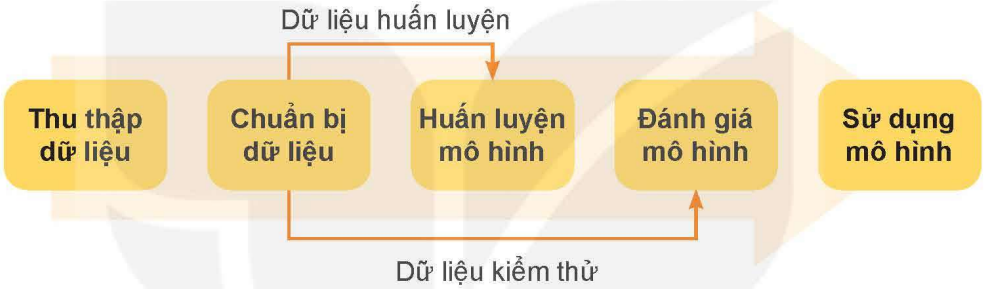
Vài ví dụ cụ thể sau đây có thể giúp em hiểu một cách sơ lược hai điểm mấu chốt trong khái niệm Học máy: "máy tính tự học từ dữ liệu" và "không cần lập trình rõ ràng".

Trong trường hợp lọc thư điện tử, việc "không cần lập trình rõ ràng" có nghĩa là không cần viết chương trình để hướng dẫn máy tính các quy tắc cụ thể, ví dụ, "một thư điện tử chứa từ X hoặc Y là thư rác". Thay vào đó, chỉ cần cung cấp cho máy tính tập dữ liệu các ví dụ về thư rác và thư hợp lệ, máy tính sử dụng dữ liệu này

để học những đặc điểm, mẫu hoặc quy luật mà nó sẽ sử dụng để đoán nhận và phân loại thư điện tử mới được gửi tới.

Tương tự như vậy, nếu muốn máy tính nhận dạng con ngựa trong hình ảnh, việc lập trình rõ ràng có thể là viết một chương trình máy tính với các mô tả như "Con vật có 4 chân cao, mặt dài, đôi tai nhọn là con ngựa". Tuy nhiên, việc mô tả tất cả đặc điểm cụ thể của con ngựa trong mọi trường hợp có thể gặp là không khả thi và không hiệu quả. Thay vào đó, chỉ cần cung cấp cho máy tính hàng nghìn hình ảnh chứa con ngựa và các con vật khác để máy tính tự học từ dữ liệu này. Máy tính tự xác định các đặc trưng từ dữ liệu, ví dụ, "Con ngựa thường có 4 chân cao, mặt dài, đôi tai nhọn" và sử dụng chúng để nhận dạng ngựa trong hình ảnh nhận được sau này.

Trong cả hai ví dụ nêu trên, máy tính không biết trước như thế nào là thư rác hoặc như thế nào là con ngựa và cách nhận dạng chúng - nó tự học từ dữ liệu mà chúng ta cung cấp. Đây cũng là điểm mấu chốt nhất trong tất cả các ứng dụng Học máy. Điều này cho phép máy tính giải quyết nhiều bài toán nhờ việc "tự học" từ dữ liệu, không đòi hỏi phải hướng dẫn trực tiếp bằng cách lập trình rõ ràng.



Hình 25.2. Quy trình Học máy

Việc xây dựng các ứng dụng Học máy có thể chia thành 5 bước cơ bản như Hình 25.2. Tuỳ theo bài toán cần giải quyết, việc *thu thập dữ liệu* để xây dựng mô hình Học máy có thể được lấy từ nhiều nguồn khác nhau, như các cơ sở dữ liệu, tệp tin hoặc thậm chí thông qua việc ghi chép trực tiếp. Thông thường, dữ liệu đó không phù hợp để có thể sử dụng được ngay. Do vậy, cần thực hiện các thao tác *chuẩn bị dữ liệu* (còn được gọi là "làm sạch dữ liệu") bao gồm việc loại bỏ dữ liệu nhiễu, bổ sung các giá trị thiếu, chuyển đổi dữ liệu sang định dạng phù hợp và giảm kích thước dữ liệu nếu cần. Cần lưu ý, đây là hai bước quan trọng, chiếm nhiều thời gian và công sức nhất của quá trình xây dựng ứng dụng Học máy. Hai bước này có thể phải thực hiện lặp đi lặp lại cho tới khi thu được bộ dữ liệu như mong muốn. Tập dữ liệu thu được thường được chia thành hai phần: *dữ liệu huấn luyện* (thường chiếm khoảng 70% đến 80%) và *dữ liệu kiểm thử*. Dữ liệu huấn luyện được dùng để *huấn luyện mô hình*, dữ liệu kiểm thử được dùng để *đánh giá mô hình*.

Tiếp theo, cần chọn thuật toán học máy phù hợp với loại bài toán và dữ liệu thu thập được. Các loại thuật toán này khá đa dạng như hồi quy tuyến tính, cây quyết định, mạng nơron,... Về mặt bản chất, thuật toán Học máy sử dụng các mô hình toán học để kết nối các đặc trưng và thông tin liên quan tới tập dữ liệu. Thực hiện thuật toán học máy trên tập dữ liệu huấn luyện, thường được gọi là *huấn luyện mô hình*, giúp máy tính học cách phân biệt giữa các mẫu thuộc các lớp dữ liệu khác nhau. Kết quả của quá trình này sẽ là một mô hình Học máy để giải quyết một bài toán cụ thể. Áp dụng mô hình đó trên tập dữ liệu kiểm thử để đánh giá hiệu suất của mô hình trong

việc dự đoán dữ liệu mới. Dựa trên kết quả đánh giá, mô hình có thể cần được cải thiện, bằng cách bổ sung thêm dữ liệu huấn luyện mới, điều chỉnh các tham số của thuật toán Học máy hoặc sử dụng các thuật toán Học máy khác. Các công việc này được gọi chung là bước *đánh giá mô hình*. Hai bước huấn luyện và đánh giá có thể được thực hiện lặp đi lặp lại cho tới khi thu được mô hình Học máy như mong muốn. Cuối cùng, *sử dụng mô hình* thu được để giải quyết vấn đề đặt ra, thực hiện dự đoán hay phân cụm trên dữ liệu mới.

Học máy là một lĩnh vực của AI nghiên cứu và phát triển các thuật toán và mô hình đem lại khả năng học cho máy tính. Nó cho phép máy tính tự động tìm hiểu từ dữ liệu và tạo ra các mô hình dự đoán hoặc quyết định dựa trên dữ liệu mà không cần phải được lập trình cụ thể.

1. Chọn phương án trả lời đúng. Học máy là:

A. Chương trình máy tính có khả năng đưa ra quyết định hay dự đoán dựa trên dữ liệu.

B. Khả năng máy tính phân tích dữ liệu thu nhận được để đưa ra dự đoán hoặc quyết định dựa trên các quy tắc được xác định rõ ràng.

C. Việc sử dụng các phương pháp và kĩ thuật cho phép máy tính học từ dữ liệu để đưa ra dự đoán hoặc quyết định mà không cần lập trình cụ thể.

D. Chương trình máy tính có khả năng tự cải thiện hiệu suất thực hiện nhiệm vụ thông qua việc cập nhật các dữ liệu mới sau khi hoàn thành nhiệm vụ đó nhiều lần.

2. Tại sao cần chia dữ liệu học máy thành hai phần: dữ liệu huấn luyện và dữ liệu kiểm tra?

2. PHÂN LOẠI VÀ VAI TRÒ CỦA HỌC MÁY TRONG THỰC TẾ

Hoạt động 2 **Tìm hiểu vai trò của Học máy trong một số công việc cụ thể**

Trong Mục 1 em đã được giới thiệu một số ứng dụng thực tế của Học máy. Hãy kể tên một vài công việc mà ngày nay không thể thiếu vai trò của Học máy.

a) Phân loại Học máy

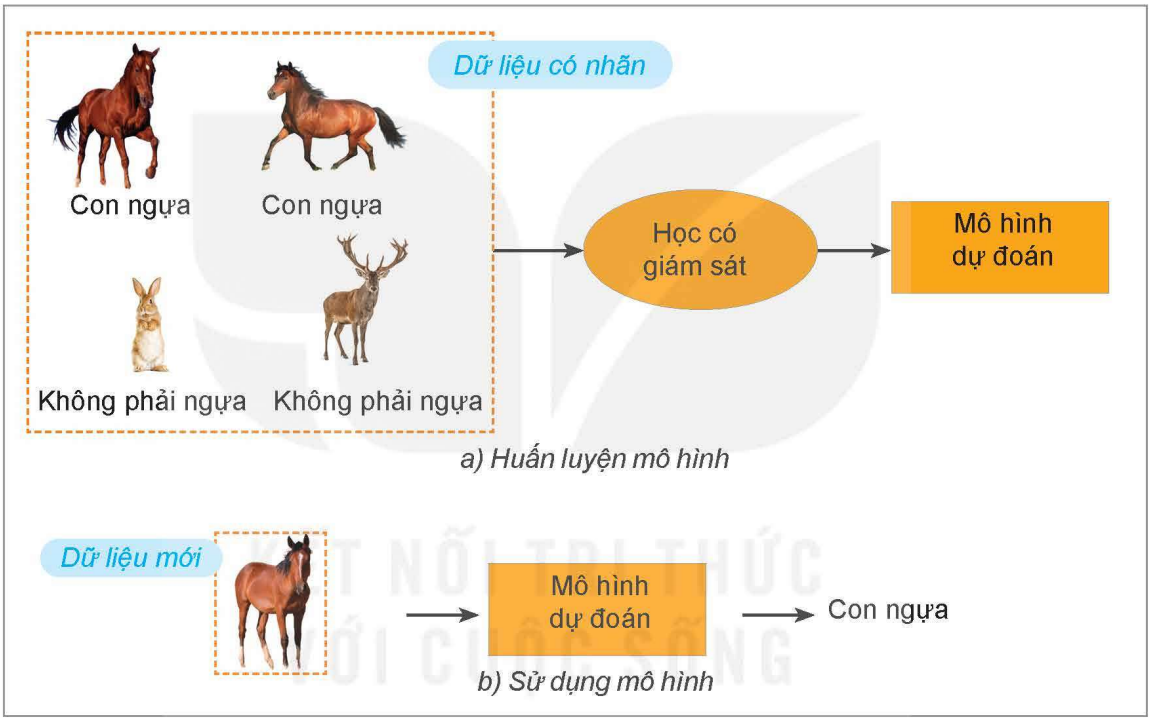
Trong Học máy, tập dữ liệu đầu vào gồm hai loại chính: *dữ liệu có nhãn* và *dữ liệu không có nhãn.* Dữ liệu được gắn kết với một nhãn hoặc một giá trị đích cụ thể được gọi là dữ liệu có nhãn, trường hợp ngược lại, là dữ liệu không có nhãn. Nhãn hoặc giá trị đích này thường chỉ ra thông tin quan trọng về đối tượng, thuộc tính hoặc phân loại mà mẫu dữ liệu đó đại diện. Việc gán nhãn dữ liệu thường được thực hiện bằng cách thủ công. Dữ liệu có nhãn đóng vai trò rất quan trọng trong quá trình huấn luyện mô hình học máy, vì nó cung cấp thông tin cần thiết cho mô hình để học và đưa ra dự đoán chính xác trên các dữ liệu mới. Việc có dữ liệu được gán nhãn đúng và đa dạng là một yếu tố quyết định để xây dựng mô hình học máy hiệu quả và đáng tin cậy.

Tương ứng với hai loại dữ liệu đầu vào nêu trên là hai phương pháp học máy cơ bản: *học có giám sát* và *học không giám sát.* Đây cũng là hai phương pháp học máy được sử dụng nhiều nhất trong thực tế để giải quyết các bài toán phân loại và phân cụm dữ liệu.

Học có giám sát

Học có giám sát là phương pháp học máy trong đó tập dữ liệu đầu vào là dữ liệu đã được gán nhãn. Trên cơ sở được "học" từ dữ liệu loại này, máy tính có khả năng mô hình hoá mối quan hệ giữa dữ liệu đầu vào với đầu ra tương ứng (pha huấn luyện mô hình). Khi đưa một dữ liệu mới chưa biết vào, máy tính sẽ thực hiện việc xác định các đặc trưng dữ liệu, từ đó đưa ra phản hồi (dự đoán) dữ liệu đó cùng loại với dữ liệu nào được gán nhãn (pha sử dụng mô hình). Hình 25.3 mô tả một hệ thống học có giám sát, với dữ liệu là các hình ảnh được gán nhãn (Con ngựa hay Không phải ngựa), để xác định xem dữ liệu mới được đưa vào là một con ngựa hay là một loại động vật khác.

Học có giám sát là phương pháp học máy được sử dụng rộng rãi nhất. Nó có nhiều ứng dụng trong thực tế như xây dựng bộ lọc thư rác, nhận dạng hình ảnh, nhận dạng chữ viết tay, nhận dạng tiếng nói,...

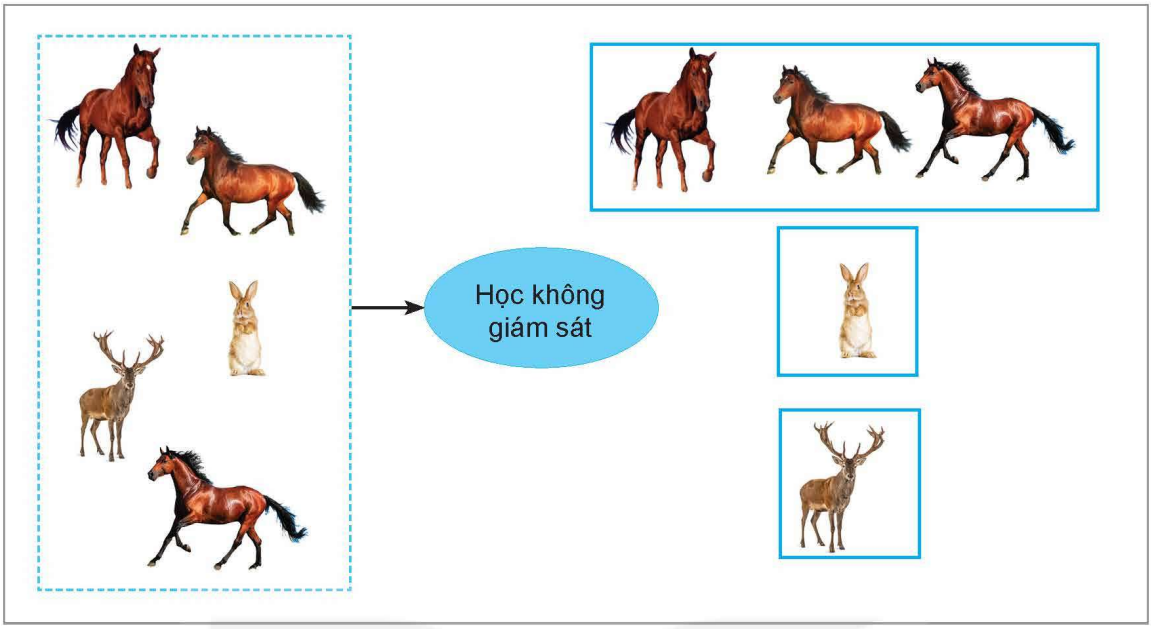


Hình 25.3. Hai pha của mô hình học có giám sát (phân loại dữ liệu)

Học không giám sát

Học không giám sát là phương pháp học máy sử dụng dữ liệu không có nhãn. Sử dụng thông tin về mối quan hệ tương tự hay khác biệt, cũng như dựa trên xác suất đồng xuất hiện của các đối tượng hoặc các biến có trong dữ liệu, các thuật toán và mô hình học trong phương pháp này sẽ thực hiện việc mô hình hoá cấu trúc hoặc mô tả các thông tin ẩn chứa trong dữ liệu.

Học không giám sát thường được ứng dụng để phân chia dữ liệu thành các nhóm dựa trên sự tương đồng của các mẫu dữ liệu. Ví dụ, trong Hình 25.4, mô hình học không giám sát thực hiện việc phân nhóm các con vật dựa trên hình ảnh của chúng. Có thể chỉ ra một số bài toán khác có thể áp dụng học không giám sát, chẳng hạn như xác định các phân khúc khách hàng dựa trên lịch sử mua hàng của họ; phát hiện bất thường trong các giao dịch thẻ tín dụng để xác định gian lận; xác định các chủ đề khác nhau hoặc xác định chủ đề chính được thảo luận trong một tập hợp các bài báo,...



Hình 25.4. Mô hình học không giám sát (phân cụm dữ liệu)

b) Vai trò của Học máy

Học máy có vai trò quan trọng trong nhiều công việc và ứng dụng thực tế. Nó hỗ trợ khai phá các loại dữ liệu đa dạng, có quy mô lớn, bao gồm cả các dữ liệu không ngừng thay đổi theo thời gian, để trích xuất được những thông tin và tri thức hữu ích. Dưới đây là một vài ví dụ cụ thể:

- *Lọc thư rác:* Trong trường hợp này, Học máy giúp xây dựng mô hình có khả năng phân loại thư điện tử là thư rác hoặc thư thường dựa trên các đặc điểm của thư gửi tới, như từ khoá, cấu trúc thư và nhiều yếu tố khác. Học máy giúp giảm thời gian và công sức của người dùng trong việc đánh dấu thư rác, đồng thời cải thiện hiệu suất lọc thư theo thời gian bằng cách học hỏi từ dữ liệu và cập nhật mô hình.
- *Chẩn đoán bệnh:* Học máy sử dụng dữ liệu về tình trạng sức khoẻ của bệnh nhân cùng kết quả xét nghiệm và các cơ sở dữ liệu bệnh lí khác để xây dựng mô hình chẩn đoán bệnh. Mô hình này còn có thể dự báo tình trạng sức khoẻ và đề xuất phương án điều trị phù hợp cho bệnh nhân. Mô hình Học máy có thể học từ hàng ngàn lần chẩn đoán cho nhiều bệnh nhân khác nhau trước đó, giúp bác sĩ đưa ra quyết định dựa trên dữ liệu một cách chính xác và nhanh chóng hơn.
- *Phân tích thị trường:* Học máy có thể phân tích dữ liệu thị trường từ nhiều nguồn khác nhau để xác định xu hướng, dự báo biến động giá cả, trợ giúp hình thành các chiến lược kinh doanh dựa trên các mô hình dự đoán. Nó giúp người đầu tư và nhà kinh doanh hiểu rõ hơn về thị trường, tăng khả năng đưa ra quyết định đầu tư dựa trên thông tin và các phân tích kĩ thuật.
- *Nhận dạng tiếng nói:* Học máy giúp xây dựng các mô hình âm thanh để biểu diễn những đặc trưng của tiếng nói, giúp máy tính có thể học và nhận dạng các biểu hiện âm thanh của từng đơn vị tiếng (phoneme), từ đó tạo ra biểu diễn số hoá của âm thanh. Những đặc điểm âm thanh cá nhân trong các mô hình âm thanh còn giúp cải thiện khả năng nhận dạng và phân biệt tiếng nói của những người nói khác nhau.

- *Nhận dạng chữ viết:* Học máy giúp xây dựng mô hình hình học cho phép xác định hình dạng, kích thước, góc xoay của các kí tự trong hình ảnh chữ viết tay. Những năm gần đây, sự phát triển của học sâu (một lĩnh vực của Học máy) cho phép học và trích xuất các đặc trưng phức tạp từ hình ảnh chữ viết tay, giúp cải thiện đáng kể khả năng nhận dạng chữ viết tay.
- *Dịch tự động:* Học máy sử dụng dữ liệu về bản dịch và bản gốc trong các ngôn ngữ khác nhau để xây dựng mô hình dịch tự động. Mô hình này có khả năng dịch văn bản, tiếng nói từ ngôn ngữ này sang ngôn ngữ khác. Khả năng dịch tự động của máy tính giúp hạn chế rào cản ngôn ngữ trong giao tiếp, phát triển hợp tác và trao đổi thông tin mọi lĩnh vực, đặc biệt trong giáo dục, đào tạo và nghiên cứu khoa học.

Trong các công việc trên, cũng như trong nhiều lĩnh vực khác, vai trò quan trọng của Học máy được thể hiện ở nhiều góc độ khác nhau: giúp xử lí một lượng lớn dữ liệu trong thời gian thực một cách nhanh chóng và hiệu quả để xác định các mẫu và xu hướng quan trọng có trong dữ liệu, tự động hoá các nhiệm vụ phức tạp mà trước đây cần sự can thiệp của con người,... Do có khả năng học từ dữ liệu, Học máy có thể giúp các chuyên gia và các nhà nghiên cứu từng bước xây dựng và bổ sung tri thức. Hơn thế nữa, nhờ khả năng không ngừng bổ sung dữ liệu và tự động cập nhật mô hình đã được huấn luyện, Học máy ngày càng có vai trò không thể thiếu trong các ứng dụng mà dữ liệu có quy mô và chủng loại đa dạng, không ngừng thay đổi theo thời gian, như sự xuất hiện các mẫu thư rác mới, các triệu chứng bệnh mới, hay các bản dịch ngôn ngữ mới,...

Hai phương pháp học máy cơ bản là học có giám sát và học không giám sát, tuỳ theo tập dữ liệu cung cấp cho mô hình học máy là dữ liệu có nhãn hay không có nhãn. Học máy giúp xử lí lượng lớn dữ liệu một cách nhanh chóng và hiệu quả, bao gồm cả các dữ liệu không ngừng thay đổi theo thời gian, trợ giúp các quá trình ra quyết định cũng như tự động hoá các nhiệm vụ phức tạp.

Vai trò quan trọng của Học máy trong các lĩnh vực khác nhau được thể hiện như thế nào?

LUYỆN TẬP

Tại sao có thể nói Học máy có vai trò không thể thiếu trong các ứng dụng mà dữ liệu không ngừng thay đổi theo thời gian? Hãy chỉ ra một vài minh hoạ cụ thể.

VẬN DỤNG

Ngoài hai phương pháp học máy cơ bản nêu trong bài, một số tài liệu còn đề cập tới phương pháp *học bán giám sát* và *học tăng cường*. Hãy tìm hiểu về các phương pháp học máy này trên Internet.