CHỦ ĐỀ F^CS

# GIẢI QUYẾT VẤN ĐỀ VỚI SỰ TRỢ GIÚP CỦA MÁY TÍNH

## F^CS1. GIỚI THIỆU HỌC MÁY VÀ KHOA HỌC DỮ LIỆU

BÀI 1

## GIỚI THIỆU VỀ HỌC MÁY

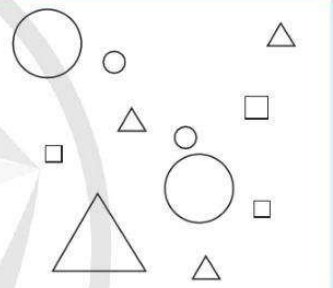
***Học xong bài này, em sẽ:***

- 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,...

Cho *Hình 1* và ba nhãn phân loại là "vuông", "tròn", "tam giác". Cần gán nhãn phân loại cho từng đối tượng trong *Hình 1*. Em hãy trả lời các câu hỏi sau:

1) Nếu con người thực hiện thì nhiệm vụ trên là dễ hay khó?

2) Theo em, máy tính có thể tự động thực hiện nhiệm vụ trên thay cho con người hay không? Lập trình để máy tính làm công việc này là dễ hay khó?



*Hình 1. Ví dụ một số đối tượng cần phân loại*

### 1 Khái niệm học máy

Bài *Giới thiệu về trí tuệ nhân tạo* (AI) ở Chủ đề A đã đề cập đến học máy như một nhánh nghiên cứu trong ngành AI nhằm làm cho máy tính có khả năng học từ dữ liệu. Bài học này sẽ giúp các em hiểu sâu hơn về học máy.

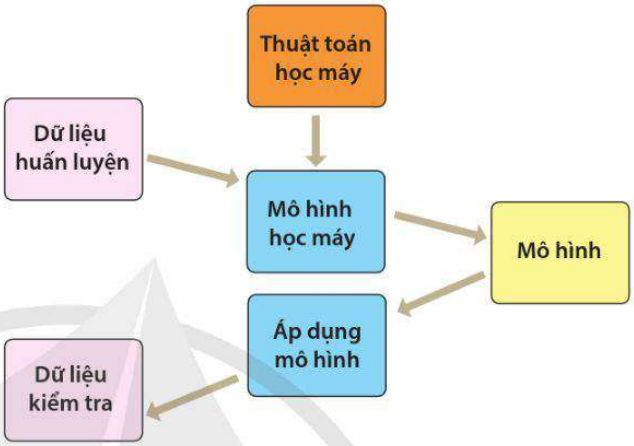
Học máy huấn luyện máy tính để nó có thể tự động phát hiện ra các mối quan hệ có trong dữ liệu. Học máy có thể giải quyết nhiều loại bài toán khác nhau, tiêu biểu là bài toán phân loại và bài toán phân cụm.

Để huấn luyện máy tính học từ dữ liệu, người ta dùng một tập dữ liệu gọi là dữ liệu huấn luyện. Máy tính thực hiện thuật toán trên tập dữ liệu huấn luyện để có khả năng tự thực hiện những nhiệm vụ tương tự trên tập dữ liệu mới. Dữ liệu huấn luyện mô tả các đối tượng trong thế giới thực. Mỗi đối tượng tương ứng với một mẫu dữ liệu.

### a) Mô hình học máy

*Thuật toán học máy*: Từ tập dữ liệu đầu vào, thuật toán học máy rút ra các thông tin liên quan tới dữ liệu, các đặc điểm chung quan trọng,... Từ đó giúp máy tính học cách phân biệt giữa các mẫu dữ liệu khác nhau hoặc nhóm các mẫu dữ liệu thành các cụm nhiều mẫu tương tự nhau.

*Mô hình học máy*: Các mô hình học máy được tạo ra từ các thuật toán học máy và trải qua quá trình huấn luyện bằng cách sử dụng dữ liệu huấn luyện (*Hình 2*). Thực hiện thuật toán học máy trên tập dữ liệu huấn luyện tức là *huấn luyện mô hình học máy*.




*Hình 2. Minh hoạ mô hình học máy*

Quá trình huấn luyện nhằm tạo ra mô hình học máy để giải quyết một bài toán cụ thể. Áp dụng mô hình cho phần dữ liệu chưa dùng trong huấn luyện để đánh giá mô hình. Mô hình được đưa vào sử dụng thực tế nếu kết quả đánh giá đáp ứng yêu cầu ứng dụng.

### b) Quy trình học máy

Tuỳ theo lĩnh vực ứng dụng và bài toán cụ thể, người ta chọn tập dữ liệu, dùng thuật toán học máy và cách đánh giá kết quả huấn luyện thích hợp.

Quy trình học máy có thể mô tả như sau:

1. *Thu thập dữ liệu*: Chọn dữ liệu phù hợp với bài toán cụ thể. Dữ liệu có thể được chọn từ nhiều nguồn, có khuôn dạng khác nhau, có thể là dữ liệu có cấu trúc ví dụ như các bản ghi trong cơ sở dữ liệu hoặc phi cấu trúc. Tập dữ liệu thu thập được là dữ liệu thô, chưa sẵn sàng để sử dụng trong quá trình huấn luyện.

2. *Chuẩn bị dữ liệu*: Làm sạch, loại bỏ nhiễu, bổ sung các giá trị thiếu và chuyển đổi dữ liệu sang một khuôn dạng phù hợp. Chia dữ liệu đã chuẩn bị thành hai phần. Một phần lớn dữ liệu được dùng làm dữ liệu huấn luyện và phần còn lại dùng làm dữ liệu để đánh giá mô hình.

3. *Xây dựng mô hình*:

– Chọn thuật toán học máy phù hợp với bài toán và dữ liệu đã chuẩn bị. Có nhiều loại thuật toán học máy như: hồi quy tuyến tính, cây quyết định, mạng nơ ron,...

– Huấn luyện mô hình để mô hình học từ dữ liệu và trở nên thích ứng với bài toán cụ thể đó.

4. *Đánh giá mô hình*: Áp dụng mô hình sau huấn luyện cho phần dữ liệu dành để đánh giá mô hình. Đối chiếu kết quả với tiêu chí đánh giá để xác định mức độ đáp ứng yêu cầu ứng dụng. Việc huấn luyện và đánh giá thường được thực hiện nhiều lần cho tới khi mô hình đạt yêu cầu mong muốn. Nếu kết quả đánh giá chưa đạt, cần tiếp tục cải thiện mô hình. Để cải thiện mô hình, có thể: phân chia lại dữ liệu huấn luyện và dữ liệu dành để đánh giá, bổ sung thêm dữ liệu mới, điều chỉnh các tham số của thuật toán học máy hoặc sử dụng thuật toán học máy khác.

5. *Triển khai ứng dụng mô hình*: Sử dụng mô hình đã được huấn luyện thành công vào ứng dụng học máy trong bài toán thực tế.

Có thể chia học máy thành hai loại chính: học có giám sát và học không giám sát.

## 2 Học có giám sát

Trong học có giám sát, tập dữ liệu huấn luyện gồm các mẫu dữ liệu được liên kết với đầu ra tương ứng, gọi là nhãn. Máy tính học để phát hiện ra mối quan hệ giữa các mẫu dữ liệu với nhãn. Sau khi học xong, máy tính có thể đưa ra dự đoán nhãn cho dữ liệu mới.

Học có giám sát có thể dùng để giải quyết nhiều loại bài toán khác nhau, trong đó có bài toán phân loại.

### a) Bài toán phân loại

**1**

Xét các bài toán sau:

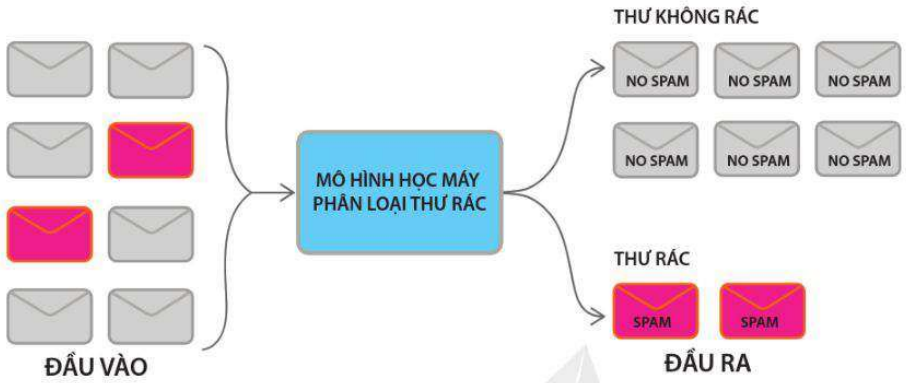
1) Hệ thống thư điện tử cần xác định các email nghi là thư rác và đánh dấu nó thuộc loại "spam" (thư rác) để cảnh báo cho người nhận. Những email còn lại thuộc loại "no spam".

2) Ngân hàng cần xác định khách hàng thuộc loại "tốt", "bình thường" hay "xấu" để quyết định hạn mức cho vay và lãi suất áp dụng.

Theo em, những bài toán trên và nhiệm vụ ở phần khởi động có những điểm chung là gì?

Có một số nhãn phân loại cho trước. Việc gán cho mỗi đối tượng một nhãn phân loại tuỳ theo các thuộc tính đặc trưng của nó là bài toán phân loại. Hai bài toán nêu trong Hoạt động 1 là bài toán phân loại. Bài toán 1 có hai nhãn phân loại là "spam" và "no spam", bài toán 2 có ba nhãn phân loại là " tốt", " bình thường" và " xấu". Học máy giúp xây dựng mô hình phân loại để phân loại thư rác, phân loại khách hàng vay tín dụng. *Hình 3* minh hoạ vai trò của học máy trong phân loại thư rác. Biểu tượng email màu đỏ thể hiện thư có các thuộc tính đặc trưng của thư rác; biểu tượng email

màu xám thể hiện thư không có những thuộc tính đặc trưng của thư rác. Mô hình phân loại thư rác đã được huấn luyện thành công bằng học có giám sát gán nhãn "spam" hay "no spam" cho thư mới nhận được.




*Hình 3. Hệ thống thư điện tử phân loại thư rác*

### b) Dữ liệu huấn luyện

Mỗi đối tượng cần phân loại được mô tả bởi một số thông tin là các thuộc tính đặc trưng của nó. Ví dụ, việc phân loại email là "spam" hay "no spam" dựa vào một số thông tin như: địa chỉ người gửi, địa chỉ người nhận, dòng tiêu đề, sự có mặt của những từ đặc trưng cho thư rác,...

Dữ liệu huấn luyện là các mẫu dữ liệu về một số email đã biết trước là thư rác và một số email khác không là thư rác.

### c) Huấn luyện và đánh giá mô hình

Máy tính được huấn luyện để sử dụng dữ liệu huấn luyện và tự dự đoán nhãn phân loại theo thuật toán học máy. Nhãn phân loại được xác định khi biết giá trị các thuộc tính đặc trưng của đối tượng. Áp dụng mô hình cho phần dữ liệu dùng để đánh giá sẽ nhận được dữ liệu kiểm tra. So sánh nhãn đã biết với nhãn do mô hình dự đoán để đánh giá mô hình. Mục tiêu huấn luyện nhằm giảm thiểu nhãn bị gán sai đến mức ngưỡng chấp nhận được.

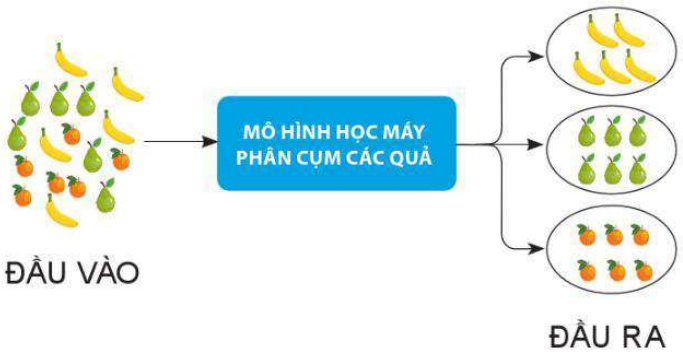
## 3 Học không giám sát

Khác với học có giám sát, học không giám sát được thực hiện với tập dữ liệu không có nhãn. Học không giám sát có thể dùng để huấn luyện máy tính giải quyết nhiều loại bài toán khác nhau, trong đó có bài toán phân cụm.

### a) Bài toán phân cụm

**2** Cho một số quả, theo em máy tính có thể trả lời được có mấy loại quả hay không?

Phân cụm là chia tập đối tượng thành các cụm dựa trên sự tương tự và khác biệt của các đối tượng. Các đối tượng trong cùng một cụm có các đặc điểm tương tự nhau. Các đối tượng trong các cụm khác nhau thì có đặc điểm khác nhau. Học không giám sát giúp xây dựng mô hình phân cụm. *Hình 4* minh hoạ vai trò của học máy trong phân cụm các quả. Mô hình phân cụm các quả đã được huấn luyện thành công bằng học không giám sát gom các quả thành ba cụm dựa vào đặc điểm các loại quả.




*Hình 4. Ứng dụng học máy phân cụm các quả*

### b) Dữ liệu huấn luyện

Tập dữ liệu huấn luyện gồm các mẫu dữ liệu mô tả các thuộc tính đặc trưng của đối tượng. Ví dụ, trường hợp phân cụm các quả, các thuộc tính đặc trưng của mỗi quả là hình dạng, kích thước, màu sắc,... Trong dữ liệu không có tên các loại quả, tức là không có nhãn kèm theo.

### c) Huấn luyện và đánh giá mô hình

Dựa trên thông tin về sự tương tự của các mẫu dữ liệu, thuật toán học máy sẽ nhóm các mẫu dữ liệu thành các cụm. Có một số cách khác nhau để phân cụm như: dựa trên mật độ các mẫu dữ liệu trong một vùng; dựa trên phân phối xác suất của các mẫu dữ liệu. Mức độ tương tự giữa hai mẫu dữ liệu có thể biểu diễn trực quan bằng "khoảng cách" giữa hai mẫu. Thuật toán học máy có thể phân cụm dựa trên khoảng cách giữa các mẫu dữ liệu và khoảng cách từ mẫu dữ liệu đến tâm cụm.

Khác với học có giám sát, kết quả phân cụm được đánh giá trực tiếp dựa vào tính chất của dữ liệu và yêu cầu phân tích dữ liệu, vì không có nhãn để kiểm tra là đúng hay sai.

## 4 Một số ứng dụng của học máy

Học máy được ứng dụng để lọc thư rác, phân loại khách hàng vay tín dụng, phân cụm các quả như đã trình bày ở trên. Sau đây là một số ứng dụng khác của học máy.

### a) Nhận dạng tiếng nói

Máy tính có thể nhận dạng tiếng nói nhờ học máy. Dữ liệu đầu vào là các đoạn tiếng nói và nhờ các thuật toán học máy sẽ xây dựng được mô hình nhận dạng tiếng nói.

Nhận dạng tiếng nói được ứng dụng trong: chuyển lời nói thành văn bản; tìm kiếm bằng lời nói, điều khiển thiết bị thông minh bằng lời nói; dịch vụ trả lời tự động, chatbot trợ lí ảo hay xác thực bằng sinh trắc học tiếng nói,... *Hình 5* minh hoạ nhận dạng tiếng nói sau đó chuyển thành văn bản trên ứng dụng Google Dịch.


*Hình 5. Nhận dạng tiếng nói trong ứng dụng Google Dịch*

### b) Nhận dạng chữ viết

Máy tính có thể nhận dạng chữ viết nhờ học máy. Dữ liệu đầu vào là các kí tự hoặc các câu đã được số hoá và nhờ các thuật toán học máy sẽ xây dựng được mô hình nhận dạng chữ viết.

*Hình 6* minh hoạ một người đang viết ghi chú bằng chữ viết tay trên điện thoại thông minh. Những ghi chú bằng chữ viết tay này sẽ được chuyển thành văn bản và lưu thành tệp văn bản.

Nhận dạng chữ viết tay có thể chia thành hai chế độ, "tĩnh" và "động". Ở chế độ tĩnh, cũng gọi là ngoại tuyến (offline), hình ảnh chữ viết tay được camera thu nhận và sau đó máy tính phân tích hình dạng chữ viết tay.

Ở chế độ động, cũng gọi là trực tuyến (online), người trực tiếp viết chữ lên tấm cảm ứng, máy tính sẽ thu nhận chữ viết cùng lúc với thao tác viết và phân tích hình dạng chữ viết kết hợp với chuyển động, áp lực,... Phân tích chữ viết tay trực tuyến có thể ứng dụng để xác thực sinh trắc học chữ kí.

*Hình 6. Nhận dạng chữ viết tay trên ứng dụng ghi chú của điện thoại*

### c) Dịch máy

Dịch máy sử dụng học máy để phân tích văn bản và dự đoán khả năng một từ hoặc cụm từ cụ thể trong ngôn ngữ nguồn sẽ là từ hoặc cụm từ tương ứng nào trong ngôn ngữ đích.

Google Dịch là một ví dụ tiêu biểu ứng dụng dịch máy. Kết hợp nhận dạng chữ viết tay, nhận dạng tiếng nói với dịch máy cung cấp nhiều tính năng và ứng dụng đa dạng như: trợ lí ảo Google Assistant có chế độ phiên dịch cho phép trò chuyện với người đối thoại nói bằng nhiều ngôn ngữ khác nhau; phiên dịch văn bản trực tiếp bằng cách hướng camera vào văn bản, người dùng có thể xem kết quả dịch ngay trên màn hình; Google Dịch có thể dịch từng từ, từng câu hay toàn bộ một trang web; người dùng Gmail cũng có thể dễ dàng dịch email sang ngôn ngữ mong muốn.

### d) Chẩn đoán bệnh

Máy tính có thể chẩn đoán bệnh nhờ học máy. Dữ liệu để chẩn đoán bệnh là các triệu chứng hoặc kết quả xét nghiệm y tế. Các nhãn phân loại là tên bệnh. Sử dụng học máy để phân tích dữ liệu có thể dự đoán tên bệnh giúp các bác sĩ chẩn đoán nhanh hơn, tốt hơn.

### e) Phân tích thị trường

Học máy không giám sát giúp xây dựng mô hình phân cụm dữ liệu khách hàng của doanh nghiệp. Dữ liệu khách hàng được phân cụm theo sự tương tự về giới tính, độ tuổi, nghề nghiệp hay về nhu cầu tiêu dùng, sở thích,... Kết quả phân cụm là các nhóm khách hàng mục tiêu thích hợp cho từng loại sản phẩm, dịch vụ. Từ đó, doanh nghiệp rút ra thông tin hữu ích để xây dựng chiến lược tiếp thị, giúp tăng doanh số, tăng thị phần, nâng cao hiệu quả hoạt động sản suất kinh doanh.

Mỗi phát biểu sau về học máy là đúng hay sai?

a) Học không giám sát sử dụng dữ liệu huấn luyện không có nhãn.

b) Học có giám sát sử dụng dữ liệu kiểm tra để đánh giá kết quả huấn luyện.

c) Học có giám sát dành cho huấn luyện máy tính phân cụm.

d) Học có giám sát và không giám sát đều giúp máy tính giải quyết cùng một bài toán như nhau.

ChatGPT là một hệ thống AI nổi tiếng có nhiều khả năng khác nhau. Hãy kể ra một vài khả năng mà theo em có sự đóng góp của học máy để phát triển hệ thống này.

**Câu 1.** Học máy là gì? Sự khác nhau giữa học có giám sát và học không giám sát là gì?

**Câu 2.** Hãy kể một vài ứng dụng cụ thể trong đó có thể sử dụng học máy để thực hiện nhiệm vụ phân loại và phân cụm.

***Tóm tắt bài học***

- Từ dữ liệu huấn luyện và thuật toán học máy xây dựng được mô hình học máy và huấn luyện mô hình. Có hai loại mô hình học máy chính: học có giám sát và học không giám sát tương ứng với dữ liệu huấn luyện là có gán nhãn và không có gán nhãn.
- Các mô hình học máy giúp phân loại hoặc phân cụm các mẫu dữ liệu và được ứng dụng trong: lọc thư rác, chẩn đoán bệnh, nhận dạng tiếng nói và chữ viết, dịch tự động, phân tích thị trường,...