Học máy (Machine Learning)

1. Tổng quát

1.1 Khái niệm

Học máy (tiếng Anh: machine learning) là một lĩnh vực của trí tuệ nhân tạo liên quan đến việc nghiên cứu và xây dựng các kĩ thuật cho phép các hệ thống "học" tự động từ dữ liệu để giải quyết những vấn đề cụ thể. Ví dụ như các máy có thể "học" cách phân loại thư điện tử xem có phải thư rác (spam) hay không và tự động xếp thư vào thư mục tương ứng. Học máy rất gần với suy diễn thống kê (statistical inference) tuy có khác nhau về thuật ngữ.

Học máy có liên quan lớn đến thống kê, vì cả hai lĩnh vực đều nghiên cứu việc phân tích dữ liệu, nhưng khác với thống kê, học máy tập trung vào sự phức tạp của các giải thuật trong việc thực thi tính toán. Nhiều bài toán suy luận được xếp vào loại bài toán NP-khó, vì thế một phần của học máy là nghiên cứu sự phát triển các giải thuật suy luận xấp xỉ mà có thể xử lý được.

Một trong những trọng tâm khác của học máy là đạt được tính phổ quát (tiếng Anh: generalization), nói cách khác là tính chất của chương trình có thể làm việc tốt với dữ liệu mà nó chưa gặp bao giờ (tiếng Anh: unseen data). Một chương trình chỉ hiệu quả với dữ liệu đã gặp nhìn chung không có nhiều tính hữu dụng.

Hình 1. Machine Learning

Mối quan hệ của học máy bao gồm:

- Học máy và big data.
- Học máy và trí tuệ nhân tạo (AI).

- Học máy và dự đoán tương lai.

Hầu hết chúng ta đều không biết rằng chúng ta đã và đang tương tác với Machine Learning mỗi ngày. Mỗi khi ta Google một cái gì đó, nghe một bài hát hoặc thậm chí chụp ảnh là ta đang sử dụng machine learning.

Lý do khiến cho machine learning là một công nghệ thú vị là bởi vì nó là một bước tiến rất xa so với các hệ thống rule-based. Một cách truyền thống, các kỹ sư sẽ lập trình các luật cho phần mềm đối với các trường hợp của dữ liệu để tìm ra đáp án cho một bài toán. Trái lại, machine learning sử dụng dữ liệu và các đáp án mẫu để tìm ra các luật đằng sau một bài toán. Cách tiếp cận này hiệu quả hơn so với cách truyền thống.

Để tìm ra luật chi phối một hiện tượng, máy tính phải trải qua một quá trình học tập (huấn luyện), thử các quy luật khác nhau và cải thiện dựa trên chính sai lầm của mình. Đó là lý do tại sao công nghệ này được gọi là machine learning.

1.2 Quy trình xây dựng mô hình machine learning

1. **Thu thập dữ liệu**: Thu thập dữ liệu để mô hình học

2. **Chuẩn bị dữ liệu**: Xử lý và đưa dữ liệu về định dạng tối ưu, trích chọn đặc trưng hoặc giảm chiều dữ liệu

3. **Huấn luyện**: Tại pha này, thuật toán machine learning thực hiện việc học thông qua các ví dụ đã được thu thập và chuẩn bị từ hai bước trên

4. **Đánh giá**: Kiểm thử mô hình để đánh giá xem chất lượng của mô hình tốt đến đâu

5. **Tinh chỉnh**: Tinh chỉnh mô hình để tối ưu hiệu quả

2. Phương pháp học máy

Phương pháp quy nạp: Máy học/phân biệt các khái niệm dựa trên dữ liệu đã thu thập được trước đó. Phương pháp này cho phép tận dụng được nguồn dữ liệu rất nhiều và sẵn có.

Phương pháp suy diễn: Máy học/phân biệt các khái niệm dựa vào các luật. Phương pháp này cho phép tận dụng được các kiến thức chuyên ngành để hỗ trợ máy tính.

Hiện nay, các thuật toán đều cố gắng tận dụng được ưu điểm của hai phương pháp này. Các nhóm giải thuật học máy:

2.1 Học có giám sát

Máy tính được xem một số mẫu gồm đầu vào (input) và đầu ra (output) tương ứng trước. Sau khi học xong các mẫu này, máy tính quan sát một đầu vào mới và cho ra kết quả.

Một mô hình tốt là mô hình có khả năng tổng quát hóa tốt dữ liệu. Trường hợp mô hình chỉ tập trung ghi nhớ các ví dụ trong tập dữ liệu huấn luyện mà không tìm ra

được quy luật tổng quát, mô hình không thể làm việc tốt trên dữ liệu tương lai. Một điểm cần lưu ý nữa đó là dữ liệu chuẩn bị cho học có giám sát cần tin cậy và khách quan. Không có dữ liệu tốt thì không có mô hình tốt.

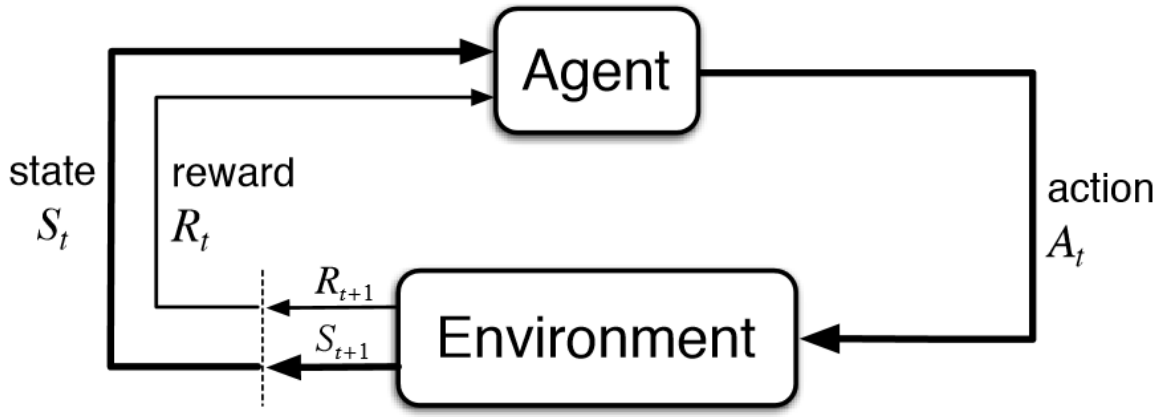
2.2 Học không giám sát

Máy tính chỉ được xem các mẫu không có đầu ra, sau đó máy tính phải tự tìm cách phân loại các mẫu này và các mẫu mới. Học nửa giám sát: Một dạng lai giữa hai nhóm giải thuật trên.

Khó khăn trong học không giám sát là việc định nghĩa bài toán. Việc không tập trung vào một mục tiêu cụ thể có thể khiến cho mô hình cho ra những kết quả mơ hồ. Tương tự như việc học chơi đàn, việc tự mày mò với cây đàn để tạo ra được những bản nhạc bất hủ sẽ khó hơn rất nhiều so với việc học với giáo viên hoặc những ví dụ cụ thể.

2.3 Học tăng cường

Thuật toán này lấy cảm hứng từ quá trình học hỏi theo cơ chế thưởng phạt ở con người. Reinforcement learning học và tích lũy kinh nghiệm để đưa ra hành động (action) sao cho với trạng thái môi trường hiện tại (state), hành động sẽ tối đa hóa phần thưởng (reward) nhận được.



Hình 2. Phương pháp học tăng cường

3. Thuật toán học máy

3.1 Support Vector Machines

Một thuật toán cố gắng xây dựng một siêu mặt phẳng trong không gian nhiều chiều để phân biệt các đối tượng ở các lớp khác nhau; Làm sao cho khoảng cách giữa 2 đối tượng khác label gần nhau nhất có khoảng cách cực đại. Ý tưởng của thuật toán cực kỳ đơn giản, nhưng mô hình này lại rất phức tạp và có hiệu quả. Thực tế, ở một số bài toán, SVM là một mô hình machine learning cho hiệu quả tốt nhất.

3.2 Mô hình xác suất (Probabilistic Models)

Các mô hình này cố gắng giải quyết bài toán bằng phân bố xác suất. Một thuật toán phổ biến nhất là phân loại Naive Bayes; Nó sử dụng lý thuyết Bayes và giả thiết các đặc trưng là độc lập. Điểm mạnh của mô hình xác suất là đơn giản nhưng hiệu quả. Đầu ra của nó không chỉ là label mà còn đi kèm xác suất thể hiện độ chính xác cho kết quả đó.

3.3 Học sâu (Deep learning)

Hiện đang là xu hướng trong machine learning dựa trên các mô hình mạng nơ ron nhân tạo(Artificial Neural Networks). Mạng nơ ron có cách tiếp cận kết nối và sử dụng ý tưởng theo cách bộ não con người làm việc. Chúng bao gồm số lượng lớn các nơ ron liên kết với nhau; được tổ chức thành các lớp(layers). Học sâu liên tục được phát triển với các cấu trúc mới sâu hơn; Nó không chỉ cố gắng học mà còn xây dựng các cấu trúc biểu diễn các đặc trưng quan trọng một cách tự động. Hai thuật toán Deep Learning có tính ứng dụng cao nhất là Recurrent Neural Network và Convolutional Neural Network.

4. Học máy trên thực tế

4.1 Khai phá dữ liệu

Khai phá dữ liệu (Data mining) là quá trình khám phá ra các thông tin có giá trị hoặc đưa ra các dự đoán từ dữ liệu. Định nghĩa này có vẻ bao quát, nhưng bạn hãy nghĩ về việc tìm kiếm thông tin hữu ích từ một bảng dữ liệu rất lớn. Mỗi bản ghi sẽ là một đối tượng cần phải học, và mỗi cột là một đặc trưng. Chúng ta có thể dự đoán giá trị của một cột của bản ghi mới dựa trên các bản ghi đã học. Hoặc là phân nhóm các bản ghi của bản. Sau đây là những ứng dụng của khai phá dữ liệu:

- *Anomaly detection*, phát hiện các ngoại lệ, ví dụ như phát hiện gian lận thẻ tín dụng. Bạn có thể phát hiện một giao dịch là khả nghi dựa trên các giao dịch thông thường của người dùng đó.
- *Association rules*, ví dụ, trong một siêu thị hay một trang thương mại điện tử. Bạn có thể khám phá ra khách hàng thường mua các món hàng nào cùng nhau. Dễ hiểu hơn, khách hàng của bạn khi mua món hàng A thường mua kèm món hàng nào? Các thông tin này rất hữu ích cho việc tiếp thị sản phẩm.
- *Grouping*, ví dụ, trong các nền tảng SaaS, người dùng được phân nhóm theo hành vi hoặc thông tin hồ sơ của họ.
- *Predictions*, dự đoán các cột giá trị (của một bản ghi mới trong database). Ví dụ, bạn có thể dự đoán giá của căn hộ dựa trên các dữ liệu về giá các căn hộ bạn đã có.

4.2 Phân tích văn bản

Phân tích văn bản (Text analysis) là công việc trích xuất hoặc phân lọi thông tin từ văn bản. Các văn bản ở đây có thể là các facebook posts, emails, các đoạn chats, tài liệu,… Một số ví dụ phổ biến là:

- *Lọc spam (Spam filtering)*, là một trong những ứng dụng phân loại văn bản được biết và sử dụng nhiều nhất. Ở đây, phân loại văn bản là xác định chủ đề cho một văn bản. Bộ lọc spam sẽ học cách phân loại một email có phải spam không dựa trên nội dung và tiêu đề của email.
- *Phân tích ngữ nghĩa (Sentiment Analysis),* học cách phân loại một ý kiến là tích cực, trung tính hay tiêu cực dựa trên nội dung văn bản của người viết.
- *Khai thác thông tin (Information Extraction),* từ một văn bản, học cách để trích xuất các thông tin hữu ích. Chẳng hạn như trích xuất địa chỉ, tên người, từ khóa,…

4.3 Trò chơi điện tử và Robot

Trò chơi điện tử (Video games) và robot (Robotics) là lĩnh vực lớn có sự góp mặt của machine learning. Nếu ta có một nhân vật cần di chuyển và tránh các chướng ngại vật trong game. Machine learning có thể học và giải quyết công việc này thay bạn. Một kỹ thuật phổ biến được áp dụng trong trường hợp này là Học tăng cường (Reinforcement learning). Ở đó, máy sẽ học tăng cường với mục tiêu là giải quyết nhiệm vụ trên. Học tăng cường là tiêu cực nếu nó va phải chướng ngại vật, là tích cực nếu nó chạm tới đích.

TÀI LIỆU THAM KHẢO

[1] Nguyễn Văn Hiếu (06-2020), "MACHINE LEARNING" <https://nguyenvanhieu.vn/machine-learning/>

[2] Nguyễn Văn Hiếu (2019), "Machine learning là gì? Tổng quan về machine learning <https://nguyenvanhieu.vn/machine-learning-la-gi/#51-mot-so-thuat-toan-machine-learning>

[3] Nguyễn Phúc Lương (28-04-2017), "Tổng quan Machine Learning" <https://viblo.asia/p/machine-learning-tong-quan-ve-machine-learning-RQqKLxaOK7z>

[4] Wikipedia, "Học máy" <https://vi.wikipedia.org/wiki/H%E1%BB%8Dc_m%C3%A1y>