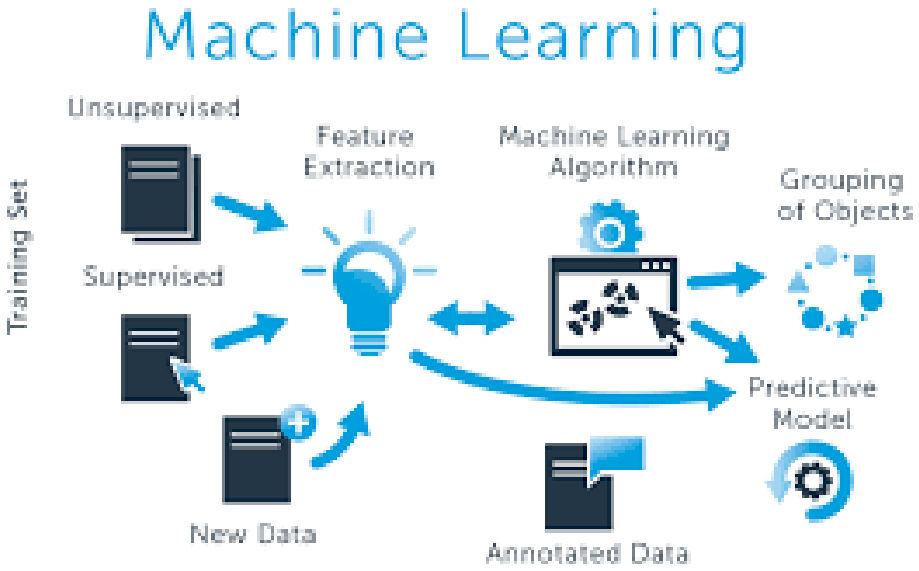
MACHINE LEARNING

1. Khái niệm

Machine learning là một lĩnh vực con của Trí tuệ nhân tạo(Artificial Intelligence) sử dụng các thuật toán cho phép máy tính có thể học từ dữ liệu để thực hiện các công việc thay vì được lập trình một cách rõ ràng.

Machine Learning là một thuật ngữ rộng để chỉ hành động bạn dạy máy tính cải thiện một nhiệm vụ mà nó đang thực hiện. Cụ thể hơn, machine learning đề cập tới bất kỳ hệ thống mà hiệu suất của máy tính khi thực hiện một nhiệm vụ sẽ trở nên tốt hơn sau khi hoàn thành nhiệm vụ đó nhiều lần. Hay nói cách khác, khả năng cơ bản nhất của machine learning là sử dụng thuật toán để phân tích những thông tin có sẵn, học hỏi từ nó rồi đưa ra quyết định hoặc dự đoán về một thứ gì đó có liên quan. Thay vì tạo ra một phần mềm với những hành động, hướng dẫn chi tiết để thực hiện một nhiệm vụ cụ thể, máy tính được "huấn luyện" bằng cách sử dụng lượng dữ liệu và các thuật toán để học cách thực hiện nhiệm vụ.



Hình 1. Tổng quan Machine Learning

Machine Learning có 3 mối quan hệ sau:

- Machine learning và trí tuệ nhân tạo.
- Machine learning và Big data.
- Machine learning và dự đoán tương lai.

Trong đó, Big data là một thuật ngữ cho việc xử lý một tập hợp dữ liệu rất lớn và phức tạp mà các ứng dụng xử lý dữ liệu truyền thống không xử lý được. Những tập dữ liệu lớn này có thể bao gồm các dữ liệu có cấu trúc, không có cấu trúc và bán cấu trúc. Có thể nói, machine learning phát triển hơn nhờ sự gia tăng của khối lượng dữ liệu của Big Data; ngược lại, giá trị của Big Data phụ thuộc vào khả năng khai thác tri thức từ dữ liệu của machine learning.

2. Cách tiếp cận của Machine Learning

Có nhiều cách tiếp cận có thể được thực hiện khi tiến hành machine learning. Chúng thường được xếp vào các nhóm dưới đây. Học có giám sát và học không giám sát là những cách tiếp cận kinh điển và được sử dụng phổ biến nhất. Học bán giám sát và học tăng cường là những cách tiếp cận mới hơn, phức tạp hơn nhưng cũng đã cho thấy những kết quả ấn tượng.

Định lý No Free Lunch là một định lý nổi tiếng trong machine learning. Nó nói rằng không có thuật toán duy nhất nào hoạt động tốt cho tất cả các nhiệm vụ. Mỗi nhiệm vụ mà bạn cố gắng giải quyết đều có những đặc điểm riêng. Do đó, có rất nhiều thuật toán và cách tiếp cận phù hợp với từng vấn đề riêng lẻ.

2.1 Học có giám sát (Supervised learning)

Trong học có giám sát, mục tiêu của mô hình là tìm ra luật để ánh xạ giữa đầu vào và đầu ra. Ví dụ, đầu vào là thông tin về thời tiết, đầu ra là số lượng người sẽ đến bãi biển và trong học có giám sát, mô hình cần tìm được mối liên hệ giữa thời tiết và lượng người đến bãi biển.

Các ví dụ đã được gán nhãn là các dữ liệu quá khứ chứa các cặp đầu vào / đầu ra. Qua quá trình huấn luyện, mô hình sẽ thử đoán giá trị của đầu ra và so sánh với nhãn chuẩn, từ đó hiệu chỉnh lại dự đoán của mình. Khi được huấn luyện đủ nhiều, mô hình sẽ bắt đầu có được những dự đoán chính xác. Đây cũng là lý do cách tiếp cận này gọi là học có giám sát.

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 (Unsupervised learning)

Trong học không giám sát, dữ liệu không được gán nhãn, nhiệm vụ của mô hình là tự mình tìm ra các mẫu ẩn nằm trong dữ liệu. Ví dụ trực quan cho cách tiếp cận này là việc xếp các đồng xu cùng loại vào cùng một đống. Dù bạn không biết đồng xu đó là tiền của nước nào, mệnh giá bao nhiêu nhưng bạn vẫn có thể nhóm các đồng xu giống nhau vào với nhau.

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.

Cách tiếp cận này có những ứng dụng thú vị. Ví dụ, ta có thể biết được những khách hàng nào có hành vi mua hàng giống nhau, những mặt hàng nào thường được

mua cùng với nhau, phát hiện bất thường trong các giao dịch. Ngoài ra, cách tiếp cận này có thể tìm ra cách biểu diễn dữ liệu hiệu quả hơn thông qua việc giảm chiều dữ liệu.

2.3 Học bán giám sát (Semi-supervised learning)

Học bán giám sát là sự pha trộn giữa học có giám sát và không giám sát. Quá trình huấn luyện không được giám sát chặt chẽ với các nhãn đầu ra cho mỗi ví dụ đầu vào. Nhưng chúng ta cũng không để mô hình sinh ra các kết quả một cách tùy tiện.

Pha trộn một lượng nhỏ dữ liệu có nhãn và một lượng lớn hơn các dữ liệu không có nhãn giúp giảm gánh nặng trong các bài toán không có nhiều dữ liệu. Do đó, ta có thể đưa machine learning vào nhiều bài toán với nhiều ứng dụng thú vị hơn.

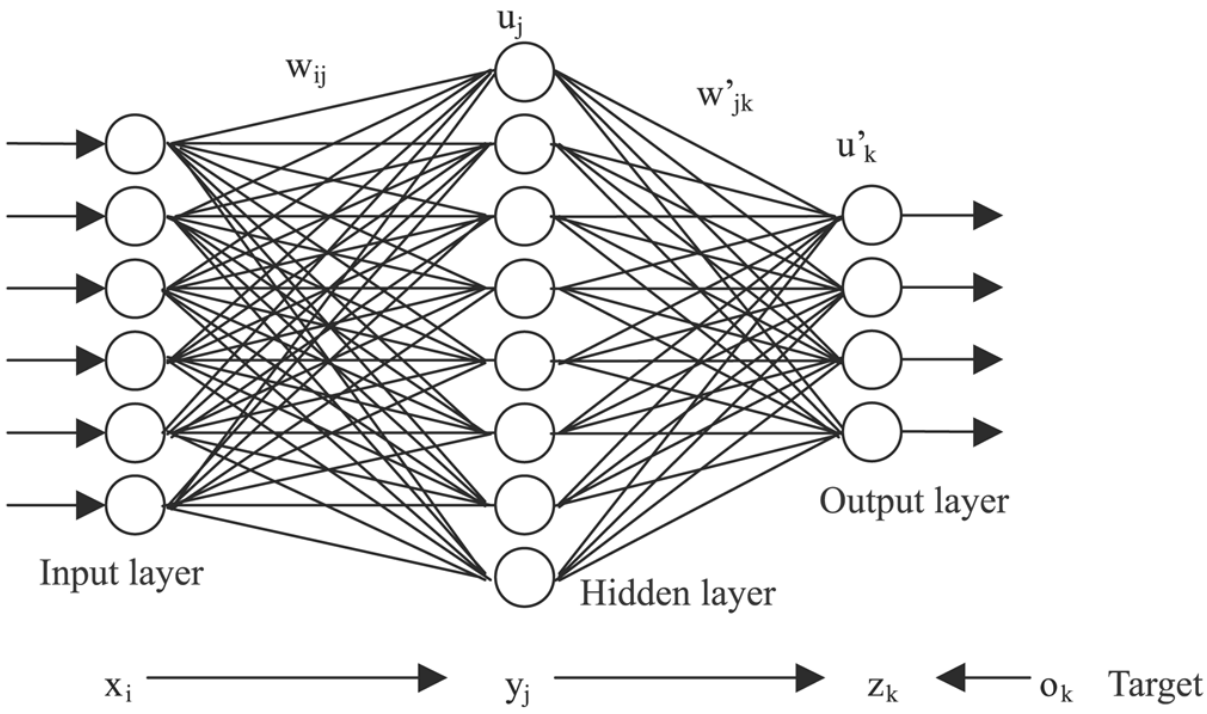
2.4 Học tăng cường (Reinforcement learning)

Học tăng cường là cách tiếp cận không sử dụng nhãn của dữ liệu mà giúp mô hình học thông qua cơ chế thưởng phạt. Ý tưởng đằng sau cách tiếp cận này là việc đưa ra những phản hồi tích cực và tiêu cực giúp tăng cường các hành vi đúng đắn. Phương pháp này có những nét chung với phương pháp "Cây gậy – Củ cà rốt" trong quản trị học.

3. Thuật toán Machine Learning

3.1 Deep learning

Thuật toán này sử dụng neural network, cấu trúc được lấy "cảm hứng" từ mạng lưới thần kinh của con người, một trong những cấu trúc sống với khả năng học hỏi cao nhất. Giống như mạng lưới thần kinh trong não bộ, Neural Network bao gồm nhiều lớp "neurons" liên kết với nhau thành một mạng lưới. Hai thuật toán Deep Learning có tính ứng dụng cao nhất là Recurrent Neural Network và Convolutional Neural Network.



Hình 2. Neural Network

3.2 Probabilistic Models

Đây là mô hình cố gắng giải quyết bài toán bằng phân bổ 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.

4. Ứng dụng thực tế

4.1 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).
- Phân tích ngữ nghĩa (Sentiment Analysis).
- Khai thác thông tin (Information Extraction).

4.2 Xử lý ảnh

Bài toán xử lý ảnh(Image Processing) giải quyết các vấn đề phân tích thông tin từ hình ảnh hay thực hiện một số phép biến đổi. Một số ví dụ là:

- Gắn thẻ hình ảnh(Image Tagging).
- Nhận dạng ký tự (Optical Character Recognition).

- Ô tô tự lái (Self-driving cars)

4.3 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:

- Phát hiện bất thường(Anomaly detection).
- Phát hiện các quy luật (Association rules).
- Gom nhóm (Grouping). Ví dụ, trong các nền tảng SaaS, người dùng được phân nhóm theo hành vi hoặc thông tin hồ sơ của họ.
- Dự đoán (Predictions). Ví dụ, bạn có thể dự đoán giá của căn hộ dựa trên các dữ liệu về giá các căn hộ bạn đã có.

TÀI LIỆU THAM KHẢO

[1] 08/02/2020, “Machine Learning và các khái niệm cơ bản” <https://trituenhantao.io/kien-thuc/machine-learning-va-cac-khai-niem-co-ban/>

[2] Nguyễn Văn Hiếu (2019), “Machine learning là gì? Tổng quan về Machine learning” <https://nguyenvanhieu.vn/machine-learning-la-gi/#1-machine-learning-la-gi>

[3] Wikipedia (13/03/2021), “Học máy” <https://vi.wikipedia.org/wiki/H%E1%BB%8Dc_m%C3%A1y>