**Machine Learning**

Machine Learning tập trung vào các ứng dụng học hỏi từ kinh nghiệm và cải thiện độ chính xác trong quá trình ra quyết định hoặc dự đoán của chúng theo thời gian.

1. **Machine Learning là gì?**

* Machine Learning là một nhánh của trí tuệ nhân tạo (AI) tập trung vào việc xây dựng các ứng dụng học hỏi từ dữ liệu và cải thiện độ chính xác của chúng theo thời gian mà không được lập trình để làm như vậy.
* Trong khoa học dữ liệu, một thuật toán là một chuỗi các bước xử lý thống kê. Trong học máy, các thuật toán được ‘trained’ để tìm ra các mẫu và tính năng trong một lượng lớn dữ liệu để đưa ra quyết định và dự đoán dựa trên dữ liệu mới. Thuật toán càng tốt thì các quyết định và dự đoán càng chính xác khi nó xử lý nhiều dữ liệu hơn.
* Ngày nay, các ví dụ về machine learning đều ở xung quanh chúng ta. Trợ lý kỹ thuật số tìm kiếm trên web và phát nhạc theo lệnh thoại của chúng ta. Các trang web giới thiệu các sản phẩm, phim và bài hát dựa trên những gì chúng ta đã mua, đã xem hoặc đã nghe trước đó, Robot hút bụi sàn nhà. Trình phát hiện thư rác ngăn các email không mong muốn đến hộp thư đến của chúng ta. Hệ thống phân tích hình ảnh y tế giúp bác sĩ phát hiện khối u mà họ có thể đã bỏ sót. Và những chiếc xe tự lái đầu tiên đang lên đường.

1. **Machine Learning hoạt động như thế nào?**

* Có bốn bước cơ bản để xây dựng một ứng dụng (hoặc mô hình) học máy. Những điều này thường được thực hiện bởi các nhà khoa học dữ liệu làm việc chặt chẽ với các chuyên gia kinh doanh mà mô hình đang được phát triển.

1. Chọn và chuẩn bị training data set.

* Training data set là tập dữ liệu đại diện cho dữ liệu mà mô hình học máy sẽ nhập vào để giải quyết vấn đề.
* Trong một số trường hợp, training data set được gắn nhãn dữ liệu — ‘tagged’ để gọi ra các tính năng và phân loại mà mô hình sẽ cần xác định. Dữ liệu khác không được gắn nhãn và mô hình sẽ cần trích xuất các tính năng đó và chỉ định phân loại của riêng nó.

1. Chọn 1 thuật toán để chạy trên training data set.

* Các loại thuật toán machine learning phổ biến để sử dụng với dữ liệu được gắn nhãn bao gồm: **Regression algorithms, Decision trees, Instance-based algorithms.**
* Các thuật toán để sử dụng với dữ liệu không được gắn nhãn bao gồm: **Clustering algorithms, Association algorithms, Neural networks.**

1. Training thuật toán để tạo ra model.

* Training thuật toán là một quá trình lặp đi lặp lại – nó liên quan đến việc chạy các biến thông qua thuật toán, so sánh kết quả đầu ra với kết quả mà nó đáng lẽ phải tạo ra, điều chỉnh trọng số và độ lệch trong thuật toán để có thể mang lại kết quả chính xác hơn và chạy lại các biến cho đến khi thuật toán trả về kết quả chính xác hầu hết thời gian. Thuật toán chính xác được đào tạo kết quả là mô hình học máy — một điểm khác biệt quan trọng cần lưu ý, bởi vì 'thuật toán' và 'mô hình' được sử dụng thay thế cho nhau một cách không chính xác.

1. Sử dụng và cải tiến model.

* Bước cuối cùng là sử dụng mô hình với dữ liệu mới và trong trường hợp tốt nhất là để mô hình cải thiện độ chính xác và hiệu quả theo thời gian. Dữ liệu mới đến từ đâu sẽ phụ thuộc vào vấn đề đang được giải quyết.

1. **Machine Learning methods**

* Phương pháp học máy được chia thành 3 loại chính:

1. Supervised machine learning.

* Supervised machine learning tự training trên tập dữ liệu được gắn nhãn. Tức là, dữ liệu được gắn nhãn với thông tin mà mô hình học máy đang được xây dựng để xác định và thậm chí có thể được phân loại theo những cách mà mô hình được cho là để phân loại dữ liệu.
* Supervised machine learning yêu cầu ít training data hơn so với các phương pháp học máy khác và giúp đào tạo dễ dàng hơn vì kết quả của mô hình có thể được so sánh với kết quả được dán nhãn thực tế. Tuy nhiên, dữ liệu được gắn nhãn đúng cách rất tốn kém để chuẩn bị và có nguy cơ trang bị quá nhiều hoặc tạo ra một mô hình gắn chặt và thiên vị với dữ liệu đào tạo đến mức nó không xử lý chính xác các biến thể trong dữ liệu mới.

1. Unsupervised machine learning.

* Unsupervised machine learning nhập dữ liệu không được gắn nhãn và sử dụng các thuật toán để trích xuất các tính năng có ý nghĩa cần thiết để gắn nhãn, sắp xếp và phân loại dữ liệu trong thời gian thực mà không cần sự can thiệp của con người. Unsupervised machine learning không phải là việc tự động hóa các quyết định và dự đoán, mà nhiều hơn về việc xác định các mẫu và mối quan hệ trong dữ liệu mà con người sẽ bỏ lỡ.

1. Semi-supervised machine learning.

* Semi-supervised machine learning cung cấp một phương tiện vui vẻ giữa supervised và unsupervised learning. Trong quá trình training, nó sử dụng một tập dữ liệu có nhãn nhỏ hơn để hướng dẫn phân loại và trích xuất tính năng từ một tập dữ liệu lớn hơn, không được gắn nhãn. Semi-supervised machine learning có thể giải quyết vấn đề không có đủ dữ liệu được gắn nhãn (hoặc không đủ khả năng gắn nhãn đủ dữ liệu) để đào tạo thuật toán học có giám sát.