Họ và tên: Mai Minh Quân

MSSV: 20225661

# Tìm hiểu về Machine Learning

Machine Learning là quá trình cho phép học các quy tắc (pattern) từ dữ liệu được cung cấp.

* Các quy tắc (pattern) được máy học có thể được sử dụng để đưa ra một số ước tính/dự đoán.

A diagram of data mining

Description automatically generated

Source: Baig M. Optimizing AI and Machine Learning Solutions. Your ultimate guide...2024

# Xác định một bài toán Machine Learning

Điều quan trọng cần chú ý là không phải tất cả các vấn đề đều cần được giải quyết bằng các bài toán Machine Learning.

Machine Learning đi kèm với những yêu cầu và thách thức riêng của nó.

* Một khối lượng lớn dữ liệu huấn luyện mà máy cần học để tạo ra mô hình.
* Tốn nhiều thời gian.

Điều kiện lý tưởng cho một giải pháp Machine Learning là một vấn đề mà bạn có đủ dữ liệu, đủ thời gian và yêu cầu độ chính xác cao hơn.

Source: Baig M. Optimizing AI and Machine Learning Solutions. Your ultimate guide...2024

# Các kiểu học máy

Chúng ta có ba kiểu học máy chính.

A diagram of a diagram of a learning process

AI-generated content may be incorrect.

|  |  |  |  |
| --- | --- | --- | --- |
| Các kiểu học máy | Học có giám sát  (Supervised Learning) | Học không giám sát  (Unsupervised Learning) | Học tăng cường  (Reinforcement Learning) |
| Định nghĩa | Máy học từ dữ liệu được cung cấp có nhãn. | Máy học từ dữ liệu được cung cấp không có nhãn. | Một tác nhân tương tác với môi trường bằng cách thực hiện các hành động và học hỏi từ các lỗi và phần thưởng. |
| Các loại bài toán | Phân loại (Classification), Hồi quy (Regression) | Phân cụm (Clustering), Giảm chiều dữ liệu (Dimensionality reduction) | Chính sách/hành động tối ưu (Optimal policy/actions) |
| Huấn luyện | Có sự giám sát bên ngoài | Không có sự giám sát | Không có sự giám sát |
| Cách tiếp cận | Giảm thiểu lỗi trên tập dữ liệu huấn luyện. | Nắm bắt các mẫu/cấu trúc trong dữ liệu. | Phương pháp thử và sai nhằm tối đa hóa kết quả. |

Ngoài ra, có nhiều quy tắc kinh nghiệm để lựa chọn cách định dạng cụ thể cho bài toán học máy đối với một nhiệm vụ/tình huống nhất định:

* Học có giám sát: Nếu giá trị mục tiêu có sẵn.
* Phân loại: Nếu giá trị mục tiêu là một danh mục (category).
* Hồi quy: Nếu giá trị mục tiêu là một đại lượng (quantity).
* Học không giám sát: Nếu nhãn mục tiêu không có sẵn.
* Phân cụm: Để nhóm các thực thể.
* Giảm chiều dữ liệu: Tinh lọc thông tin thành một tập hợp các đặc trưng (feature) nhỏ hơn.
* Học tăng cường: Để xác định hành động tốt nhất trong một tình huống.

Source: Baig M. Optimizing AI and Machine Learning Solutions. Your ultimate guide...2024

# Kết hợp các kiểu học máy.

Nhược điểm của các phương pháp riêng lẻ:

* Phương pháp học có giám sát thuần túy gặp khó khăn với vấn đề mất cân bằng lớp nghiêm trọng.
* Phương pháp học không giám sát ít bị ảnh hưởng bởi vấn đề này nhưng học không giám sát có thể tạo ra các cụm nghi ngờ là điểm ngoại lai.
* Do đó, cả hai phương pháp đều có những nhược điểm riêng.

Giải pháp kết hợp:

* Để khắc phục những nhược điểm trên, chúng ta có thể kết hợp hai phương pháp.
* Cụ thể, kết quả gán cụm từ phương pháp học không giám sát được sử dụng làm biến dự báo trong dữ liệu huấn luyện cho mô hình phân loại (học có giám sát).
* Đặc trưng mới "Cụm” (Cluster) này cung cấp thông tin hữu ích, giúp mô hình học có giám sát hoạt động tốt hơn, ngay cả khi dữ liệu bị mất cân bằng lớp.
* Việc thêm đặc trưng này giúp giảm thiểu vấn đề các đặc trưng bị chi phối bởi thông tin từ lớp chiếm đa số.

Source: Baig M. Optimizing AI and Machine Learning Solutions. Your ultimate guide...2024

# Huấn luyện mô hình

* Tổng quan về huấn luyện mô hình học có giám sát:

Dữ liệu huấn luyện được cung cấp cho mô hình. Mô hình sẽ tạo ra các dự đoán. Các dự đoán này được so sánh với các giá trị y thực tế từ tập dữ liệu nhãn. Mục tiêu của học có giám sát là so sánh kết quả đầu ra của mô hình với giá trị thực tế, từ đó điều chỉnh các tham số. Quá trình lặp lại bao gồm việc phân tích sai lệch và điều chỉnh mô hình, được gọi là huấn luyện mô hình.

A black square with white text

AI-generated content may be incorrect.

* Để đánh giá hiệu quả của mô hình, chúng ta chia toàn bộ tập dữ liệu thành ba phần:
* Tập dữ liệu huấn luyện (Training data set)
* Tập dữ liệu này được sử dụng để huấn luyện mô hình.
* Mô hình sẽ tạo ra các dự đoán dựa trên dữ liệu huấn luyện.
* Chúng ta tính toán "loss" giữa các dự đoán và giá trị thực tế.
* "Loss" là một đại lượng số học thể hiện sự khác biệt giữa dự đoán và thực tế.
* Dựa trên "loss", chúng ta điều chỉnh các tham số của mô hình. Quá trình này được gọi là "huấn luyện".

A black square with white text

AI-generated content may be incorrect.

* Tập dữ liệu kiểm chứng (Validation data set)
* Tập dữ liệu này được sử dụng để kiểm tra hiệu suất của mô hình sau quá trình huấn luyện.
* Sau mỗi lần huấn luyện, chúng ta đưa tập dữ liệu kiểm chứng vào mô hình để tính toán "loss".
* "Loss" trên tập dữ liệu kiểm chứng giúp chúng ta đánh giá khả năng tổng quát hóa của mô hình đối với dữ liệu mới.
* Điểm quan trọng: "Loss" từ tập dữ liệu kiểm chứng không được sử dụng để điều chỉnh trực tiếp các tham số của mô hình. Điều này khác với quá trình huấn luyện.

A black square with white text

AI-generated content may be incorrect.

* Tập dữ liệu kiểm thử (Testing data set)
* Sau khi hoàn thành quá trình huấn luyện và lựa chọn mô hình tốt nhất chúng ta sử dụng tập dữ liệu kiểm thử để đánh giá hiệu suất cuối cùng.
* Tập dữ liệu kiểm thử giúp xác định khả năng tổng quát hóa thực sự của mô hình đã chọn.
* Tập dữ liệu kiểm thử là bước kiểm tra cuối cùng, để xác định xem mô hình có hoạt động tốt trên dữ liệu mới hay không.

A black square with white text

AI-generated content may be incorrect.

Source: Machine Learning for Everybody – Full Course – freeCodeCamp.org

# Các mô hình học máy

Một vài mô hình học máy cơ bản

* K-Nearest Neighbors
* Naive Bayes
* Linear Regression
* Logistic Regression
* Support Vector Machine
* Neural Network

Source: Machine Learning for Everybody – Full Course – freeCodeCamp.org