* 1. **Machine learning**
     1. ***Khái niệm***

Machine learning (ML) [8] 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.

Trong cuốn sách của mình [9] , Tom Mitchell đã định nghĩa machine learning: “*A computer program is said to learn to perform a****task T****from****experience E****, if its performance at task T, as measured by a****performance metric P****, improves with experience E over time”*

Tạm dịch: “*Một chương trình máy tính được cho là học để thực hiện một****nhiệm vụ T****từ****kinh nghiệm E****, nếu hiệu suất thực hiện công việc T của nó được đo bởi****chỉ số hiệu suất P****và được cải thiện bởi kinh nghiệm E theo thời gian*”

Định nghĩa trên được dùng phổ biến và cho ta hình dung một các rõ ràng các phần của một thuật toán ML. Nó gồm 3 yếu tố cơ bản: nhiệm vụ T, kinh nghiệm E, chỉ số hiệu suất P. Dựa vào các yếu tố này mà người ta phân loại có ML theo nhiều các khác nhau, ứng dụng nhiều lĩnh vực trên thực tế

* + 1. ***Ứng dụng***

Trên thực tế, hiện nay Machine learning đang là lĩnh vực cho thấy sự tiến bộ trong tiến trình của Trí tuệ nhân tạo. Nó đang là một chủ đề nóng và có khả năng làm cho máy móc trở nên thông minh hơn. Một số lĩnh vực phổ biến mà ML góp mặt như:

* *Xử lí 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), giống như Facebook, một thuật toán tự động phát hiện khuôn mặt của bạn và bạn bè trên những bức ảnh. Về cơ bản, thuật toán này học từ những bức ảnh mà bạn tự gắn thẻ cho mình trước đó.
* **Nhận dạng ký tự** (Optical Character Recognition), là một thuật toán chuyển dữ liệu trên giấy tờ, văn bản thành dữ liệu số hóa. Thuật toán phải học cách nhận biết ảnh chụp của một ký tự là ký tự nào.
* **Ô tô tự lái** (Self-driving cars), một phần cơ chế sử dụng ở đây là xử lý ảnh. Một thuật toán machine learning giúp phát hiện các mép đường, biển báo hay các chướng ngại vật bằng cách xem xét từng khung hình video từ camera.
* *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,…
* *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 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.
* **Phát hiện các quy luật** (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.
* **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), 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.
* *Trò chơi điện tử & Robot*

*T*rò 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.

Một thành tựu gần đây nhất là cỗ máy Alpha Go của Google DeepMind đã đánh bại kỳ thủ cờ vậy số 1 thế giới. Trong khi cờ vây là một trò chơi có không gian trạng thái cực kỳ lớn.

* + 1. ***Phân loại***

Học máy có thể được phân làm 3 loại: Học có dám sát (Supervied Learning), Học không dám sát (Unsupervised Learning), Học tăng cường (Reinforcement Learning). Ngoài khác nhau về cách thuật toán sử dụng dữ liệu và loại vấn đề mà chúng giải quyết thì khác biệt lớn giữa các loại thuật toán này là cách mà chúng ta cung cấp tập dữ liệu huấn luyện cho mô hình.

***\* Học có giám sát:***

Trong học có giám sát, thuật toán machine learning thực hiện quá trình “**chuyển dữ liệu đầu vào thành đầu ra mong muốn”.**

Đầu vào được gọi là input, đầu ra còn được gọi là  nhãn (label). Tập dữ liệu huấn luyện được gọi là training set; Mỗi mẫu dữ liệu trong tập huấn luyện được gọi là training example.

Thuật toán học có giám sát cần học cách để chuyển đổi mỗi input (đầu vào) thành nhãn (kết quả) tương ứng. Do vậy, mọi mẫu dữ liệu huấn luyện trong tập dữ liệu huấn luyện cần phải biết trước nhãn của nó.

Tùy thuộc vào loại đầu ra mong muốn, chúng ta tiếp tục chia nhỏ học có giám sát gồm:

* Phân loại (Classification): Khi đầu ra mong muốn của chúng ta là một tập hữu hạn và rời rạc. Chẳng hạn như là giá trị cảm xúc của khách hàng: tích cực, tiêu cực, trung tính.
* Hồi quy (Regression):Khi đầu ra mong muốn là một dải giá trị liên tục. Về mặt bản chất hồi quy là bài toán gán nhãn cho dữ liệu thực, biểu diễn và dự đoán đầu ra dựa trên tổng quát hóa các dữ liệu từ đầu vào để tìm ra một hàm dự đoán. Chẳng hạn xây dựng hàm dự đoán giá nhà, giá cổ phiếu theo thời gian hoặc các biến đầu vào khác.

Nếu ta chú ý có thể thấy rằng, thực chất bài toán phân lớp cũng là một trường hợp đặc biệt của bài toán hồi quy, khi giá trị các dự đoán đầu ra thay vì các lớp thì là các giá trị rời rạc đại diện cho lớp.

Học có giám sát là thuật toán phổ biến nhất trong các thuật toán machine learning. Hạn chế khi sử dụng thuật toán này là chúng ta cần cung cấp dữ liệu có gán nhãn. Trong nhiều trường hợp, để có được dữ liệu gán nhãn này rất tốn rất nhiều chi phí. Chẳng hạn, nếu ta cần 10.000 review có nhãn để huấn luyện mô hình; Việc này sẽ cần con người đọc từng review và gán nhãn thủ công; Điều này rất tốn thời gian và công sức. Đây cũng là một rào cản của ML: xây dựng các tập dữ liệu gán nhãn chất lượng.

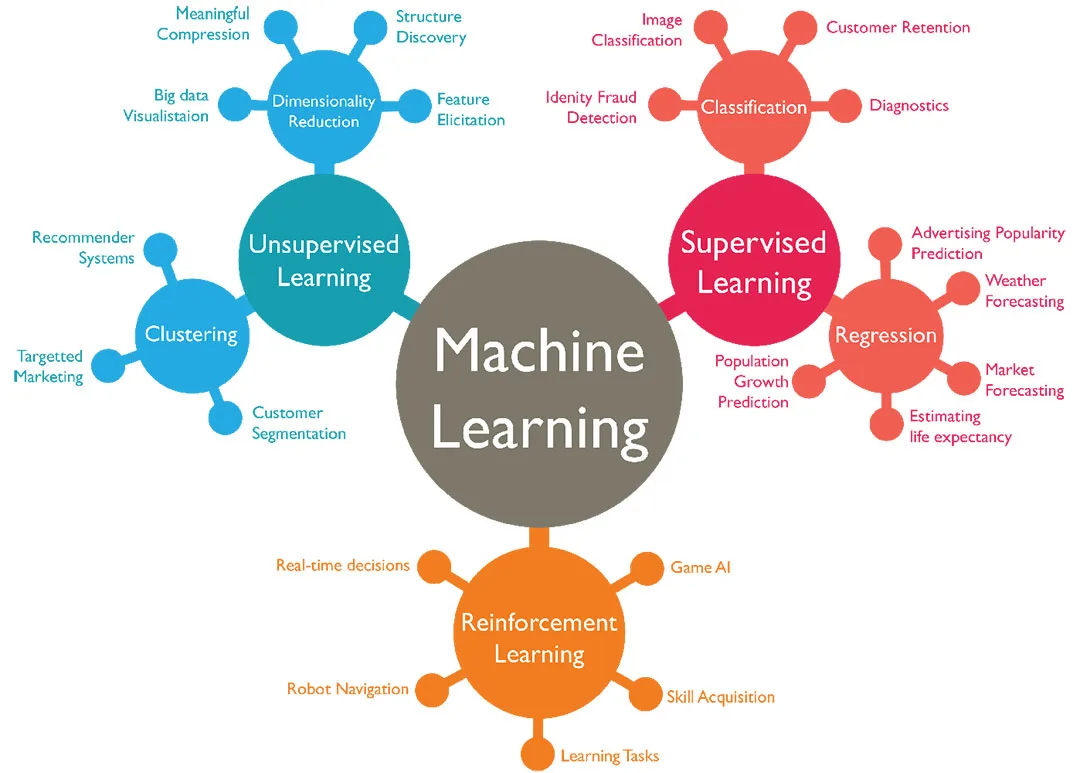
***\*Học không giám sát***

Trong học không giám sát, các mẫu dữ liệu trong học không giám sát chỉ cần input (đầu vào) mà không cần nhãn (đầu ra). Nó được sử dụng nhiều trong việc khám phá cấu trúc và mối quan hệ của dữ liệu. Một thuật toán điển hình là bài toán phân cụm (Clustering). Nó học cách để tìm các mẫu dữ liệu tương tự nhau và nhóm vào thành các cụm.

***\*Học tăng cường***

Học tăng cường là nhiệm vụ học thông qua thử và sai, mục tiêu của nó là đưa ra quyết định.Đây là kiểu bài toán mà máy sẽ được học một hàm mục tiêu sao cho khả năng thích nghi (được nhận “phần thưởng” nhiều nhất và ít bị “phạt” nhất) theo môi trường. Chẳng hạn rẽ trái được thưởng 1 điểm còn rẽ phải trừ 1 điểm. Việc học thích nghi theo dữ liệu môi trường nhằm tối ưu hóa một mục tiêu được đặt ra.

Khác với hoc có giám sát, trong học tăng cường không có các cặp dữ liệu vào/kết quả đúng, các hành động gần tối ưu cũng không được đánh giá đúng sai một cách tường minh.



Hình 1.10: Phân loại Machine learning [10]

* 1. **Bài toán dự đoán dựa trên mô hình Machine learning**

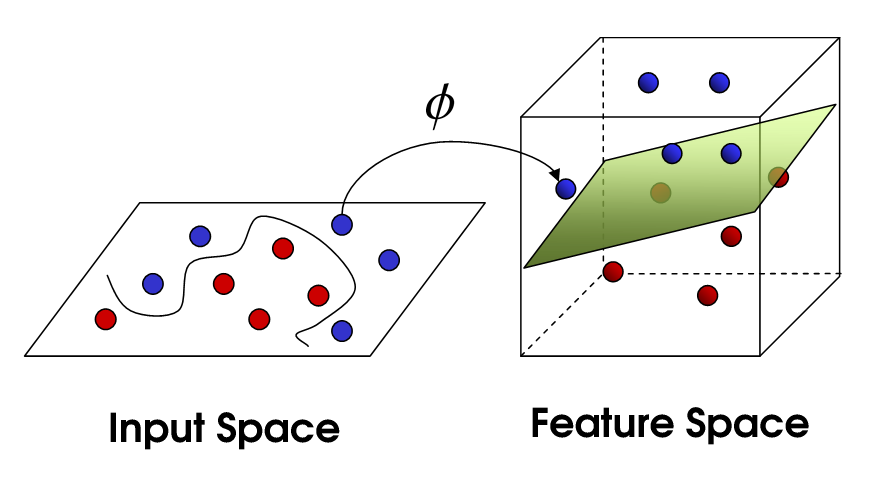
Bài toán dự đoán (prediction) trong Machine learning được hiểu là bài toán gán nhãn, tức là quá trình từ một lượng thông tin đầu vào, thuật toán học máy có thể tính toán cho ra dự đoán về đầu ra (lable) mong muốn. Để làm được điều này nó cần nó bộ dữ liệu huấn luyện được gán nhãn trước. Như vậy học có dám sát được đề xuất để giải quyết vấn đề này. Có 2 bài toán điển hình đã được nói ở phần trên đó là phân lớp (Classification) và hồi quy (Regression). Đối với đầu ra xác định như trong bài toán dự đoán ngành nghề của chúng ta là các nhóm ngành thì thuật toán phân lớp được đề xuất. Sau đây là một số thuật toán phân lớp phổ biến được nghiên cứu trong đề tài:

* + 1. ***Thuật toán SVM***

SVM (Support Vector Machine) [11] là một thuật toán học máy có giám sát, có thể sử dụng cho cả hai chương trình hồi quy và phân loại. Tuy nhiên thường được dùng chủ yếu và hiệu quả hơn trong bài toán phân loại.

Quy trình điển hình của thuật toán: Đầu tiên mỗi mục dữ liệu là được vẽ trong không gian n chiều, trong đó n là số đối tượng và giá trị của mỗi đối tượng là giá trị của tọa độ cụ thể đó. Bước tiếp theo là phân loại bằng cách lấy siêu mặt phẳng tuyến tính (*linear* *hyper-plane*) để phân tách thành hai lớp riêng biệt. Một yếu tố làm nên hiệu quả của SVM đó là việc sử dụng thuật toán hạt nhân (***Kernel function)*** khiến cho các phương pháp chuyển không gian trở nên linh hoạt hơn. Nói một cách đơn giản nó thực hiện một số biết đổi dữ liệu phức tạp, sau đó tìm ra quá trình tách dữ liệu dựa trên các nhãn hoặc đầu ra mà chúng ra đã xác định trước.

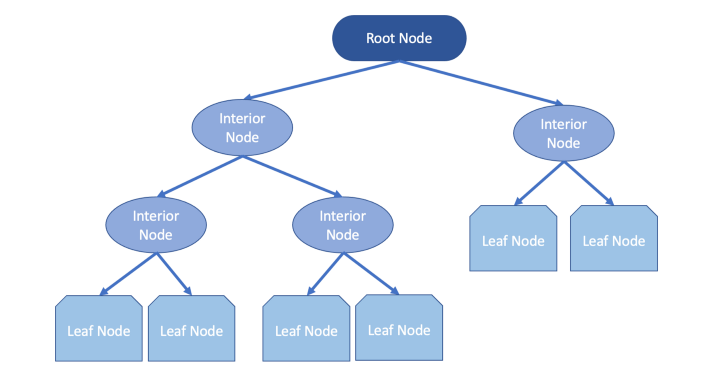
Để tính toán sự tối ưu bằng toán học, trong SVM sử dụng thuật ngữ ***Margin.*** *Margin*là khoảng cách giữa siêu phẳng đến 2 điểm dữ liệu gần nhất tương ứng với 2 phân lớp. SVM cố gắng tối ưu thuật toán bằng các tìm cách maximize giá trị *margin* này, từ đó tìm ra siêu phẳng đẹp nhất để phân 2 lớp dữ liệu.



Hình 1.11: Mô tả thuật toán phân lớp SVM

* + 1. ***Thuật toán Decision tree***

Cây quyết định (Decision Tree) [12] một trong những thuật toán phổ biến của Machine Learning thuộc nhánh học có dám sát. Decision Tree ra đời từ những năm 1975 từ một tác giả có tên Ross Quinlan. Nó sử dụng một loạt các quyết định tuần tự được thực hiện để đạt được kết quả cụ thể. Nó được dùng trong cả bài toán phân lớp và hồi quy. Cây quyết định là cây mà mỗi nút (node) biểu diễn một đặc trưng (tính chất), mỗi nhánh (branch) biểu diễn một quy luật (rule) và mỗi lá (leaf) biểu biễn một kết quả.



Hình 1.12: Cấu trúc cây quyết định

Hãy nhìn vào cấu trúc cây quyết định ở trên, chúng ta có thể thấy nó bắt đầu với một nút gốc, các thuộc tính tốt nhất trở thành các nút bên trong, tức là các nút quyết định. Sau đó, các nút bên trong kiểm tra một điều kiện và thực hiện quyết định, chia không gian mẫu thành hai. Các nút lá thể hiện một phân lớp, khi bản ghi đến nút lá, thuật toán sẽ gán nhãn của lá tương ứng. Quá trình này được gọi là phân vùng đệ quy của không gian mẫu. Cây quyết định sử dụng một số hàm chi phí để chọn cách phân chia tốt nhất. Tôi đang cố gắng tìm thuộc tính / tính năng tốt nhất hoạt động tốt nhất trong việc phân loại dữ liệu đào tạo. Quá trình này được lặp lại cho đến khi đạt được nút lá và do đó, được gọi là tách nhị phân đệ quy. Khi thực hiện quy trình này, tất cả các giá trị đều được xếp thẳng hàng và cây sẽ kiểm tra các phần tách khác nhau và chọn một giá trị trả về chi phí thấp nhất, làm cho đây là một cách tiếp cận tham lam.

Có 2 thuật toán cây quyết định phổ biến nhất, được phân loại dựa trên tham số đánh giá sự phân lớp, đó là:

* *ID3 (Iterative Dichotomiser 3):* dùng Entropy Function và Information Gain
* Entropy:Được phát triển bởi Claude Shannon, entropy cung cấp cho chúng ta các thước đo về sự lộn xộn của dữ liệu. Giá trị entropy cao cho thấy khả năng dự đoán ít hơn, hãy coi entropy của một đối tượng là lượng thông tin trong đối tượng đó. Cây quyết định hoạt động để tối đa hóa độ tinh khiết của các lớp khi thực hiện tách, cung cấp sự rõ ràng hơn của các lớp trong các nút lá. Entropy được tính toán trước và sau mỗi lần tách. Nếu entropy tăng lên, một lần tách khác sẽ được thử hoặc nhánh của cây sẽ dừng lại, tức là cây hiện tại có entropy thấp nhất. Nếu entropy giảm, sự phân chia sẽ được giữ nguyên. Công thức tính toán entropy của toàn bộ tập dữ liệu:



Trong đó:

- pi: tần suất xuất hiện của các mẫu trong lớp Ci với I = {1,2,…,m}

- S: số lượng tập huấn luyện

- Si: số các mẫu S trong lớp Ci

* Information Gain:Thông tin thu được sử dụng entropy như một thước đo tạp chất. Đó là sự khác biệt về entropy từ trước đến sau khi phân tách, và sẽ cho chúng ta một con số về độ không chắc chắn đã giảm đi bao nhiêu. Nó cũng là tiêu chí chính được sử dụng trong thuật toán cây phân loại ID3. Để tính toán mức tăng thông tin:



Trong đó:

- S là tập hợp ban đầu với thuộc tính A. Các giá trị của v tương ứng là các giá trị của thuộc tính A.

- Sv bằng tập hợp con của tập S mà có thuộc tính A mang giá trị v.

- |SV| là số phần tử của tập SV.

- |S| là số phần tử của tập S.

* *CART* *(Classification and Regression Trees)*: dùng Gini Index (Classification)

Chỉ số Gini được đề xuất bởi nhà khoa học *Corrado Gin***,** hàm này cho chúng ta biết các nút lá trên cây “thuần khiết” như thế nào. Tạp chất gini sẽ luôn có giá trị từ 0 đến 0,5, giá trị càng cao thì nhóm càng bị rối loạn. Công thức tính tạp chất gini:



trong đó (pi) là xác suất thuộc nhóm thứ i.