

 Machine Learning là một lĩnh vực nhỏ của Khoa Học Máy Tính, nó có khả năng tự học hỏi dựa trên dữ liệu đưa vào mà không cần phải được lập trình cụ thể.

Có hai cách phổ biến phân nhóm các algorithm của Machine learning. Một là dựa trên phương thức học (learning style), hai là dựa trên chức năng (function) (của mỗi thuật toán).

1. Phân nhóm dựa trên sự giống nhau về learning style của các algorithm

Theo phương thức học, các thuật toán Machine Learning thường được chia làm 4 nhóm: **Supervise learning,** **Unsupervised learning**, **Semi-supervised** **lerning** và **Reinforcement learning**. *Có một số cách phân nhóm không có Semi-supervised learning hoặc Reinforcement learning.*

**Supervised Learning (Học có giám sát):**

Supervised learning là thuật toán dự đoán đầu ra (outcome) của một dữ liệu mới (new input) dựa trên các cặp (*input, outcome*) đã biết từ trước. Cặp dữ liệu này còn được gọi là **(*data, label*)**, tức (*dữ liệu, nhãn*). Supervised learning là nhóm phổ biến nhất trong các thuật toán Machine Learning.

Một cách toán học, Supervised learning là khi chúng ra có một tập hợp biến đầu vào X={x1,x2,…,xN} và một tập hợp nhãn tương ứng Y={y1,y2,…,yN}, trong đó xi,yi là các vector. Các cặp dữ liệu biết trước (xi,yi)∈X×Y được gọi là tập *training data* (dữ liệu huấn luyện). Từ tập traing data này, chúng ta cần tạo ra một hàm số ánh xạ mỗi phần tử từ tập X sang một phần tử (xấp xỉ) tương ứng của tập Y:

yi≈f(xi),  ∀i=1,2,…,N

Mục đích là xấp xỉ hàm số f thật tốt để khi có một dữ liệu x mới, chúng ta có thể tính được nhãn tương ứng của nó y=f(x).

Thuật toán supervised learning còn được tiếp tục chia nhỏ ra thành hai loại chính:

* **Classification**

Một bài toán được gọi là *classification* nếu các *label* của *input data* được chia thành một số hữu hạn nhóm. Ví dụ: Gmail xác định xem một email có phải là spam hay không; các hãng tín dụng xác định xem một khách hàng có khả năng thanh toán nợ hay không.

* **Regression**

Nếu *label* không được chia thành các nhóm mà là một giá trị thực cụ thể. Ví dụ: một căn nhà rộng x m2, có y phòng ngủ và cách trung tâm thành phố z km sẽ có giá là bao nhiêu?

Gần đây Microsoft có một ứng dụng dự đoán giới tính và tuổi dựa trên khuôn mặt. Phần dự đoán giới tính có thể coi là thuật toán **Classification**, phần dự đoán tuổi có thể coi là thuật toán **Regression**. *Chú ý rằng phần dự đoán tuổi cũng có thể coi là****Classification****nếu ta coi tuổi là một số nguyên dương không lớn hơn 150, chúng ta sẽ có 150 class (lớp) khác nhau.*

**Unsupervised Learning (Học không giám sát)**

Một cách toán học, Unsupervised learning là khi chúng ta chỉ có dữ liệu vào X mà không biết *label* Y tương ứng.

Các bài toán Unsupervised learning được tiếp tục chia nhỏ thành hai loại:

* **Clustering (phân nhóm)**

Một bài toán phân nhóm toàn bộ dữ liệu XX thành các nhóm nhỏ dựa trên sự liên quan giữa các dữ liệu trong mỗi nhóm. Ví dụ: phân nhóm khách hàng dựa trên hành vi mua hàng. Điều này cũng giống như việc ta đưa cho một đứa trẻ rất nhiều mảnh ghép với các hình thù và màu sắc khác nhau, ví dụ tam giác, vuông, tròn với màu xanh và đỏ, sau đó yêu cầu trẻ phân chúng thành từng nhóm. Mặc dù không cho trẻ biết mảnh nào tương ứng với hình nào hoặc màu nào, nhiều khả năng chúng vẫn có thể phân loại các mảnh ghép theo màu hoặc hình dạng.

* **Association**

Là bài toán khi chúng ta muốn khám phá ra một quy luật dựa trên nhiều dữ liệu cho trước. Ví dụ: những khách hàng nam mua quần áo thường có xu hướng mua thêm đồng hồ hoặc thắt lưng; những khán giả xem phim Spider Man thường có xu hướng xem thêm phim Bat Man, dựa vào đó tạo ra một hệ thống gợi ý khách hàng (Recommendation System), thúc đẩy nhu cầu mua sắm.

**Semi-Supervised Learning (Học bán giám sát)**

**Reinforcement Learning (Học Củng Cố)**

1. Phân nhóm dựa trên sự giống nhau giữa các function của các algorithm

Regression Algorithms

* Linear Regression
* Logistic Regression
* Stepwise Regression

Classification Algorithms

* Linear Classifier
* Support Vector Machine (SVM)
* Kernel SVM
* Sparse Representation-based classification (SRC)

Instance-based Algorithms

* k-Nearest Neighbor (kNN)
* Learning Vector Quantization (LVQ)

Regularization Algorithms

* Ridge Regression
* Least Absolute Shrinkage and Selection Operator (LASSO)
* Least-Angle Regression (LARS)

Bayesian Algorithms

* Naive Bayes
* Gaussian Naive Bayes

Clustering Algorithms

* k-Means clustering
* k-Medians
* Expectation Maximization (EM)

Artificial Neural Network Algorithms

* Perceptron
* Softmax Regression
* Multi-layer Perceptron
* Back-Propagation

Dimensionality Reduction Algorithms

* Principal Component Analysis (PCA)
* Linear Discriminant Analysis (LDA)

Ensemble Algorithms

* Boosting
* AdaBoost
* Random Forest