1. Học máy có giám sát (supervised learning)

Học có giám sát (supervised learning) là một kỹ thuật của ngành học máy nhằm mục đích xây dựng một hàm f từ dữ tập dữ liệu huấn luyện (Training data). Dữ liệu huấn luyện bao gồm các cặp đối tượng đầu vào và đầu ra mong muốn. Đầu ra của hàm f có thể là một giá trị liên tục hoặc có thể là dự đoán một nhãn phân lớp cho một đối tượng đầu vào.

Nhiệm vụ của chương trình học có giám sát là dự đoán giá trị của hàm f cho một đối tượng đầu vào hợp lệ bất kì, sau khi đã xét một số mẫu dữ liệu huấn luyện (nghĩa là các cặp đầu vào và đầu ra tương ứng). Để đạt được điều này, chương trình học phải tổng quát hóa từ các dữ liệu sẵn có để dự đoán được những tình huống chưa gặp phải theo một cách hợp lý.

### Cách giải một bài toán học có giám sát

* Bước 1: Xác định loại của các dữ liệu huấn luyện: Trước tiên ta cần phải quyết định xem loại dữ liệu nào sẽ được sử dụng làm dữ liệu huấn luyện. Ta có thể chọn dữ liệu một kí tự viết tay đơn lẻ, toàn bộ một từ viết tay, hay toàn bộ một dòng chữ viết tay, …
* Bước 2: Thu thập tập dữ liệu huấn luyện. Khi thu thập tập dữ liệu huấn luyện cần phải đảm bảo được sự đặc trưng cho thực tế sử dụng của hàm chức năng. Do đó tập các dữ liệu đầu vào và đầu ra tương ứng phải được thu thập từ các chuyên gia hoặc từ việc đo đạc tính toán.
* Bước 3: Xác định việc biễu diễn các đặc trưng đầu vào cho hàm mục tiêu cần tìm. Độ chính xác của mục tiêu phụ thuộc rất lớn vào các đối tượng đầu vào được biểu diễn như thế nào. Đa số các đối tượng đầu vào được chuyển đổi thành một véc tơ đặc trưng chứa các đặc trưng cơ bản của đối tượng đó. Chú ý số lượng các đặc trưng không được lớn quá, để tránh sự bùng nổ tổ hợp tuy nhiên nó phải đủ lớn để đảm bảo dự đoán chính xác đầu ra.
* Bước 4: Xác định cấu trúc của hàm mục tiêu cần tìm và giải thuật học tương  
  ứng. Ví dụ, ta có thể sử dụng mạng nơ-ron nhân tạo, cây quyết định, …
* Bước 5: Hoàn thiện và thiết kế chương trình. Tiến hành chạy giải thuật học với tập dữ liệu huấn luyện thu thập được. Ta có thể điều chỉnh các tham số của giải thuật học bằng cách tối ưu hóa hiệu năng trên một tập con của tập huấn luyện, (gọi là tập kiểm chứng -validation set) của tập huấn luyện hay thông qua kiểm chứng chéo (cross-validation). Sau đó ta tiến hành đo đạc hiệu năng của giải thuật trên một tập dữ liệu kiểm tra độc lập với tập huấn luyện.
* A picture containing text, screenshot, font, diagram

  Description automatically generated

1. Học máy không giám sát (unsupervised learning)

Học không giám sát là một lớp mô hình học sử dụng một thuật toán để mô tả hoặc trích xuất ra các mối quan hệ tiềm ẩn trong dữ liệu. Khác với học có giám sát, học không giám sát chỉ thực thi trên dữ liệu đầu vào không cần các thuộc tính nhãn, hoặc mục tiêu của việc học. Tức là không hề được cung cấp trước một kiến thức nào trước trừ dữ liệu. Các dữ liệu không được "hướng dẫn" trước như trong trường hợp học có giám sát. Các thuật toán cần học được từ dữ liệu mà không hề có bất cứ sự hướng dẫn nào.

1. Học máy bán giám sát (Semi - supervised learning)

Trong học bán giám sát, một số dữ liệu quan sát dược gán nhãn và một số dữ liệu lại không được gán nhãn (số lượng này thường lớn hơn nhiều so với dữ liệu được gán nhãn). Các dữ liệu này được kết hợp với nhau để xây dựng các mô hình hiệu quả hơn.

Mục tiêu của mô hình học bán giám sát là tận dụng hiệu quả của toàn bộ dữ liệu có được, không chỉ các dữ liệu đã được gán nhãn mà cả các dữ liệu chưa được gán nhãn. Thông thường, điều này sẽ làm tăng tính chính xác của mô hình xây dựng được. Để làm được điều này, mô hình học bán giám sát thường kết hợp cả các thuật toán học không giám sát lẫn có giám sát. Các mô hình học không giám sát như phân cụm hoặc xấp xỉ phân phối được sử dụng để tìm ra các mẫu và nhóm dữ liệu, và các nhóm dữ liệu này sau đó sẽ được sử dụng để gán nhãn dựa trên các thuật toán học có giám sát. Học bán giám sát rất hiệu quả trong nhiều bài toán thực tế, khi tài nguyên sử dụng để gán nhãn dữ liệu rất tốn kém. Chẳng hạn trong các bài toán thị giác máy tính (dữ liệu ảnh), xử lý ngôn ngữ tự nhiên (dữ liệu văn bản), nhận dạng giọng nói (dữ liệu âm thanh). Học bán giám sát có thể làm tăng độ chính xác cho các mô hình học có giám sát bằng cách tận dụng thông tin từ các dữ liệu không gán nhãn. Các dữ liệu này vốn rất phổ biến trên môi trường Internet và có thể dễ dàng kiếm được hơn là số lượng hạn chế các dữ liệu có gán nhãn

**1. Training Set là gì?**

Training Set là là một tập dữ liệu dùng để huấn luyện cho mô hình của thuật toán Machine Learning. Đây thường là một tập dữ liệu có kích thước lớn, được dùng để training trong quá trình huấn luyện máy học.

Trong mô hình Hồi quy tuyến tính (Linear Regression), các điểm trong tập huấn luyện được sử dụng để tìm ra hàm số hay đường phù hợp nhất mô tả quan hệ giữa đầu vào và đầu ra của tập dữ liệu huấn luyện bằng cách sử dụng một số phương pháp tối ưu hóa như công thức nghiệm hoặc các thuật toán tối ưu gần đúng như gradient descent hay stochastic gradient descent.

Training Set bao gồm input và output tương ứng, trong đó output thường được gọi là nhãn (label hoặc target). Các thuật toán nói chung sẽ tìm cách tối ưu sai số dự đoán trên tập huấn luyện này đến mức đủ tốt. Trong trường hợp overfitting sai số dự đoán của mô hình trên tập huấn luyện có thể rất thấp, thậm chí = 0%.

**2. Testing Set là gì?**

Nói một cách dễ hiểu, testing set là tập dữ liệu dùng để test sau khi máy tính đã học xong. Một mô hình máy học sau khi được huấn luyện, sẽ cần phải được kiểm chứng xem nó có đạt hiệu quả không. Mô hình máy học cũng vậy, sau mỗi quá trình huấn luyện, các mô hình này sẽ được kiểm chứng độ chính xác, nếu đáp ứng được yêu cầu thì là đạt, không thì cần phải xem xét lại. Và để kiểm nghiệm được độ chính xác của mô hình này, người ta dùng tập Testing set. Khác với Training set, Testing set chỉ gồm các giá trị input (TD, GC, YKNTK, TI, và RRCN) mà không có giá trị output (YDM).

=> Testing set là tập các giá trị input và được dùng để kiểm thử độ chính xác của những mô hình máy học sau khi được huấn luyện.

**3. Validation Set là gì?**

Trước khi nắm được Validation Set bạn cần tìm hiểu khái niệm Overfitting. Giống như training set, validation set cũng bao gồm các cặp giá trị input và output tương ứng. Tuy nhiên, input & output không dùng để huấn luyện trong quá trình máy học mà sử dụng để kiểm tra độ chính xác của mô hình máy học trong quá trình huấn luyện. Vậy sự khác nhau giữa công dụng Testing Set và Validation Set là gì ?  Testing được dùng để kiểm thử sau quá trình huấn luyện, còn validation set được sử dụng để kiểm thử trong quá trình huấn luyện. Chính vì vậy, thuật ngữ overfitting (hiện tượng mô hình dự đoán quá khớp với tập training set, dẫn đến dự đoán không hiệu quả đối với tập testing set.) cần phải nắm rõ trong quá trình sử dụng. Thông thường, người ta ngầm cho rằng Validation set mà có độ chính xác cao thì tập Testing set cũng có độ chính xác cao. Do đó, quá trình training chủ yếu dựa trên kết quả của tập Validation set chứ không phải tập training set.

=> Validation set là tập các giá trị input đi kèm với giá trị output và được dùng để kiểm thử độc chính xác của mô hình máy học trong quá trình huấn luyện.