Chương 1

Giới thiệu

Cuốn sách này giúp giải thích cách hoạt động của các mô hình học máy được huấn luyện theo phương pháp học có giám sát (supervised learning). Mặc dù cuốn sách có bao gồm nhiều công thức toán học, nhưng tôi tin rằng việc hiểu được ý tưởng chung đằng sau các phương pháp sẽ giúp bạn dễ dàng tiếp cận cuốn sách này hơn là quá quan tâm chi tiết đến các công thức đó. Một lưu ý nữa cuốn sách này không dành cho các bạn mới học Machine Learning hoặc chưa có kinh nghiệm làm việc thực tế với Machine Learning, vì thế ta có thể coi cuốn sách này như là Machine Learning nâng cao. Từ đó, nếu thấy cần thiết, hãy bắt đầu với những khóa học trực tuyến giới thiệu về Machine Learning của Andrew Ng hoặc tìm đọc cuốn Machine Learning cơ bản (Vũ Hữu Tiệp).

Ngày càng nhiều phương pháp mới được tìm ra nhằm giải thích tính nhân quả của các mô hình Machine Learning. Do vậy, việc bao hàm tất cả các phương pháp trong một cuốn sách là không thể, tuy nhiên ta sẽ tìm hiểu những kiến thức cơ bản nhất về học máy khả diễn giải và các phương pháp nổi tiếng trong tài liệu này.

Cuốn sách bắt đầu bằng định nghĩa về tính khả diễn giải (interpretability), ta sẽ cùng thảo luận tại sao ta cần quan tâm đến tính khả diễn giải trong Machine Learning. Hầu hết các mô hình và phương pháp trong cuốn sách này sử dụng các dữ liệu trong chương 3. Để giải thích một mô hình, cách đơn giản nhất là bản thân chính mô hình đó là khả diễn giải, ví dụ như những dạng mô hình đơn giản như hồi quy tuyến tính hay cây quyết định. Bên cạnh đó, một cách tiếp cận khác là ta sử dụng những công cụ phân tích kiểu mẫu (model-agnostic interpretation tools), nghĩa là công cụ của chúng ta có thể

được áp dụng trên bất kì mô hình Machine Learning nào để phân tích tính nhân quả. Ví dụ như một công cụ có thể vừa giải thích được cây quyết định, mạng neuron, hay hồi quy tuyến tính. Cuối cuốn sách, ta sẽ nhìn lại và cùng nhau dự đoán tương lai của học máy khả diễn giải.

1.1 Machine Learning là gì?

Machine learning (hay học máy) là một tập hợp các phương pháp mà máy tính dùng để tạo ra và cải thiện các dự đoán (prediction) dựa trên dữ liệu. Ví dụ, khi ta dự đoán giá của một căn nhà, máy tính sẽ học các đặc tính (patterns) của các căn nhà đã được bán thành công trước đó.

Tài liệu này tập trung vào học máy có giám sát (supervised Machine Learning), là khi ta đã có dữ liệu đầu vào (các thông tin về ngôi nhà như diện tích, vị trí, ...) và các kết quả mong muốn cho trước (giá bán), việc huấn luyện (training) mô hình sẽ giúp ta có thể dự đoán giá của các ngôi nhà trong tương lai. Mục đích cuối của việc học có giám sát là ta có thể ánh xạ (map) một đầu vào với các đặc tính (diện tích ngôi nhà, vị trí, vân vân) tới một đầu ra (giá bán). Nếu đầu ra của ta là các là dữ liệu kiểu hạng mục (categorical data), bài toán (task) sẽ là phân loại (classification), ví dụ như 1 bài toán phân loại chó, mèo, người từ một tập ảnh đầu vào. Nếu giá trị đầu ra là các giá trị số, bài toán được đưa về hồi quy, ví dụ như dự đoán giá nhà, giá cổ phiếu, ... Các mô hình Machine Learning học bằng việc ước lượng và thay đổi các trọng số (weights) trong mạng neuron hoặc kiến trúc (structures) trong các mô hình dạng cây (trees). Việc cập nhật các mô hình này tương ứng với việc tối ưu hóa (minize) một hàm mục tiêu (objective function) được định nghĩa từ ban đầu. Trong ví dụ dự đoán giá nhà, hàm mục tiêu là chênh lệch giữa giá nhà thực tế và giá trị mà mô hình trả về. Sau quá trình huấn luyện, mô hình được dùng để dự đoán giá của các căn nhà trong tương lai.

Dự đoán giá nhà, gợi ý sản phẩm, phát hiện biển báo, hay chấm điểm tín dụng đều có thể được giải quyết bằng Machine Learning. Mặc dù các bài toán là khác nhau, nhưng hướng tiếp cận chung sau đây sẽ được áp dụng:

1. Thu thập dữ liệu nhiều nhất có thể.
2. Huấn luyện mô hình với dữ liệu thu được.
3. Sử dụng mô hình đã được huấn luyện trên dữ liệu mới.

Trong những tác vụ như chơi cờ (Go Chess) hay dự báo thời tiết, Machine Learning đang và đã thể hiện sự vượt trội so với con người. Thêm vào đó, Machine Learning còn có thể thực hiện công việc một cách nhanh chóng và được triển khai rộng rãi. Một ví dụ rõ ràng đó là các camera giám sát thông minh (intelligent surveillance camera) ở Trung Quốc đều có tích hợp công nghệ trí tuệ nhân tạo để có thể giám sát người dân. Trong khi đó, việc đào tạo cho con người thực hiện thuần thục một công việc cụ thể có thể mất vài tuần, thậm chí là vài năm.

Tuy nhiên, một vấn đề lớn của việc sử dụng Machine Learning đó là ta không thể hiểu được các mô hình học từ dữ liệu như thế nào. Việc xây dựng một hàm ánh xạ giữa đầu vào và đầu ra của một mô hình được thực hiện trong quá trình huấn luyện, đầu ra cuối cùng có sự đóng góp của tất cả các phần tử trong mạng. Do đó, việc diễn giải cơ chế hoạt động của một mô hình bằng công thức toán học gần như là điều không thể trong hầu hết các trường hợp.

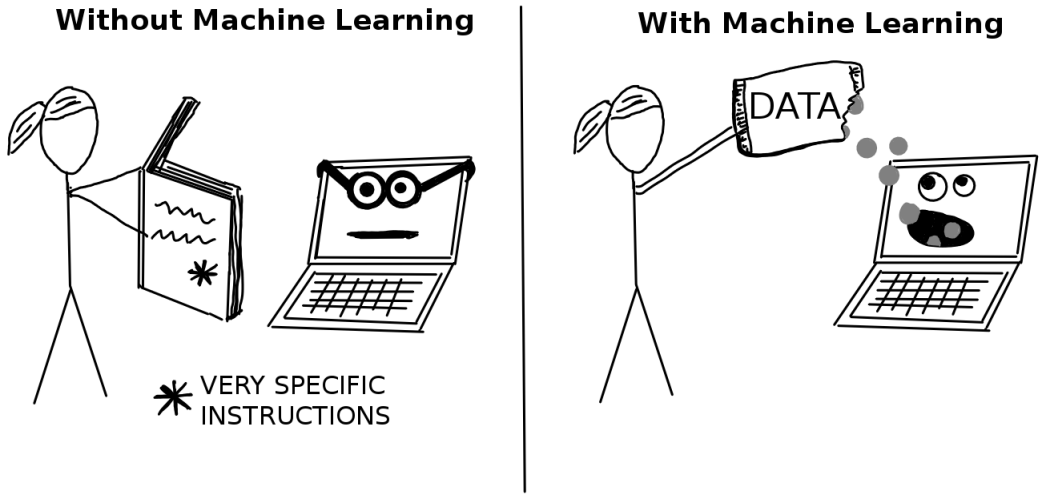
Hiện nay, các mô hình giành chiến thắng trong các cuộc thi thường là các mô hình hợp (ensemble models), dẫn đến việc cấu trúc của chúng trở nên cực kỳ phức tạp. Thậm chí nếu từng mô hình con trong mô hình hợp là khả diễn giải, việc giải thích vẫn sẽ rất khó khăn.

1.2 Thuật ngữ

Để tránh nhầm lẫn trong quá trình học cuốn sách này, một số thuật ngữ cơ bản sẽ được định nghĩa như sau.

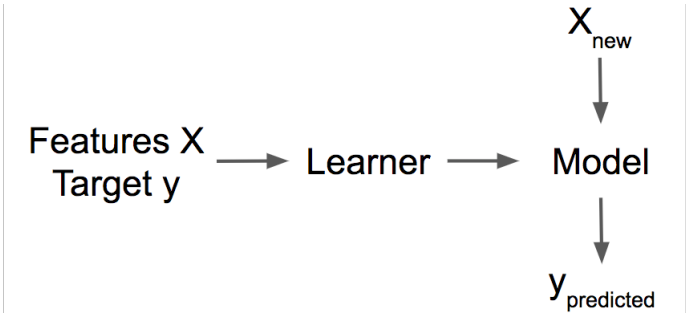
Thuật toán là một tập các quy tắc mà máy tính tuân theo để đạt được một mục tiêu nhất định. Một thuật toán có thể được định nghĩa bởi đầu vào, đầu ra và các bước cần thiết để biến đổi đầu vào thành đầu ra.

Machine Learning hay học máy là một tập hợp các phương pháp cho phép máy tính học từ dữ liệu cho trước để tạo ra dự đoán sau này (ví dụ như dự đoán bệnh ung thư, giá bán hàng, giá cổ phiếu, ...). Học máy có thể được coi là một sự chuyển đổi của công việc lập trình khi mà thay vì con người phải chỉ rõ các lệnh cần thực hiện cho máy tính thì ta sẽ để máy tính tự học từ dữ liệu.



Một bộ học (learner) hay một thuật toán Machine Learning là chương trình được dùng để học một mô hình từ dữ liệu.

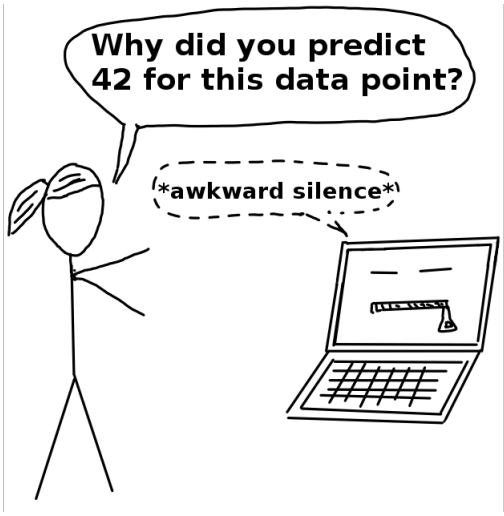
Mô hình Machine Learning là một chương trình máy tính thực thi tác vụ ánh xạ dữ liệu đầu vào và đầu ra. Mô hình này có thể là một tập hợp của các trọng số của mô hình tuyến tính hay một mạng neuron. Ngoài ra, trong tùy từng nhiệm vụ cụ thể, một mô hình còn có thể được gọi như một bộ dự đoán (predictor), bộ phân loại (classifier), hay mô hình hồi quy (regression model). Trong các công thức tính toán, một mô hình học máy sẽ được kí hiệu là $\hat{f}$ hoặc $\hat{f}(x)$.



Hình 1.1: Một bộ học xây dựng mô hình từ các dữ liệu huấn luyện được gắn nhãn. Mô hình sau đó sẽ được sử dụng trong việc dự đoán

Mô hình hộp đen (Black box model) là một hệ thống mà ta không thể biết được cơ chế hoạt động khi quan sát từ bên ngoài. Hộp đen trong Machine Learning có thể được hiểu là các mô hình bao gồm các trọng số mà với việc phân tích các trọng số này ta không thể hiểu được cơ chế làm việc

của chúng. Trái lại, ta có mô hình hộp trắng (White box model), ám chỉ các mô hình khả diễn giải.



Học máy khả diễn giải là lĩnh vực liên quan tới các mô hình khả diễn giải và phương pháp nhằm giải thích cơ chế hoạt động các mô hình Machine Learning cho người dùng.

Tập dữ liệu là một bảng dữ liệu mà từ đó máy móc có thể học được. Tập dữ liệu bao gồm các đặc trưng và mục tiêu cho việc dự đoán. Khi sử dụng để tạo ra một mô hình, tập dữ liệu được gọi là dữ liệu huấn luyện.

Thể hiện hay một điểm dữ liệu hay một quan sát là một hàng trong tập dữ liệu. Một thể hiện bao gồm các đặc trưng $x^{(i)}$ và đầu ra mong muốn (nếu có) y_i.

Đặc trưng là các đầu vào sử dụng cho việc dự đoán hoặc phân loại. Một đặc trưng là một cột trong tập dữ liệu. Trong tài liệu này, các đặc trưng được giả thiết là khả diễn giải, nghĩa là bản thân chúng đơn giản và dễ hiểu, ví dụ như nhiệt độ trong ngày hay chiều cao của một người. Sự khả diễn giải của các đặc trưng là một giả thiết quan trọng, vì nếu đặc trưng khó hiểu, việc hiểu mô hình được xây dựng từ đặc trưng đó sẽ thậm chí còn khó khăn hơn. Ma trận của tất cả các đặc trưng được gọi là X và $x^{(i)}$ đại diện cho một điểm dữ liệu. Vector của một đặc trưng cho tất cả các điểm dữ liệu là x_j và giá trị đặc trưng thứ j của thể hiện thứ i là $x_j^{(i)}$.

Mục tiêu là thông tin mà mô hình muốn dự đoán. Trong các công thức toán học, mục tiêu được ký hiệu là y hoặc y_i cho thể hiện thứ i.

Một bài toán học máy là tổ hợp của một tập dữ liệu với các đặc trưng và mục tiêu. Tùy thuộc vào kiểu mục tiêu, bài toán có thể là phân loại, hồi quy, phân tích, gộp nhóm, hoặc phát hiện ngoại lai.

Dự đoán là giá trị đầu ra mà mô hình học máy "đoán" dựa trên các đặc trưng được cho. Trong cuốn sách này, dự đoán của mô hình được ký hiệu là $\hat{f}(x^{(i)})$ hoặc $\hat{y}$.