Cuốn sách "Machine Learning Cơ Bản – Học Máy cho Người Mới Bắt Đầu" nhằm mục tiêu giúp người đọc nắm vững nền tảng của học máy (machine learning) từ cơ bản, bao gồm các thuật toán cốt lõi, quy trình xây dựng mô hình, và cách ứng dụng vào thực tế. Sách không yêu cầu kiến thức chuyên sâu trước về trí tuệ nhân tạo (AI) hay toán học phức tạp.

**Chương 1: Giới thiệu về Machine Learning**

Machine Learning (Học máy) là phương pháp giúp máy tính có khả năng học từ dữ liệu và đưa ra quyết định mà không cần được lập trình chi tiết từng bước. Trong học máy, thay vì viết code hướng dẫn từng tình huống, ta cung cấp cho máy dữ liệu và để nó tự học ra quy luật.

Ví dụ đơn giản: nếu bạn muốn máy phân biệt giữa ảnh chó và mèo, bạn không cần chỉ rõ tai, đuôi hay hình dạng – bạn chỉ cần cung cấp nhiều ảnh đã được gắn nhãn (mèo hoặc chó), thuật toán học máy sẽ tự tìm ra đặc điểm giúp phân biệt.

Machine Learning thường được chia thành ba nhóm chính:

* **Học có giám sát (Supervised Learning)**: học từ dữ liệu có nhãn.
* **Học không giám sát (Unsupervised Learning)**: tìm quy luật từ dữ liệu chưa gắn nhãn.
* **Học tăng cường (Reinforcement Learning)**: học qua hành động và phản hồi (thưởng/phạt).

Các ứng dụng của học máy rất phổ biến: từ lọc email spam, nhận diện giọng nói, gợi ý phim, đến chẩn đoán y tế và xe tự lái.

Trong các chương tiếp theo, chúng ta sẽ cùng tìm hiểu các mô hình học máy cơ bản như hồi quy tuyến tính, cây quyết định, KNN và mạng nơ-ron đơn giản – từ lý thuyết đến ví dụ thực hành bằng Python.

**Chương 2: Hồi Quy Tuyến Tính (Linear Regression)**

Hồi quy tuyến tính là một trong những thuật toán cơ bản nhất trong học máy, dùng để **dự đoán giá trị liên tục** dựa trên dữ liệu đầu vào. Ví dụ: dự đoán giá nhà dựa trên diện tích, số phòng ngủ,…

Giả sử bạn có một tập dữ liệu gồm diện tích nhà và giá tương ứng. Mục tiêu của hồi quy tuyến tính là tìm đường thẳng tốt nhất (gọi là “đường hồi quy”) sao cho **khoảng cách giữa dự đoán và thực tế là nhỏ nhất**.

Công thức cơ bản:

ini

Sao chépChỉnh sửa

y = w \* x + b

Trong đó:

* x là đầu vào (diện tích nhà),
* y là đầu ra (giá nhà),
* w là hệ số góc (trọng số),
* b là hệ số chệch (bias).

Thuật toán sẽ tìm ra w và b tối ưu bằng cách **giảm thiểu sai số** thông qua kỹ thuật như **Gradient Descent** – một phương pháp tối ưu hóa giúp tìm điểm cực tiểu của hàm mất mát (loss function).

Hồi quy tuyến tính tuy đơn giản nhưng mang tính nền tảng, thường được dùng để kiểm tra mô hình trước khi dùng các thuật toán phức tạp hơn.

**Chương 3: Hồi Quy Logistic (Logistic Regression)**

Hồi quy logistic không còn dùng để dự đoán giá trị liên tục, mà là để **phân loại** – ví dụ: phân loại email là spam hoặc không spam.

Mặc dù tên là "hồi quy", đây là một thuật toán **phân loại nhị phân**, đầu ra chỉ là 0 hoặc 1 (hoặc xác suất nằm giữa 0 và 1). Mô hình dùng một hàm gọi là **sigmoid** để biến đổi đầu ra thành giá trị xác suất:

Sao chépChỉnh sửa

sigmoid(z) = 1 / (1 + e^(-z))

Trong đó, z = w \* x + b. Sau khi tính được xác suất, ta có thể đưa ra quyết định:

* Nếu xác suất > 0.5 → phân loại là 1 (ví dụ: là spam),
* Nếu ≤ 0.5 → phân loại là 0 (không phải spam).

Hồi quy logistic được dùng nhiều trong các bài toán thực tế như:

* Chẩn đoán bệnh (có bệnh/không có bệnh),
* Nhận diện khách hàng có rời bỏ dịch vụ không (churn),
* Phân loại văn bản, ảnh,.