**Tóm Tắt Kiến Thức Đã Học**

Chương 1 – Giới thiệu về Machine Learning với Python

* + Machine Learning với Python nên tập trung vào thực hành, tránh học quá sa đà lý thuyết.
  + Mục tiêu: Xây dựng mô hình dự đoán ứng dụng được trong thực tế.
  + Các bước trong một dự án Machine Learning:
* Xác định bài toán
* Phân tích dữ liệu
* Tiền xử lý dữ liệu
* Đánh giá thuật toán
* Tối ưu kết quả
* Trình bày kết quả
  + Cuốn sách đi theo hướng "học qua dự án" thay vì lý thuyết nặng.

Chương 2 – Hệ Sinh Thái Python cho Machine Learning

* + Python là ngôn ngữ đa mục đích, dễ đọc, dễ học, lý tưởng cho Machine Learning.
  + SciPy ecosystem bao gồm:
    - NumPy: Làm việc với mảng dữ liệu.
    - Matplotlib: Trực quan hóa dữ liệu.
    - Pandas: Tổ chức và phân tích dữ liệu.
  + scikit-learn là thư viện Machine Learning chính, hỗ trợ nhiều thuật toán và công cụ.
  + Cách cài đặt môi trường:
    - Cài Python (nên dùng Python 3.8+).
    - Cài đặt SciPy stack: scipy, numpy, matplotlib, pandas.
    - Cài scikit-learn.
    - Hoặc dùng Anaconda để cài đặt nhanh tất cả.
  + Lợi ích: Có thể làm R&D và triển khai production trên cùng một nền tảng Python.
* Chương 3 – Khóa Học Cấp Tốc Python và SciPy
  + Ôn lại các cấu trúc dữ liệu cơ bản trong Python: Tuple, List, Dictionary, Hàm.
  + Hiểu rõ cú pháp điều kiện (if-else), vòng lặp for và while.
  + Làm việc với NumPy:
    - Tạo mảng, truy cập phần tử, thực hiện phép toán vector.
  + Dùng Matplotlib để trực quan hoá dữ liệu với biểu đồ đường và biểu đồ điểm.
  + Dùng Pandas để tạo Series và DataFrame:
    - Truy xuất dữ liệu bằng chỉ số, tên cột.
    - In dữ liệu và hiểu cách tổ chức dạng bảng.
* Chương 4 – Cách Nạp Dữ Liệu Machine Learning
  + Hiểu định dạng CSV là phổ biến trong Machine Learning.
  + Biết các vấn đề thường gặp khi load CSV: header, comment, delimiter, quote.
  + Làm quen với tập dữ liệu Pima Indians – dự đoán tiểu đường.
  + Ba cách nạp dữ liệu CSV:
* Sử dụng thư viện chuẩn csv → cần xử lý thô, đổi sang NumPy.
* Sử dụng NumPy → dùng loadtxt, nhanh, gọn.
* Sử dụng Pandas → linh hoạt, dễ gán tên cột, dễ khám phá.
  + Kết luận: nên dùng Pandas cho hầu hết dự án thực tế.

Chương 5 – Hiểu Dữ Liệu Qua Thống Kê Mô Tả

• Học cách mô tả và khám phá dữ liệu trước khi huấn luyện Machine Learning.

• 7 kỹ thuật mô tả dữ liệu chính:

1. Peek At Your Data:

- Dùng head() để xem 20 dòng đầu tiên.

- Phát hiện sớm giá trị thiếu hoặc bất thường.

2. Dimensions of Your Data:

- Dùng shape để kiểm tra số lượng dòng và cột.

- Đánh giá mức độ đủ hay thiếu dữ liệu.

3. Data Types:

- Dùng dtypes để kiểm tra kiểu dữ liệu từng cột.

- Xác định các cột cần chuyển đổi kiểu.

4. Descriptive Statistics:

- Dùng describe() để xem thống kê: trung bình, min, max, phần tư.

- Giúp hiểu được phân bố và đặc tính tổng quan của dữ liệu.

5. Class Distribution:

- Dùng groupby().size() để kiểm tra số lượng mẫu ở mỗi lớp.

- Phát hiện dữ liệu mất cân bằng.

6. Correlations Between Attributes:

- Dùng corr() để tính ma trận tương quan Pearson.

- Phát hiện biến tương quan cao có thể ảnh hưởng đến mô hình.

7. Skew of Univariate Distributions:

- Dùng skew() để xem độ lệch phân phối.

- Phát hiện cần chuẩn hóa hoặc biến đổi dữ liệu.

Lời khuyên:

• Không chỉ tạo thống kê mà phải đọc kỹ và suy nghĩ về ý nghĩa của số liệu.

• Ghi chú lại các phát hiện để hỗ trợ cho quá trình xử lý và chọn mô hình sau này.

**Chương 6 – Hiểu Dữ Liệu Qua Trực Quan Hóa**

• Học cách sử dụng trực quan hóa để hiểu dữ liệu Machine Learning nhanh và sâu hơn.

• 5 kỹ thuật trực quan hóa chính đã học:

1. Histograms:

- Biểu đồ tần suất chia dữ liệu thành các nhóm (bins).

- Phát hiện phân phối chuẩn, lệch, hoặc phân phối bất thường.

- Phát hiện giá trị ngoại lai.

2. Density Plots:

- Biểu đồ mật độ làm mượt histogram thành đường cong liên tục.

- Giúp phân tích dạng phân phối rõ ràng hơn.

3. Box and Whisker Plots:

- Biểu đồ hộp thể hiện trung vị, phần tư, và độ lan rộng của dữ liệu.

- Dễ dàng phát hiện outliers và độ lệch.

4. Correlation Matrix Plot:

- Ma trận hiển thị mức độ tương quan giữa các biến.

- Giúp nhận biết biến nào liên quan chặt chẽ với nhau.

5. Scatter Plot Matrix:

- Ma trận các biểu đồ phân tán giữa từng cặp biến.

- Tìm kiếm mối quan hệ tuyến tính hoặc phi tuyến giữa các biến.

**Lời khuyên:**

• Không chỉ tạo biểu đồ mà cần đọc và phân tích ý nghĩa mỗi hình.

• Ghi chú lại các hiện tượng quan sát được như outliers, phân phối lệch, tương quan cao.

• Trực quan hóa dữ liệu là bước cực kỳ quan trọng trước khi xử lý dữ liệu và huấn luyện mô hình.