# **Artificial Intelligence Tutorial for Beginners**

## 

## **21.Linear Regression Demo**

Điều chúng ta sẽ làm ở đây là xây dựng một mối quan hệ tuyến tính giữa nhiệt độ tối đa và nhiệt độ tối thiểu vào một ngày cụ thể. Chúng ta sẽ dự báo thời tiết. Nhiệm vụ của chúng ta là dự đoán nhiệt độ tối đa, lấy nhiệt độ tối thiểu làm đặc trưng đầu vào. Tôi sẽ cố gắng giúp bạn hiểu hồi quy tuyến tính qua ví dụ này. Hãy xem cách nó hoạt động thực tế.

Trước khi bắt đầu, tôi muốn giới thiệu một chút về tập dữ liệu. Tập dữ liệu của chúng ta được lưu trong đường dẫn này và có tên là weather.csv. Tập dữ liệu này chứa các dữ liệu về điều kiện thời tiết được ghi nhận hàng ngày tại các trạm khí tượng khác nhau trên thế giới. Thông tin bao gồm lượng mưa, tuyết, nhiệt độ, tốc độ gió, và các điều kiện thời tiết như bão hoặc các điều kiện thời tiết xấu khác. Vì vậy, bước đầu tiên trong bất kỳ ví dụ nào là nhập tất cả các thư viện cần thiết.

Chúng ta sẽ bắt đầu ví dụ bằng cách nhập tất cả các thư viện cần thiết. Sau đó, chúng ta sẽ đọc dữ liệu vào biến dataset, sử dụng hàm read.csv vì tập dữ liệu của chúng ta có định dạng CSV. Tiếp theo, tôi sẽ hiển thị cách tập dữ liệu trông như thế nào và xem xét kỹ hơn về tập dữ liệu. Đầu tiên, chúng ta sẽ chạy ví dụ và xem kết quả. Chúng ta nhận được một số biểu đồ mà tôi sẽ giải thích sau.

Đầu tiên, chúng ta sẽ in kích thước của tập dữ liệu, và kết quả là khoảng 366 hàng và 22 cột. 22 cột này đại diện cho các biến dự đoán, có nghĩa là chúng ta có 22 biến dự đoán để dự báo các điều kiện thời tiết vào một ngày cụ thể. Mục tiêu chính là dự báo thời tiết bằng cách sử dụng một tập hợp các biến dự đoán.

Giờ đây, tôi sẽ vẽ các điểm dữ liệu trên biểu đồ 2D để hiểu dữ liệu và xem liệu có thể tìm thấy mối quan hệ giữa các biến hay không. Ở đây chúng ta sẽ sử dụng nhiệt độ tối thiểu và nhiệt độ tối đa để phân tích. Khi bạn nhìn vào biểu đồ này, nó hiển thị mối quan hệ giữa nhiệt độ tối thiểu và tối đa. Nhiệt độ tối đa là biến phụ thuộc mà chúng ta sẽ dự đoán (đây là biến y), còn nhiệt độ tối thiểu là biến độc lập (biến x).

Biểu đồ cho thấy một mối quan hệ tuyến tính giữa hai biến này, mặc dù có một số điểm ngoại lệ. Dữ liệu của chúng ta rất tuyến tính nên có thể sử dụng hồi quy tuyến tính để giải quyết vấn đề này.

Tiếp theo, tôi sẽ kiểm tra nhiệt độ trung bình và tối đa, đây là một phần của phân tích dữ liệu để hiểu rõ hơn về dữ liệu. Nhiệt độ tối đa trung bình nằm trong khoảng từ 28 đến 32 độ.

Sau khi hiểu rõ hơn về tập dữ liệu và mối quan hệ tuyến tính giữa biến đầu vào và biến đầu ra, chúng ta sẽ tiến hành chia dữ liệu. Đây là quá trình chia tập dữ liệu thành tập huấn luyện và kiểm thử, vì chúng ta chỉ sử dụng hai biến (nhiệt độ tối thiểu và nhiệt độ tối đa) để làm đơn giản. Chúng ta sẽ lấy 80% dữ liệu cho tập huấn luyện và 20% còn lại cho tập kiểm thử để mô hình có thể huấn luyện tốt hơn.

Giờ đến lúc huấn luyện mô hình. Chúng ta sẽ nhập lớp hồi quy tuyến tính (LinearRegression) từ thư viện sklearn, tạo một đối tượng của lớp này, sau đó sử dụng phương thức fit cùng với dữ liệu huấn luyện. Phương pháp này tìm giá trị tối ưu cho hằng số intercept và slope, điều này giúp mô hình hồi quy tuyến tính tìm ra đường thẳng tốt nhất cho dữ liệu.

Hệ số chặn intercept của chúng ta là khoảng 14.56 và hệ số coefficient là 0.82. Điều này có nghĩa là với mỗi thay đổi một đơn vị của nhiệt độ tối thiểu, nhiệt độ tối đa sẽ thay đổi khoảng 0.82 đơn vị.

Sau khi huấn luyện mô hình, chúng ta sẽ thực hiện dự đoán trên tập kiểm thử và so sánh kết quả dự đoán với giá trị thực tế. Kết quả khá sát nhau. Có một số trường hợp giá trị dự đoán khác biệt nhiều, điều này có thể cải thiện bằng cách huấn luyện mô hình với nhiều dữ liệu hơn.

Chúng ta cũng có thể xem kết quả này dưới dạng biểu đồ đường, cho thấy sự so sánh giữa giá trị thực tế và giá trị dự đoán. Một số dự đoán khá gần với giá trị thực tế, trong khi một số khác có sự khác biệt nhất định. Các ngoại lệ trong dữ liệu đầu vào cũng có thể ảnh hưởng đến kết quả đầu ra.

Cuối cùng, chúng ta sẽ vẽ một đường thẳng biểu diễn mối quan hệ tuyến tính giữa biến đầu vào và biến đầu ra. Ví dụ, nếu có một điểm dữ liệu mới với giá trị x = 20, bạn có thể sử dụng đường thẳng này để dự đoán rằng với nhiệt độ tối thiểu là 20, nhiệt độ tối đa sẽ khoảng 25.

Bước cuối cùng là đánh giá hiệu suất của mô hình. Với các mô hình hồi quy, ba chỉ số đánh giá thường được sử dụng là mean absolute error, mean squared error, và root mean square error. Những lỗi này giúp hiểu rõ hơn về sai số của mô hình.

Thư viện scikit-learn đã tích hợp sẵn các hàm để tính toán những giá trị này. Kết quả cho thấy mean absolute error là khoảng 3.5, mean squared error là 17.01, và root mean squared error là 4.12. Những giá trị lỗi này cho thấy độ chính xác của mô hình chưa hoàn toàn chính xác nhưng vẫn đưa ra nhiều dự đoán tốt và một mối quan hệ tuyến tính ổn định.

Có nhiều cách để cải thiện hiệu suất, chẳng hạn như điều chỉnh tham số hoặc sử dụng nhiều dữ liệu hơn để huấn luyện mô hình. Tuy nhiên, mục tiêu ở đây là giúp bạn hiểu cách hoạt động của hồi quy tuyến tính.

Hy vọng mọi người đã nắm rõ hơn về hồi quy tuyến tính thông qua ví dụ này. Nếu còn bất kỳ thắc mắc nào về hồi quy tuyến tính, vui lòng để lại ở phần bình luận, chúng tôi sẽ cố gắng giải đáp.