**1. Ý tưởng chính của Linear Regression**

Linear Regression nhằm tìm ra mối quan hệ tuyến tính giữa một hoặc nhiều biến độc lập (**Independent Variables** - Features) và biến phụ thuộc (**Dependent Variable** - Target). Mô hình giả định rằng sự thay đổi trong biến mục tiêu có thể được giải thích bởi sự thay đổi tuyến tính trong các biến đầu vào.

**Linear Regression đơn biến (Simple Linear Regression)**

Đối với một biến độc lập, phương trình của Linear Regression có dạng:

* : Biến mục tiêu cần dự đoán (Target).
* : Biến độc lập (Feature).
* ​: Hệ số chặn (Intercept), là giá trị khi .
* : Hệ số hồi quy (Slope), đại diện cho độ thay đổi của khi thay đổi 1 đơn vị.
* : Phần dư (Residual), thể hiện sai số giữa giá trị dự đoán và giá trị thực tế.

**Linear Regression đa biến (Multiple Linear Regression)**

Khi có nhiều biến độc lập, phương trình tổng quát của mô hình là:

Trong đó:

* ​: Các biến độc lập.
* ​: Các hệ số tương ứng với từng biến độc lập.

Ví dụ: Nếu bạn muốn dự đoán giá nhà dựa trên diện tích, số phòng ngủ, và khoảng cách tới trung tâm thành phố, bạn có thể sử dụng mô hình hồi quy tuyến tính với 3 biến độc lập.

**2. Quá trình huấn luyện mô hình**

Linear Regression tìm cách tối ưu các tham số ​ sao cho tổng sai số giữa các giá trị dự đoán và giá trị thực tế là nhỏ nhất. Điều này thường được thực hiện bằng cách sử dụng **Ordinary Least Squares (OLS)**, tức là tối thiểu hóa hàm bình phương sai số (Mean Squared Error - MSE):

* ​: Giá trị thực tế của mẫu thứ i.
* ​: Giá trị dự đoán của mẫu thứ i.
* nnn: Số lượng mẫu dữ liệu.

Mục tiêu là tìm các giá trị ​ sao cho MSE nhỏ nhất, hay nói cách khác, các hệ số này sẽ tạo ra đường hồi quy tối ưu nhất để dự đoán.

**3. Các giả định trong Linear Regression**

Linear Regression đưa ra một số giả định về dữ liệu, bao gồm:

1. **Tính tuyến tính (Linearity)**: Mối quan hệ giữa biến độc lập và biến phụ thuộc là tuyến tính.
2. **Độc lập của phần dư (Independence of Residuals)**: Các giá trị phần dư phải độc lập với nhau.
3. **Phân phối chuẩn của phần dư (Normality of Residuals)**: Phần dư cần tuân theo phân phối chuẩn, nghĩa là mô hình dự đoán chính xác cho đa số các điểm dữ liệu.
4. **Homoscedasticity**: Phần dư phải có phương sai đồng nhất, nghĩa là sai số không phụ thuộc vào giá trị dự đoán của mô hình.
5. **Không có đa cộng tuyến (No Multicollinearity)**: Các biến độc lập không nên có mối tương quan mạnh với nhau, nếu không sẽ gây ra vấn đề đa cộng tuyến (Multicollinearity), làm cho mô hình kém chính xác.

**4. Cách giải quyết đa cộng tuyến**

Nếu phát hiện có sự đa cộng tuyến giữa các biến độc lập, bạn có thể sử dụng một số biện pháp sau:

* **Bỏ bớt biến**: Loại bỏ các biến có tương quan mạnh với nhau.
* **Sử dụng Ridge Regression hoặc Lasso Regression**: Đây là các biến thể của Linear Regression, thêm vào một số điều chỉnh để giảm ảnh hưởng của đa cộng tuyến.

**5. Ưu và nhược điểm của Linear Regression**

**Ưu điểm:**

* **Dễ hiểu và dễ triển khai**: Đây là mô hình đơn giản, dễ giải thích và hiểu mối quan hệ giữa các biến.
* **Hiệu quả với dữ liệu nhỏ và tuyến tính**: Linear Regression hoạt động tốt với dữ liệu có quy mô nhỏ và có mối quan hệ tuyến tính.
* **Tính toán nhanh**: So với các mô hình phức tạp khác, Linear Regression có chi phí tính toán thấp.

**Nhược điểm:**

* **Giả định tuyến tính**: Linear Regression không hoạt động tốt khi mối quan hệ giữa biến độc lập và biến phụ thuộc không phải là tuyến tính.
* **Nhạy cảm với outliers**: Mô hình có thể bị ảnh hưởng nghiêm trọng bởi các ngoại lệ (outliers).
* **Đa cộng tuyến**: Khi các biến độc lập có sự tương quan mạnh, mô hình có thể trở nên không ổn định.
* **Không mô hình hóa được quan hệ phi tuyến**: Linear Regression không có khả năng mô hình hóa các mối quan hệ phức tạp, phi tuyến tính.

**6. Ví dụ cụ thể**

Giả sử bạn có một dataset về giá nhà và muốn xây dựng mô hình dự đoán giá nhà dựa trên diện tích và số phòng ngủ:

| **Diện tích (m²)** | **Số phòng ngủ** | **Giá nhà (triệu đồng)** |
| --- | --- | --- |
| 50 | 2 | 800 |
| 70 | 3 | 1200 |
| 90 | 3 | 1600 |
| 120 | 4 | 2000 |

* Biến độc lập: Diện tích, Số phòng ngủ.
* Biến phụ thuộc: Giá nhà.

Mô hình Linear Regression sẽ cố gắng tìm các hệ số ​ để ước lượng giá nhà dựa trên diện tích và số phòng ngủ. Ví dụ, phương trình dự đoán có thể là:

Nếu căn nhà có diện tích 100 m² và 3 phòng ngủ, giá dự đoán sẽ là:

**7. Mở rộng**

Linear Regression là mô hình cơ bản, nhưng bạn có thể mở rộng nó thành các phiên bản mạnh hơn như:

* **Ridge Regression**: Thêm regularization L2 để chống overfitting.
* **Lasso Regression**: Thêm regularization L1, giúp loại bỏ các biến ít quan trọng.
* **Polynomial Regression**: Mô hình hóa mối quan hệ phi tuyến bằng cách thêm các bậc cao của biến độc lập vào mô hình.

**Kết luận:**

Linear Regression là mô hình hữu ích trong các bài toán dự đoán giá trị liên tục với mối quan hệ tuyến tính. Tuy nhiên, để sử dụng hiệu quả, cần kiểm tra kỹ các giả định của mô hình, và mở rộng hoặc điều chỉnh khi gặp phải các vấn đề phức tạp như đa cộng tuyến hay quan hệ phi tuyến.