

Regression

1. Khái niệm

Linear Regression Model: Là mô hình vẽ **một đường thẳng** khớp tốt nhất với dữ liệu huấn luyện.

Khác biệt giữa Regression và Classification:

- **Regression:** Dự đoán giá trị **liên tục** (số thực) như giá nhà.
- **Classification:** Dự đoán **nhãn phân loại** (như mèo hoặc chó).

Dữ liệu và Ký hiệu Chuẩn:

- **Training set:** Tập dữ liệu gồm các cặp (x, y) dùng để huấn luyện.
- **x:** Đầu vào (ví dụ: diện tích nhà).
- **y:** Đầu ra (ví dụ: giá nhà).
- **m:** Tổng số ví dụ huấn luyện (ở ví dụ này $m = 47$).
- $(x^{(i)}, y^{(i)})$: Ví dụ huấn luyện thứ i (không phải mũ, chỉ là chỉ số hàng).

Trong một mô hình Supervised Learning:

- Dữ liệu huấn luyện gồm **input features (x)** và **output targets (y)**.
- Thuật toán học từ tập huấn luyện để tạo ra **một hàm $f(x)$** (gọi là **model**).

Dự đoán:

- Với input mới **x**, mô hình **dự đoán** output gọi là **\hat{y}** (y-hat).
- **y:** Giá trị thực (true label).
- **\hat{y} :** Giá trị dự đoán (predicted label).
- **Trong mô hình Linear Regression:**
 - Hàm được chọn: **$f(x) = wx + b$** (hàm tuyến tính, đường thẳng).
 - **w** và **b** là các tham số mô hình (số thực).

Univariate Linear Regression: "Univariate" = chỉ có **một biến đầu vào** (ví dụ: diện tích nhà).

2. Cost function

Cost function được sử dụng để đo độ tốt của mô hình và tối ưu mô hình, ta có

- Mô hình dự đoán: $\hat{y}^{(i)} = f(x^{(i)}) = wx^{(i)} + by^{(i)} = f(x^{(i)}) = wx(i) + b$
- **w** và **b** gọi là **parameters** (tham số), đôi khi cũng gọi là **coefficients** hoặc **weights**.

Khi so sánh giá trị dự đoán $\hat{y}^{(i)}$ với giá trị thực $y^{(i)}$, ta có **Error** (lỗi) = $\hat{y}^{(i)} - y^{(i)}$.

- **Xây dựng Cost Function:**

- Tính **biên phương sai số** cho mỗi điểm: $(\hat{y}^{(i)} - y^{(i)})^2$.
- Tổng sai số bình phương trên toàn bộ tập dữ liệu:

$$\sum_{i=1}^m (\hat{y}^{(i)} - y^{(i)})^2.$$

- Lấy **trung bình** để không phụ thuộc vào số lượng mẫu:

$$\frac{1}{2m} \sum_{i=1}^m (\hat{y}^{(i)} - y^{(i)})^2.$$

- Theo quy ước, thêm chia 2 để đơn giản hóa đạo hàm sau này:

$$J(w, b) = \frac{1}{2m} \sum_{i=1}^m (\hat{y}^{(i)} - y^{(i)})^2$$

hoặc viết gọn:

$$J(w, b) = \frac{1}{2m} \sum_{i=1}^m (f(x^{(i)}) - y^{(i)})^2$$

- **Vai trò của J(w,b):**

- **J lớn** → Mô hình dự đoán kém.
- **J nhỏ** → Mô hình dự đoán tốt.

- **Kết luận:**

- Mục tiêu: **Tìm w và b sao cho J(w,b) nhỏ nhất.**

