Buổi 2

Huấn luyện mô hình:

- Có bộ dữ liệu có nhãn
- Xử lý dữ liệu
- Chia tập huấn luyện
- Huấn luyện mô hình (Hồi quy tuyến tính):
    - Dự đoán giá nhà: y^ = wX + b
    - Thực tế giá nhà là y
    - ⇨ Huấn luyện mô hình thực chất là tìm hệ số w và b phù hợp để y^ gần y nhất, hay chính là sai số giữa y^ và y là nhỏ nhất
    - ⇨ Lúc đấy, ta gọi w1, b1 là hệ số tối ưu
    - Sau khi huấn luyện, ta có y^:
        - Khi có một ngôi nhà, với các đặc trưng cho trước x (giống với X)
            - Ví dụ: ban đầu X có 3 đặc điểm về số phòng ngủ, diện tích nhà, số phòng ăn.
                - Khi đó, x truyền vào mô hình cũng phải có số phòng ngủ, diện tích nhà, số phòng ăn.
        - Nó sẽ dự đoán giá nhà, bằng công thức:
            - Giá nhà: y^ = w1x + b1
- Đánh giá mô hình

Đánh giá mô hình:

✅ 1. MAE – Mean Absolute Error

🔹 Công thức:

$$\text{MAE} = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i|$$

- y_i: Giá trị thực tế tại điểm thứ i
- $\hat{y}_i$: Giá trị dự đoán tại điểm thứ i
- n: Tổng số điểm dữ liệu

🔹 Ý nghĩa:

- MAE là **trung bình sai số tuyệt đối** giữa giá trị thực và dự đoán.
- Mỗi sai số được lấy trị tuyệt đối ⇒ không phân biệt sai trên hay sai dưới.
- **Đơn vị giống với đơn vị đầu ra** (ví dụ: nếu đầu ra là triệu đồng thì MAE cũng là triệu đồng).

📌 **Ưu điểm**: Dễ hiểu, không phạt quá nặng với các sai số lớn.
📌 **Nhược điểm**: Không nhạy với outliers.

✅ 2. MSE – Mean Squared Error

🔹 Công thức:

$$\text{MSE} = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$$

🔹 Ý nghĩa:

- Là **trung bình bình phương sai số**.
- Do sai số được bình phương, nên **các sai số lớn sẽ bị phóng đại**.
- Dùng phổ biến trong bài toán tối ưu (Gradient Descent) vì MSE là **hàm liên tục và khả vi**.

📌 **Ưu điểm**: Nhấn mạnh các lỗi lớn, tốt khi bạn muốn mô hình **không được phép mắc lỗi nghiêm trọng**.
📌 **Nhược điểm**: Đơn vị là **bình phương của đơn vị đầu ra**, hơi khó diễn giải.

✅ 3. R² Score – Coefficient of Determination

🔹 Công thức:

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}$$

- $\bar{y}$: Trung bình của toàn bộ giá trị thực tế

🔹 Ý nghĩa:

- R^2 đo **tỷ lệ phương sai** của dữ liệu thực được mô hình giải thích.
- R^2 nằm trong khoảng từ:
  - **0 đến 1**: càng gần 1 thì mô hình càng tốt.
  - Có thể < 0 nếu mô hình **dự đoán tệ hơn cả trung bình cộng**.

📌 **Ưu điểm**: Không phụ thuộc vào đơn vị dữ liệu, rất trực quan để so sánh giữa các mô hình.
📌 **Nhược điểm**: Không cho biết mức độ sai số cụ thể.

✅ 4. **MAPE – Mean Absolute Percentage Error**

🔹 Công thức:

$$\text{MAPE} = \frac{100\%}{n} \sum_{i=1}^{n} \left| \frac{y_i - \hat{y}_i}{y_i} \right|$$

🔹 Ý nghĩa:

- Là **tỷ lệ phần trăm sai số tuyệt đối trung bình**.
- Dễ hiểu với người không chuyên (ví dụ: "dự đoán sai lệch trung bình 8%").
- Thường dùng trong kinh doanh, tài chính (do tỷ lệ % dễ diễn giải).

📌 **Ưu điểm**: Diễn giải đơn giản, trực quan.
📌 **Nhược điểm**:

- Không dùng được nếu $y_i = 0$ (chia cho 0).
- Nhạy cảm nếu y_i rất nhỏ ⇒ sai số % cao không hợp lý.

📌 Khi MAPE = 10%, ta hiểu là:

Trung bình, mô hình của bạn dự đoán sai lệch khoảng 10% so với giá trị thực tế.

Diễn giải dễ nhớ:

- Nếu giá nhà thực tế là **1 tỷ**, mô hình có thể dự đoán **dao động trong khoảng 900 triệu – 1.1 tỷ**.
- Nếu điểm thi thực tế là **8 điểm**, mô hình dự đoán có thể sai lệch khoảng **±0.8 điểm**.

MAPE chỉ là **giá trị trung bình** của phần trăm sai số, nên:

❗ **Có thể có những lần dự đoán sai lệch đến 20%, thậm chí hơn**, nhưng tổng thể trung bình vẫn chỉ là 10%.

Feature Engineering

1. Feature Engineering là gì?

Feature Engineering là quá trình:

- **Tạo mới**,
- **Chuyển đổi**,
- **Chọn lọc**,

các biến đầu vào (feature) để **giúp mô hình học tốt hơn từ dữ liệu**.

2. Các kỹ thuật Feature Engineering cơ bản trong hồi quy

❖ **Tạo Feature mới từ Feature cũ (Feature Creation)**

❖ **Biến đổi phi tuyến (Nonlinear Transformation)**

❖ Xử lý biến phân loại (Categorical Encoding)

❖ **Chuẩn hóa (Scaling)**

❖ Chọn lọc đặc trưng (Feature Selection)

❖ **Xử lý giá trị thiếu (Missing Values)**

```
Dữ liệu ban đầu:
   longitude  latitude  housing_median_age  total_rooms  total_bedrooms  \
0    -122.23     37.88                41.0        880.0           129.0
1    -122.22     37.86                21.0       7099.0          1106.0
2    -122.24     37.85                52.0       1467.0           190.0
3    -122.25     37.85                52.0       1274.0           235.0
4    -122.25     37.85                52.0       1627.0           280.0

   population  households  median_income  median_house_value ocean_proximity
0       322.0       126.0         8.3252            452600.0        NEAR BAY
1      2401.0      1138.0         8.3014            358500.0        NEAR BAY
2       496.0       177.0         7.2574            352100.0        NEAR BAY
3       558.0       219.0         5.6431            341300.0        NEAR BAY
4       565.0       259.0         3.8462            342200.0        NEAR BAY
LinearRegression - MAE: 50413.4333081006, MSE: 4802173538.60416
  - R² Score: 0.6488
  - MAPE: 28.60%
```