Báo cáo: Dự đoán giá nhà California bằng Mạng Nơron Hồi quy Nhiều lớp (MLP)

1. Mục tiêu bài toán.

Bài toán yêu cầu xây dựng mô hình học sâu để dự đoán giá trị trung vị của nhà tại các vùng ở California, sử dụng tập dữ liệu California Housing từ sklearn.datasets.

2. Dữ liệu và tiền xử lý.

2.1. Nguồn dữ liệu

Tên file: Week3/data/california_housing.csv được tải từ thư viện sklearn.datasets.

2.2. Tiền xử lý dữ liệu

- **Kiểm tra dữ liệu thiếu:** Không phát hiện giá trị thiếu (NaN) trong tập dữ liệu.
- **Chuẩn hóa dữ liệu:** Dữ liệu đầu vào được chuẩn hóa bằng StandardScaler để đưa các đặc trưng về phân phối chuẩn (mean=0, std=1).
- **Tách dữ liệu:** Chia tập train-test theo tỷ lệ 80% - 20% bằng train_test_split.

3. Mô hình MLP (Multi-layer Perceptron).

3.1. Cấu trúc mô hình

- Mô hình gồm 3 lớp tuyến tính với số lượng neurons khác nhau ở từng lớp được xen kẽ các hàm activation (ReLU)
- MLP(
  Linear(8 → 64) → ReLU →
  Linear(64 → 32) → ReLU →
  Linear(32 → 1)
)

3.2. Thông số mô hình

- Hàm mất mát: MSELoss (Mean Squared Error)
- Bộ tối ưu: Adam
- Batch size: 64

4. Huấn luyện mô hình và ghi log

- Mỗi cấu hình được chạy 5 lần với khởi tạo khác nhau.
- Thư viện log: Wandb
- Log các chỉ số: train_loss, test_loss, MAE, RMSE, R^2
- Biểu đồ được lưu trực tuyến trên trang Wandb.

5. Các cấu hình siêu tham số và kết quả

5.1. Cấu hình thử nghiệm

Cấu hình	Learning rate	Epochs
Config 1	0.001	200
Config 2	0.0005	200
Config 3	0.0001	250
Config 4	0.001	450
Config 5	0.01	300

5.2. Kết quả trung bình sau 5 lần chạy

Cấu hình	MAE(± std)	RMSE (± std)	R^2 (± std)
Config 1	0.3450 ± 0.0029	0.5138 ± 0.0065	0.7985 ± 0.0052
Config 2	0.3481 ± 0.0044	0.5163 ± 0.0049	0.7966 ± 0.0038
Config 3	0.3676 ± 0.0026	0.5390 ± 0.0012	0.7783 ± 0.0010
Config 4	0.3399 ± 0.0047	0.5137 ± 0.0061	0.7986 ± 0.0048
Config 5	0.3531 ± 0.0042	0.5350 ± 0.0049	0.7816 ± 0.0040

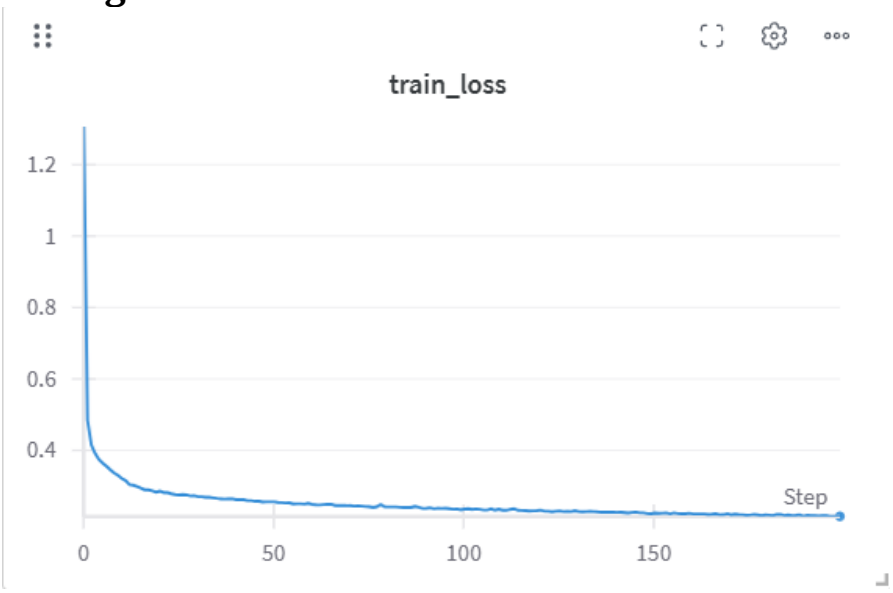
5.3. Kết luận

- **Config 4** cho kết quả tốt nhất về **MAE** và **R^2**.
- Tuy nhiên, **Config 1** gần tương đương nhưng ít epochs hơn → có thể lựa chọn để tối ưu thời gian.

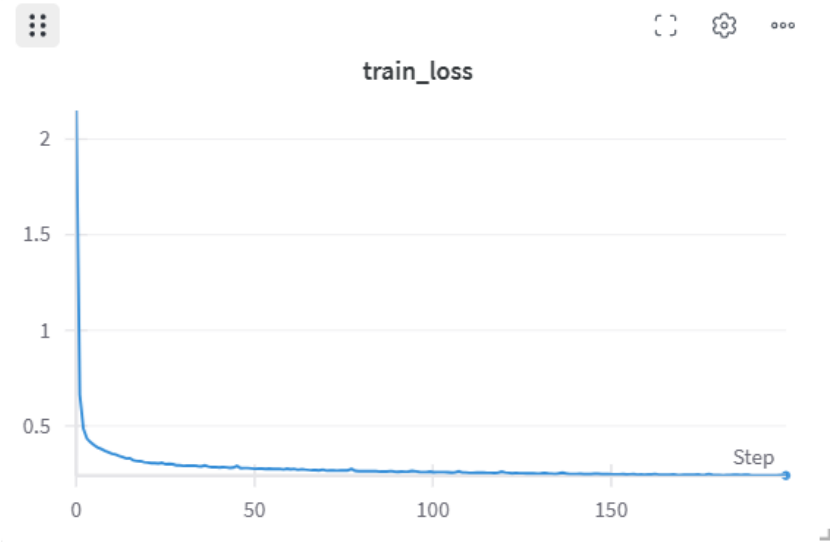
6. Biểu đồ log quá trình huấn luyện

Biểu đồ được log bằng wandb. Dưới đây là các liên kết (ví dụ minh họa):

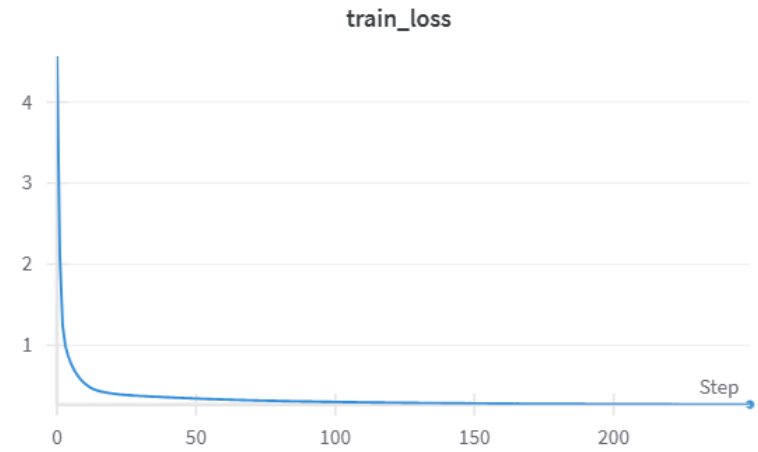
Config 1



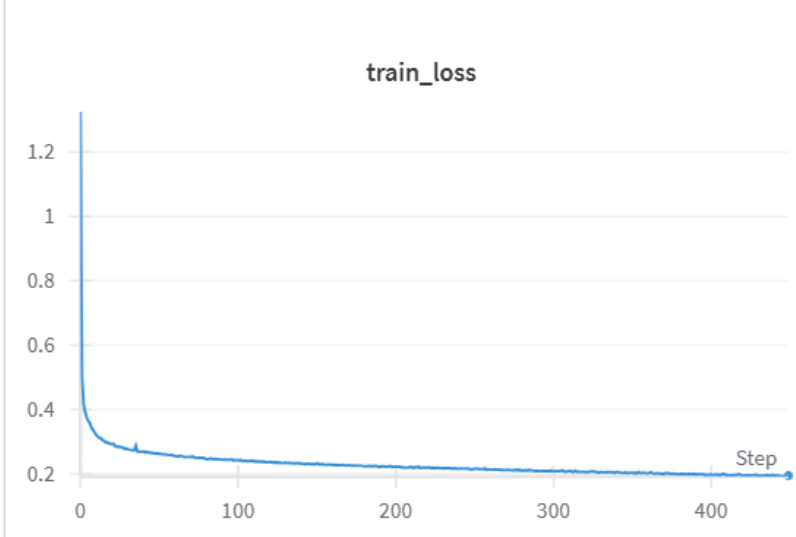
Config 2



Config 3



Config 4



Config 5

Biểu đồ bao gồm:

- Quá trình giảm train_loss
- Biến động test_loss, MAE, RMSE
- Chỉ số R^2 tăng dần và ổn định qua các epoch

7. Tổng kết.

- Mô hình MLP cho kết quả dự đoán tốt với R^2 gần 0.80.
- Việc lựa chọn learning rate và số epoch ảnh hưởng rõ rệt đến độ chính xác.
- Wandb giúp quá trình thử nghiệm và đánh giá mô hình trở nên trực quan và dễ dàng hơn.