Sử dụng Linear Regression để dự đoán giá nhà ở USA.

1. Thu thập dữ liệu :

- Em thu thập dữ liệu từ trang web :

https://www.kaggle.com/harlfoxem/housesalesprediction

- Gồm 21600 samples. Mỗi sample gồm 20 đặc trưng :

Id, date, Bedrooms, bathrooms, sqft_living , sqft_lot, floors,waterfront, view, condition,graden, sqft_above, sqft_basement, yr_built, yr_renovated, zipcode, lat, long,sqft_living15, sqft_lot15.

- Thuộc tính price là giá trị đầu ra dự đoán.

2. Chuẩn hóa dữ liệu.

- Thực hiện loại bỏ các thuộc tính Id, date và zipcode.

- Chia bộ dữ liệu thành 3 phần : train_set : valid_set : Test_set = 3 : 1 : 1.

- Chuẩn hóa dữ liệu với công thức :

$$X' = \frac{X - \min(X)}{\max(X) - \min(X)}$$

- Trong đó min(X), max(X) sử dụng từ bộ train_set để chuẩn hóa cho valid_set và test_set.

3. Tìm kiếm tham số mô hình.

- Tìm tham số mô hình dựa vào công thức :

$$W = (XX^T + \lambda \mathrm{I})^{-1} Xy$$

- Tìm tham số mô hình sử dung GD và SGD : Em sử dụng train_set để tìm kiếm tham số mô hình W, sử dụng tập valid_set để tìm learning_rate cho kết quả lỗi

thấp nhất. Với learning_rate [0 : 0.01 : 1]. Sau khi chạy thu được best_learning_rate = 0.65.

- Tìm tham số mô hình : Sử dụng thư viện sklearn để so sánh kết quả.

4. Hàm đánh giá lỗi :

- Em sử dụng hàm Mean Square Error :

$$Loss = \frac{1}{2N}\sum(y_{real} - y_{pred})^2$$

5. Kết quả thu được :

Method	MSE
Sử dụng công thức :	16231720653.072
Gradient Descent :	18113611298.706
SGD :	62200452688.807
Dùng thư viện :	16232337097.738

- Kết quả thu được từ thư viện và công thức cho kết quả tốt nhất và tương tự nhau về tham số mô hình.

- GD cho kết quả tốt hơn hẳn so với SGD cho tham số mô hình tốt với learning rate =0.65.