# Assignment:

T1. Viết mã giả của thuật toán tìm phương trình hồi quy tuyến tính.

T2. Đánh giá độ phức theo thời gian của thuật toán tìm phương trình hồi quy tuyến tính.

T3. Với tập dữ liệu về Việt Nam tùy chọn, dùng MS Excel, ngôn ngữ R và Python, hãy tìm phương trình hồi quy tuyến tính.

# Submission:

## Task and member assignment table:

|  |  |  |
| --- | --- | --- |
| **Member** | | **Task** |
| Id | Name |
| 240104026 | Huỳnh Quốc Bữu | * **T2.** Đánh giá độ phức theo thời gian của thuật toán tìm phương trình hồi quy tuyến tính * **T3.4.** Tìm phương trình hồi quy tuyến tính bằng R * **T3.5.** Tìm phương trình hồi quy tuyến tính bằng Python |
| 240104039 | Nguyễn Đức Linh | * **T1.** Viết mã giả của thuật toán tìm phương trình hồi quy tuyến tính * **T3.1.** Chọn tập dữ liệu về Việt Nam * **T3.2.** Tiền xử lý dữ liệu * **T3.3.** Tìm phương trình hồi quy tuyến tính bằng Excel |

## References sources:

|  |  |
| --- | --- |
| **Sources** | **References** |
| Tập dữ liệu về Việt Nam:  vietnam-housing-dataset-2024 | https://www.kaggle.com/datasets/nguyentiennhan/vietnam-housing-dataset-2024/data |
| **Github-Code-Team11** | https://github.com/buuhq-it/uit-msc-data-analyst-assignments/tree/main/hw1 |

## Solution:

### **T1. Viết mã giả của thuật toán tìm phương trình hồi quy tuyến tính:**

1. **Mô tả bài toán**

Trong trường hợp tổng quát, mô hình hồi quy tuyến tính đa biến (multiple linear regression) được biểu diễn dưới dạng toán học:

Trong đó:

* là biến phụ thuộc.
* là hệ số tự do.
* là hệ số hồi quy (coefficients) tương ứng với các biến độc lập .
* là thành phần sai số.

Một mẫu (sample) trong (1) có thể biểu diễn là:

Để thuận tiện hơn trong tính toán, chúng ta đặt

, **,** ,

Khi đó, phương trình hồi quy tuyến tính đa biến có thể được biểu diễn dạng ma trận là:

Hàm mất mát (loss function) là hàm biểu thị độ sai lệch giữa giá trị dự đoán và giá trị thực tế. Hàm mất mát được sử dụng phổ biến nhất trong hồi quy tuyến tính là hàm trung bình độ lỗi bình phương (Mean Squared Error – MSE). Với là giá trị dự đoán của , phương trình hàm mất mát được biểu diễn dạng

Sau khi có hàm mất mát, mục tiêu của bài toán hồi quy tuyến tính trở thành tìm hệ số sao cho hàm mất mát nhỏ nhất.

nhỏ nhất

1. **Mã giả**

**Input:**

- Tập dữ liệu gồm n mẫu.

Mỗi mẫu i (với i = 1..n) gồm:

+ Vector đặc trưng

+ Nhãn/giá trị thực

- Số chiều (k) của vector đặc trưng.

**Process:**

- Khởi tạo ma trận X (k+1 cột) và vector y (n×1):

+ Với mỗi mẫu i:

- X[i,0] = 1 # cột 1 dùng cho hệ số tự do β0

-

-

...

-

-

2. Tính ma trận kích thước (p×p) với p = k+1.

3. Tính ma trận kích thước (p×1).

4. Tính nghịch đảo (giả sử là khả nghịch).

5. Tính vector hệ số hồi quy β:

β =

**Output:**

- Các hệ số hồi quy

### **T2.** Đánh giá độ phức theo thời gian của thuật toán tìm phương trình hồi quy tuyến tính.

* Tính Phép nhân là giữa ma trận p×n và n×p, do đó chi phí thời gian thường được đánh giá
* là phép nhân giữa ma trận pxn với nx1, chi phí O(np)
* có kích thước pxp sấp sỉ

### **T3.** Với tập dữ liệu về Việt Nam tùy chọn, dùng MS Excel, ngôn ngữ R và Python, hãy tìm phương trình hồi quy tuyến tính.

#### **T3.1** Chọn tập dữ liệu về Việt Nam

1. Chọn tập dữ liệu về giá nhà tại Việt Nam để phân tích
2. Nguồn dữ liệu: https://www.kaggle.com/datasets/nguyentiennhan/vietnam-housing-dataset-2024/data
3. File dữ liệu: vietnam\_housing\_dataset.csv
4. Giới thiệu tập dữ liệu:

Bộ dữ liệu này chứa thông tin về các bất động sản nhà ở tại Việt Nam. Nó bao gồm các đặc điểm chi tiết của từng bất động sản như vị trí, đặc điểm vật lý, tình trạng pháp lý và nội thất, cùng với giá bán. Bộ dữ liệu này rất hữu ích cho việc phân tích bất động sản, xây dựng mô hình dự đoán giá và phân tích xu hướng thị trường.

Nguồn thu thập dữ liệu: Dữ liệu được thu thập (crawled) từ trang web [batdongsan.vn](https://batdongsan.vn).

1. Mô tả các cột dữ liệu:

* Address (Địa chỉ): Địa chỉ đầy đủ của bất động sản, bao gồm tên dự án (nếu có), đường, phường, quận và thành phố.
* Area (Diện tích): Tổng diện tích của bất động sản, đơn vị tính là mét vuông (m²).
* Frontage (Mặt tiền): Chiều rộng mặt trước của bất động sản, đơn vị tính là mét (m).
* Access Road (Đường vào): Chiều rộng của con đường dẫn vào bất động sản, đơn vị tính là mét (m).
* House Direction (Hướng nhà): Hướng chính mà mặt trước của căn nhà quay về (ví dụ: Đông, Tây, Nam, Bắc).
* Balcony Direction (Hướng ban công): Hướng mà ban công quay về.
* Floors (Số tầng): Tổng số tầng của căn nhà.
* Bedrooms (Số phòng ngủ): Số lượng phòng ngủ trong bất động sản.
* Bathrooms (Số phòng tắm): Số lượng phòng tắm trong bất động sản.
* Legal Status (Tình trạng pháp lý): Trạng thái pháp lý của bất động sản, ví dụ như có sổ hồng, đang trong hợp đồng mua bán, v.v.
* Furniture State (Tình trạng nội thất): Tình trạng trang bị nội thất, ví dụ như đầy đủ nội thất, nội thất cơ bản hoặc chưa có nội thất.
* Price (Giá bán): Giá của bất động sản, đơn vị tính là **tỷ đồng Việt Nam (VND)**.

#### **T3.2** Tiền xử lý dữ liệu

* Lọc bỏ dữ liệu rỗng hoặc NA
* Trích xuất thông tin Provine từ trường Address
* Chuẩn hóa định dạng dữ liệu
* Encode các cột text (chưa làm)

**Input:** vietnam\_housing\_dataset.csv 🡪 **Preprocessing-code (python)** 🡪 **Output:** housing\_data\_preprocessor.xlsx

**Preprocessing-code (python):** Github-Code-Team11/hw1-preprocessor.py

A computer screen shot of a program

AI-generated content may be incorrect.

Output: Github-Code-Team11/datasets/ vietnam\_housing\_dataset.csv

#### **T3.3** Tìm phương trình hồi quy tuyến tính bằng Excel

* Lưu file housing\_data\_preprocessor.xlsx 🡪 **hw1-regression-excel.xlsx** để chạy hồi qui
* Chọn X: ['Area', 'Frontage', 'Access Road', 'Floors', 'Bedrooms', 'Bathrooms']
* Chọn Y: ['Price']

A table with numbers and letters

AI-generated content may be incorrect.

A screenshot of a computer

AI-generated content may be incorrect.

* Kết quả hồi quu tuyến tính:

là cột **Coefficients** như hình bên dưới

A screenshot of a spreadsheet

AI-generated content may be incorrect.

* Excel file: Github-Code-Team11/ hw1-regression-excel.xlsx

#### **T3.4** Tìm phương trình hồi quy tuyến tính bằng R

* **Code:** Github-Code-Team11/hw1-regression-R.R

A computer screen shot of a computer code

AI-generated content may be incorrect.

* **Kết quả:**

là cột **Coefficients [Estimate]** như hình bên dưới

A screenshot of a computer

AI-generated content may be incorrect.

* **Kết luận:** giống với kết quả chạy bằng excel

#### **T3.5:** Tìm phương trình hồi quy tuyến tính bằng Python

* **Code:** Github-Code-Team11/hw1-regression-python.py

A computer screen with text and symbols

AI-generated content may be incorrect.

* **Kết quả:**

là cột **coef** như hình bên dưới

A screenshot of a computer

AI-generated content may be incorrect.

* **Kết luận:** giống với kết quả chạy bằng excel và R