

TRƯỜNG ĐẠI HỌC THỦY LỢI

**KHOA CÔNG NGHỆ THÔNG TIN**

BÀI TẬP LỚN

HỌC PHẦN: HỌC MÁY

# ĐỀ TÀI: **Paris Housing Price Prediction**

Giáo viên hướng dẫn: Nguyễn Huy Đức

Sinh viên/nhóm sinh viên thực hiện:

1. Nguyễn Gia Bảo, lớp 62TH-VA
2. Lê Đạt Anh, lớp 62TH-VA
3. Nguyễn Đức Anh, lớp 62TH-VA

**Hà Nội, năm 2022**

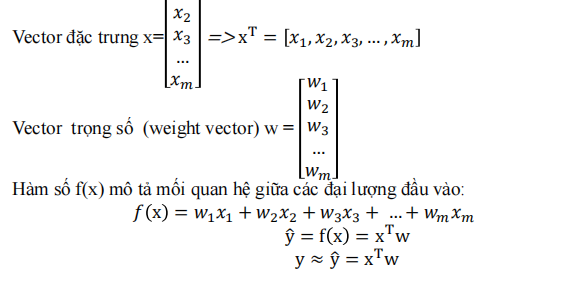
## Phần 1: Tổng quan

### *Mô tả bài toán*

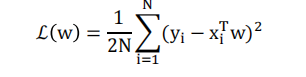
* + Tên bài toán: Paris Housing Price Prediction (Dự đoán giá nhà ở Paris)
  + Input: Dữ liệu sử dụng để dự đoán
  + Ouput: Giá dự đoán của ngôi nhà dựa vào dữ liệu trên
  + tóm tắt công việc thực hiện của bài toán.

### *Phương pháp học máy*

*\** Linear Regression:



+)hàm mất mát:



+)nghiệm của bài toán:

- Tìm giá trị tối ưu của w có thể được thực hiện thông qua việc giải phương trình đạo hàm của ℒ w theo w bằng không

-Đạo hàm theo w của hàm ℒ w là:



- Giải phương pháp đạo hàm bằng không:



- Nếu ma trận XXT khả nghịch thì phương trình trên có nghiệm duy nhất là:



- Nếu ma trận XXT không khả nghịch thì nghiệm của phương trình có thể xác

định dựa vào giả nghịch đảo:



\*PCA:

+ Phương pháp *Principal Component Analysis* (PCA):

- Dựa trên quan sát rằng dữ liệu thường không phân bố ngẫu nhiên trong không gian mà thường phân booss gần các đường/mặt đặc biệt nào đó.

- PCA xem xét một trường hợp đặt biệt khi các mặt đặc biệt đó có dạng tuyến tính là các không gian con

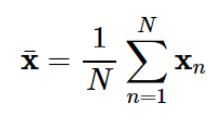
+ Định nghĩa:

-PCA là phương pháp đi tìm một hệ trực chuẩn mới sao cho trong hệ này, các thành phần quan trọng nhất nằm trong K thành phần đầu tiên.

-Để cho đơn giản trong tính toán, PCA sẽ tìm một hệ trực chuẩn để làm cơ sở mới.

+ Các bước thực hiên PCA:

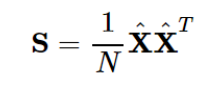
B1: Tính vector kỳ vọng của toàn bộ dữ liệu:



B2: Trừ mỗi điểm dữ liệu đi vector kỳ vọng của toàn bộ dữ liệu:



B3: Tính ma trận hiệp phương sai S:



B4: Tính các trị riêng và vector riêng của ma trận S , sắp xếp chúng theo thứ tự giảm dần của giá trị riêng.

B5: Chọn vector riêng ứng với trị riêng lớn nhất để xây dựng ma trận Uk có

các cột tạo thành một hệ trực giao. vectors này là các thành phần chính.

B6:Chiếu dữ liệu ban đầu đã chuẩn hoá xuống không gian con tìm được. Dữ

liệu mới chính là toạ độ của các điểm dữ liệu trên không gian mới.

## Phần 2: Thực nghiệm

### *Mô tả tập dữ liệu của bài toán*

* + Dữ liệu gồm những chiều thông tin gì (mỗi mẫu (vertor) dữ liệu có những thông tin gì), có bao nhiêu mẫu dữ liệu (ít nhất là 100 vector dữ liệu). Mô tả nhãn lớp của dữ liệu.
  + Toàn bộ dữ liệu của các thuộc tính đều dưới dạng số. Các nhãn lớp của tập dữ liệu:
    - * squareMeters – diện tích
      * numberOfRooms – số phòng
      * hasYard – có sân (1 = true, 0 = false)
      * hasPool – có bể bơi (1 = true, 0 = false)
      * floors – số tầng
      * cityCode – code zip của thành phố
      * cityPartRange – chất lượng dân cư xung quanh
      * numPrevOwners – số chủ từng ở ngôi nhà
      * made – thời điểm hoàn tất xây dựng (năm)
      * isNewBuilt – đã được tân trang (1 = true, 0 = false)
      * hasStormProtector – có chống bão (1 = true, 0 = false)
      * basement – diện tích tầng hầm
      * attic – diện tích gác mái
      * garage – điện tích garare
      * hasStorageRoom – có phòng kho (1 = true, 0 = false)
      * hasGuestRoom – có phòng cho khách (1 = true, 0 = false)
      * price – giá trị ngôi nhà
  + Mô tả ma trận dữ liệu (X), nhãn lớp (Y)
    - X =

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 75523 | 3 | 0 | 1 | 63 | 9373 | 3 | 8 | 2005 | 0 | 1 | 4313 | 9005 | 956 | 0 | 7 |
| 80771 | 39 | 1 | 1 | 98 | 39381 | 8 | 6 | 2015 | 1 | 0 | 3653 | 2436 | 128 | 1 | 2 |
| 55712 | 58 | 0 | 1 | 19 | 34457 | 6 | 8 | 2021 | 0 | 0 | 2937 | 8852 | 135 | 1 | 9 |
| ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... |
| 1440 | 84 | 0 | 0 | 49 | 18412 | 6 | 10 | 1994 | 1 | 0 | 8485 | 2024 | 278 | 1 | 6 |

* + - Y =

|  |
| --- |
| 7559081.5 |
| 8085989.5 |
| 5574642.1 |
| ... |
| 146708.4 |

* + Chia tập dữ liệu thành 2 phần: 70% dùng để huấn luyện mô hình, 30% dùng để kiểm tra.

### *Phân tích kết quả của chương trình*

* + Tỷ lệ dự đoán đúng: 0.999999577176331
  + Tỷ lệ dự đoán sai : 4.2282366896628787e-07

## Kết luận:

## Đề bài: dự đoán giá tiền nhà ở Paris

- Input: Một tập dữ liệu của các thông tin bài toán

- Ouput: Tập dữ liệu dự đoán được dựa vào dữ liệu đầu vào

- Công việc: Người dùng sẽ phải nhập thông tin vào giao diện người dùng để dự đoán giá tiền của ngôi nhà

+ Đọc tập dữ liệu

+ sử dụng PCA để giảm số chiều của X

+ chia tập dữ liệu X\_new thành 70% train và 30% valid

+ tìm dữ liệu có tỉ lệ tốt nhất rồi gán cho mô hinh dự đoán Linear Regression

- Huấn luyện mô hình và dự đoán

- Tỉ lệ cao nhất khi dùng PCA cho dự đoán đúng là : 56%

## Tài liệu tham khảo

* + Các tài liệu được tham khảo trong báo cáo:

+ scikit-learn

+ machine learning cơ bản