

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

1. **giới thiệu về học máy**

**-** Học máy (Machine Learning): nghiên cứu và xây dựng các kĩ thuật cho phép các

hệ thống “học” tự động từ dữ liệu để giải quyết những vấn đề cụ thể

- Học máy có giám sát: Mô hình học có giám sát là mô hình học trên dữ liệu có dán nhãn, tức là mục tiêu của bài toán machine learning cần học đã được gán nhãn sẵn trong dữ liệu huấn luyện

\***ưu điểm**:

- Học tập có giám sát cho phép thu thập dữ liệu và tạo ra dữ liệu đầu ra từ những kinh nghiệm trước đó.

- Giúp tối ưu hóa các tiêu chí hiệu suất với sự trợ giúp của kinh nghiệm.

- ML có giám sát giúp giải quyết nhiều loại vấn đề tính toán trong thế giới thực.

**\*nhược điểm:**

- Phân loại dữ liệu lớn có thể là một thách thức.

- Đào tạo cho việc học có giám sát cần rất nhiều thời gian tính toán, vì vậy, nó đòi hỏi rất nhiều thời gian.

- Học máy không giám sát: Học không giám sát là một lớp mô hình học sử dụng một thuật toán để mô tả hoặc trích xuất ra các mối quan hệ tiềm ẩn trong dữ liệu. Khác với học có giám sát, học không giám sát chỉ thực thi trên dữ liệu đầu vào không cần các thuộc tính nhãn, hoặc mục tiêu của việc học.

\***ưu điểm**:

- Nó có thể thấy những gì tâm trí con người không thể hình dung.

- Nó được sử dụng để đào các mẫu ẩn có tầm quan trọng tối đa trong ngành và có các ứng dụng rộng rãi trong thời gian thực.

- Có sự phức tạp ít hơn so với nhiệm vụ học tập có giám sát. Ở đây, không ai được yêu cầu giải thích các nhãn liên quan và do đó nó giữ độ phức tạp ít hơn.

- Nó dễ dàng hơn một cách hợp lý để có được dữ liệu không nhãn.

**\*nhược điểm:**

**-** Nó tốn kém hơn vì nó có thể đòi hỏi sự can thiệp của con người để hiểu các mô hình và tương quan chúng với kiến thức miền

- Không phải lúc nào cũng chắc chắn rằng các kết quả thu được sẽ hữu ích vì không có thước đo nhãn hoặc đầu ra để xác nhận tính hữu dụng của nó.

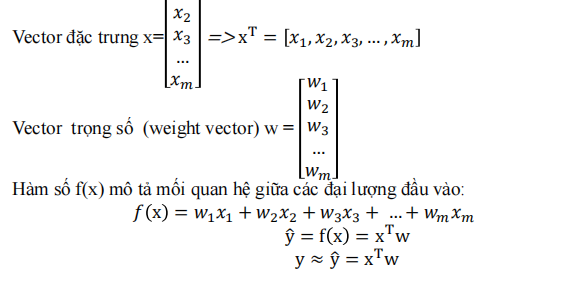
- Kết quả thường có độ chính xác thấp hơn

### *Mô tả bài toán*

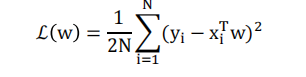
* + Tên bài toán: Paris Housing Price Prediction (Dự đoán giá nhà ở Paris)
  + Input: Dữ liệu sử dụng để dự đoán
  + Ouput: Giá dự đoán của ngôi nhà dựa vào dữ liệu trên

### *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:

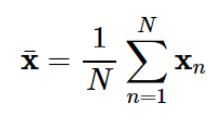
+ Định nghĩa:

-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:

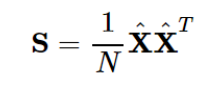
B1: Tính vector kỳ vọng của toàn bộ dữ liệu:



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.

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

- Tỉ lệ cao nhất khi dùng PCA cho dự đoán đúng là : 56,2666%

## 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

+ Dùng phương pháp Phân tích thành phần chính để lựa chọn tập các thuộc tính tốt nhất cho bài toán.

+ Từ tập training data và test data ban đầu, sử dụng các thành phần chính tốt nhất đã chọn để tạo ra tập training data và test data mới.

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

***3. Mô tả các chức năng của chương trình***

- chương trình hiển thị giao diện để người dùng nhập số liệu cụ thể cho từng thông tin được đưa ra để dự đoán

## Tài liệu tham khảo

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

+ scikit-learn

+ machine learning cơ bản