**TRƯỜNG ĐẠI HỌC THỦY LỢI**

**KHOA CÔNG NGHỆ THÔNG TIN**

****

**BÀI TẬP LỚN**

HỌC PHẦN: HỌC MÁY

**ĐỀ TÀI: ỨNG DỤNG THUẬT TOÁN LOGISTIC REGRESSION VÀ THUẬT TOÁN SUPPORT VECTOR MACHINE ĐỂ DỰ ĐOÁN GIÁ NHÀ Ở VIỆT NAM**

Giáo viên hướng dẫn: Nguyễn Thị Kim Ngân

Sinh viên/nhóm sinh viên thực hiện: Nhóm 15

1. Lâm Văn Thái, lớp 60TH1

2. Phạm Thị Phương Nga, lớp 60TH1

3. Đặng Thị Thúy, lớp 60TH1

**Hà Nội, năm 2021**

**Phần 1: Tổng quan**

# **1.Mô tả bài toán**

- Tên bài toán: Ứng dụng thuật toán Logistic regression và thuật toán Support Vector Machine để dự đoán giá nhà ở Việt Nam.

- Input (11 trá trị ): gồm

* BEDROOMS ( PHÒNG NGỦ )
* BATHROOMS ( PHÒNG TẮM )
* STORIES (TỒN TẠI)
* MAINROAD ( ĐƯỜNG CHÍNH )
* GUESTROOM ( PHÒNG KHÁCH )
* BASEMENT (TẦNG HẦM)
* HOTWATERHEATING (BÌNH NÓNG LẠNH)
* AIRCONDITIONING (ĐIỀU HÒA NHIỆT ĐỘ)
* PARKING (BÃI ĐẬU XE)
* PREFAREA (KHO)
* FURNITURETATUS (NỘI THẤT)
* Output: dự đoán giá nhà
* PRICE\_SEGMENT (TẦM GIÁ)

- Tóm tắt công việc thực hiện của bài toán:

# Bước 1: Thu thập dữ liệu bài toán:Dataset: <https://www.kaggle.com> .

# Bước 2: Xác định tập dữ liệu của bài toán:

* Input (11 trá trị ): gồm
* BEDROOMS ( PHÒNG NGỦ )
* BATHROOMS ( PHÒNG TẮM )
* STORIES (LỊCH SỬ LÂU ĐỜI)
* MAINROAD ( ĐƯỜNG CHÍNH )
* GUESTROOM ( PHÒNG KHÁCH )
* BASEMENT (TẦNG HẦM)
* HOTWATERHEATING (BÌNH NÓNG LẠNH)
* AIRCONDITIONING (ĐIỀU HÒA NHIỆT ĐỘ)
* PARKING (BÃI ĐẬU XE)
* PREFAREA (KHO)
* FURNITURETATUS (NỘI THẤT)

+Output: dự đoán giá nhà

* PRICE\_SEGMENT (TẦM GIÁ)

# Bước 3: Mô tả ma trận dữ liệu (X), nhãn lớp (Y)

# Bước 4: Chia tập dữ liệu thành 2 phần: 70% dùng để huấn luyện mô hình, 30% dùng để kiểm tra sự phù hợp của mô hình.

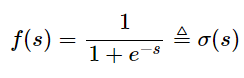
# Bước 5: Dùng tập dữ liệu kiểm tra để so sánh tỷ lệ mẫu được phân chia lớp đúng của hai thuật toán Logistic regression và Support vector machine trên python.

# **2. Phương pháp học máy**

## **2.1 Phương pháp Logistic Regression**

* ***Sigmoid function***

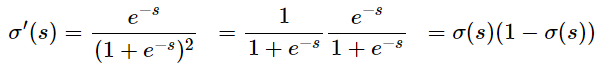
Trong số các hàm số có 3 tính chất nói trên thì hàm *sigmoid*:



được sử dụng nhiều nhất, vì nó bị chặn trong khoảng (0,1)Thêm nữa:



Đặc biệt hơn nữa:



hàm *tanh*



Hàm số này nhận giá trị trong khoảng (−1,1) nhưng có thể dễ dàng đưa nó về khoảng (0,1)có thể chứng minh được:



### ***Xây dựng hàm mất mát***

giả sử rằng xác suất để một điểm dữ liệu x rơi vào class 1 là f(wTx) và rơi vào class 0 là 1−f(wTx). Với mô hình được giả sử như vậy, với các điểm dữ liệu training (đã biết đầu ra y), ta có thể viết như sau:



trong đó

P(yi=1|xi;w)được hiểu là xác suất xảy ra sự kiện đầu ra yi=1 khi biết tham số mô hình w và dữ liệu đầu vào xi.Mục đích của chúng ta là tìm các hệ số w sao cho

f(wTxi) càng gần với 1 càng tốt với các điểm dữ liệu thuộc class 1 và càng gần với 0 càng tốt với những điểm thuộc class 0.

Ký hiệu zi=f(wTxi) và viết gộp lại hai biểu thức bên trên ta có:



Biểu thức này là tương đương với hai biểu thức (1) và (2) ở trên vì khi yi=1

, phần thứ hai của vế phải sẽ triệt tiêu, khi yi=0, phần thứ nhất sẽ bị triệt tiêu! Chúng ta muốn mô hình gần với dữ liệu đã cho nhất, tức xác suất này đạt giá trị cao nhất.

Xét toàn bộ training set với X=[x1,x2,…,xN]∈Rd×N và y=[y1,y2,…,yN], chúng ta cần tìm w để biểu thức sau đây đạt giá trị lớn nhất:



ở đây, ta cũng ký hiệu X,y như các biến ngẫu nhiên

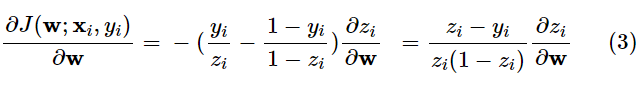


### ***Tối ưu hàm mất mát***

Hàm mất mát với chỉ một điểm dữ liệu (xi,yi) llà:



với đạo hàm

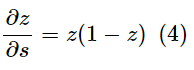


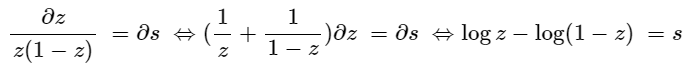
Để cho biểu thức này trở nên *gọn* và *đẹp* hơn, chúng ta sẽ tìm hàm z=f(wTx)

sao cho mẫu số bị triệt tiêu. Nếu đặt s=wTx, chúng ta sẽ có:



Một cách trực quan nhất, ta sẽ tìm hàm số z=f(s)sao cho:







### Công thức cập nhật cho logistic sigmoid regression

Tới đây, bạn đọc có thể kiểm tra rằng:

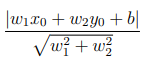




## **2.2 Phương pháp Support Vector Machine**

* Support vector machine (SVM) là một trong những thuật toán phân lớp phổ biến và hiệu quả. Ý tưởng đứng sau SVM khá đơn giản, nhưng để hiểu được cách tìm nghiệm của nó, chúng ta cần một chút kiến thức về tối ưu và duality.
* Khoảng cách từ một điểm tới một siêu mặt phẳng

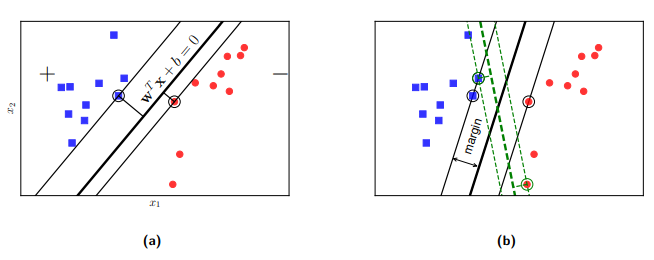
+Trong không gian hai chiều, khoảng cách từ một điểm có toạ độ (x0, y0) tới đường thẳng có phương trình w1x + w2y + b = 0 được xác định bởi:



+Trong không gian ba chiều, khoảng cách từ một điểm có toạ độ (x0, y0, z0) tới một mặt phẳng có phương trình w1x + w2y + w3z + b = 0 được xác định bởi:



* Giả sử rằng các cặp dữ liệu trong tập huấn luyện là (x1, y1),(x2, y2), . . . ,(xN , yN ) với vector xi ∈ R d thể hiện đầu vào của một điểm dữ liệu và yi là nhãn của điểm dữ liệu đó, d là số chiều của dữ liệu và N là số điểm dữ liệu. Giả sử rằng nhãn của mỗi điểm dữ liệu được xác định bởi yi = 1 hoặc yi = −1 giống như trong PLA.



Ta quan sát thấy một điểm quan trọng như sau. Với cặp dữ liệu (xn, yn) bất kỳ, khoảng cách từ điểm đó tới mặt phân chia là



Điều này có thể được nhận thấy vì theo giả sử ở trên, yn luôn cùng dấu với phía của xn. Từ đó suy ra yn cùng dấu với (wT xn +b), vì vậy tử số luôn là một đại lượng không âm. Với mặt phân chia này, margin được tính là khoảng cách gần nhất từ một điểm (trong cả hai lớp, vì cuối cùng margin của cả hai lớp sẽ như nhau) tới mặt đó, tức là:

Bài toán tối ưu của SVM chính là việc tìm w và b sao cho margin này đạt giá trị lớn nhất:



Xác định lớp cho một điểm dữ liệu mới Sau khi đã tìm được mặt phân cách wT x + b = 0, nhãn của bất kỳ một điểm nào sẽ được xác định đơn giản bằng:



**Phần 2: Thực nghiệm**

# **1.Mô tả tập dữ liệu của bài toán**

* Dữ liệu bài toán gồm: ma trận dữ liệu (X), nhãn lớp (Y)
* Ma trận x là vector đầu vào tập hợp các mẫu dữ liệu, mỗi mẫu dữ liệu là một vector mô tả giá nhà nhỏ hơn, lớn hơn hoặc bằng 7 tỷ là nhãn lớp y (giá trị thực của đầu ra) gồm có 11 thông tin dữ liệu:
* BEDROOMS ( PHÒNG NGỦ )
* BATHROOMS ( PHÒNG TẮM )
* STORIES (LỊCH SỬ LÂU ĐỜI)
* MAINROAD ( ĐƯỜNG CHÍNH )
* GUESTROOM ( PHÒNG KHÁCH )
* BASEMENT (TẦNG HẦM)
* HOTWATERHEATING (BÌNH NÓNG LẠNH)
* AIRCONDITIONING (ĐIỀU HÒA NHIỆT ĐỘ)
* PARKING (BÃI ĐẬU XE)
* PREFAREA (KHO)
* FURNITURETATUS (NỘI THẤT)
* Bài toán có 150 mẫu dữ liệu.
* Dữ liệu bài toán:

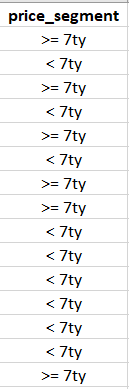
Dataset: <https://www.kaggle.com> .

- Mô tả ma trận dữ liệu (X), nhãn lớp (Y)

+Ma trận dữ liệu X: demo 15 mẫu dữ liệu ma trận X



+Ma trận nhãn lớp Y: demo 15 mẫu dữ liệu nhãn lớp Y



- Chia tập dữ liệu thành 2 phần: 70% dùng để huấn luyện mô hình, 30% dùng để kiểm tra sự phù hợp của mô hình.

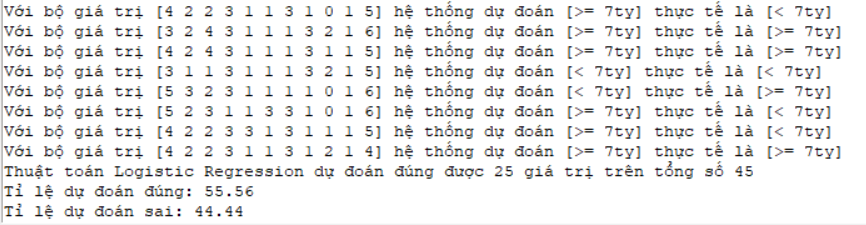
* Tập dữ liệu có 150 mẫu dữ liệu chia: 105 mẫu dữ liệu dùng để huấn luyện mô hình, 45 mẫu dữ liệu dùng để kiểm tra sự phù hợp của mô hình.
* Dữ liệu tập huấn luyện (Training).
* Ma trận xtrain là vector đầu vào tập hợp các điểm dữ liệu, mỗi mẫu dữ liệu là một vetor mô tả giá nhà nhỏ hơn, lớn hơn hoặc bằng 7 tỷ là nhãn lớp ytrain (giá trị thực của đầu ra) gồm có 11 thông tin dữ liệu:BEDROOMS ( PHÒNG NGỦ ),BATHROOMS ( PHÒNG TẮM ), STORIES (LỊCH SỬ LÂU ĐỜI), MAINROAD ( ĐƯỜNG CHÍNH ), GUESTROOM ( PHÒNG KHÁCH ), BASEMENT (TẦNG HẦM), HOTWATERHEATING (BÌNH NÓNG LẠNH), AIRCONDITIONING (ĐIỀU HÒA NHIỆT ĐỘ), PARKING(BÃI ĐẬU XE), PREFAREA(KHO), FURNITURETATUS(NỘI THẤT).
* Tập dữ liệu huấn luyện gồm 105 mẫu dữ liệu.
* Dữ liệu tập kiểm tra (Testing)
* Ma trận xtest là vector đầu vào tập hợp các điểm dữ liệu, mỗi mẫu dữ liệu là một vector mô tả giá nhà nhỏ hơn, lớn hơn hoặc bằng 7 tỷ là nhãn lớp ytest (giá trị thực của đầu ra) gồm có 11 thông tin dữ liệu: BEDROOMS ( PHÒNG NGỦ ), BATHROOMS (PHÒNG TẮM), STORIES (LỊCH SỬ LÂU ĐỜI), MAINROAD ( ĐƯỜNG CHÍNH ), GUESTROOM ( PHÒNG KHÁCH ), BASEMENT (TẦNG HẦM), HOTWATERHEATING (BÌNH NÓNG LẠNH), AIRCONDITIONING (ĐIỀU HÒA NHIỆT ĐỘ), PARKING (BÃI ĐẬU XE), PREFAREA (KHO), FURNITURETATUS (NỘI THẤT).
* Tập dữ liệu kiểm tra gồm 45 mẫu dữ liệu.

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

- Tỷ lệ dự đoán đúng và Tỷ lệ dự đoán sai

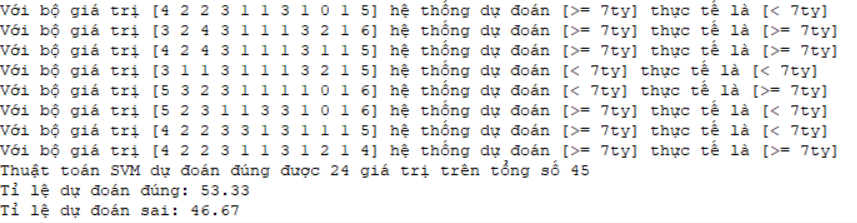
* Phương pháp Logistic Regression

….



* Phương pháp Support Vector Machine

….



- So sánh và phân tích kết quả của 2 phương pháp Logistic Regression và Support Vector Machine: từ kết quả của hai thuật toán trên ta thấy tỷ lệ dự đoán đúng kết quả của thuật toán Logistic Regression đúng nhiều hơn so với thuật toán Support Vector Machine nên với mô hình bài toán sử dụng thuật toán Logistic Regression là tối ưu hơn.

**Kết luận**

- Các nội dung chính mà bài tập lớn làm được:

* Xây dựng được mô hình bài toán dự đoán giá nhà từ 2 phương pháp Logistic Regression và Support Vector Machine.
* Tìm được phương pháp tối ưu xây dựng mô hình bài toán phù hợp.
* Tạo hai giao diện nhập số liệu mới và hiển thị kết quả dự đoán sử dụng hai thuật toán trên.

**Tài liệu tham khảo**

* Tham khảo slide “Học máy” của cô TS.Nguyễn Thị Kim Ngân
* Tài liệu file: <https://github.com/tiepvupsu/ebookMLCB> của thầy Vũ Hữu Tiệp
* Dataset:[https://www.kaggle.com/](https://www.kaggle.com/goyalshalini93/car-price-prediction-linear-regression-rfe/data?fbclid=IwAR0x2sXrpPkU5PBKFhvSrWrJL344T-Kwyx5Qtunm_z4TOkW3xHsPJ7Slg3s)