TRƯỜNG ĐẠI HỌC THỦY LỢI

**KHOA CÔNG NGHỆ THÔNG TIN**

A blue and white logo

Description automatically generated with low confidence

**BÀI TẬP LỚN**

HỌC PHẦN: HỌC MÁY

**ĐỀ TÀI: dự đoán Tuyển sinh sau đại học**

**từ góc độ của Ấn Độ.**

Giáo viên hướng dẫn: Nguyễn Thị Kim Ngân

Nhóm sinh viên thực hiện: Nhóm 7

1. Phạm Thanh Lam 63CNTT1 2151061166

2. Trần Đăng Trung 63CNTT1 2151062885

3. Nghiêm Đức Quân 63CNTT1 2151060200

4. Nguyễn Thị Giang 63CNTT1 2151062758

**Hà Nội, năm 2023**

**Phần 1: Tổng quan**

***1.Mô tả bài toán***

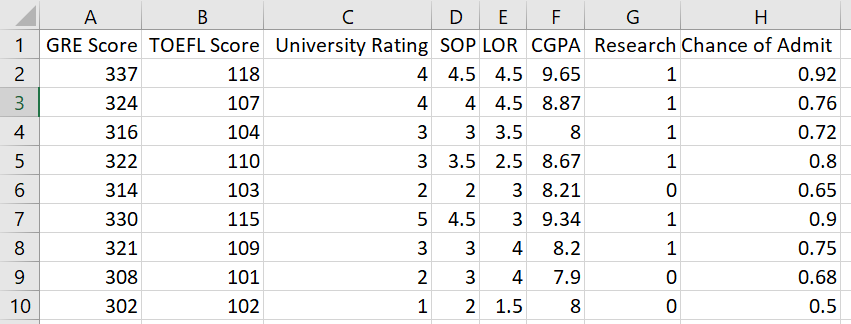
* Tên bài toán: dự đoán Tuyển sinh sau đại học từ góc độ của Ấn Độ.
* Input: vector gồm 7 trường: GRE Score, TOEFL Score, University Rating, SOP, SOP, CGPA, Research
* Ouput: Chance of Admit tỉ lệ trúng tuyển vào trường Đại học
* tóm tắt công việc thực hiện của bài toán:

1. Đọc dữ liệu
2. Chia tập dữ liệu (70% dùng để huấn luyện mô hình, 30% dùng để kiểm tra sự phù hợp của mô hình)

* Khai báo mô hình:
  + Dùng hàm Linear Regression để huấn luyện mô hình
  + Dùng hàm Lasso để huấn luyện mô hình
  + Dùng hàm Ridge để huấn luyện mô hình
* In ra kết quả

1. K-Fold Cross Validation

* Chia tập dữ liệu thành k-1 phần, 1 phần còn lại dùng để kiểm tra
* Xây dựng hàm tính trung bình lỗi y dự đoán và y thực tế
* Chọn ra mô hình tốt nhất
* In ra kết quả
* Minh họa:



2***. Phương pháp học máy***

*Hồi quy tuyến tính:*

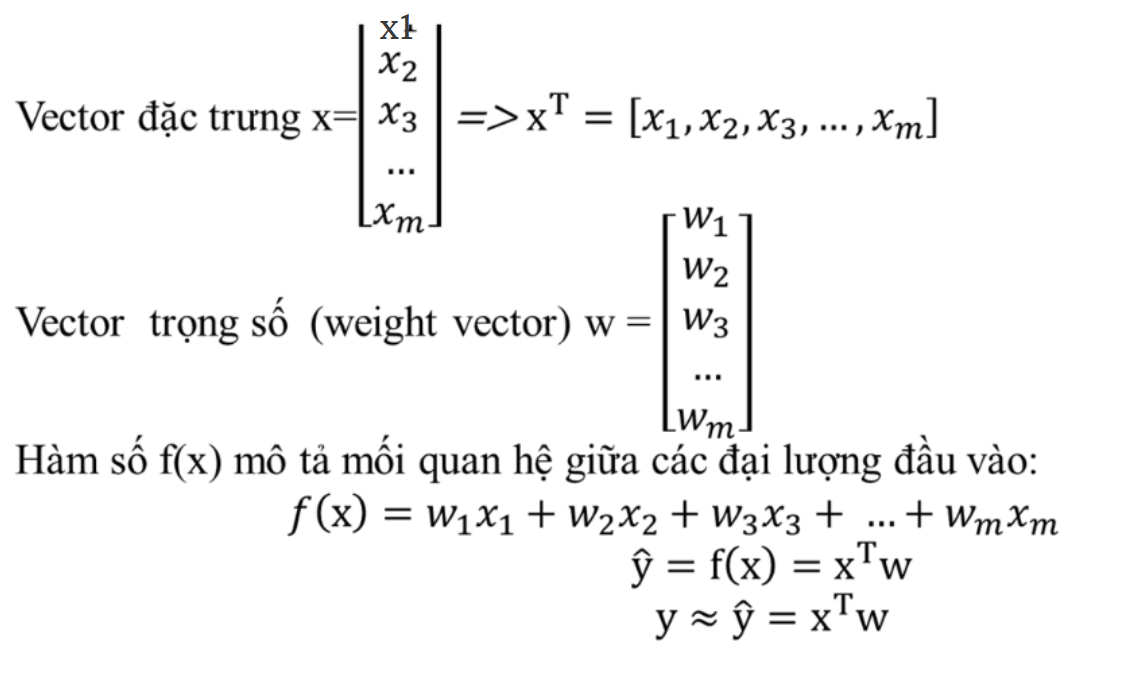
- Hồi quy tuyến tính là một kỹ thuật phân tích dữ liệu dự đoán giá trị của dữ liệu không xác định bằng cách sử dụng một giá trị dữ liệu liên quan và đã biết khác. Nó mô hình toán học biến không xác định hoặc phụ thuộc và biến đã biết hoặc độc lập như một phương trình tuyến tính

- Input:Cho tập dữ liệu huấn luyện gồm N mẫu.Mỗi mẫu là 1 cặp(xi,yi):

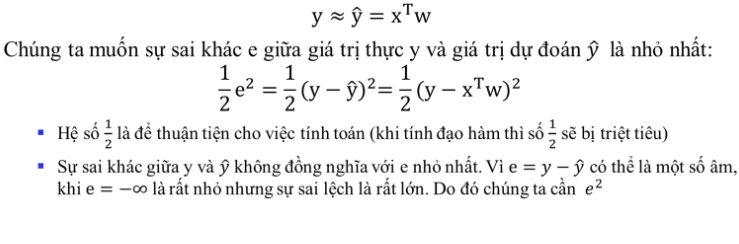
* xi: vector đặc trưng
* yi: giá trị của vector đặc trưng xi

- Output: Hàm tuyến tính có dạng f(xi)=wxi + w0

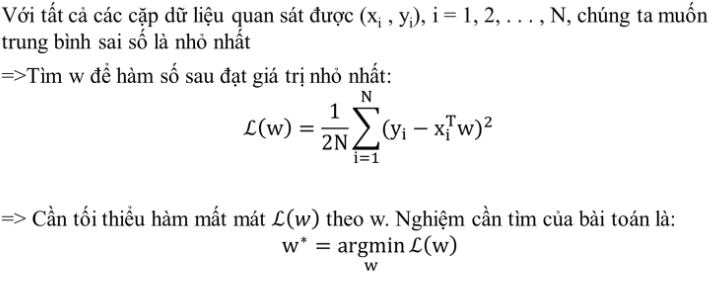
-Method(cách thực hiện)**:** Cần tìm hệ số w của hàm f(xi) sao cho trung bình sai số giữa yi và f(xi) là nhỏ nhất. Nghĩa là, tìm w để hàm số sau đạt giá trị nhỏ nhất

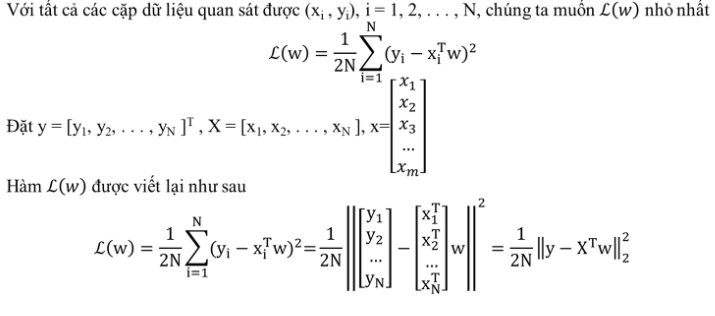


* Sai số dự đoán:

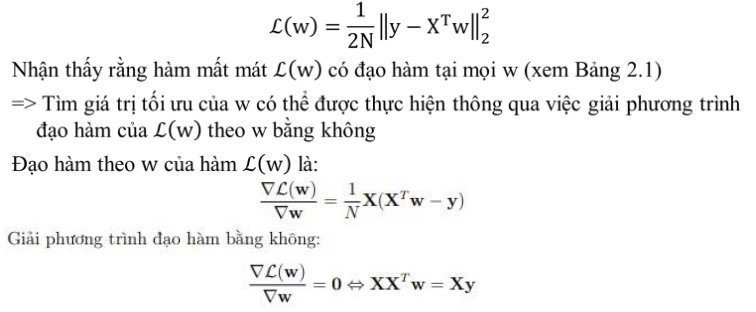


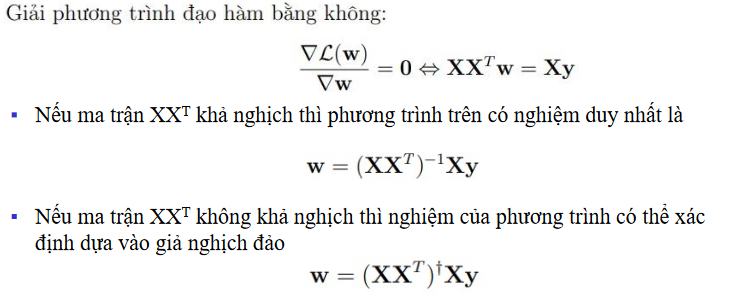
* Hàm mất mát:





* Nghiệm cho bài toán Hồi quy tuyến tính



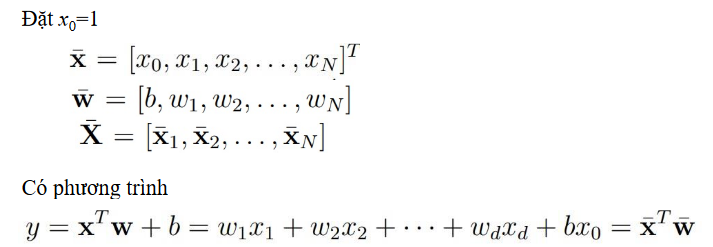


* Bias trick:

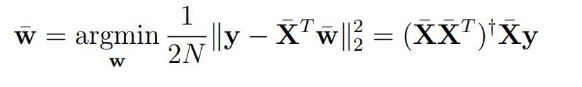
Thêm hệ số b vào sẽ khiến cho mô hình linh hoạt hơn một chút bằng cách bỏ ràng

buộc đường thẳng quan hệ giữa đầu ra và đầu vào luôn đi qua gốc toạ độ

y = xT w + b



Nghiệm



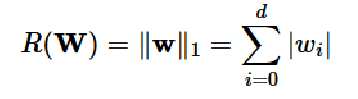
* Đánh giá phương pháp:

Trong máy học, các chương trình máy tính gọi là thuật toán phân tích các tập dữ liệu lớn và làm việc ngược từ dữ liệu đó để tính toán phương trình hồi quy tuyến tính. Các nhà khoa học dữ liệu đầu tiên sẽ đào tạo thuật toán trên các tập dữ liệu đã biết hoặc được dán nhãn và sau đó sử dụng thuật toán để dự đoán các giá trị chưa biết. Dữ liệu thực tế phức tạp hơn

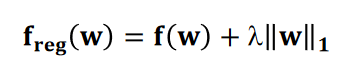
*Lasso:*

* Norm 1 là tổng các trị tuyệt đối của tất cả các phần tử. Đã chứng minh được rằng tối thiểu norm 1 sẽ dẫn tới nghiệm có nhiều phần tử bằng 0

Khi chọn



* Nghiệm w tìm được của bài toán tối thiểu freg(w) có xu hướng rất nhiều phần tử bằng không. Hàm LASSO regression

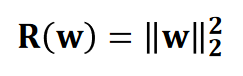


* Các thành phần khác không của w tương ứng với các đặc trưng quan trọng trong dự đoán đầu ra
* Các thành phần bằng không của w tương ứng với các đặc trưng được coi là ít quan trọng

=> LASSO regression giúp lựa chọn đặc trưng hữu ích cho mô hình

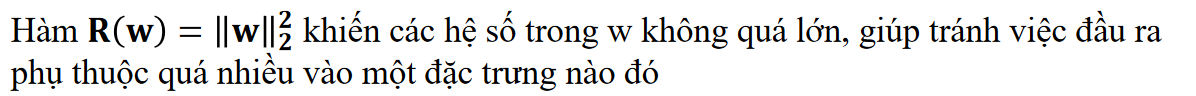
*Ridge:*

Khi chọn R(w) là Norm 2



Ta có hàm Ridge regularization





*k-fold cross validation:*

* **Cross validation** là một kỹ thuật lấy mẫu để đánh giá mô hình học máy trong trường hợp dữ liệu không được dồi dào cho lắm
* **K-Fold CV** là một phương pháp để đánh giá model một cách chính xác khi chúng ta train model nhưng có quá ít dữ liệu.
* **K** là một số nguyên dương được chọn trước (thông thường là 5 hoặc 10, nhưng có thể thay đổi tùy thuộc vào bài toán cụ thể).
* **K-fold cross validation:** Chia training set thành K tập con không giao nhau, có kích thước gần bằng nhau. X={X1, X2, …., XK}

Tại mỗi lần kiểm thử:

▪ 1 tập con được lấy ra làm validation set

▪ K-1 tập còn lại được dùng để xây dựng mô hình

▪ Mô hình cuối được xác định dựa trên trung bình của các train error và validation error (CV(w) là nhỏ nhất)

Tìm k mô hình

Lần 1: Validation set: X1, Training set: X\X1={X2, X3, …, Xk}= > M1

Lần 2: Validation set: X2, Training set: X\X2={X1, X3, …, Xk}=>M2

Lần k: Validation set: Xk, Training set: X\Xk={X1, X2, …, Xk-1}=>Mk

Nhược điểm

▪ Số lượng mô hình cần huấn luyện tỉ lệ thuận với k

▪ Nếu trong bài toán, có nhiều tham số cần xác định, khoảng giá trị của mỗi tham

số rộng, thì việc huấn luyện nhiều mô hình là khó khả thi

* Input:
* **Dữ liệu huấn luyện:** Tập dữ liệu ban đầu được chia thành K phần bằng nhau (gọi là "folds"). Mỗi phần sẽ được sử dụng như tập kiểm tra trong một lượt kiểm tra còn lại được sử dụng để huấn luyện mô hình.
* **Số lượng folds (K)**: Đây là số phần tập dữ liệu được chia. Thông thường, giá trị K thường được chọn là 5 hoặc 10, nhưng cũng có thể sử dụng các giá trị khác tùy thuộc vào kích thước của tập dữ liệu và nguồn tài nguyên có sẵn.
* **Mô hình machine learning:** Mô hình mà bạn muốn đánh giá.
* **Phép đo đánh giá (metric):** Đây là phép đo được sử dụng để đánh giá hiệu suất của mô hình trên mỗi fold.
* **Phương thức chia folds:** Có thể sử dụng các phương thức chia folds khác nhau như stratified, random, ... Tùy thuộc vào đặc điểm cụ thể của tập dữ liệu.
* Output:

=>Kết quả của phương pháp K-fold cross-validation là một tập hợp các giá trị đánh giá hiệu suất của mô hình trên các lượt kiểm tra khác nhau. Thông thường, các giá trị này bao gồm các phép đo đánh giá như Mean Squared Error (MSE), accuracy, F1-score, và nhiều phép đo khác tùy thuộc vào bài toán cụ thể.

* Ý tưởng của phương pháp K-fold cross-validation là sử dụng lại dữ liệu huấn luyện để đánh giá hiệu suất của mô hình. Thay vì chia dữ liệu thành hai phần (tập huấn luyện và tập kiểm tra) một cách cố định, ta chia dữ liệu thành K phần nhỏ bằng nhau (gọi là "folds").

**Đánh giá phương pháp:** Lý thuyết K-fold cross validation: Trong xác thực chéo, tập huấn luyện được chia thành k tập con có kích thước gần bằng nhau và không giao nhau. Tại mỗi lần thử, một trong k tập con đó được lấy ra làm tập xác thực, k − 1 tập con còn lại được coi là tập huấn luyện. Như vậy, với mỗi bộ tham số mô hình, ta có k mô hình khác nhau. Sai số huấn luyện và sai số xác thực được tính là trung bình cộng của các giá trị tương ứng trong k mô hình đó. Cách làm này có tên gọi là xác thực chéo k-fold (k-fold cross validation).

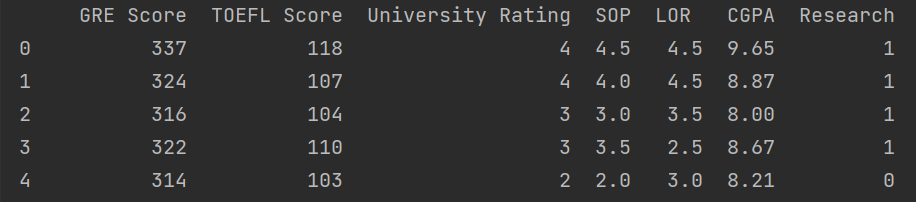
* **K-fold cross-validation** là một phương pháp quan trọng trong machine learning để đánh giá hiệu suất của mô hình. Nó bao gồm việc chia dữ liệu thành K phần bằng nhau, huấn luyện và đánh giá mô hình K lần, mỗi lần sử dụng một phần khác nhau làm tập kiểm tra. Kết quả cuối cùng là một tập hợp các kết quả đánh giá, giúp đánh giá hiệu suất tổng thể của mô hình và tránh overfitting hoặc underfitting.
* Ưu điểm: K-Fold CV sẽ giúp chúng ta đánh giá một model đầy đủ và chính xác hơn khi chúng ta có một tập dữ liệu không lớn. Để sau đó chúng ta đưa ra quyết định model đó có phù hợp với dữ liệu, bài toán hiện tại hay không để mà đưa ra next action.
* Nhược điểm: Số lượng mô hình cần huấn luyện tỉ lệ thuận với k. Nếu trong bài toán, có nhiều tham số cần xác định, khoảng giá trị của mỗi tham số rộng, thì việc huấn luyện nhiều mô hình là khó khả thi

**Phần 2: Thực nghiệm**

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

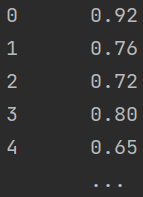
|  |  |
| --- | --- |
| Nhãn | Ý nghĩa |
| GRE Score | Điểm GRE |
| TOEFL Score | Điểm TOEFL |
| University Rating | Xếp hạng đại học |
| SOP | Quy trình thao tác chuẩn |
| LOR | Thư giới thiệu |
| CGPA | Điểm trung bình đại học |
| Research | Kinh nghiệm nghiên cứu |
| Chance of Admit | Cơ hội được nhận |

* Vector X gồm 8 nhãn: GRE Score, TOEFL Score, University Rating, SOP, SOP, CGPA, Research, Chance of Admit. Có 400 mẫu dữ liệu.
* Mô tả ma trận dữ liệu (X):

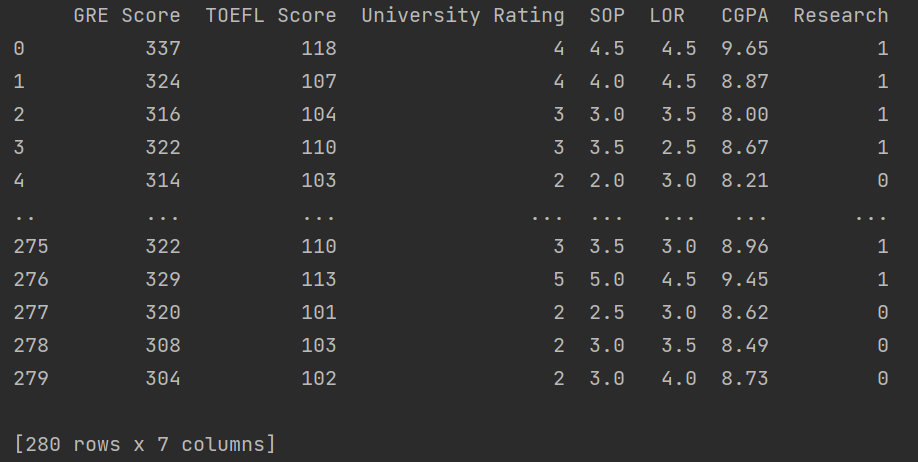


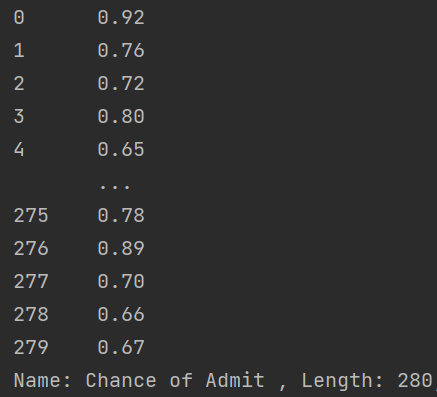
* Nhãn lớp (Y):

Chance of Admit



* 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.





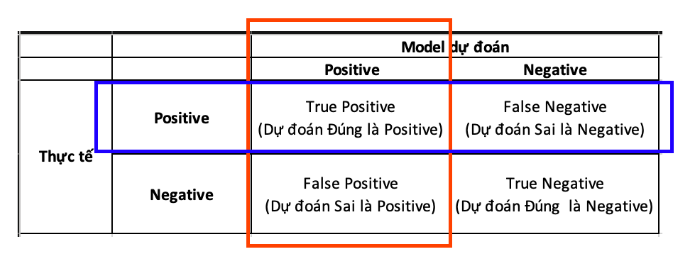
=>280 mẫu

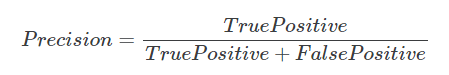
***2. Phân tích kết quả của chương trình***

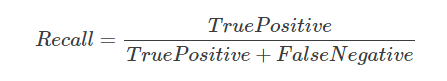
* Dùng tập dữ liệu kiểm tra để so sánh kết quả dự đoán của mô hình với kết quả thực tế, để đánh giá sự phù hợp của mô hình. Sử dụng các độ đo R2, NSE, MAE, RMSE (xem trong Chương 8. Đánh giá chất lượng mô hình học máy, phần dành cho bài toán Hồi quy).

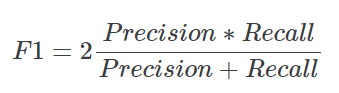
|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Linear Regression | Ridge | Lasso | K-fold CV |
| R2 | 0.828695781 | 0.828133877 | 0.276696381 | 0.812776284 |
| NSE | 0.828695781 | 0.828133877 | 0.276696381 | 0.800678304 |
| MAE | 0.041628850 | 0.041721682 | 0.090676584 | 0.044573673 |
| RMSE | 0.056897328 | 0.056990568 | 0.116914442 | 0.059482371 |

* Lựa chọn mô hình tốt nhất cho bài toán (mô hình có độ đo R2 cao nhất): mô hình Hồi quy tuyến tính (có độ đo R2 gần bằng 1 nhất)









**Kết luận**

* Hiểu được về Hồi quy tuyến tính và phương pháp K-Fold để hoàn thành bài toán dự đoán Tuyển sinh sau đại học từ góc độ của Ấn Độ.

**Tài liệu tham khảo:**

* File dữ liệu .csv trong Graduate Admission 2 trong trang web kaggle: https://www.kaggle.com/datasets/mohansacharya/graduate-admissions
* Slide “Học máy” trong “PDF bai giang” của cô Nguyễn Thị Kim Ngân
* Tài liệu tham khảo học máy cơ bản:
  + Machine learning cơ bản:<https://machinelearningcoban.com>
  + Scikt-learn: <https://scikit-learn.org/stable/modules/cross_validation.html>
  + k-fold cross validation Mì AI: https://www.miai.vn/2021/01/18/k-fold-cross-validation-tuyet-chieu-train-khi-it-du-lieu/