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 chất lượng rượu vang Vinho Verde**

Giáo viên hướng dẫn: Nguyễn Thị Kim Ngân

Nhóm sinh viên thực hiện: Nhóm 6

1. Nguyễn Thị Mây, lớp 62TH1

2. Nguyễn Thị Tươi, lớp 62TH1

3. Nguyễn Hồng Thương, lớp 62TH1

**Hà Nội, ngày 21 tháng 09 năm 2022**

Mục lục

[Phần 1: Tổng quan 3](#_Toc114867381)

[1.Mô tả bài toán 3](#_Toc114867382)

[2. Phương pháp học máy 3](#_Toc114867383)

[a) Phương pháp Hồi quy tuyến tính: 3](#_Toc114867384)

[b) k-fold cross validation. 4](#_Toc114867385)

[Phần 2: Thực nghiệm 5](#_Toc114867386)

[1.Mô tả tập dữ liệu của bài toán 5](#_Toc114867387)

[2. Phân tích kết quả của chương trình 6](#_Toc114867388)

[3. Kết luận 6](#_Toc114867389)

[4. Tài liệu tham khảo 6](#_Toc114867390)

# Phần 1: Tổng quan

## 1.Mô tả bài toán

* Tên bài toán: Dự đoán chất lượng rượu vang Vinho Verde
* Input:
* Dữ liệu có liên quan đến các biến thể màu trắng của rượu vang “Vinho Verde” của Bồ Đào Nha:

1. Độ axit cố định
2. Độ axit dễ bay hơi
3. Axit citric
4. Đường dư
5. Clorua
6. Lưu huỳnh đioxit tự do
7. Tổng lưu huỳnh đioxit
8. Tỷ trọng
9. pH
10. Sunfat
11. Rượu

* Output:
* Chất lượng rượu (được dự đoán trên thang điểm từ 0 đến 10).
* Tóm tắt công việc:
* Đọc tập dữ liệu của bài toán
* Chia dữ liệu thành 2 phần: 70% cho tập train, 30% cho tập test
* Sử dụng phương pháp k-fold cross validation chia dữ liệu của tập train thành k phần để tìm mô hình tối ưu nhất
* Huấn luyện mô hình và dự đoán nhãn

## 2. Phương pháp học máy

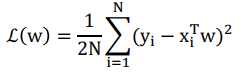
1. Phương pháp Hồi quy tuyến tính:

* là một phương pháp thống kê để hồi quy dữ liệu với biến phụ thuộc có giá trị liên tục trong khi các biến độc lập có thể có một trong hai giá trị liên tục hoặc là giá trị phân loại. Nói cách khác "Hồi quy tuyến tính" là một phương pháp để dự đoán biến phụ thuộc (Y) dựa trên giá trị của biến độc lập (X). Nó có thể được sử dụng cho các trường hợp chúng ta muốn dự đoán một số lượng liên tục.
* Input: tập dữ liệu được gán nhãn
* Output: tìm ra đường tuyến tính (tìm ra các hệ số của đg tuyến tính)
* Cách thực hiện:

x = : vector đặc trưng chứa dữ liệu đầu vào.

w = : vector trọng số.

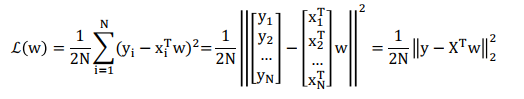
* Xây dựng hàm mất mát:
* Với tất cả các cặp dữ liệu quan sát được (xi,yi), I = 1, 2, …, N, chúng t muốn trung bình sai số là nhỏ nhất.
* Tìm w để hàm số sau nhỏ nhất:



* Cần tối thiểu hàm mất mát theo w. Nghiệm cần tìm của bài toán là:

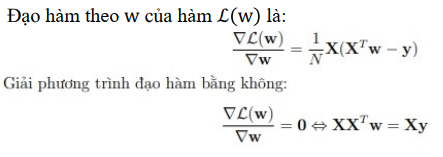


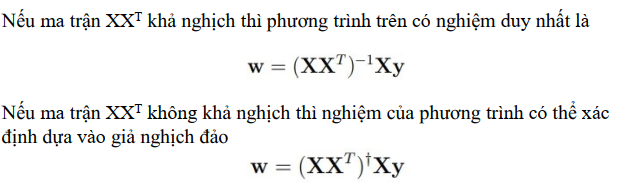
* + Đặt y = [y1,y2,…,yn]T , X = [x1,x2, …, xn], x = =
  + Hàm mất mát được viết lại như sau:



* Tìm nghiệm của bài toán tối ưu:

Tìm giá trị tối ưu của w bằng cách giải phương trình đạo hàm của hàm mất mát theo w bằng 0.





1. k-fold cross validation.

* Input: là vector đặc trưng.
* Output: số thực dương.
* Cách thực hiện:

**Bước 1:** chia tập dữ liệu thành 2 phần: tập dữ liệu huấn luyện và tập dữ liệu test.

**Bước 2:** Chia toàn bộ tập dữ liệu huấn luyện thành k phần (phương pháp k-fold cross validation).

**Bước 3:**

Chọn ngẫu nhiên k-1 phần làm training data, 1 phần còn lại làm test data.

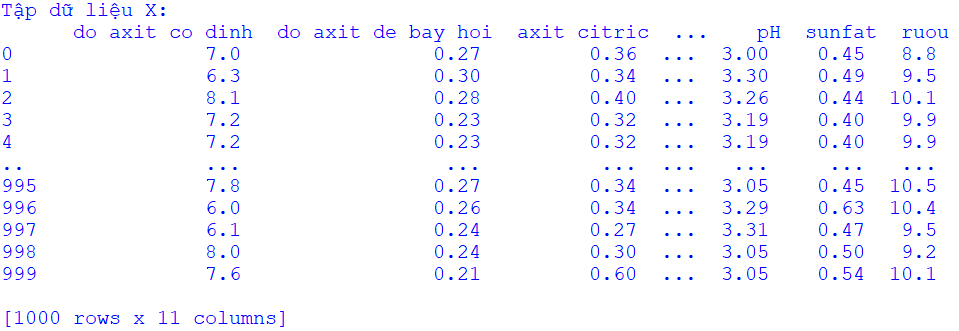
Sử dụng phương pháp học máy đã lựa chọn trên tập training data và test data để xây dựng và đánh giá mô hình (bước này được làm k lần)

**Bước 4:** Chọn mô hình có (train error + validation error) là nhỏ nhất.

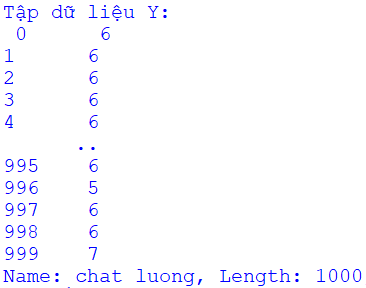
# Phần 2: Thực nghiệm

## 1.Mô tả tập dữ liệu của bài toán

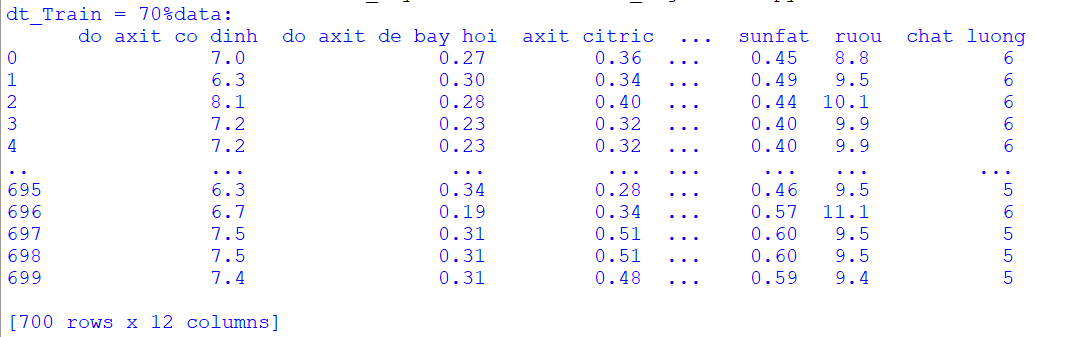
* Số lượng vector dữ liệu(số lượng mẫu rượu Vinho Verde): 1000
* Mỗi vector dữ liệu(mỗi mẫu rượu Vinho Verde) gồm 11 thông tin: Độ axit cố định, Độ axit dễ bay hơi, Axit citric, Đường dư, Clorua, Lưu huỳnh đioxit tự do, Tổng lưu huỳnh đioxit, Tỷ trọng, pH, Sunfat, Rượu.
* Nhãn lớp của dữ liệu (thông tin cần dự đoán) là chất lượng rượu, mỗi vector dữ liệu sẽ tương ứng với 1 giá trị chất lượng rượu (được đánh giá trên thang điểm 0 đến 10).
* Tập dữ liệu X [11,1000] là 1 ma trận dữ liệu đầu vào với mỗi hàng của X là 1 vector dữ liệu gồm các thông tin: Độ axit cố định, Độ axit dễ bay hơi, Axit citric, Đường dư, Clorua, Lưu huỳnh đioxit tự do, Tổng lưu huỳnh đioxit, Tỷ trọng, pH, Sunfat, Rượu.



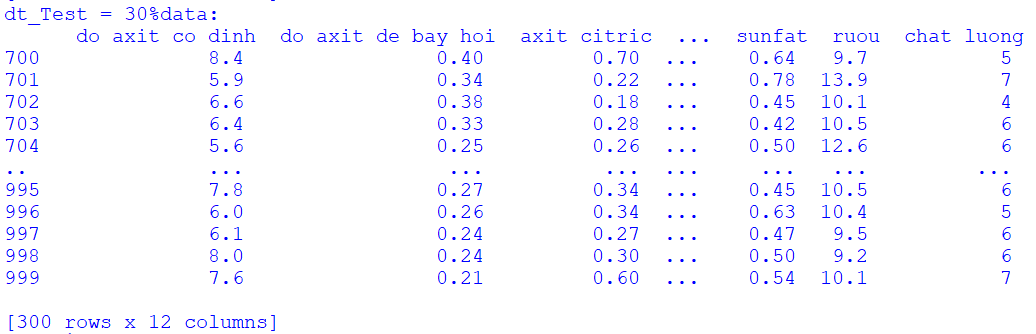
* Nhãn lớp Y là 1 vector có số dòng là 1000 và có 1 cột là “chất lượng”.



* 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.
  + 70% Dữ liệu sd để huấn luyện:



* + 30% Dữ liệu sd để kiểm tra sự phù hợp của mô hình:



## 2. Phân tích kết quả của chương trình

a) thuật toán hồi quy tuyến tính:

* Mức độ phù hợp của mô hình: 21%

b) thuật toán k-fold cross validation: (với giá trị sai khác chấp nhận được <10% giá trị thực tế)

* Tỷ lệ dự đoán đúng: 54%
* Tỷ lệ dự đoán sai: 46%

## 3. Kết luận

* Tóm lược các nội dung chính mà bài tập lớn làm được:

+ thực hiện bài toán dự đoán trên 2 thuật toán hồi quy tuyến tính (linear\_regression) và xác thực chéo (k-fold cross validation).

+ Đưa ra các bước cơ bản nhất để xây dựng 1 mô hình học máy dựa trên các thuật toán.

## 4. Tài liệu tham khảo

* <https://archive.ics.uci.edu/ml/datasets/Wine+Quality>
* <https://machinelearningcoban.com/2016/12/28/linearregression/>