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 KẾT QUẢ CUỐI KỲ**

**MÔN HỌC MÁY CỦA SINH VIÊN DỰA VÀO CÂY QUYẾT ĐỊNH**

Giáo viên hướng dẫn: PGS.TS Nguyễn Quang Hoan

Sinh viên thực hiện:

1. Nguyễn Thị Thanh Mai – 62TH5 - 2051063842

**Hà Nội, năm 2023**

**Mục Lục**

[**Phần 1: Tổng quan** 2](#_Toc124251845)

[**1.Mô tả bài toán** 2](#_Toc124251846)

[**2. Phương pháp học máy** 3](#_Toc124251847)

[**Phần 2: Thực nghiệm** 4](#_Toc124251848)

[**1.Mô tả tập dữ liệu của bài toán** 4](#_Toc124251849)

[**2. Phân tích kết quả của chương trình** 5](#_Toc124251850)

[**Kết luận** 7](#_Toc124251851)

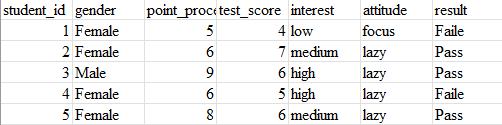
[**Tài liệu tham khảo** 8](#_Toc124251852)

# **Phần 1: Tổng quan**

## **1.Mô tả bài toán**

* Tên bài toán: Dự đoán kết quả cuối kỳ môn học máy của sinh viên dựa vào quá trình học của sinh viên bằng thuật toán cây quyết định
* Input: Các cột dữ liệu: student\_id, gender, point\_process, test\_score, interest, attitude
* Ouput: Cột dữ liệu cuối cùng: result

Ví dụ mẫu minh họa:



* Tóm tắt công việc thực hiện của bài toán:

+ Chuẩn bị dữ liệu: Chia tập dữ liệu Result.csv thành 2 phần training data dùng để huấn luyện mô hình và testing data dùng để kiểm thử hiệu năng của mô hình.

+ Huấn luyện dữ liệu: Sử dụng thư viện

## **2. Phương pháp học máy**

Phương pháp lựa chọn:

* Ý tưởng phương pháp: dùng cây quyết định để dự đoán kết quả cuối kỳ môn học máy của sinh viên
* Cách thực hiện của phương pháp

+ Bước 1: Chuẩn bị dữ liệu, chuẩn hóa dữ liệu và chia dữ liệu thành 2 phần training và testting

+ Bước 2: Huấn luyện

+ Bước 3: Test

**Bước 1**: Tính toán hệ số Entropy của biến mục tiêu S có N phần tử với Nc phần tử thuộc lớp c cho trước:

H(S) : entropy trên toàn bộ tập dữ liệu S

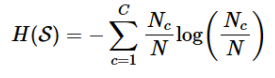
S: tập dữ liệu huấn luyện(gồm cả x và y)

N: tổng số mẫu dữ liệu huấn luyện

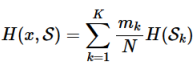
Nc: số các mẫu dữ liệu trong tập S đc gán nhãn c

C: tổng số nhãn dữ liệu khác nhau trong tập S

c: chỉ số của nhãn dữ liệu



**Bước 2**: Tính hàm số Entropy tại mỗi thuộc tính: với thuộc tính x, các điểm dữ liệu trong S được chia ra K child node S1, S2, …, SK với số điểm trong mỗi child node lần lượt là m1, m2 ,…, mK , ta có:



**Bước 3**: Chỉ số Gain Information được tính bằng:

**G(x, S) = H(S) – H(x,S)**

* Ưu điểm của phương pháp:

+ Mô hình sinh ra các quy tắc dể hiểu cho người dùng, tạo ra bộ luật với mỗi nhánh lá là 1 luật của cây

+ Dữ liệu đầu vào có thể là dữ liệu missing, không cần chuẩn hóa hoặc tạo biến giả

+ Có thể làm việc với cả dữ liệu số và dữ liệu phân loại

+ Có thể xác thực mô hình bằng cách sử dụng các kiểm tra thống kê

+ Có khả năng là làm việc với dữ liệu lớn

* Nhược điểm của phương pháp

+ Phụ thuộc rất lớn vào dữ liệu, thậm chí với 1 sự thay đổi nhỏ trong bộ dữ liệu, cấu trúc mô hình có thể thay đổi hoàn toàn

+ Hay gặp vấn đề overfitting

* Cách đánh giá phương pháp: đánh giá mô hình dựa trên các độ đo accurary

# **Phần 2: Thực nghiệm**

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

\* Dữ liệu gồm những chiều thông tin:

- Mỗi mẫu (vertor) dữ liệu có những thông tin bao gồm: student\_id (mã sinh viên), gender (giới tính), point\_process (điểm quá trình), test\_score (điểm kiểm tra), interest (độ hứng thú), attitude (thái độ), result (kết quả)

- Bao gồm: 250 dữ liệu

- Mô tả ma trận dữ liệu (X): là thông tin, nhân tố có thể ảnh hưởng đến kết quả cuối kỳ của sinh viên

Gồm các cột:

+ student\_id(mã sinh viên): dạng số

+ gender (giới tính): dạng chữ (female, male)

+ point\_process (điểm quá trình): dạng số

+ test\_score (điểm kiểm tra): dạng số

+ interest (độ hứng thú): dạng chữ (low, medium, high)

+ attitude (thái độ): dạng chữ (focus, lazy)

+ result (kết quả): dạng chữ (faile, pass)

Mô tả nhãn lớp (Y): là cột result (kết quả) là kết quả học tập cuối kỳ của sinh viênbao gồm faile, pass

* Chia tập dữ liệu thành 2 phần: 70% (175 dòng đầu) dùng để huấn luyện mô hình, 30% (75 dòng cuối) dùng để kiểm thử mô hình.

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

* Mô tả chi tiết cách đánh giá chọn mô hình

+ Khởi tạo biến đếm mô hình ban đầu stt=1, tổng tỷ lệ dự đoán cao nhất của mô hình max = 0. Nếu mô hình có sum>max (là mô hình tốt), thì lưu lại tổng tỷ lệ dự đoán cao nhất của mô hình hiện tại (max=sum)

+ Huấn luyện mô hình tốt nhất hiện tìm được dựa vào hàm fit (đào tạo mô hình và có thể đưa ra dự đoán dựa trên dữ liệu mới)

+ Lưu lại tập train và tập test của mô hình tốt nhất

+ Sau đó tăng stt lên 1 để duyệt mô hình tiếp theo

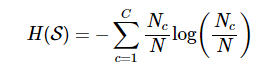
* Ý nghĩa các tham số: entropy

+ thước đo tính ngẫu nhiên của thông tin đang được xử lý. Entropy càng cao, càng khó rút ra bất kỳ kết luận nào từ thông tin đó

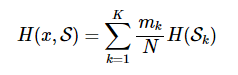
* Cách sử dụng các tham số: entropy

+ GS ta đang làm việc với một non-leaf node với các điểm dữ liệu tạo thành một tập 𝑆 với số phần tử là |𝑆| = 𝑁.

+ Giả sử thêm rằng trong số 𝑁 điểm dữ liệu này, 𝑁𝑐, 𝑐 = 1,2, …, 𝐶 điểm thuộc vào class 𝑐. Xác suất để mỗi điểm dữ liệu rơi vào một class 𝑐 được xấp xỉ bằng 𝑁𝑐 𝑁 (maximum likelihood estimation). Entropy tại node này :



+ GS thuộc tính được chọn là 𝑥. Dựa trên 𝑥, các điểm dữ liệu trong 𝑆 được phân ra thành 𝐾 child node 𝑆1, 𝑆2, …, 𝑆𝐾 với số điểm trong mỗi child node lần lượt là 𝑚1, 𝑚2, …, 𝑚𝐾. Ta định nghĩa:



+ Ta định nghĩa information gain dựa trên thuộc tính x:



- Trong ID3, tại mỗi node, thuộc tính được chọn được xác định dựa trên:



*Điều kiện dừng:*

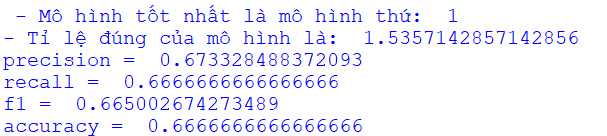
* Để tránh overfitting, một trong số các phương pháp sau có thể được sử dụng. Tại một node, nếu một trong số các điều kiện sau đây xảy ra, ta không tiếp tục phân chia node đó và coi nó là một leaf node:
* Nếu node đó có entropy bằng 0, tức mọi điểm trong node đều thuộc một class.
* Nếu node đó có số phần tử nhỏ hơn một ngưỡng nào đó. Trong trường hợp này, ta chấp nhận có một số điểm bị phân lớp sai để tránh overfitting. Class cho leaf node này có thể được xác định dựa trên class chiếm đa số trong node.
* Nếu khoảng cách từ node đó đến root node đạt tới một giá trị nào đó. Việc hạn chế chiều sâu của tree này làm giảm độ phức tạp của tree và phần nào giúp tránh overfitting.
* Nếu tổng số leaf node vượt quá một ngưỡng nào đó.
* Nếu việc phân chia node đó không làm giảm entropy quá nhiều (information gain nhỏ hơn một ngưỡng nào đó).
* Kết quả đánh giá (bảng tổng hợp).

|  |  |  |
| --- | --- | --- |
| Lần lặp | Dùng thư viện | |
| train accuracy | validation accuracy |
| Lần 1 | 0.878 | 0.657 |
| Lần 2 | 0.878 | 0.628 |
| Lần 3 | 0.878 | 0.628 |
| Lần 4 | 0.921 | 0.571 |
| Lần 5 | 0.9 | 0.485 |

* Mô tả các đánh giá hiệu năng của mô hình đã được chọn.

+ Tỉ lệ đúng của mô hình là: 1.5

* Dùng độ đo: accuracy
* Cách tính toán: Tỉ lệ giữa số điểm được dự đoán đúng và tổng số điểm trong tập dữ liệu kiểm thử
* Kết quả:



# **Kết luận**

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

+ Khai báo thư viện

+ Đọc dữ liệu

+ Chuẩn hóa dữ liệu: đưa dữ liệu dạng chuỗi về dạng số

+ Chia dữ liệu thành 2 phần: 70% train, 30% test

+ Đánh giá dữ liệu bằng cross-validation: chia dữ liệu thành 5 phần (k=5)

+ Duyệt trên từng mô hình chia được để tìm ra mô hình tốt nhất

+ Thực hiện tính giá trị dự đoán trên dữ liệu dt\_Test

+ Đánh giá mô hình trên độ đo Accuracy

* Khó khăn hạn chế, hướng phát triển

+ Khó khăn trong việc tìm tài liệu

+ Khó khăn trong việc code

* Hướng phát triển

+ Đọc kỹ hiểu rõ đề tài dữ liệu của nhóm

+ Đọc và hiểu sâu các thuật toán các phương pháp cần dùng trong bài toán

# **Tài liệu tham khảo**

<https://machinelearningcoban.com/2018/01/14/id3/>

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