**NGHIÊN CỨU HỌC MÁY**

**I. PHƯƠNG PHÁP RIDGE REGRESSION CHO BÀI TOÁN DỰ ĐOÁN**

**1.1 Một số khái niệm cần làm rõ trước.**

- Học máy (ML – Machine Learning): Học máy là lĩnh vực của trí tuệ nhân tạo (AI) tập trung vào việc phát triển các thuật toán và kỹ thuật cho phép máy tính học từ dữ liệu mà không cần được lập trình cụ thể.

- Học giám sát (SL – Supervised Learning): Học giám sát là một phương pháp học máy, trong phương pháp này sẽ ứng dụng các nhóm thuật toán giám sát như hồi quy, cây quyết định, … để xây dựng, huấn luyện mô dự đoán dựa trên tập dữ liệu huấn luyện (training set) để có thể tự đưa ra dự đoán với độ chính xác cao.

- Tập dữ liệu huấn luyện (trainning set) là tập dữ liệu được chuẩn bị có gán nhãn sẵn (nhãn là mục tiêu cần dự đoán) để mô hình có thể tự tìm hiểu quy luật và học theo.

- Tập dữ liệu kiểm tra (testing set) là tập dữ liệu không có nhãn được dùng để kiểm thử, đánh giá mô hình.

- Overfitting: xảy ra khi mô hình học máy quá phức tạp và quá dễ nhớ dữ liệu huấn luyện, dẫn đến việc hiệu suất dự đoán giảm đi khi đối mặt với dữ liệu mới. Điều này thường xảy ra khi mô hình quá sát với dữ liệu huấn luyện, bao gồm cả noise và biến động không quan trọng.

- Underfitting: xảy ra khi mô hình quá đơn giản và không thể nắm bắt được cấu trúc của dữ liệu. Khi mô hình underfit, nó không thể đưa ra dự đoán chính xác trên cả dữ liệu huấn luyện và dữ liệu mới.

- Regularization: là một kỹ thuật được sử dụng để giảm overfitting bằng cách thêm một thuật toán hoặc hạt nhân vào quá trình huấn luyện mô hình. Các phương pháp regularization bao gồm L1 (Lasso) và L2 (Ridge) regularization.

- Hàm mất mát (loss function): là một khái niệm quan trọng trong học máy, được sử dụng để đo lường sự khác biệt giữa giá trị dự đoán của mô hình và giá trị thực tế của dữ liệu huấn luyện. Mục tiêu của việc huấn luyện mô hình là tối thiểu hóa hàm mất mát, tức là làm cho dự đoán của mô hình gần giống với giá trị thực tế nhất có thể.

- Cross-validation: là một kỹ thuật đánh giá hiệu suất của mô hình bằng cách chia tập dữ liệu thành các phần nhỏ, huấn luyện mô hình trên một phần và đánh giá trên phần còn lại. Điều này giúp đánh giá khả năng tổng quát hóa của mô hình.

**1.2 Ridge Regression**

Ridge Regression, còn được gọi là L2 regularization, là một phần mở rộng của hồi quy tuyến tính, nó mang đến thuật ngữ chính quy hóa để giảm độ phức tạp của mô hình và giúp ngăn chặn việc overfitting.

Một cách dễ hiểu, Ridge Regression giúp giảm thiểu tổng bình phương của phần dư và giá trị bình phương của tham số được chia tỷ lệ theo một hệ số (lambda hoặc α). Thuật ngữ chính quy hóa này, λ, kiểm soát cường độ ràng buộc đối với các hệ số và hoạt động như một tham số điều chỉnh.

Regularization là một kỹ thuật được sử dụng trong học máy để penalize các mô hình phức tạp nhằm bảo vệ chúng khỏi bị trang bị quá mức.

A blue background with white text

Description automatically generated

\* Ridge Regression trong Sckit-Learn:

<https://scikit-learn.org/stable/modules/generated/sklearn.linear_model.Ridge.html>

**II. PHÂN TÍCH CÁC YẾU TỐ QUYẾT ĐỊNH ĐẾN ĐIỂM THI THỰC TẾ**

**2.1 Phân tích các yếu tố con người**

Đối tượng nghiên cứu của đề tài tập trung vào đối tượng học sinh, bao gồm các thí sinh đã tham gia luyện thi trên trang web. Phân loại nhóm đối tượng này sẽ giúp chúng ta hiểu rõ hơn về đặc điểm và yếu tố ảnh hưởng đến hiệu suất học tập và điểm thi của sinh viên.

*2.1.1. Độ Tuổi và Trình Độ Học Vấn:*

*- Độ Tuổi:* Thí sinh trong nhóm này thường có độ tuổi từ 16 đến 18 tuổi, tương ứng với độ tuổi học trung học phổ thông ở Việt Nam. Việc hiểu rõ độ tuổi của sinh viên giúp chúng ta đánh giá được mức độ trưởng thành và nhu cầu học tập của họ.

*- Trình Độ Học Vấn:* Phần lớn thí sinh trong nhóm này là học sinh cấp ba, đang chuẩn bị cho kỳ thi THPT Quốc gia. Trình độ học vấn này có thể ảnh hưởng đến cách tiếp cận và hiệu quả học tập của sinh viên.

*2.1.2. Mục Tiêu Học Tập:* mục tiêu đặt ra của sĩ tử có thể ảnh hưởng đến điểm số của chính sĩ tử đó. Là một bước quan trọng trong hành trình học tập của họ.Mục tiêu này thường tạo ra áp lực lớn và tăng động lực cho sĩ tử trong quá trình ôn tập.

*2.1.3. Thói Quen Học Tập:* Thí sinh có thói quen học tập, ôn luyện đúng cách sẽ giúp cải thiện thành tích đáng kể. Thu thập thông tin về thói quen học của thí sinh (về thói quen đọc sách, thói quen giải đề, thói quen tự học, …) sẽ giúp đưa ra dự đoán chính xác hơn.

**2.2 Các yếu tố tương quan ảnh hưởng đến điểm thi**

Có nhiều yếu tố ảnh hưởng trực tiếp đến điểm thi của mỗi thí sinh, việc đạt được thành tích thế nào phản ánh sự cố gắng luyện tập của chính thí sinh đó.

Trên trang Web luyện thi sẽ tập trung phân tích lịch sử ôn luyện của thí sinh đó trên chính trang web để đưa ra dự đoán cho thí sinh đó.

Các dữ liệu ôn luyện của người dùng sẽ được lưu lại để phục vụ quá trình thu thập dữ liệu “Testing Set” cho mô hình để đưa ra dự đoán cho thí sinh.

Ta có danh sách các yếu tố ảnh hưởng đến điểm thi dự đoán như bảng dưới:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | **Trường** | **Giải thích** | **Hệ số** | **Note** |
| 1 | Age | Tuổi của thí sinh |  |  |
| 2 | Grade | Khối lớp hiện tại |  |  |
| 3 | Class | Khối thi |  |  |
| 4 | TimeLearnPerDay | Thời gian tự học hàng ngày |  |  |
| 5 | TimeOnPerDay | Thời gian trên web hàng ngày |  |  |
| 6 | TotalExam | Tổng số đề đã giải |  |  |
| 7 | TotalExamSubject1 | Tổng số đề môn thi 1 đã giải |  |  |
| 8 | TotalExamSubject2 | Tổng số đề môn thi 2 đã giải |  |  |
| 9 | TotalExamSubject3 | Tổng số đề môn thi 3 đã giải |  |  |
| 10 | AvgExamPerDay | Số đề giải mỗi ngày |  |  |
| 11 | AvgScoreExamSubject1 | Điểm trung bình giải đề môn thi 1 |  |  |
| 12 | AvgScoreExamSubject2 | Điểm trung bình giải đề môn thi 2 |  |  |
| 13 | AvgScoreExamSubject2 | Điểm trung bình giải đề môn thi 3 |  |  |
| 14 | MinScoreExamSubject1 | Điểm thấp nhất giải đề môn thi 1 |  |  |
| 15 | MinScoreExamSubject2 | Điểm thấp nhất giải đề môn thi 2 |  |  |
| 16 | MinScoreExamSubject3 | Điểm thấp nhất giải đề môn thi 3 |  |  |
| 17 | MaxScoreExamSubject1 | Điểm cao nhất giải đề môn thi 1 |  |  |
| 18 | MaxScoreExamSubject2 | Điểm cao nhất giải đề môn thi 2 |  |  |
| 19 | MaxScoreExamSubject3 | Điểm cao nhất giải đề môn thi 3 |  |  |
| 20 | FirstExamScore | Điểm đề giải đầu tiên |  |  |
| 21 | LastExamScore | Điểm giải đề gần đây nhất |  |  |
| 22 | AvgExamTime | Thời gian giải đề trung bình |  |  |
| 23 | AvgExamTimeSubject1 | Thời gian giải đề trung bình môn thi 1 |  |  |
| 24 | AvgExamTimeSubject2 | Thời gian giải đề trung bình môn thi 2 |  |  |
| 25 | AvgExamTimeSubject3 | Thời gian giải đề trung bình môn thi 3 |  |  |
| 26 | ExamAccuracy | Tỉ lệ giải đề đúng |  |  |
| 27 | TotalLateSubmissions | Số lần nộp bài muộn |  |  |
| 28 | LastMonthExamScore | Điểm giải đề trong 1 tháng gần đây |  |  |
| 29 | LastWeekExamScore | Điểm giải đề trong 1 tuần gần đây |  |  |
| 30 | ExamCompletionRate | Tỉ lệ hoàn thành giải đề |  |  |
| 31 | TimeOnLessonPage | Tổng thời gian trên trang lesson |  |  |
| 32 | TimeOnLessonPagePerDay | Thời gian trên trang lesson hằng ngày |  |  |
| 33 | TimeOnLessonInClass | Thời gian trên trang lesson của môn trong khối thi |  |  |
| 34 | TimeOnLessonOutClass | Thời gian trên trang lesson của môn ngoài khối thi |  |  |
| 35 | TotalPractice | Tổng bài tập đã làm |  |  |
| 36 | TotalQuestionDone | Tổng số câu đã làm |  |  |
| 37 | PracticeAccuracy | Tỉ lệ làm bài tập đúng |  |  |
| 38 | TotalQuestionInClass | Tổng số câu trong khối thi |  |  |
| 39 | TotalQuestionCorrect | Tổng số câu làm đúng |  |  |
| 40 | TotalArena | Tổng số phòng thi đã tham gia |  |  |
| 41 | AvgArenaScore | Điểm trung bình khi tham gia phòng thi |  |  |
| 42 | MinArenaScore | Điểm thấp nhất khi tham gia phòng thi |  |  |
| 43 | MaxArenaScore | Điểm cao nhất khi tham gia phòng thi |  |  |
| 44 | FirstArenaScore | Điểm lần đầu tham gia phòng thi |  |  |
| 45 | LastArenaScore | Điểm lần tham gia phòng thi gần nhất |  |  |
| 46 | LastWeekArenaScore | Điểm trung bình tham gia phòng thi tuần gần nhất |  |  |
| 47 | TotalArenaLastWeek | Số lần tham gia phòng thi tuần gần nhất |  |  |
| 48 | ArenaCompletionRate | Tỉ lệ hoành thành phòng thi |  |  |
| 49 | GoalScore | Điểm số đặt ra |  |  |
| 50 | GoalTime | Thời gian luyện tập đặt ra |  |  |

**2.3 Cách thức phân tích yếu tố quyết định đến điểm thi thực tế**

\* Ứng dụng phương pháp Ridge Regression cho bài toán

- Hàm dự đoán: `predict\_score = w1 \* Age + w2 \* Grade + w3 \* TimeLearnPerDay + ... + wn \* GoalScore + b`

Trong đó:

* + predict\_score là điểm thi dự đoán của học sinh.
  + w1, w2, ..., wn là các hệ số của các trường dữ liệu cần học.
  + b là thuật ngữ chặn.
  + Age, Grade, TimeLearnPerDay, ..., GoalScore là các thuộc tính dùng để huấn luyện mô hình.

- Hàm mục tiêu: J(w, b) = 1/2 \* sum((y - predict\_score)^2) + lambda \* sum(w^2)

Điểm khác biệt chính giữa Ridge Regression và Hồi Quy Tuyến Tính nằm ở hàm mục tiêu. Hàm mục tiêu của Ridge Regression có thêm thành phần phạt L2 (regularization term) để hạn chế các hệ số quá lớn, giúp giảm thiểu overfitting:

J(w, b) là hàm mục tiêu cần tối ưu hóa.

y là điểm thi thực tế của thí sinh thu được từ hàm dự đoán

lambda là tham số điều chỉnh mức độ phạt.

sum(w^2) là tổng bình phương của các hệ số.

\* Quá trình học Ridge Regression tương tự như Hồi Quy Tuyến Tính, bao gồm các bước sau:

B1. Khởi tạo: Khởi tạo các hệ số w và b với giá trị ban đầu (thường là giá trị ngẫu nhiên nhỏ).

B2. Tính toán điểm thi dự đoán: Sử dụng công thức dự đoán để tính toán điểm thi dự đoán predict\_score cho mỗi học sinh trong tập dữ liệu.

B3. Tính toán lỗi dự đoán: Tính toán lỗi dự đoán (y - predict\_score)^2 cho mỗi học sinh.

B4. Tính toán giá trị hàm mục tiêu: Sử dụng công thức hàm mục tiêu để tính toán giá trị J(w, b) cho tập dữ liệu.

B5. Cập nhật hệ số: Sử dụng thuật toán tối ưu hóa (như gradient descent) để điều chỉnh các hệ số w và b theo hướng giảm thiểu giá trị hàm mục tiêu.

B6. Lặp lại: Lặp lại bước 2 đến 5 cho đến khi đạt được độ hội tụ (tức là giá trị hàm mục tiêu không thay đổi đáng kể) hoặc đạt đến số lần lặp tối đa.

**III. PHÂN TÍCH CÁC YẾU TỐ TÁC ĐỘNG ĐẾN HIỆU SUẤT ÔN TẬP**

**IV. PHƯƠNG PHÁP BIỂU DIỄN TRI THỨC CHO VIỆC DỰ ĐOÁN**

**4.1 Mô hình biểu diễn các yếu tố tương quan**

**4.2 Cơ sở tri thức cho bài toán dự đoán**

**4.3 Tổ chức cơ sở tri thức**

**V. THIẾT KẾ VÀ THỬ NGHIỆM**

**5.1 Thiết kế mô hình**

**5.2 Thử nghiệm & đánh giá**