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 DOANH SỐ DỰA TRÊN CÁC PHƯƠNG PHÁP QUẢNG CÁO**

Sinh viên/nhóm sinh viên thực hiện:

1. Nguyễn Quốc Dũng, 2151062737

2. Từ Khánh Linh, 2151061191

3. Vũ Hoàng Hiệp,2151062764

4. Nguyễn Tiến Anh, 2151062708

**Hà Nội, năm 2023**

**Phần 1: Tổng quan**

***1.Mô tả bài toán***

* Tên bài toán: dự đoán doanh thu dựa trên các phương pháp quảng cáo
* Input: bảng thống kê số liệu sản phẩm được quảng cáo theo từng phương pháp
* Ouput: doanh số thu được (sale), chỉ số độ đo R2, NSE cho từng thuật toán
* Tóm tắt công việc thực hiện của bài toán.: sử dụng 4 phương pháp hồi quy tuyến tính, Lasso, Ridge và k-fold cross validation để dự đoán doanh số bán hàng

2***. Phương pháp học máy***

* Trình bày lý thuyết của 2 phương pháp Hồi quy tuyến tính, Lasso, Ridge và k-fold cross validation.
* **Lineer regression**: "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
* **Lasso**: Lasso Regression (Least Absolute Shrinkage and Selection Operator Regression) là một phương pháp hồi quy tuyến tính được sử dụng trong thống kê và học máy để giảm quá khớp (overfitting) và thực hiện lựa chọn biến đầu vào (feature selection). Lasso là một biến thể của hồi quy tuyến tính thông thường, nhưng nó sử dụng một loại regularization được gọi là L1 regularization. Trong Lasso Regression, mục tiêu là tối ưu hóa hàm mất mát (loss function) bao gồm hai thành phần chính:

+ **Hàm mất mát bình phương** (Squared Loss): Như trong hồi quy tuyến tính thông thường, hàm mất mát này đo sự sai lệch giữa dự đoán của mô hình và giá trị thực tế. Mục tiêu là làm cho sai số này nhỏ nhất.

+ **L1 Regularization Term:** Đây là thành phần quan trọng của Lasso. Nó được sử dụng để kiểm soát sự phức tạp của mô hình bằng cách thêm tổng giá trị tuyệt đối của các trọng số (biến đầu vào) vào hàm mất mát.

* **Ridge:** Thuật toán Ridge Regression là một phương pháp hồi quy tuyến tính được sử dụng trong thống kê và học máy để giảm quá khớp (overfitting) trong mô hình hồi quy tuyến tính thông thường. Ridge Regression là một biến thể của hồi quy tuyến tính, nhưng nó sử dụng một loại regularization được gọi là L2 regularization để kiểm soát mức độ phức tạp của mô hình. Điểm chính của Ridge Regression là thêm một thành phần L2 regularization vào hàm mất mát của hồi quy tuyến tính thông thường.
* **K-fold:** Phương pháp "k-fold cross-validation" là một kỹ thuật phân chia tập dữ liệu thành các tập con riêng biệt để đánh giá hiệu suất của một mô hình máy học hoặc thuật toán. Kỹ thuật này giúp đánh giá khái quát khả năng tổng quan của mô hình và tránh tình trạng overfitting hoặc underfitting. Cách thực hiện k-fold cross-validation:

**+** Chia dữ liệu: Bước đầu tiên là chia tập dữ liệu thành k phần con (fold) có kích thước tương đối bằng nhau. Các phần con này thường không giao nhau và tạo thành một phân chia ngẫu nhiên của dữ liệu.

+ Lặp lại quá trình đánh giá: Tiếp theo, mô hình hoặc thuật toán sẽ được đào tạo và đánh giá k lần. Trong mỗi lần, một phần con trong tập dữ liệu được sử dụng để đánh giá mô hình, trong khi các phần còn lại được sử dụng để đào tạo mô hình.

+ Tính toán kết quả: Sau mỗi lần lặp, kết quả đánh giá (như độ chính xác, sai số, F1-score, và nhiều thước đo khác) được tính toán và lưu lại. Các kết quả này sau đó có thể được trung bình hoặc kết hợp để tạo ra một đánh giá tổng quan về hiệu suất của mô hình.

**Phần 2: Thực nghiệm**

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

* Dữ liệu gồm 200 dòng gồm 4 cột: TV, Radio, Newspaper, Sales
* Ma trận dữ liệu (X) gồm: 3 cột TV, Radio, Newspaper ,
* Nhãn lớp (Y) là cột Sales

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

***-*** Đối với mỗi trường hợp dữ liệu khác nhau, các thuật toán sẽ cho ra những hiệu quả nhất định. Thuật toán linear regression sẽ hợp hơn đối với những dữ liệu có dung lượng thấp, tốc độ nhanh. Trong khi đó K- fold lại cho kết quả chính xác hơn và phù hợp hơn cho những dự đoán với số lượng dữ liệu lớn

**Kết luận**

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

+ Củng cố kiến thức về 4 thuật toán hồi quy tuyến tính

+ So sánh độ hiệu quả của các thuật toán trong một trường hợp nhất định

**Tài liệu tham khảo**

* Chat GPT
* Studocu
* Youtube chanel VjetGolang