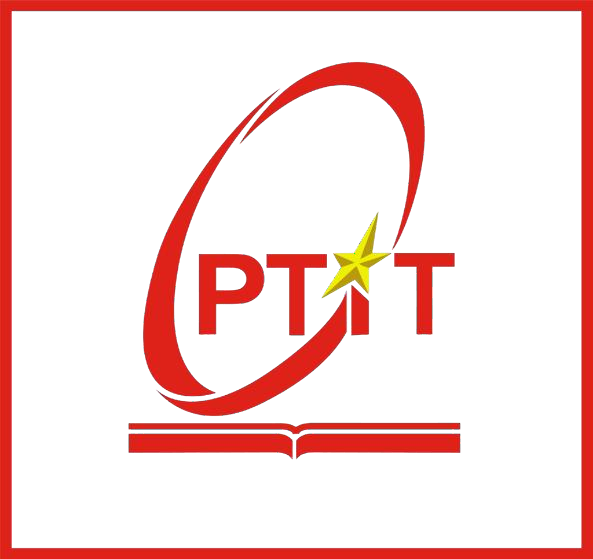
**HỌC VIỆN CÔNG NGHỆ BƯU CHÍNH VIỄN THÔNG**

**KHOA CÔNG NGHỆ THÔNG TIN 2**

----------



**BÁO CÁO CUỐI KỲ**

***Môn học: Nhập môn khoa học dữ liệu***

***Giảng viên hướng dẫn:*** ***Tân Hạnh***

***Sinh viên thực hiện: Trần Ngọc Khánh Văn***

***Mã sinh viên: N20DCCN084***

***Lớp: D20CQCNHT01-N***

**Hồ Chí Minh, 2024**

*LỜI MỞ ĐẦU*

Em xin gửi lời cảm ơn sâu sắc và chân thành tới thầy Tân Hạnh vì đã có những buổi dạy rất hay và giúp em học được thêm nhiều kiến thức có ích. Trong thời gian học với thầy em đã nhận được sự dạy dỗ, hướng dẫn nhiệt tình của thầy làm cho em có cái nhìn khách quan hơn cũng như là hiểu rõ hơn về khoa học dữ liệu ngành mà em đang muốn theo đuổi trong hiện tại. Em xin chân thành cảm ơn!

Em xin gửi lời cảm ơn cảm ơn ban lãnh đạo Học Viện Công nghệ Bưu Chính Viễn Thông cơ sở tại thanh phố Hồ Chí Minh đã dành thời gian quý báu của mình để trả lời các phiếu trắc nghiệm, tìm kiếm và cung cấp tư liệu, tư vấn, giúp đỡ chúng tôi hoàn thành bài tập cuối kỳ này.

***Hồ Chí Minh, ngày 12 tháng 01 năm 2024***

# Phần I: Nội dung thực hiện

1. ***Phương pháp hồi quy tuyến tính đa biến***

Phương pháp hồi quy tuyến tính đa biến hay còn được gọi là Multiple Regression Linear 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 được phát triển từ phưởng pháp hồi quy tuyến tính (Linear Regression).

Trên thực tế phương pháp hồi quy tuyến tính đa biến được dùng trong máy học để phân tích các tập dữ liệu lớn. 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.

Trong phân tích hồi quy tuyến tính đa biến, tập dữ liệu chứa một biến phụ thuộc và nhiều biến độc lập. Hàm đường hồi quy tuyến tính thay đổi để bao gồm nhiều yếu tố như sau:

Y= β0\*X0 + β1X1 + β2X2+…… βnXn+ ε

Với:

Y là biến phụ thuộc

X0, X1, X2,….., Xn là các biến độc lập

β0, β1, β2,….., βn là các tham số

ε là các biến ngẫu nhiên gọi là số hạng sai số

1. ***Phương pháp hồi quy logistics***

Phương pháp hồi quy logistics hay còn được biết đến là Logistic Regression là một kỹ thuật phân tích dữ liệu sử dụng toán học để tìm ra mối quan hệ giữa hai yếu tố dữ liệu. Sau đó, kỹ thuật này sử dụng mối quan hệ đã tìm được để dự đoán giá trị của những yếu tố đó dựa trên yếu tố còn lại. Dự đoán thường cho ra một số kết quả hữu hạn, như có hoặc không.

Hồi quy logistic là một kỹ thuật quan trọng trong lĩnh vực trí tuệ nhân tạo và máy học (AI/ML). Mô hình ML là các chương trình phần mềm có thể được đào tạo để thực hiện các tác vụ xử lý dữ liệu phức tạp mà không cần sự can thiệp của con người. Mô hình ML được xây dựng bằng hồi quy logistic có thể giúp các tổ chức thu được thông tin chuyên sâu hữu ích từ dữ liệu kinh doanh của mình. Họ có thể sử dụng những thông tin chuyên sâu này để phân tích dự đoán nhằm giảm chi phí hoạt động, tăng độ hiệu quả và đổi chỉnh quy mô nhanh hơn. Ví dụ: doanh nghiệp có thể khám phá các mẫu hình cải thiện khả năng giữ chân nhân viên hoặc tạo ra thiết kế sản phẩm mang về nhiều lợi nhuận hơn.

Một giải pháp để phân loại là hồi quy logistic. Thay vì điều chỉnh một đường thẳng hoặc siêu phẳng, mô hình hồi quy logistic sử dụng hàm logistic để ép đầu ra của một phương trình tuyến tính giữa 0 và 1. Hàm logistic được định nghĩa là:

Logistic()

1. ***Cách chia tập dữ liệu***
   1. *Tại sao nên chia tập dữ liệu*

Trước khi tìm hiểu về cách chia tập dữ liệu thì ta cần nói đến hai vẫn đề quan trọng trong máy học đó là overfitting và underfitting. Với các mô hình underfitting cho kết quả tệ trên cả dữ liệu quan sát sử dụng để huấn luyện lẫn dữ liệu chưa được quan sát trong thực tế, còn overfitting cho kết quả rất (quá) tốt trên dữ hiệu huấn luyện nhưng lại quá kém khi sử dụng trong thực tế.

Trong machine learning, chúng ta cần kiểm thử để dự đoán khả năng hoạt động hiệu quả của mô hình trên thực tế. Có rất nhiều cách để kiểm thử và đánh giá hiệu năng của mô hình. Cách đơn giản nhất là dùng thử nhưng điều này mang lại rủi ro và tốn kém quá cao. Vì vậy cách chia tập dữ liệu ra là cách hiệu quả nhất.

* 1. Cách chia tập dữ liệu

Tập huấn luyện là tập dữ lệu dùng cho việc huấn luyện mô hình. Các thuật toán máy học sẽ học các mô hình từ tập này.

Tập kiểm thử chính là tập dữ liệu dùng cho việc xác định độ chính xác hoặc sai số của mô hình máy đã huấn luyện từ tập dữ liệu huấn luyện. Chúng ta biết nhãn thực của mọi điểm trong tập hợp dữ liệu kiểm thử này, nhưng chúng ta sẽ tạm thời giả vờ như không biết và đưa các giá trị đầu vào của tập vào mô hình dự đoán để nhận kết quả dự đoán đầu ra. Sau đó chúng ta có thể nhìn vào các nhãn thực và so sánh nó với kết quả dự đoán của các đầu vào tương ứng này và xem liệu mô hình có dự đoán đúng hay không. Việc tính tổng trung bình của toàn bộ các lỗi này chúng ta có thể tính toán được lỗi dự đoán trên tập kiểm thử.

Có 3 cách chia dữ liệu thường được sử dụng đó là:

- Hold-out

-Train-Validation-Test Split

- Cross Validation

Ở bài tập này mục đích của chúng ta là kiểm chứng mô hình được huấn luyện vậy nên phương pháp chia **hold-out** chính là phương pháp phù hợp nhất. Theo phương pháp hold-out chúng ta sẽ dùng 80% tập dữ liệu để huấn luyện và 20% còn lại dùng để kiểm thử vì đây là cách chia hợp lý nhất. Nếu tập kiểm thử quá nhỏ thì việc kiểm thử có thể không đáng tin cậy và có độ bất ổn cao. Dựa trên mức độ “may mắn” khi chia dữ liệu mà các độ đo này có thể nhảy lên cao hoặc rất thấp

# Phần II: Chương trình Python thực hiện việc huấn luyện và kiểm thử