TRƯỜNG ĐẠI HỌC THỦY LỢI

KHOA CÔNG NGHỆ THÔNG TIN

NGHIÊN CỨU KHOA HỌC

ỨNG DỤNG MÔ HÌNH HỒI QUY TUYẾN TÍNH TRONG DỰ BÁO DÒNG CHẢY TẠI TÂN CHÂU HỖ TRỢ CẢNH BÁO LŨ LỤT

Giảng viên hướng dẫn: **PGS.TS Nguyễn Thanh Tùng**

Nhóm sinh viên thực hiện:

Nguyễn Danh Thành-1851061855

Phùng Hữu Hưởng-1851061398

Hà Nội, tháng 5 năm 2021

TÓM TẮT ĐỀ TÀI

Biến đổi khí hậu ngày càng nghiêm trọng dẫn đến các sự kiện thời tiết cực đoan như hạn hán, cháy rừng, bão, lũ lụt,.. diễn ra với tần suất tăng dần và không thể lường trước. Đặc biệt là các sự kiện lũ lụt diễn ra gần đây tại sông Trường Giang (Trung Quốc, tháng 7-2020), Hà Tĩnh (tháng 7-2020), khu vực miền núi phía Bắc (19-8-2020), miền Trung (Huế tháng 10-2020) đã yêu cầu cấp bách một giải pháp dự báo, cảnh báo, thông báo nguy cơ xảy ra lũ lụt tại địa phương.

Hiện nay, có nhiều phương pháp dự báo khác nhau như dự báo bằng hệ chuyên gia, dự báo bằng phương trình hồi quy, dự báo bằng chuỗi thời gian… Nhưng dự báo bằng phương pháp hồi quy tuyến tính được ứng dụng rộng rãi trong nhiều lĩnh vực nhất là kinh doanh và y học, dự báo tình hình các hiện tượng thiên nhiên như lũ lụt và hạn hán, nó là cơ sở khoa học rõ ràng và mang lại kết quả quan trọng với độ chính xác cao. Mô hình hồi quy tuyến tính đưa ra các phương pháp ước lượng, kiểm định giả thiết và dự báo. Thuật ngữ "*Hồi quy*" được nhà nghiên cứu *Francis Galton* sử dụng lần đầu tiên vào cuối thế kỉ XIX trong một nghiên cứu "Tại sao có sự ổn định chiều cao trung bình của dân số". Từ đó trở đi, vấn đề hồi quy được quan tâm nhiều hơn và được nghiên cứu sâu hơn. Trong đó mô hình hồi quy tuyến tính được xem là nền tảng, là cơ sở để xây dựng các đường hồi quy khác. Để hiểu rõ hơn về mô hình hồi quy cụ thể là mô hình hồi quy đa biến và ứng dụng trong dự báo chuỗi thời gian, chúng tôi đã lựa chọn đề tài "*Ứng dụng mô hình hồi quy tuyến tính trong dự báo dòng chảy tại Tân Châu hỗ trợ cảnh báo lũ lụt*".

CÁC MỤC TIÊU CHÍNH

1. Tìm hiểu về mô hình hồi quy tuyến tính (LR - Linear Regression)
2. Tìm hiểu về phần mềm và ngôn ngữ Python
3. Ứng dụng mô hình hồi quy tuyến tính trong dự báo dòng chảy tại Tân Châu hỗ trợ cảnh báo lũ lụt

KẾT QUẢ DỰ KIẾN

1. *Mô hình học máy sử dụng Python dự báo dòng chảy tại Tân Châu hỗ trợ cảnh báo lũ lụt bằng phương pháp hồi quy tuyến tính.*
2. *Viết báo cáo và tổng kết.*

Chương 1 Giới thiệu chung

I. Lý do chọn đề tài

Lũ lụt là một hiện tượng thời tiết phức tạp, để lại nhiều thiệt hại nặng nề về tài sản và tính mạng tại khu vực trực tiếp chịu thiên tai. Theo thống kê của Tổ chức Hợp tác và Phát triển Kinh tế (Organization for Economic Co-operation and Development – OECD), mỗi năm toàn thế giới chịu thiệt hại hơn 40 tỉ đô la Mỹ, ảnh hưởng đến xấp xỉ 250 triệu người, tần suất xuất hiện lũ lụt đã tăng gần gấp đôi trong giai đoạn 2000-2009 so với thập kỷ trước đó và số lần lũ lụt trong khoảng thời gian 2010 đến 2013 nhiều hơn tổng số lần lũ lụt của cả thập niên 80. Việt Nam là một trong những nước có tần suất lũ lụt cao; các khu vực vùng Đồng bằng sông Cửu Long, Đồng bằng duyên hải miền Trung, Đồng bằng Bắc Bộ nhiều năm gần đây chứng kiến nhiều trận lũ lịch sử đỉnh điểm là trận lũ miền Trung diễn ra vào tháng 10 và 11 của năm 2020.

Vì thế việc xác định dòng chảy và đánh giá về ngập lụt kịp thời và chính xác ở Việt Nam để có thể hỗ trợ các nỗ lực cứu trợ và phục hồi thiên tai cũng như phục vụ công tác thống kê những địa phương chịu ảnh hưởng là rất cần thiết và cấp bách. Xây dựng mô hình dự báo dòng chảy đáp ứng tối ưu việc cập nhật, xác định và đánh giá về lũ lụt xảy ra trong suốt thời gian lũ lụt từ đó có những giải pháp cứu trợ, tiếp cận kịp thời khi xuất hiện những trận lũ lụt gây ra hơn là các đánh giá dự báo thông thường. Với sự phát triển của lĩnh vực công nghệ thông tin, trí tuệ nhân tạo mà đặc biệt trong lĩnh vực học máy, học sâu những năm gần đây đang thực sự được quan tâm và ứng dụng vào thực tiễn các lĩnh vực, ngành nghề trong cuộc sống. Bài toán xây dựng mô hình hồi quy tuyến tính dự báo dòng chảy là một minh chứng cho sự phát triển của công nghệ thông tin dựa trên các phương pháp, thuật toán đã có thể xác định, cập nhật tính ưu việt trong công tác dự báo, đánh giá với mức độ chính xác cao nhằm góp phần thiết thực vào thực tiễn trong lĩnh vực quản lý tài nguyên nước và phòng chống thiên tai tại Việt Nam. Nghiên cứu xây dựng mô hình hồi quy tuyến tính dự báo dòng chảy có vai trò vô cùng quan trọng trong việc đánh giá tác động của trận lũ lịch sử khi các tài liệu đầu vào để đánh giá theo phương pháp truyền thống còn nhiều hạn chế tại thời điểm tức thời, thời gian thực khi lũ xuất hiện, đây là một ý nghĩa quan trọng trong quản lý tài nguyên nước, phòng chống thiên tai cho khu vực nghiên cứu. Vì những lý

do nêu trên, đề tài nghiên cứu của đồ án: "Ứng dụng mô hình hồi quy tuyến tính trong dự báo dòng chảy tại Tân Châu hỗ trợ cảnh báo lũ lụt" được đề xuất là rất cần thiết.

II. Mục tiêu đề tài

Dựa trên lý do chọn đề tài, trong nghiên cứu này có các mục tiêu chính sau:

- Tiếp cận cơ sở lý thuyết về hồi quy tuyến tính (Linear Regression).
- Ứng dụng mô hình hồi quy tuyến tính trong dự báo dòng chảy tại Tân Châu hỗ trợ cảnh báo lũ lụt, bằng ngôn ngữ Python.
- Đánh giá kết quả từ mô hình đạt được, từ đó đưa ra kết luận và kiến nghị.

Chương 2 Kiến thức cơ sở

I. Tổng quan về hồi quy tuyến tính trong học máy

Học máy, là một tập hợp con của Trí tuệ nhân tạo (AI), đã và đang đóng một vai trò chi phối trong cuộc sống hàng ngày của chúng ta. Các kỹ sư và nhà phát triển khoa học dữ liệu làm việc trong các lĩnh vực khác nhau đang sử dụng rộng rãi các thuật toán học máy để làm cho nhiệm vụ của họ trở nên đơn giản hơn và cuộc sống dễ dàng hơn. Ví dụ: một số thuật toán học máy nhất định cho phép Google Maps tìm ra tuyến đường nhanh nhất đến các điểm đến của chúng tôi, cho phép Tesla sản xuất ô tô không người lái, giúp Amazon tạo ra gần 35% thu nhập hàng năm của họ, AccuWeather nhận dự báo thời tiết của 3,5 triệu địa điểm trong tuần trong nâng cao, Facebook để tự động phát hiện khuôn mặt và đề xuất thẻ, v.v.

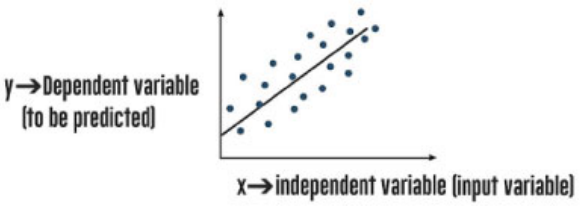
Trong thống kê và học máy, hồi quy tuyến tính là một trong những thuật toán phổ biến nhất và được hiểu rõ. Hầu hết những người đam mê khoa học dữ liệu và những người cuồng học máy đều bắt đầu cuộc hành trình của họ với các thuật toán hồi quy tuyến tính. Trong bài viết này, chúng ta sẽ xem xét cách hoạt động của thuật toán hồi quy tuyến tính và cách nó có thể được sử dụng hiệu quả trong các dự án máy học của bạn để xây dựng các mô hình tốt hơn.

Hồi quy tuyến tính là một trong những thuật toán học máy trong đó kết quả được dự đoán bằng cách sử dụng các tham số đã biết có tương quan với kết quả đầu ra.

Nó được sử dụng để dự đoán các giá trị trong một phạm vi liên tục thay vì cố gắng phân loại chúng thành các loại. Các tham số đã biết được sử dụng để tạo ra một độ dốc liên tục và không đổi được sử dụng để dự đoán kết quả hoặc điều chưa biết.

1. Vấn đề hồi quy là gì?

Phần lớn các thuật toán học máy thuộc danh mục học có giám sát. Đây là quá trình mà một thuật toán được sử dụng để dự đoán một kết quả dựa trên các giá trị đã nhập trước đó và kết quả được tạo ra từ chúng. Giả sử chúng ta có một biến đầu vào 'x' và một biến đầu ra 'y' trong đó y là một hàm của x (y = f {x}). Học có giám sát đọc giá trị của biến đã nhập 'x' và biến kết quả là 'y' để có thể sử dụng các kết quả đó để sau này dự đoán dữ liệu đầu ra có độ chính xác cao của 'y' từ giá trị đã nhập của 'x'. Vấn đề hồi quy là khi biến kết quả chứa giá trị thực hoặc giá trị liên tục. Nó cố gắng vẽ đường phù hợp nhất từ dữ liệu thu thập được từ một số điểm.



Ví dụ, ***cái nào trong số này là một bài toán hồi quy?***

- Tôi sẽ chi tiêu bao nhiêu xăng nếu tôi lái xe cho 100 dặm?
- Quốc tịch của một người là gì?
- Tuổi của một người là bao nhiêu?
- Hành tinh nào gần Mặt trời nhất?

Dự đoán lượng xăng phải tiêu và tuổi của một người là những bài toán hồi quy. Dự đoán quốc tịch là phân loại và hành tinh gần Mặt trời nhất là rời rạc.

2. Hồi quy tuyến tính là gì?

Giả sử chúng ta có một tập dữ liệu chứa thông tin về mối quan hệ giữa 'số giờ đã học' và 'điểm đạt được'. Một số học sinh đã được quan sát và ghi lại giờ học cùng với điểm của họ. Đây sẽ là dữ liệu đào tạo của chúng tôi. Mục tiêu của chúng tôi là

thiết kế một mô hình có thể dự đoán điểm nếu số giờ nghiên cứu được cung cấp. Sử dụng dữ liệu huấn luyện, một đường hồi quy thu được sẽ cho sai số tối thiểu. Phương trình tuyến tính này sau đó được sử dụng để áp dụng cho một dữ liệu mới. Nghĩa là, nếu chúng tôi đưa số giờ học của một sinh viên làm đầu vào, mô hình của chúng tôi sẽ có thể dự đoán điểm của họ với sai số tối thiểu.

- **Giả thuyết về hồi quy tuyến tính**

Mô hình hồi quy tuyến tính có thể được biểu diễn bằng phương trình sau:

$$Y = \theta_0 + \theta_1 x_1 + \theta_2 x_2 + ... + \theta_n x_n$$

Ở đây,

+ Y là giá trị dự đoán
+ θ_0 là thuật ngữ thiên vị.
+ θ_1,…, θn là các thông số mô hình
+ x_1, x_2,…, xn là các giá trị của đối tượng.

Giả thuyết trên cũng có thể được biểu diễn bằng

$$Y = \theta^T x$$

Trong đó, θ là vectơ tham số của mô hình bao gồm số hạng thiên vị θ_0; x là vectơ đặc trưng với $x_0 = 1$

Y (pred) = b0 + b1*x

Các giá trị b0 và b1 phải được chọn sao cho sai số là nhỏ nhất. Nếu tổng sai số bình phương được lấy làm thước đo để đánh giá mô hình, thì mục tiêu là thu được một đường giúp giảm lỗi tốt nhất.

$$\text{Error} = \sum_{i=1}^{n} (actual\_output - predicted\_output) ** 2$$

Nếu chúng ta không bình phương sai số, thì các điểm tích cực và tiêu cực sẽ triệt tiêu lẫn nhau.

Đối với mô hình có một công cụ dự đoán,

- **Khám phá 'b1'**

Nếu b1 > 0, thì x (dự đoán) và y (mục tiêu) có mối quan hệ cùng chiều. Tức là tăng x sẽ làm tăng y.

Nếu b1 <0, thì x (dự đoán) và y (mục tiêu) có mối quan hệ nghịch biến. Tức là tăng x sẽ giảm y.

- **Khám phá 'b0'**

Nếu mô hình không bao gồm x = 0, thì dự đoán sẽ trở nên vô nghĩa khi chỉ có b0 . Ví dụ: chúng tôi có một tập dữ liệu liên quan đến chiều cao (x) và cân nặng (y). Lấy x = 0 (tức là chiều cao bằng 0), sẽ làm cho phương trình chỉ có giá trị b0 là hoàn toàn vô nghĩa vì trong thời gian thực chiều cao và cân nặng không bao giờ có thể bằng không. Kết quả là do việc xem xét các giá trị mô hình nằm ngoài phạm vi của nó.

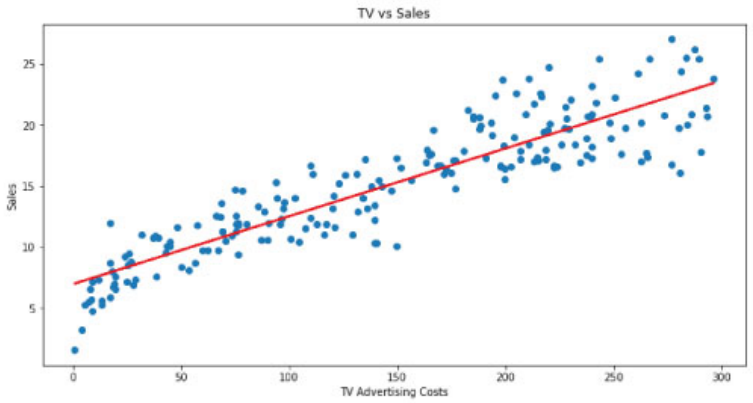
Nếu mô hình bao gồm giá trị 0, thì 'b0 ' sẽ là giá trị trung bình của tất cả các giá trị được dự đoán khi x = 0. Tuy nhiên, việc đặt 0 cho tất cả các biến dự báo thường là không thể.

Giá trị của b0 đảm bảo rằng phần dư sẽ bằng 0 trung bình. Nếu không có số hạng 'b0 ', thì hồi quy sẽ buộc phải vượt qua điểm gốc. Cả hệ số hồi quy và dự đoán sẽ bị sai lệch.

3. Hồi quy tuyến tính hoạt động như thế nào?

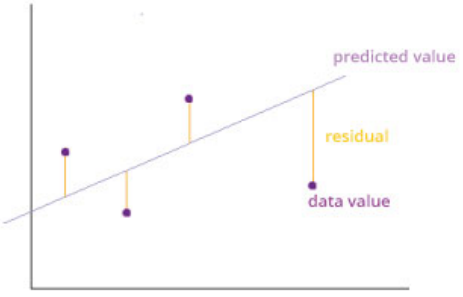
Hãy xem xét một kịch bản mà hồi quy tuyến tính có thể hữu ích: giảm cân. Chúng ta hãy xem xét rằng có mối liên hệ giữa lượng calo bạn nạp vào và cân nặng của bạn; phân tích hồi quy có thể giúp bạn hiểu mối liên hệ đó. Phân tích hồi quy sẽ cung cấp cho bạn một mối quan hệ có thể được trực quan hóa thành biểu đồ để đưa ra dự đoán về dữ liệu của bạn. Ví dụ: nếu bạn đã tăng cân trong vài năm qua, nó có thể dự đoán bạn sẽ nặng bao nhiêu trong mười năm tới nếu bạn tiếp tục tiêu thụ cùng một lượng calo và đốt cháy chúng với tốc độ tương tự.

Mục tiêu của phân tích hồi quy là tạo ra một đường xu hướng dựa trên dữ liệu bạn đã thu thập được. Sau đó, điều này cho phép bạn xác định xem các yếu tố khác ngoài lượng calo tiêu thụ có ảnh hưởng đến cân nặng của bạn hay không, chẳng hạn như số giờ bạn ngủ, áp lực công việc, mức độ căng thẳng, loại bài tập bạn làm, v.v. Trước khi tính đến, chúng ta cần để xem xét các yếu tố và thuộc tính này và xác định liệu có mối tương quan giữa chúng hay không. Hồi quy tuyến tính sau đó có thể được sử dụng để vẽ một đường xu hướng sau đó có thể được sử dụng để xác nhận hoặc phủ nhận mối quan hệ giữa các thuộc tính. Nếu thử nghiệm được thực hiện trong một khoảng thời gian dài, dữ liệu mở rộng có thể được thu thập và kết quả có thể được đánh giá chính xác hơn. Ở phần cuối của bài viết này, chúng tôi sẽ xây dựng một mô hình giống như hình dưới đây, tức là xác định một đường phù hợp nhất với dữ liệu.



- **Làm thế nào để chúng tôi xác định đường phù hợp nhất?**

Đường phù hợp nhất được coi là đường mà sai số giữa các giá trị dự đoán và các giá trị quan sát là nhỏ nhất. Nó còn được gọi là đường ***hồi quy*** và sai số còn được gọi là ***phần dư*** . Hình dưới đây cho thấy phần dư. Nó có thể được hình dung bằng các đường thẳng đứng từ giá trị dữ liệu quan sát đến đường hồi quy.



4. Khi nào sử dụng hồi quy tuyến tính?

Sức mạnh của Linear Regression nằm ở sự đơn giản của nó, có nghĩa là nó có thể được sử dụng để giải quyết các vấn đề trên nhiều lĩnh vực khác nhau. Lúc đầu, dữ liệu thu thập được từ các quan sát cần được thu thập và vẽ biểu đồ dọc theo một đường thẳng. Nếu sự khác biệt giữa giá trị dự đoán và kết quả là gần như nhau, chúng ta có thể sử dụng hồi quy tuyến tính cho bài toán.

- Các giả định trong hồi quy tuyến tính

Nếu bạn định sử dụng hồi quy tuyến tính cho vấn đề của mình thì có một số giả định bạn cần xem xét:

- Mối quan hệ giữa các biến phụ thuộc và độc lập phải gần như tuyến tính.
- Dữ liệu là tương đồng, có nghĩa là phương sai giữa các kết quả không được quá nhiều.
- Các kết quả thu được từ một lần quan sát không được ảnh hưởng bởi các kết quả thu được từ lần quan sát trước đó.
- Phần còn lại nên được phân phối bình thường. Giả thiết này có nghĩa là hàm mật độ xác suất của các giá trị thặng dư được phân phối chuẩn tại mỗi giá trị độc lập.

Bạn có thể xác định xem dữ liệu của mình có đáp ứng các điều kiện này hay không bằng cách vẽ biểu đồ và sau đó thực hiện một chút tìm hiểu cấu trúc của nó.

- Một số thuộc tính của dòng hồi quy

Dưới đây là một số tính năng mà đường hồi quy có:

- Hồi quy đi qua giá trị trung bình của biến độc lập (x) cũng như giá trị trung bình của biến phụ thuộc (y).
- Đường hồi quy giảm thiểu tổng "Bình phương phần dư". Đó là lý do tại sao phương pháp hồi quy tuyến tính được gọi là "Bình phương nhỏ nhất thông thường (OLS)". Chúng ta sẽ thảo luận chi tiết hơn về Quảng trường Bình thường Ít nhất ở phần sau.

- B 1 giải thích sự thay đổi của Y với sự thay đổi của x một đơn vị. Nói cách khác, nếu chúng ta tăng giá trị của 'x', nó sẽ dẫn đến sự thay đổi giá trị của Y.

5. Hiệu suất mô hình(Model Performance)

Sau khi xây dựng mô hình, nếu chúng ta thấy sự khác biệt về giá trị của dữ liệu dự đoán và thực tế là không nhiều thì được coi là mô hình tốt và có thể sử dụng để đưa ra các dự đoán trong tương lai. Số tiền mà chúng tôi coi là "không nhiều" hoàn toàn phụ thuộc vào tác vụ bạn muốn thực hiện và tỷ lệ phần trăm biến động trong dữ liệu có thể được xử lý. Dưới đây là một số công cụ đo lường mà chúng tôi có thể sử dụng để tính toán lỗi trong mô hình

R – Square (R2)

$$R^2 = \frac{TSS - RSS}{TSS}$$

Tổng bình phương (TSS - Total Sum of Squares): tổng bình phương (TSS) là một đại lượng xuất hiện như một phần của cách trình bày kết quả tiêu chuẩn của một phân tích như vậy. Tổng bình phương là thước đo mức độ thay đổi của một tập dữ liệu xung quanh một số trung tâm (như giá trị trung bình). Tổng Bình phương Tổng cho biết có bao nhiêu sự thay đổi trong biến phụ thuộc.

$$TSS = \Sigma (Y - Mean[Y])2$$

Tổng bình phương còn lại (RSS - Residual Sum of Squares): Tổng bình phương còn lại cho bạn biết mức độ biến thiên của biến phụ thuộc mà mô hình của bạn không giải thích được. Nó là tổng của sự khác biệt bình phương giữa Y thực tế và Y dự đoán.

$$RSS = \Sigma (Y - f[Y])2$$

(TSS - RSS) đo lường mức độ thay đổi trong phản hồi được giải thích bằng cách thực hiện hồi quy.

Các thuộc tính của R2

- R2 luôn nằm trong khoảng từ 0 đến 1.
- R2 của 0 nghĩa là không có mối tương quan giữa biến phụ thuộc và biến độc lập.
- R2 của 1 có nghĩa là biến phụ thuộc có thể được dự đoán từ biến độc lập mà không có bất kỳ sai số nào.
- R2 giữa 0 và 1 cho biết mức độ mà biến phụ thuộc có thể dự đoán được. R 2 bằng 0,20 có nghĩa là có 20% phương sai trong Y có thể dự đoán được từ X; R 2 là 0,40 có nghĩa là 40% có thể dự đoán được; và như thế.

Root Mean Square Error (RMSE)

Root Mean Square Error (RMSE) là độ lệch chuẩn của các phần dư (sai số dự đoán). Công thức tính RMSE là:

$$R^2 = \{(1/N) * \sum [(x_i - mean(x)) * (y_j - mean(y))] / (\sigma_x * \sigma_y) \}^2$$

Trong đó N: Tổng số quan sát

Khi các quan sát chuẩn hóa được sử dụng làm đầu vào RMSE, có mối quan hệ trực tiếp với hệ số tương quan. Ví dụ, nếu hệ số tương quan là 1, RMSE sẽ là 0, bởi vì tất cả các điểm nằm trên đường hồi quy (và do đó không có sai số).

Mean Absolute Percentage Error (MAPE)

Có một số hạn chế nhất định đối với việc sử dụng RMSE, vì vậy các nhà phân tích thích MAPE hơn RMSE đưa ra sai số về tỷ lệ phần trăm để các mô hình khác nhau có thể được xem xét cho nhiệm vụ và xem chúng hoạt động như thế nào. Công thức tính MAPE có thể được viết như sau:

$$RMSE = \sqrt{\frac{\sum(Y_{Actual} - Y_{Predicted})^2}{N}}$$

Trong đó N: Tổng số quan sát

6. Lựa chọn tính năng(Feature Selection)

Lựa chọn đối tượng là lựa chọn tự động các thuộc tính cho dữ liệu của bạn có liên quan nhất đến mô hình dự đoán mà bạn đang làm việc. Nó tìm cách giảm số lượng các thuộc tính trong tập dữ liệu bằng cách loại bỏ các tính năng không cần thiết cho việc xây dựng mô hình. Lựa chọn tính năng không loại bỏ hoàn toàn một thuộc tính được xem xét cho mô hình, thay vào đó, nó tắt đặc tính cụ thể đó và hoạt động với các tính năng ảnh hưởng đến mô hình.

Phương pháp lựa chọn tính năng hỗ trợ sứ mệnh của bạn để tạo ra một mô hình dự đoán chính xác. Nó giúp bạn bằng cách chọn các tính năng sẽ cung cấp cho bạn độ chính xác tốt hoặc tốt hơn trong khi yêu cầu ít dữ liệu hơn. Phương pháp lựa chọn tính năng có thể được sử dụng để xác định và loại bỏ các thuộc tính không cần thiết, không liên quan và dư thừa khỏi dữ liệu không góp phần vào độ chính xác của mô hình hoặc thậm chí có thể làm giảm độ chính xác của mô hình. Có ít thuộc tính hơn là mong muốn vì nó làm giảm độ phức tạp của mô hình và một mô hình đơn giản hơn sẽ dễ hiểu, dễ giải thích và dễ làm việc hơn.

Thuật toán lựa chọn tính năng:

- **Phương pháp Bộ lọc(Filter Method): Phương pháp** này liên quan đến việc ấn định điểm số cho các tính năng riêng lẻ và xếp hạng chúng. Các tính năng có rất ít hoặc hầu như không có tác động sẽ bị loại bỏ khỏi việc xem xét trong khi xây dựng mô hình.
- **Phương thức Wrapper(Wrapper Method):** Phương thức gói khá giống với phương thức Bộ lọc ngoại trừ thực tế là nó xem xét các thuộc tính trong một nhóm tức là một số thuộc tính được lấy và kiểm tra xem chúng có tác động đến mô hình hay không và nếu không thì một tổ hợp khác được áp dụng.
- **Phương pháp nhúng(Embedded Method): Phương pháp** nhúng là tốt nhất và chính xác nhất trong tất cả các thuật toán. Nó tìm hiểu các tính năng ảnh hưởng đến mô hình trong khi mô hình đang được xây dựng và chỉ xem xét các tính năng đó. Loại phương pháp lựa chọn tính năng nhúng phổ biến nhất là phương pháp chính quy hóa.

7. Chức năng ước lượng(Cost Function)

Hàm Cost giúp tìm ra các ô tốt nhất có thể được sử dụng để vẽ đường phù hợp nhất cho các điểm dữ liệu. Vì chúng tôi muốn giảm sai số của giá trị kết quả, chúng tôi thay đổi quá trình tìm ra kết quả thực tế thành một quá trình có thể giảm sai số giữa giá trị dự đoán và giá trị thực tế.

$$minimize \frac{1}{n}\sum_{i=1}^{n}(pred_i - y_i)^2$$

$$J = \frac{1}{n}\sum_{i=1}^{n}(pred_i - y_i)^2$$

Ở đây, J là hàm chi phí.

Hàm trên được thực hiện ở định dạng này để tính toán sai lệch giữa các giá trị dự đoán và các giá trị được vẽ trên đồ thị. Chúng tôi lấy bình phương của tổng của tất cả các điểm dữ liệu và chia nó cho tổng số điểm dữ liệu. Hàm chi phí J này còn được gọi là hàm bình phương sai số (MSE). Sử dụng hàm MSE này, chúng ta sẽ dự đoán các giá trị sao cho giá trị MSE ổn định ở cực tiểu, làm giảm hàm chi phí.

8. Xuống dốc(Gradient Descent)

Gradient Descent là một thuật toán tối ưu hóa giúp các mô hình học máy tìm ra đường dẫn đến giá trị tối thiểu bằng cách sử dụng các bước lặp lại. Gradient descent được sử dụng để thu nhỏ một chức năng để nó cho kết quả đầu ra thấp nhất của chức năng đó. Chức năng này được gọi là Chức năng Mất mát. Hàm mất mát cho chúng ta biết có bao nhiêu lỗi do mô hình học máy tạo ra so với kết quả thực tế. Mục đích của chúng tôi là giảm hàm chi phí càng nhiều càng tốt. Một cách để đạt được một hàm chi phí thấp là bằng quá trình giảm độ dốc. Độ phức tạp của một số phương trình gây khó khăn cho việc sử dụng, đạo hàm riêng của hàm chi phí đối với tham số được xem xét có thể cung cấp giá trị hệ số tối ưu.

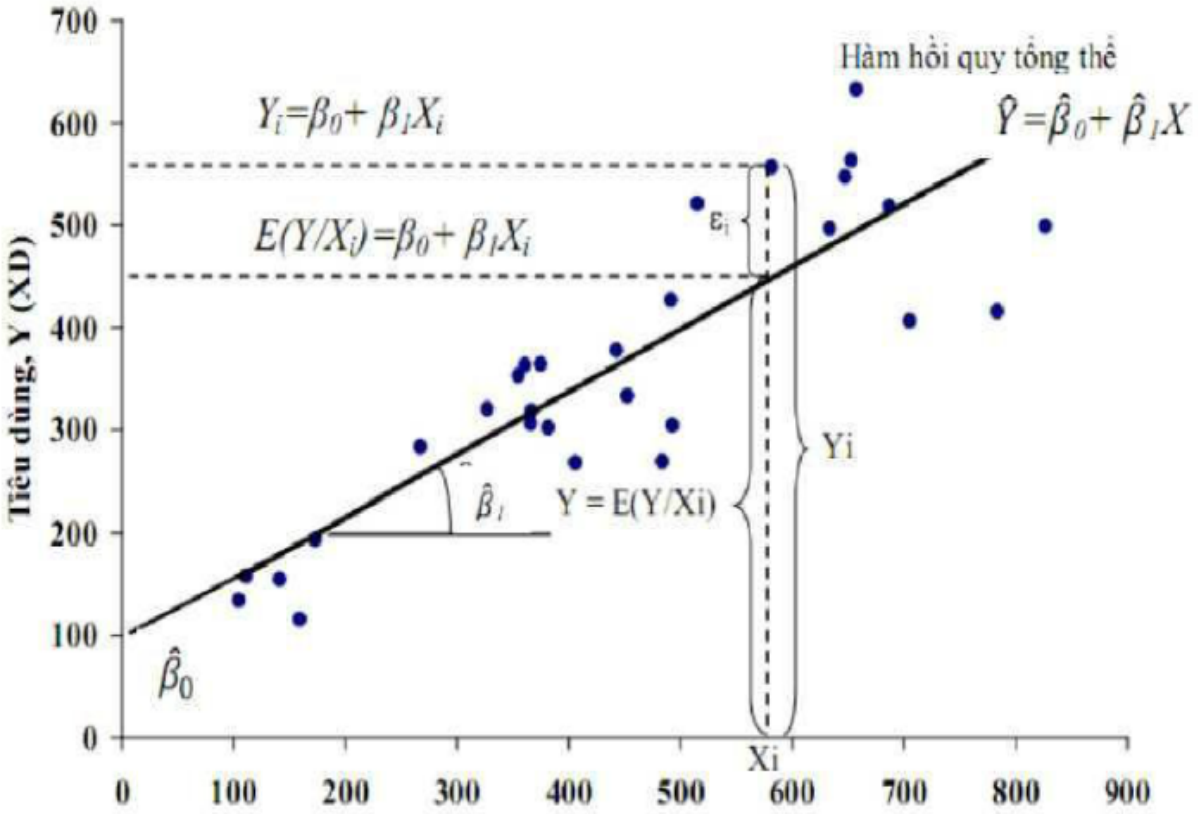
II. Mô hình hồi quy tuyến tính đơn giản

Mô hình hồi quy tuyến tính đơn giản là giữa một biến phụ thuộc Y và một biến độc lập X. Mối quan hệ giữa X và Y là tuyến tính. Mô hình hồi quy tuyến tính được viết như sau:

$$Y = \beta_0 + \beta_1 X + \varepsilon$$

Trong đó, β_0 là giá trị chặn (Intercept) và β_1 là độ dốc (slope) của mô hình, ε là sai số ngẫu nhiên, ε là một biến số theo luật phân phối chuẩn với trung bình 0 và phương sai σ^2.

β_0, β_1 là hai giá trị không biết chính xác, do vậy từ giá trị X mà ta thu thập được phải ước tính các hệ số của mô hình là β_0, β_1, σ^2.



$$y = \beta_0 + \beta_1 x$$

Với y là biểu thị cho giá trị dự đoán Y, x = X.

1. Ước tính các tham số β_0, β_1

$(x_j\ y_j),\ (x_2\ y_2),\ (x_3\ y_3)..\ (x_n\ y_n)$ là n cặp quan sát. Mục đích của hồi quy tuyến tính là ước tính các tham số β_0, β_1 của mô hình hồi quy tuyến tính sao cho biểu thị đúng các cặp dữ liệu mà ta quan sát được, $y = \beta_0 + \beta_1 x$ với i =

1...n. Ta có $e_i = y_i - y$ là sai số (residual) thứ i. Đây là sự khác biệt giữa giá trị quan sát thứ i là giá trị thứ i được dự đoán bằng mô hình hồi quy tuyến tính. Ta gọi tổng bình phương của phần dư là ESS (explained sum of squares).

$$\text{ESS} = e_1^2 + e_2^2 + \ldots + e_n^2$$

$$\text{ESS} = (y_1 - \beta_0 - \beta_1 x_1)^2 + (y_2 - \beta_0 - \beta_1 x_2)^2 + \ldots + (y_n - \beta_0 - \beta_1 x_n)^2$$

Phương pháp bình quân tối thiểu (the least squares) chọn β_0, β_1 sao cho ESS đạt giá trị minimize. Các hệ số ước tính của mô hình hồi quy tuyến tính được tính theo phương pháp bình phương tối thiểu.

$$\frac{\partial}{\partial \beta_0} \sum_{i=1}^{n} [y_i - (\beta_0 + \beta_1 x_i)]^2 = 0$$

$$\frac{\partial}{\partial \beta_1} \sum_{i=1}^{n} [y_i - (\beta_0 + \beta_1 x_i)]^2 = 0$$

Lấy vi phân từng phần theo β_0, β_1 ta có:

$$\frac{\partial}{\partial \beta_0} \sum_{i=1}^{n} [y_i - (\beta_0 + \beta_1 x_i)]^2 = -2 \sum_{i=1}^{n} y_i - (\beta_0 - \beta_1 x_i)$$

$$\frac{\partial}{\partial \beta_1} \sum_{i=1}^{n} [y_i - (\beta_0 + \beta_1 x_i)]^2 = -2 \sum_{i=1}^{n} (y_i - \beta_0 + \beta_1 x_i) x_i$$

Xây dựng hệ phương trình ta có:

$$\sum_{i=1}^{n} y_i = n\beta_0 + \beta_1 \sum_{i=1}^{n} x_1$$

$$\sum_{i}^{n} y_i x_i = \beta_0 \sum_{i=1}^{n} x_i + \beta_1 \sum_{i=1}^{n} x_i^2$$

Giải hệ phương trình trên ta được:

$$\beta_1 = \frac{\sum_i^n y_i x_i - \sum_{i=1}^n x_i \sum_{i=1}^n y_i}{n\sum_{i=1}^n x_i^2 - (\sum_{i=1}^n x_i)^2} = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^n (x_i - \bar{x})^2}$$

$$\beta_0 = \frac{\sum_{i=1}^n x_i^2 \sum_{i=1}^n y_i - \sum_{i=1}^n x_i \sum_{i=1}^n x_i y_i}{n\sum_{i=1}^n x_i^2 - (\sum_{i=1}^n x_i)^2} = \beta_0 = \bar{y} - \beta_1 \bar{x}$$

Trong đó, $\bar{y} = \frac{1}{n}\sum_{i=1}^n$ và $\bar{x} = \frac{1}{n}\sum_{i=1}^n x_i$

2. Đánh giá độ chính xác của mô hình hồi quy tuyến tính

Để đánh giá sự phù hợp của mô hình hồi quy tuyến tính ta tìm hiểu hai khái niệm là sai số chuẩn RSE (residual standard error) và hệ số xác định R^2(R squares).

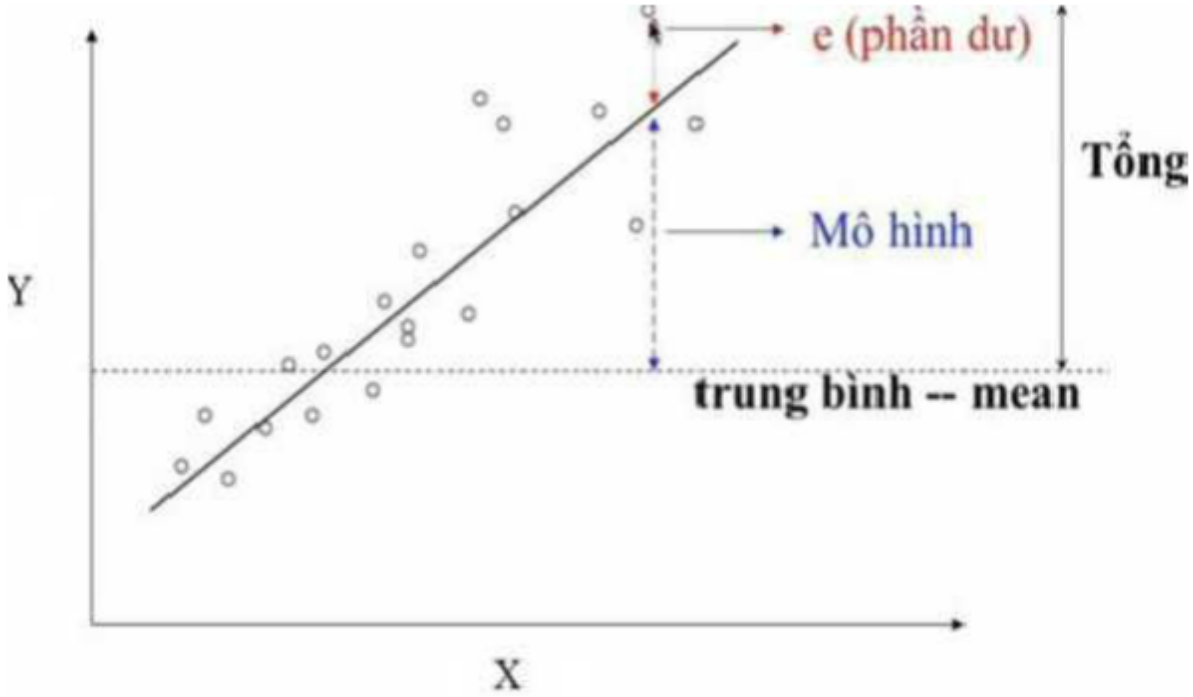
- Sai số chuẩn RSE (s hay σ^2) là ước tính độ lệch chuẩn hay phương sai của phần dư, đó là giá trị trung bình của các giá trị quan sát so với đường hồi quy, được tính theo công thức sau:

$$RSE = \sqrt{\frac{1}{n-2} ESS} = \frac{1}{n-2}\sum_{i=1}^n (y_i - y_i)^2$$

Sai số chuẩn được coi là giải pháp để đánh giá sự đúng đắn của mô hình hồi quy tuyến tính, khi đó $y_i \approx y_i$, sai số càng nhỏ thì giá trị dự báo càng gần với giá trị quan sát, nghĩa là mô hình hồi quy là phù hợp.

- Hệ số R^2

  Một câu hỏi được đặt ra là làm thế nào chúng ta đo lường mức độ phù hợp của hàm hồi quy tìm được cho dữ liệu mẫu. Thước đo độ phù hợp của mô hình đối với dữ liệu là R^2. Để có cái nhìn trực quan về R^2, chúng ta xem xét đồ thị:

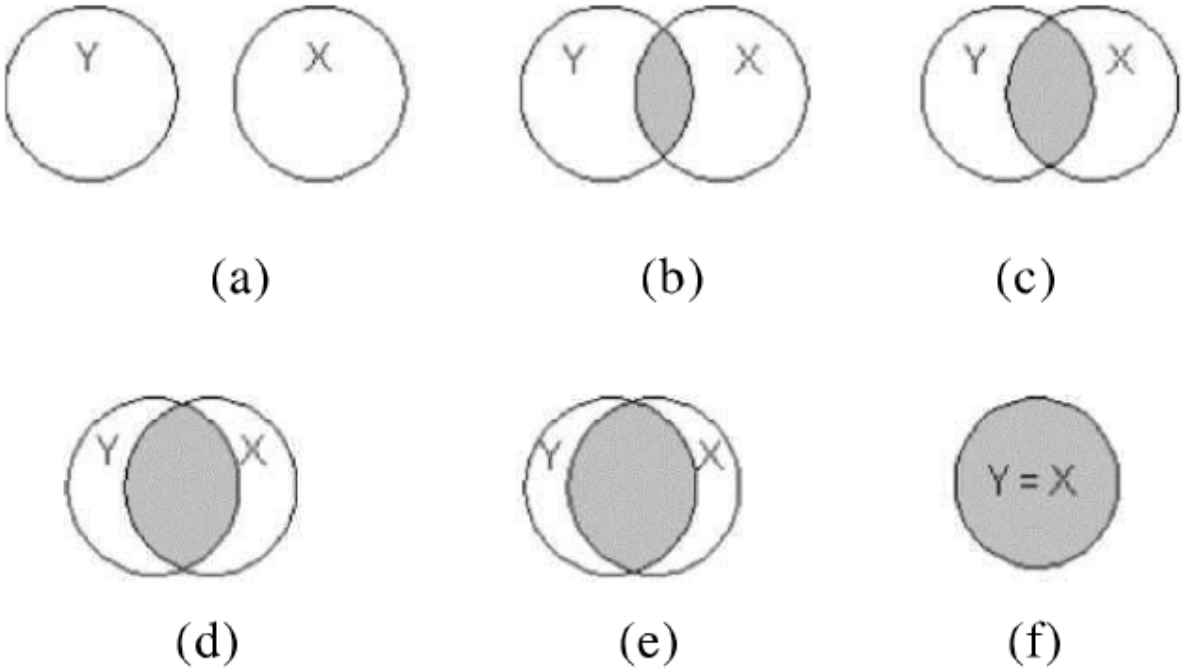


R^2 được tính theo công thức sau: $ESS = \sum_{i=1}^{n}(y_i - y_i)^2$

$$R^2 = \frac{TSS - ESS}{TSS} = 1 - \frac{ESS}{TSS}$$

Trong đó, $TSS = \sum_{i=1}^{n}(y_i - \bar{y})^2$ là tổng bình phương (total sum of squares), là số các biến thiên của các giá trị y_i so với giá trị trung bình.

$ESS = e_1^2 + e_2^2 + ... + e_1^2$ là tổng bình phương của phần dư. ESS là tổng số biến thiên không giải thích được sau khi thực hiện hồi quy. Do đó, hiệu của TSS và ESS biểu thị lượng biến thiên giải thích được bằng mô hình hồi quy. R^2 là tỷ lệ biến đổi trong Y có thể giải thích được bằng X, có giá trị $0 \leq R^2 \leq 1$. Nếu R^2gần giá trị 1 cho thấy tỷ lệ lớn các biến thiên đã được giải thích bằng hồi quy, do vậy mô hình hồi quy tuyến tính được đưa ra là phù hợp với dữ liệu. Khi R^2 gần 0 chỉ ra rằng hồi quy không giải thích được nhiều sự biến thiên và mô hình hồi quy tuyến tính là không phù hợp hoặc lỗi σ^2 là cao hoặc cả hai trường hợp trên. Thể hiện R^2theo phương pháp đồ thị Venn, hay là Ballentine như sau:



Phương pháp Ballentine với R^2, (a) $R^2 = 0$, (f) $R^2 = 1$

Vòng tròn Y tượng trưng cho biến thiên trong biến phụ thuộc Y và vòng tròn X tượng trưng cho biến thiên trong biến độc lập X. Vùng chồng lên nhau của hai vòng tròn (vùng tối) chỉ rõ phạm vi mà độ biến thiên trong Y được giải thích bởi biến thiên trong X (cho là theo hướng hồi quy các bình phương tối thiểu thông thường OLS). Phạm vi vùng chồng lên càng lớn, độ biến thiên trong Y được giải thích bởi X càng lớn. R^2đơn giản là đại lượng đo bằng số cho vùng tối này. Trong hình, khi ta di chuyển từ trái sang phải, vùng tối tăng dần nghĩa là tỷ lệ biến thiên trong Y được giải thích bởi X tăng dần.

III. Mô hình hồi quy tuyến tính đa biến

Mô hình hồi quy tuyến tính đơn giản là một giải pháp hữu ích để dự báo trên cơ sở một biến dự báo duy nhất. Tuy nhiên trong thực tế chúng ta thường có nhiều hơn một yếu tố dự báo. Một giải pháp có thể được đưa ra là sử dụng n mô hình hồi quy đơn giản cho n biến, tuy nhiên cách tiếp cận của mỗi mô hình tuyến tính đơn giản không hoàn toàn thỏa mãn. Trước hết, nó không rõ ràng và nó không là duy nhất cho mỗi phương tiện truyền thông vì mỗi phương tiện truyền thông gắn với một hàm hồi quy riêng. Thứ hai, một trong n hàm hồi quy bỏ qua (n - 1) phương tiện truyền thông khác khi thực hiện phân tích tương quan. Do vậy, thay vì sử dụng hồi quy tuyến tính đơn giản riêng biệt cho từng dự báo, một cách tiếp cận tốt hơn là mở rộng mô hình hồi quy

tuyến tính đơn giản để nó chứa nhiều hơn một yếu tố dự báo là mô hình hồi quy tuyến tính đa biến. Dạng tổng quát của mô hình hồi quy đa biến:

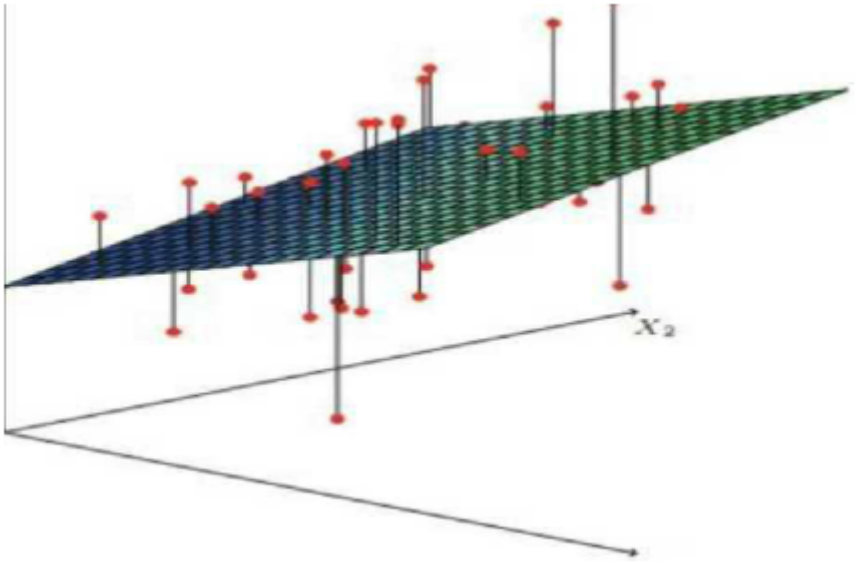
$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + ... + \beta_p X_p + \varepsilon$$

Trong đó, Xj là các biến dự báo thứ j, và Bj là các hệ số của mô hình đa biến.

1. Ước tính các hệ số hồi quy của mô hình hồi quy tuyến tính đa biến

Cũng giống như mô hình hồi quy tuyến tính đơn giản, các tham số $\beta_0, \beta_1, \ldots, \beta_p$ không biết được, do vậy phải ước tính các hệ số $\beta_0, \beta_1, \ldots, \beta_p$ và sử dụng công thức để dự báo sau:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_p x_p$$



Mô hình ba chiều, gồm hai biến dự báo và một biến phụ thuộc

Các tham số được ước tính giống phương pháp bình phương tối thiểu đã được trình bày trong mô hình hồi quy tuyến tính đơn giản. Chúng ta chọn $\beta_0, \beta_1, \ldots, \beta_p$ sao cho tổng bình phương của phần dư là nhỏ nhất.

$$\mathrm{ESS} = \sum_{i=1}^{n} (y_i - y_i)^2$$

$$= \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_{i1} - \beta_2 x_{i2} - ... \beta_p x_p)^2$$

Vì có nhiều biến dự báo nên giá trị của X là một ma trận n x phần tử được viết gọn lại Y = Xβ + ε trong đó:

$$X = \begin{bmatrix} 1 & x_{11} & \dots & x_{1p} \\ 1 & x_{21} & \dots & x_{2p} \\ \dots & \dots. & \dots & \dots. \\ 1 & x_{n1} & \dots & x_{np} \end{bmatrix}; \qquad Y = \begin{bmatrix} y_1 \\ y_2 \\ \dots \\ y_n \end{bmatrix}; \qquad \beta = \begin{bmatrix} \beta_1 \\ \beta_2 \\ \dots \\ \beta_p \end{bmatrix}$$

Áp dụng phương pháp bình phương tối thiểu để ước tính các giá trị β sao cho sai số là nhỏ nhất.

$$X^T X \beta = X^T y$$

$$\beta = (X^T X)^{-1} X^T y$$

2. Đánh giá mức độ phù hợp của mô hình hồi quy tuyến tính đa biến

Tương tự mô hình hồi quy tuyến tính đơn giản, ta cũng sử dụng hệ số xác định R^2 để đánh giá sự phù hợp của mô hình hồi quy tuyến tính đa biến

$$R^2 = \frac{RSS}{TSS} = 1 - \frac{ESS}{TSS}$$

Trong đó:

$$TSS = \sum_{i=1}^{n} (y_i - \bar{y})^2$$

$$ESS = \sum_{i=1}^{n} (y_i - y_i)^2$$

RSS = TSS - ESS

$0 \le R \le 1$, R^2 càng gần giá trị 1 thì sự phù hợp của mô hình càng cao và ngược lại, R^2 càng gần giá trị 0 thì sự phù hợp của mô hình càng thấp.

Chương 3 Kết quả thực nghiệm

I. Mô tả dữ liệu

Dữ liệu được dùng trong thực nghiệm là số liệu đo đạc khí tượng – thuỷ văn khu vực thượng lưu lưu vực sông Mê Công gồm 480 bản ghi.

Dữ liệu training từ 1980 - 2015, dữ liệu test từ 2016 - 2019.

II. Tham số mô hình và phương pháp đánh giá

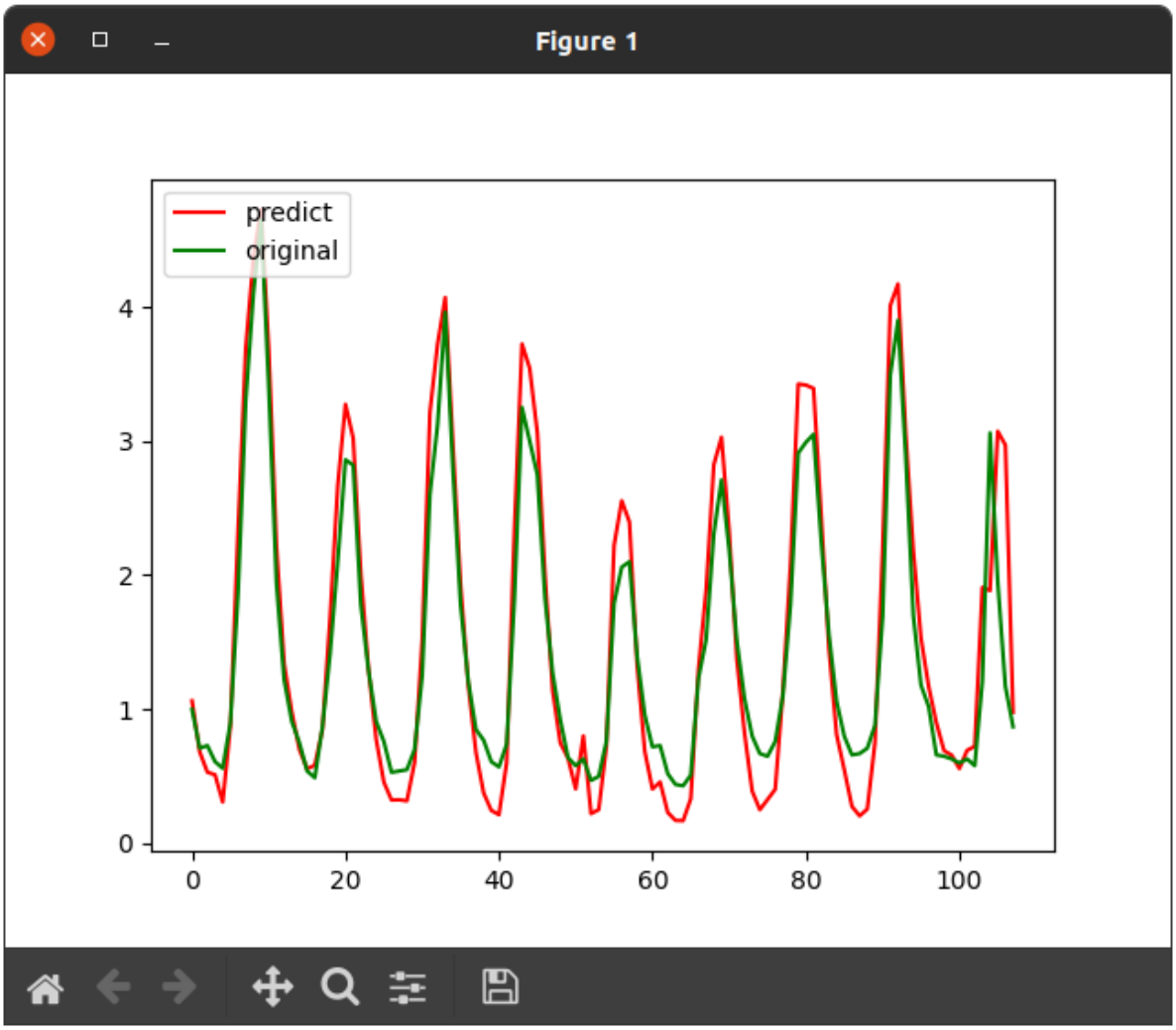
Chúng tôi dùng căn bình phương sai số (Root mean square error - RMSE), sai số tuyệt đối (mean absolute error - MAE) và hệ số xác định bội (coefficient of determination - R^2) để đánh giá tính hiệu quả của mô hình:

$$RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(Yi - \hat{Y}i)^2} \quad ; \quad MAE = \frac{1}{N}\sum_{i=1}^{N}|Yi - \hat{Y}i|$$

$$R^2 = 1 - \sum_{i=1}^{N}(Yi - \hat{Y}i) / \sum_{i=1}^{N}(Yi - \bar{Y}i)$$

Trong đó: Y_i, $\hat{Y}_i$, $\bar{Y}_i$ chỉ giá trị thực, giá trị dự đoán và giá trị trung bình của mẫu thứ i tương ứng. Mô hình cho kết quả tốt là mô hình đạt được sai số RMSE và MAE nhỏ. Giá trị R^2 cao là một dấu hiệu cho thấy mối liên hệ giữa các biến giải thích và biến số SHL (biến đích đo sự hài lòng) chặt chẽ. Giá trị R^2 càng cao cho thấy mô hình sử dụng để phân tích có khả năng giải thích càng tốt các khác biệt về mực nước . Giá trị MAE càng thấp thì khả năng dự báo của mô hình càng cao.

Kết quả dự đoán mực nước tại Tân Châu



Kết quả mô hình LR dự đoán mực nước tại Tân Châu trên dữ liệu kiểm thử

Mô hình hồi quy	R^2	RMSE	MAE
y = 0.573*X1 + 0.251*X2 - 0.184*X3 - 0.178*X4 + 0.173*X5 - 0.147*X6 - 0.144*X7 + 0.099*X8 + 0.056*X9 - 0.045*X10 - 0.036*X11 - 0.03*X11	0.905	0.375	0.281

Chương 4 Kết luận

Chúng tôi đã trình bày mô hình hồi quy tuyến tính đa biến dự đoán dòng chảy tại Tân Châu. Chỉ tiêu đánh giá dựa trên phương pháp đánh giá R^2, RMSE và MAE. Kết quả thực nghiệm cho thấy mô hình hồi quy dễ cài đặt và dễ sử dụng, nhưng khả năng còn chưa tối ưu so với các mô hình khác. Trong tương lai, chúng tôi sẽ áp dụng kết quả nghiên cứu mở rộng, áp dụng cùng với các mô hình học máy khác để tìm ra mô hình tối ưu cùng kết quả mong muốn có độ chính xác cao nhất.

TÀI LIỆU THAM KHẢO

[1]https://www.knowledgehut.com/blog/data-science/linear-regression-for-machine-learning

[2]https://machinelearningcoban.com/2016/12/28/linearregression/

[3]https://cafedev.vn/tu-hoc-ml-hoi-quy-nhieu-tuyen-tinh-bang-python/

[4]https://www.geeksforgeeks.org/linear-regression-python-implementation/