TRƯỜNG ĐẠI HỌC KINH TẾ - LUẬT
KHOA TOÁN – KINH TẾ

BÁO CÁO NGHIÊN CỨU HỌC MÁY

PHƯƠNG PHÁP CHÍNH QUY HÓA: HỒI QUY RIDGE VÀ HỒI QUY LASSO

GVHD: TS. Phạm Hoàng Uyên
Th.S Lương Thanh Quỳnh
SVTH: Nguyễn Đức Hoài An
Lê Nguyễn Hiếu Nghĩa
Nguyễn Trần Quỳnh Như

TP.HCM THÁNG 4/2023

PHỤ LỤC

1. Giới thiệu

1.1. Regressio – Hồi quy và mục đích của hồi quy

Các bài toán phổ biến trong Học máy (Machine Learning) thường được sử dụng để giải quyết các bài toán phức tạp có thể điểm qua:

Bài toán Phân Loại (Classification): trong bài toán này, mô hình được xây dựng cần phải xác định được lớp/nhãn (class/label) của một điểm dữ liệu trong số C nhãn khác nhau. Ví dụ về bài toán thực tế: Cần phân loại email rác thì mục đích chính là xác nhận xem email mới trong hộp thư đến có phải là email rác hay không, phép đánh giá chính là tỉ lệ email rác trên email thường được xác định đúng, và với kinh nghiệm là cặp các (email, nhãn) thu thập được trước đó.

Bài toán Phân Cụm (Clustering): Bài toán này sẽ chia dữ liệu X thành các cụm nhỏ dựa trên sự liên quan giữa các dữ liệu trong mỗi cụm. Ta có ví dụ như sau: Phân cụm khách hàng dựa trên hành vi mua hàng. Dựa trên việc mua bán và theo dõi của người dùng trên một trang web thương mại điện tử, mô hình có thể phân người dùng vào các cụm theo sở thích mua hàng. Từ đó, mô hình có thể quảng cáo các mặt hàng mà người dùng có thể quan tâm.

Bài toán Hồi Quy (Regression): Bài toán hồi quy hay bài toán tiên lượng là một bài toán trong thống kê, kinh tế lượng cũng như học máy. Một trong những mục tiêu của bài toán là xác định mối quan hệ giữa một biến phụ thuộc (Dependent Variable) hay biến mục tiêu (Target Variable) ngoài ra còn được gọi là biến đầu ra (Output Variable) đối với một hoặc nhiều biến độc lập (Independent Variables) hay biến đặc trưng (Feature Variables) ngoài ra còn được gọi là biến đầu vào (Input Variables). Mục đích của bài toán hồi quy là xây dựng mô hình dự đoán giá trị của biến mục tiêu dựa trên các giá trị của biến đặc trưng

Hồi quy tuyến tính (Linear Regression) là một thuật toán mà đầu ra là một hàm số tuyến tính của đầu vào. Hồi quy tuyến tính đa biến có thể được mô tả bằng phương trình:

$$Y_i = \beta_1 + \beta_2 X_{i2} + \beta_3 X_{i3} + \cdots + \beta_k X_{ik} + \epsilon_i$$

Và nếu ta có n quan sát, hàm hồi quy tuyến tính đa biến được tổng quát như sau:

$$\begin{aligned} Y_1 &= \beta_1 + \beta_2 X_{12} + \beta_3 X_{13} + \cdots + \beta_k X_{1k} + \epsilon_1 \\ Y_2 &= \beta_1 + \beta_2 X_{22} + \beta_3 X_{23} + \cdots + \beta_k X_{2k} + \epsilon_2 \\ &\vdots \qquad\qquad\qquad\qquad\qquad\qquad \vdots \\ Y_n &= \beta_1 + \beta_2 X_{n2} + \beta_3 X_{n3} + \cdots + \beta_k X_{nk} + \epsilon_n \end{aligned}$$

Khi viết về dạng ma trận, ta được như sau:

$$Y = \begin{pmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{pmatrix}, \beta = \begin{pmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_n \end{pmatrix}, \epsilon = \begin{pmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{pmatrix}, X = \begin{pmatrix} 1 & X_{12} & X_{13} & \dots & X_{1k} \\ 1 & X_{22} & X_{23} & \dots & X_{2k} \\ \vdots & \vdots & \vdots & \dots & \vdots \\ 1 & X_{n2} & X_{n3} & \dots & X_{nk} \end{pmatrix}$$

Và hàm hồi quy mẫu (ước lượng) sẽ có dạng:

$$\widehat{Y_i} = \widehat{\beta_1} + \widehat{\beta_2} X_{i2} + \widehat{\beta_3} X_{i3} + \cdots + \widehat{\beta_k} X_{ik}$$

Từ đó, ta có mối quan hệ giữa giá trị thực tế và giá trị ước lượng là:

$$Y_i = \widehat{Y_i} + e_i$$

Khi viết về dạng ma trận, ta được:

$$Y_i = \widehat{Y_i} + e_i$$

1.2. OLS – Phương pháp Bình phương cực tiểu

Mỗi mô hình Machine Learning được mô tả bởi bộ các tham số mô hình (model parameter). Công việc của một thuật toán machine learning là đi tìm các tham số mô hình tối ưu cho mỗi bài toán. Việc đi tìm các tham số mô hình có liên quan mật thiết đến các phép đánh giá. Mục đích chính là đi tìm các tham số mô hình sao cho các phép đánh giá đạt kết quả cao nhất. Trong bài toán phân loại, kết quả tốt có thể được hiểu là khi có ít điểm dữ liệu bị phân loại sai. Trong bài toán hồi quy, kết quả tốt là khi sự sai lệch giữa đầu ra dự đoán và đầu ra thực sự là nhỏ.

Quan hệ giữa một phép đánh giá và các tham số mô hình được mô tả thông qua một hàm số gọi là hàm mất mát (loss function hoặc cost function). Hàm số này thường có giá trị nhỏ khi phép đánh giá cho kết quả tốt và ngược lại. Ở đây, phép đánh giá (evaluation metric) có thể được hiểu là một chỉ số đánh giá mức độ hiệu quả của một mô hình dựa trên dữ liệu đầu vào, điều này có nghĩa phép đánh giá sẽ có mối quan hệ với các tham số mô hình cũng như giá trị dự đoán. Có nhiều phép đánh giá khác nhau, tùy vào mục đích khác nhau và mô hình khác nhau, ví dụ như có các phép đánh giá: Độ chính xác (Accuracy), Độ nhạy(Recall), Độ đặc hiệu (Precision),…Việc đi tìm các tham số mô hình sao cho phép đánh giá trả về kết quả tốt tương đương với việc tối thiểu hàm mất mát. Như vậy, việc xây dựng một mô hình machine learning chính là việc đi giải một bài toán tối ưu.Quá trình đó được coi là quá trình learning của machine.

Tập hợp các tham số mô hình được ký hiệu bằng θ, hàm mất mát của mô hình được ký hiệu là $L(\theta)$ hoặc $J(\theta)$. Bài toán đi tìm tham số mô hình tương đương với bài toán tối thiểu hàm mất mát:

$$\theta^* = \underset{\theta}{\operatorname{argmin}}\, L(\theta)$$

Trong đó, ký hiệu $\underset{\theta}{\operatorname{argmin}}\, L(\theta)$ được hiểu là giá trị của θ để hàm số $L(\theta)$ đạt giá trị nhỏ nhất.

Mục đích trong hồi quy mong muốn nhận được là sai số giữa giá trị thực tế và giá trị ước lượng là nhỏ hoặc rất nhỏ, từ đây, phương pháp bình phương cực tiểu OLS (Ordinary Least Squares) được tiếp cận.

Sai số giữa giá trị thực tệ và giá trị nhận được biểu hiện như sau:

$$\sum e^2 = \sum (y_i - \widehat{y_i})^2 = RSS$$

Phương pháp OLS trong hồi quy đa biến sẽ có phương trình:

$$e^T e = \sum_{i=1}^{n} e_i^2 = y^T y - 2\hat{\beta} X^T y + \hat{\beta}^T X^T X \hat{\beta} \longrightarrow min$$

Vậy hàm mất mát của hồi quy đa biến sẽ có dạng:

$$L(\beta) = \sum_{i=1}^{n} \left(y_i - \beta_1 - \sum_{j=1}^{k} x_{ij} \beta_j \right)^2$$

Và ước lượng của các tham số dựa vào vào các phép toán ma trận, khi XX^Tkhả nghịch:

$$\widehat{\beta_\iota} = (XX^T)^{-1} Xy$$

1.3. Overfitting - Vấn đề quá khớp trong hồi quy

Quá khớp (overfitting) là một hiện tượng không mong muốn thường gặp trong việc xây dựng một mô hình hồi quy trong machine learning. Hiện tượng này rất phổ biến nên cần phải nắm được các kỹ thuật cần thiết để khắc phục hoặc né tránh hiện tượng này.

Hiện tượng này xảy ra khi mô hình quá cồng kềnh, phức tạp, hoặc quá nhiều tham số, dẫn đến hiệu suất dự đoán kém, vì nó sẽ phản ứng quá mạnh với các biến động nhỏ trong dữ liệu huấn luyện. Trong thống kê và học máy, một trong những nhiệm vụ phổ biến nhất làm phù hợp một mô hình với một tập dữ liệu huấn luyện (training data), với mục tiêu đưa ra các dự đoán đáng tin cậy về dữ liệu thử nghiệm không được xác định.

Một mô hình được coi là tốt nếu cả training error và test error đều thấp. Nếu training error thấp nhưng test error cao, ta nói mô hình bị quá khớp (overfitting). Nếu training test cao và test error cao, ta nói mô hình bị chưa khớp (underfitting).

1.4. Regularization – Phương pháp chính quy hóa

Từ đây, kết hợp với hai vấn đề đã đề cập nếu XX^T không khả nghịch, tức ta sẽ phải kết hợp mô hình machine learning với hàm mất mát và vấn đề quá khớp. Ta sẽ sử dụng hai kỹ thuật phổ biến giúp tránh các vấn đề này là validation và regularization. Đặc biệt nhóm sẽ chú trọng về phương pháp regularization gồm hai mô hình Ridge Regression và LASSO Regression (Least Absolute Shrinkage and Selection Operator).

Regularization là một kỹ thuật phổ biến giúp tránh quá khớp theo hướng làm giảm độ phức tạp của mô hình. Việc giảm độ phức tạp này có thể khiến lỗi huấn luyện tăng lên nhưng lại làm tăng tính tổng quát của mô hình. Dưới đây là một vài kỹ thuật kiểm soát.

2. Ridge Regression – Hồi quy Độ dốc

2.1. Định nghĩa

Hồi quy Ridge là một phương pháp ước tính tham số phổ biến được sử dụng để giải quyết vấn đề đa cộng tuyến thường phát sinh trong hồi quy đa biến. Giảm thiểu các hệ số trong mô hình hồi quy và tránh tình trạng quá khớp (overfitting). Khi số lượng biến đầu vào (predictor variables) lớn hơn số lượng quan sát (observations), hệ số ước lượng sẽ trở nên không ổn định và có thể gây ra hiện tượng overfitting.

Về bản chất, hồi quy Ridge tối ưu song song hai thành phần bao gồm tổng bình phương phần dư và thành phần hiệu chỉnh. Hàm mục tiêu của Hồi quy Ridge được xác định bằng cách cộng thêm một phần tử giá trị bình phương của hệ số vào trong hàm mục tiêu của mô hình hồi quy. Cụ thể, hàm mục tiêu của Ridge Regression có dạng:

$$L(\beta) = \sum_{i=1}^{n}\left(y_i - \beta_0 - \sum_{j=1}^{k} x_{ij}\beta_j\right)^2 + \lambda\sum_{j=1}^{k}\beta_j^2 = RSS + \lambda\sum_{j=1}^{k}\beta_j^2$$

Trong đó, $L(\beta)$ là hàm mục tiêu của, RSS là tổng bình phương sai số của mô hình hồi quy, λ là tham số được gọi là tham số hiệu chỉnh, là tổng bình phương của các hệ số.

Ridge sử dụng thuật toán hiệu chỉnh tham số nhằm thu nhỏ hệ số hồi quy biến kém tính giải thích, thậm chí nén về bằng 0 (Mangal & Holm, 2018). Tương tự, các biến tương quan cao, chứa đựng cùng thông tin giải thích sẽ bị lược bỏ, hệ số hồi quy bị nén về 0 (Doan & Kalita, 2015). Ridge phù hợp với dữ liệu tiềm ẩn tương quan cao giữa biến giải thích. Kết quả hồi quy Ridge còn được dùng như một bước sàng lọc biến, xây dựng mô hình học máy tối ưu (Mangal & Holm, 2018).

2.2. Tuning parameter – Vai trò của tham số (λ) trong Hồi quy Ridge

Tham số λ được gọi là tuning parameter (tham số hiệu chỉnh) ($\lambda > 0$), điều chỉnh độ phức tạp của mô hình, kiểm soát độ lớn của thành phần điều chỉnh tác động lên hàm mất mát.

Trong trường hợp λ rất lớn, hầu như tất cả các tham số mô hình suy giảm về 0 và được gọi là hiện tượng phù hợp dưới mức (underfitting). Khi λ rất nhỏ, hồi quy Ridge trở thành hồi quy tuyến tính thông thường. Điều này dẫn đến hiện tượng quá khớp (overfitting).

Khi giá trị của tham số hiệu chỉnh λ tăng, các giá trị của hệ số ước lượng sẽ giảm, do đó mô hình sẽ trở nên đơn giản hơn và ít bị ảnh hưởng bởi các giá trị nhiễu (noise) trong dữ liệu. Tuy nhiên, giá trị của tham số hiệu chỉnh cần được lựa chọn sao cho đủ lớn để tránh overfitting nhưng đủ nhỏ để giữ lại các giá trị hệ số quan trọng trong mô hình

2.3. Điểm tối ưu và điểm bất lợi của Hồi quy Ridge

2.3.1. Điểm tối ưu

Hồi quy Ridge giúp *giảm overfittin*g trong mô hình hồi quy tuyến tính bằng cách giới hạn độ lớn của các hệ số ước lượng. Việc giảm overfitting này có thể dẫn đến việc tăng khả năng dự báo chính xác của mô hình.

Việc giới hạn độ lớn của các hệ số ước lượng *giúp giảm độ phức tạp của mô hình*, điều này có thể giúp giảm chi phí tính toán và giúp các mô hình trở nên dễ hiểu hơn.

Thích hợp cho các bộ dữ liệu lớn: Hồi quy Ridge thường hoạt động tốt trên các bộ dữ liệu lớn, do đó nó có thể hữu ích cho việc phân tích, tính toán ước lượng.

2.3.2. Điểm bất lợi

Không thể loại bỏ hoàn toàn các biến không quan trọng: Mặc dù Hồi quy Ridge giúp giảm độ lớn của các hệ số ước lượng, nhưng nó không thể loại bỏ hoàn toàn các biến không quan trọng khỏi mô hình.

Việc tìm ra giá trị thích hợp cho tham số hiệu chỉnh λ là một vấn đề quan trọng và không phải lúc nào cũng dễ dàng.

Không giải quyết được vấn đề đa cộng tuyến: Hồi quy Ridge có thể giảm tác động của đa cộng tuyến, nhưng không thể giải quyết hoàn toàn vấn đề này.

Lỗi trong tập dữ liệu training có thể lớn hơn hồi quy OLS.

2.4. Ứng dụng của Hồi quy Độ dốc

Dự báo giá cổ phiếu: Hồi quy Ridge được sử dụng để dự báo giá cổ phiếu trong thị trường tài chính. Việc điều chuẩn giúp giảm thiểu ảnh hưởng của nhiễu và giúp tăng tính ổn định của mô hình dự báo. (Toại, T. K., Võ, H. T. X., & Võ, H. M. (2021). Áp dụng hồi quy Ridge và mạng nơron nhân tạo để dự báo giá ICO sau sáu tháng.)

Phân tích dữ liệu y tế: Hồi quy Ridge có thể được sử dụng để phân tích các bộ dữ liệu y tế. (Xuân, T. T., Nhân, T. V., Tùng, H. Đ. T., Hải, T. N., & Hưng, T. Đ. Tổng quan ứng dụng học máy trong dự đoán nguy cơ đa di truyền hướng tới y học cá thể hóa.)

Dự báo thời tiết: Hồi quy Ridge cũng được sử dụng trong các mô hình dự báo thời tiết để giảm thiểu ảnh hưởng của nhiễu và tăng tính ổn định của mô hình. (Huy, N. H., & Giang, H. T. T. Hướng tiếp cận hồi quy mới cho dự báo tốc độ gió.)
Phân tích dữ liệu tài chính: Hồi quy Ridge cũng có thể được sử dụng trong phân tích dữ liệu tài chính. (Xuân, P. T. T., & Trung, N. Đ. Tác động trực tiếp của tín dụng công nghệ đến bất bình đẳng thu nhập).

3. Lasso Regression

3.1. Định nghĩa

LASSO (Least Absolute Shrinkage Selection Sperator) là một phương pháp để ước lượng các tham số của mô hình hồi quy tuyến tính được đề xuất bởi Tibshirani (1996). Mục tiêu của LASSO là cực tiểu tổng bình phương các sai số với ràng buộc là tổng trị tuyệt đối của các tham số ước lượng trong mô hình nhỏ hơn một hằng số. Vì bản chất của ràng buộc này, phương pháp hồi quy LASSO có xu hướng thu nhỏ các tham số và tạo ra một số các tham số chính xác bằng không và từ đó đưa ra sự lựa chọn chính xác một tập hợp con của các tham số hồi quy mà không cần kiểm định giả thuyết, do đó không cần dùng P-value; đồng thời thể hiện sự ổn định mô hình hồi quy ngay cả trong trường hợp có đa cộng tuyến giữa các biến giải thích.

Phương pháp LASSO là cũng phương pháp hồi quy tuyến tính đa biến có hiệu chỉnh mô hình, trong phương pháp này các hệ số $\widehat{\beta_j}$ ($j = \overline{1;k}$) được ước tính dựa trên bài toán tìm cực trị của hàm:

$$S(\hat{\beta}) = \sum_{i=1}^{n}\left(y_i - \widehat{\beta_0} - \sum_{j=1}^{k} x_{ij}\widehat{\beta_J}\right)^2$$

Với điều kiện ràng buộc $\left\|\hat{\beta}\right\|_1 \leq s$

Trong đó $\left\|\hat{\beta}\right\|_1 = \sum_{j=1}^{k}\left|\widehat{\beta_J}\right|$ là chuẩn l_1 của vector $\hat{\beta}$ và s là một hằng số lớn hơn 0.

Từ đó, Bài toán cực trị có điều kiện tương đương bài toán Lagrange:

$$L(\beta) = \sum_{i=1}^{n}\left(y_i - \beta_0 - \sum_{j=1}^{k}\beta_j x_{ij}\right)^2 + \lambda\sum_{j=1}^{k}|\beta_j| = RSS + \lambda\sum_{j=1}^{k}|\beta_j|$$

Trong đó, λ là nhân tử Largrange dùng để điều chỉnh mô hình, chuẩn l_1 được dùng cho việc dự đoán các tham số.

Tuy nhiên, $\left\|\hat{\beta}\right\|_1 = \sum_{j=1}^{k}\left|\widehat{\beta_J}\right|$ vì là hàm lồi (nhưng không phải là hàm lồi nghiêm ngặt nên có thể có nhiều hơn một nghiệm) nhưng không khả vi. Do đó, không có công thức nghiệm cụ thể cho bài toán LASSO. Rõ ràng, hồi quy LASSO phụ thuộc vào tham số hiệu chỉnh λ để xác định các hệ số nào sẽ có giá trị bằng không. Tuy nhiên, chúng ta không thể sử dụng các đạo hàm riêng để tìm ra phương án tối ưu của bài toán Lagrange . Có một cách trực tiếp để xác định tham số λ đó là sử dụng phương pháp Cross-validation. Một cách thường được sử dụng của phương pháp Cross-validation là chia tập training ra k tập con không có phần tử chung, có kích thước gần bằng nhau. Tại mỗi lần kiểm thử , được gọi là run, một trong số k tập con được lấy ra làm validata set. Mô hình sẽ được xây dựng dựa vào hợp của k–1 tập con còn lại. Cách làm này còn có tên gọi là k-fold cross validation. Cuối cùng, chúng ta sẽ chọn λ nào cung cấp cho chúng ta trung bình bình phương của các train error và validation error nhỏ nhất, nghĩa là $MSE = \frac{\sum_{i=1}^{n} e_i^2}{n}$, trong đó e_i là chênh lệch giữa giá trị dự báo và giá trị thực tế

3.2. Tuning parameter – Vai trò của tham số λ trong Hồi quy Lasso

Tham số λ được gọi là tuning parameter (tham số hiệu chỉnh), có vai trò quyết định độ lớn của hệ số ước lượng trong mô hình.

Tham số càng lớn thì các hệ số ước lượng sẽ càng gần bằng 0, dẫn đến việc giảm số lượng biến độc lập ảnh hưởng đến kết quả dự đoán. Tuy nhiên, nếu giá trị quá lớn, thì sẽ có quá ít biến độc lập được giữ lại và mô hình sẽ trở nên quá đơn giản và thiếu khả năng dự đoán. Ngược lại, nếu giá trị quá nhỏ, mô hình sẽ bị overfitting, tức là mô hình quá phức tạp và quá khớp dữ liệu huấn luyện. Do đó, là rất quan trọng trong việc cân bằng giữa độ chính xác và độ đơn giản của mô hình.

Cách chọn giá trị tốt nhất thường được thực hiện bằng cách sử dụng các kỹ thuật như cross - validation để tìm ra giá trị lambda tối ưu nhất cho mô hình.

3.3. Điểm tối ưu và điểm bất lợi của Hồi quy Lasso

3.3.1. Điểm tối ưu

Giúp xác định các biến quan trọng: Hồi quy Lasso giúp xác định các biến độc lập quan trọng trong mô hình hồi quy tuyến tính. Cho phép ước lượng các hệ số có ảnh hưởng cao đến biến phụ thuộc và loại bỏ các hệ số không có ảnh hưởng

Khả năng xử lý các mô hình với nhiều biến: Hồi quy Lasso rất hữu ích trong việc xử lý các mô hình có nhiều biến độc lập. Việc loại bỏ các biến không quan trọng giúp giảm số lượng biến và đơn giản hóa mô hình.

Thích hợp cho các bộ dữ liệu lớn: Hoạt động tốt trên các bộ dữ liệu có kích thước lớn và có số lượng biến đầu vào lớn.

3.3.2. Điểm bất lợi

Nó có thể dẫn đến một số biến quan trọng bị loại bỏ khỏi mô hình nếu chúng có sự tương quan mạnh với các biến khác trong mô hình. Điều này có thể ảnh hưởng đến tính toàn vẹn của mô hình và dẫn đến kết quả dự đoán không chính xác.

Việc chọn giá trị thích hợp cũng là một thách thức, và nếu không chọn đúng giá trị thì có thể dẫn đến mô hình quá đơn giản hoặc quá phức tạp, ảnh hưởng đến hiệu suất dự đoán của mô hình.

Hồi quy LASSO không phù hợp cho các bộ dữ liệu có các biến đầu vào không phải là độc lập tuyến tính.

3.4. Ứng dụng của Hồi quy Lasso

Feature selection: có thể sử dụng để chọn các đặc trưng quan trọng và loại bỏ các đặc trưng không quan trọng trong mô hình. Điều này giúp giảm chiều dữ liệu và cải thiện hiệu suất của mô hình.

Dự báo trong kinh tế: ứng dụng trong phân tích dữ liệu về kinh tế để phân tích thị trường 1 cách có hiệu quả (Nguyễn, Đ. T. (2021). Hiệu quả trong dự báo giá dầu thô: Một so sánh giữa mô hình VAR, mô hình LASSO và mô hình LSTM.)

Mô hình hóa thời gian: sử dụng để mô hình hóa dữ liệu thời gian. (Nguyễn, Đ. T., Lê, H. A., & Đinh, T. P. A. (2021). Dự báo tăng trưởng kinh tế và lạm phát Việt Nam: một so sánh giữa mô hình Var, Lasso và MLP.).

Nghiên cứu y học: Hồi quy Lasso có thể được sử dụng để xác định các biến đầu vào quan trọng trong các nghiên cứu y học. Nó có thể giúp tìm ra các yếu tố quan trọng nhất ảnh hưởng đến kết quả của các cuộc thử nghiệm y học. (Vân, N. Q., & Hùng, T. Đ. (2022). Hồi quy LASSO và ứng dụng trong phân tích dữ liệu ung thư vú. TNU Journal of Science and Technology, 227(08), 433-440.\).

Phân tích dữ liệu môi trường: ứng dụng dự báo các chỉ số môi trường. (Nam, T. X., & Tùng, N. T. DEEP LEARNING: Ứng dụng cho dự báo lưu lượng nước đến hồ chứa Hòa Bình).

4. So sánh Ridge Regression và Lasso Regression

4.1. Điểm tương đồng

Cả hồi quy Ridge và hồi quy Lasso đều sử dụng các thành phần điều chỉnh để giới hạn giá trị của hệ số trong mô hình hồi quy. Điều này giúp giảm thiểu overfitting và giúp tăng tính ổn định của mô hình.

Tính toán đơn giản: Cả hai kỹ thuật đều có thể được tính toán bằng phương pháp gradient descent hoặc theo cách tính toán đóng (closed-form solution).

Cả hai mô hình hồi quy đều có thể được điều chỉnh thông qua tham số hiệu chỉnh λ trong hàm mất mát.

4.2. Điểm khác biệt chính

4.2.1. Mục tiêu của phương pháp

Ridge Regression tập trung vào việc giảm thiểu tổng bình phương của các hệ số ước lượng, trong khi Lasso Regression tập trung vào việc giảm tổng giá trị tuyệt đối của các hệ số ước lượng.

4.2.2. Cách xử lý vấn đề quá khớp

Ridge Regression giải quyết vấn đề quá khớp bằng cách giới hạn giá trị của hệ số ước lượng thông qua tham số , làm giảm độ lớn của các hệ số ước lượng.

Lasso Regression giải quyết vấn đề quá khớp bằng cách áp đặt hệ số ước lượng bằng 0 cho những biến không quan trọng, loại bỏ chúng khỏi mô hình.

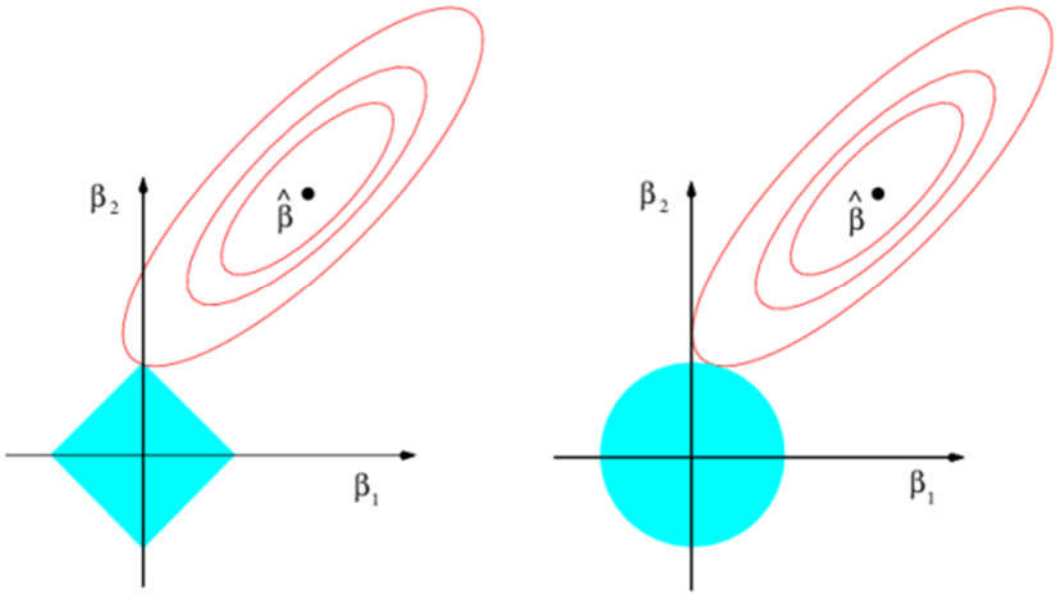
4.2.3. Sự lựa chọn của biến đầu vào

Ridge Regression giúp giảm thiểu tác động của các biến đầu vào có độ tương quan cao trên kết quả dự đoán bằng cách giảm độ lớn của hệ số ước lượng.

Lasso Regression thường chọn ra một tập con của các biến quan trọng để đưa vào mô hình, loại bỏ những biến không quan trọng.

Phương pháp LASSO và Ridge có thể được minh họa như sau:

Lasso Regression Rigde Regression



$|\beta_1| + |\beta_2| \leq s$ $\beta_1^2 + \beta_2^2 \leq s$

Lasso có thể đặt các hệ số bằng 0, trong khi hồi quy Ridge tương tự bề ngoài thì không thể. Điều này là do sự khác biệt về hình dạng của ranh giới ràng buộc của chúng. Cả hồi quy lasso và Ridge đều có thể được hiểu là cực tiểu hóa cùng một hàm mục tiêu.

Nhưng đối với các ràng buộc khác nhau: $\|\hat{\beta}\|_1$ cho Lasso và $\|\hat{\beta}\|_2^2$ cho Ridge. Hình vẽ cho thấy vùng giới hạn được xác định bởi chuẩn l_1 là một hình vuông được xoay sao cho các góc của nó nằm trên các trục (nói chung là đa giác chéo), trong khi vùng được xác định bởi chuẩn l_2 là một hình tròn (nói chung là n -sphere), invariant và do đó, không có góc. Như đã thấy trong hình, một đối tượng lồi nằm tiếp tuyến với ranh giới, chẳng hạn như đường được hiển thị (với điểm ước lượng và các khoảng ước lượng xung quanh) , có khả năng gặp một góc (hoặc tương đương ở chiều cao hơn) của một siêu khối, trong đó một số thành phần của giống hệt nhau bằng 0, trong khi trong trường hợp n - sphere, các điểm trên ranh giới mà một số thành phần của bằng 0 không phân biệt được với các đối tượng khác và đối tượng lồi không còn khả năng tiếp xúc với điểm mà tại đó một số thành phần của bằng không hơn một mà không có cái nào trong số chúng.

5. Kết luận

5.1. Mục đích chính của hai mô hình

Hồi quy Ridge và LASSO là hai phương pháp quan trọng trong học máy để giải quyết vấn đề overfitting trong quá trình huấn luyện mô hình. Mục đích chính của việc sử dụng hai phương pháp này là tìm ra một mô hình hồi quy tối ưu với độ chính xác cao trên dữ liệu mới.

Cả Ridge và LASSO đều là những kỹ thuật dựa trên việc giới hạn lượng thông tin được sử dụng trong quá trình huấn luyện mô hình hồi quy. Tuy nhiên, cách thức giới hạn này lại khác nhau giữa hai phương pháp. Ridge sử dụng regularization L2 để giới hạn các hệ số trong mô hình, trong khi LASSO sử dụng regularization L1.

Việc sử dụng hồi quy Ridge và LASSO có thể giúp giảm thiểu vấn đề overfitting, nâng cao khả năng dự đoán và giảm độ phức tạp của mô hình. Ngoài ra, hai phương pháp này còn có thể giúp xác định các biến quan trọng trong mô hình bằng cách loại bỏ các biến không quan trọng và giữ lại các biến quan trọng.

Ví dụ, trong việc dự đoán giá nhà, có thể sử dụng hồi quy Ridge và LASSO để loại bỏ các biến không quan trọng như kích thước sân vườn hay khoảng cách đến trung tâm thành phố và giữ lại các biến quan trọng như số phòng ngủ, số phòng tắm và diện tích căn nhà để tối ưu hóa mô hình dự đoán.

Tóm lại, mục đích chính của việc sử dụng hồi quy Ridge và LASSO là giảm thiểu overfitting, nâng cao khả năng dự đoán và giảm độ phức tạp của mô hình, cũng như giúp xác định các biến quan trọng trong mô hình.

5.2. Định hướng

Hồi quy Elastic Net là một phương pháp học máy kết hợp giữa hồi quy Ridge và hồi quy Lasso, cung cấp một giải pháp linh hoạt và hiệu quả hơn để xử lý vấn đề overfitting và lựa chọn biến trong mô hình.

Một trong những đóng góp chính của hồi quy Elastic Net là giải quyết vấn đề có thể xảy ra với hồi quy Ridge và hồi quy Lasso khi số lượng biến đầu vào lớn và chúng tương quan cao với nhau. Trong trường hợp này, hồi quy Ridge có thể không đưa ra kết quả tốt vì nó cần phải giữ tất cả các biến, trong khi hồi quy Lasso có thể loại bỏ những biến quan trọng. Hồi quy Elastic Net giúp giải quyết vấn đề này bằng cách kết hợp cả hai phương pháp và sử dụng một tham số điều chỉnh để điều tiết tỷ lệ giữa hồi quy Ridge và hồi quy Lasso.

Hồi quy Elastic Net cũng có nhiều lợi ích như giảm thiểu overfitting, tạo ra mô hình ổn định và xác định các biến quan trọng trong mô hình. Bằng cách kết hợp cả hai phương pháp, nó giúp giảm thiểu số lượng biến không quan trọng và đồng thời giữ lại những biến có tác động lớn đến kết quả dự đoán. Kết quả là mô hình có thể dự đoán tốt hơn trên dữ liệu mới và độ chính xác cao hơn.

Cuối cùng, hồi quy Elastic Net cũng cải thiện tốc độ huấn luyện mô hình. Khi số lượng biến đầu vào lớn, thời gian huấn luyện mô hình có thể rất lâu. Tuy nhiên, hồi quy Elastic Net có thể giảm số lượng biến đầu vào và giới hạn giá trị của các hệ số, giúp tăng tốc độ huấn luyện mô hình.

Tóm lại, hồi quy Elastic Net là một phương pháp học máy hiệu quả, cung cấp giải pháp linh hoạt để xử lý vấn đề overfitting và lựa chọn biến trong mô hình. Nó giúp cải thiện độ chính xác của dự đoán, tăng tốc độ huấn luyện mô hình.

TÀI LIỆU THAM KHẢO

[1] James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). An introduction to statistical learning (Vol. 112, p. 18). New York: springer.

[2] McDonald, G. C. (2009). Ridge regression. Wiley Interdisciplinary Reviews: Computational Statistics, 1(1), 93-100.

[3] Toại, T. K., Hạnh, V. T. X., & Huân, V. M. (2023). Áp dụng hồi quy Ridge và mạng nơron nhân tạo để dự báo giá ICO sau sáu tháng. Tạp chí Khoa học Đại học Mở Thành phố Hồ Chí Minh-Kinh tế và Quản trị Kinh doanh, 18(4).

[4] Phạm, U. H., & Võ, U. T. L. (2019). Hồi quy LASSO kết hợp với Hồi quy Ridge trong phân tích kinh tế. *International conference for Young researchers in Economics & Business 2019 Icyreb 2019*