Thuyết trình LR

Cân nhắc về Multiple linear regression

Multiple linear regression là một phương pháp hồi quy được sử dụng để dự đoán một biến phụ thuộc từ nhiều biến độc lập. Phương pháp này dựa trên giả định rằng mối quan hệ giữa biến phụ thuộc và biến độc lập là tuyến tính.

Khi sử dụng Multiple linear regression, có một số cân nhắc quan trọng cần lưu ý, bao gồm:

. Overfitting

Slide

**Multicollinearity (đa cộng tuyến)**

Giới thiệu

Multicollinearity là hiện tượng các biến độc lập trong mô hình hồi quy phụ thuộc tuyến tính lẫn nhau. Điều này có nghĩa là một biến độc lập có thể được giải thích hoàn toàn bằng một hoặc nhiều biến độc lập khác.

Hậu quả

Đa cộng tuyến có thể gây ra những hậu quả sau:

Hệ số hồi quy không đáng tin cậy: Giá trị của hệ số hồi quy có thể bị phóng đại hoặc thu hẹp, dẫn đến kết luận không chính xác về tác động của một biến đối với biến mục tiêu.

Sai số chuẩn của hệ số hồi quy tăng lên: Điều này làm cho các ước lượng không đáng tin cậy hơn.

Tầm quan trọng thống kê của các biến độc lập giảm xuống: Điều này có thể khiến các nhà nghiên cứu loại bỏ các biến độc lập có giá trị thực sự.

Các phương pháp phát hiện đa cộng tuyến

Có hai phương pháp chính để phát hiện đa cộng tuyến:

Tương quan theo cặp: Kiểm tra mối tương quan theo cặp giữa các cặp biến độc lập khác nhau có thể mang lại những hiểu biết hữu ích trong việc phát hiện đa cộng tuyến. Nếu hai biến độc lập có mối tương quan cao (trên 0,8), thì chúng có thể có mối quan hệ đa cộng tuyến.

Hệ số lạm phát phương sai (VIF): VIF là một thước đo tổng quát hơn về mức độ đa cộng tuyến. VIF được tính cho mỗi biến độc lập và được định nghĩa như sau:

VIF = 1 / (1 - R^2)

Trong đó, R^2 là hệ số tương quan bình phương giữa biến độc lập đó và các biến độc lập còn lại.

Các giá trị VIF cao (trên 10) cho thấy có thể có đa cộng tuyến. Tuy nhiên, không có một giá trị VIF cụ thể nào được coi là ngưỡng để xác định đa cộng tuyến. Giá trị VIF phụ thuộc vào số lượng biến độc lập trong mô hình.

Các phương pháp xử lý đa cộng tuyến

Có một số phương pháp để xử lý đa cộng tuyến, bao gồm:

Xóa bỏ một hoặc nhiều biến độc lập: Đây là phương pháp đơn giản nhất, nhưng có thể dẫn đến mất thông tin.

Thay đổi cách đo lường các biến độc lập: Ví dụ, biến độc lập là tỷ lệ có thể được chuyển đổi thành biến độc lập là số lượng.

Sử dụng phương pháp hồi quy khác: Ví dụ, hồi quy Ridge hoặc hồi quy Lasso có thể giúp giảm thiểu tác động của đa cộng tuyến.

**Overfitting và Underfitting**

(Slide)

SỰ ĐÁNH ĐỔI PHƯƠNG SAI THIÊN VỊ

Hình ảnh này thể hiện mối quan hệ giữa độ phức tạp của mô hình và độ chính xác của mô hình.

Trên hình ảnh, trục x thể hiện độ phức tạp của mô hình, trong khi trục y thể hiện độ chính xác của mô hình. Có thể thấy rằng, khi độ phức tạp của mô hình tăng lên, độ chính xác của mô hình cũng tăng lên. Tuy nhiên, khi độ phức tạp của mô hình tiếp tục tăng lên, độ chính xác của mô hình bắt đầu giảm xuống.

Điều này là do mô hình bắt đầu học các đặc điểm nhiễu trong dữ liệu huấn luyện. Các đặc điểm nhiễu là những đặc điểm không liên quan đến biến mục tiêu và chỉ xuất hiện ngẫu nhiên trong dữ liệu huấn luyện.

Mục tiêu của việc xây dựng mô hình học máy là tìm ra điểm tối ưu trên đường cong độ phức tạp và phương sai. Điểm tối ưu là điểm mà độ chính xác trên dữ liệu thử nghiệm cao nhất.

Trong hình ảnh, điểm tối ưu là điểm mà đường cong độ phức tạp và phương sai cắt nhau. Ở điểm này, mô hình có độ chính xác cao trên dữ liệu huấn luyện và dữ liệu thử nghiệm, đồng thời ít bị overfitting.

**OVERFITTING**

Overfitting là hiện tượng mô hình quá phù hợp với dữ liệu huấn luyện, dẫn đến việc mô hình hoạt động kém trên dữ liệu thử nghiệm.

Cách khắc phục Overfitting

Để khắc phục Overfitting, có thể thực hiện một số biện pháp sau:

Giảm độ phức tạp của mô hình: Có thể giảm số lượng tham số của mô hình hoặc sử dụng các kỹ thuật như regularization.

Tăng kích thước của tập dữ liệu huấn luyện: Tập dữ liệu huấn luyện lớn hơn sẽ giúp mô hình học được các mối quan hệ quan trọng hơn và giảm khả năng học các đặc điểm nhiễu.

Sử dụng các kỹ thuật kiểm soát Overfitting: Có một số kỹ thuật kiểm soát Overfitting, chẳng hạn như regularization, cross-validation và bootstrapping.

**UNDERFITTING**

Underfitting là hiện tượng mô hình không phù hợp với dữ liệu huấn luyện, dẫn đến việc mô hình hoạt động kém trên cả dữ liệu huấn luyện và dữ liệu thử nghiệm.

Để khắc phục Underfitting, có thể thực hiện một số biện pháp sau:

Xác thực chéo

Tăng độ phức tạp của mô hình: Có thể tăng số lượng tham số của mô hình hoặc sử dụng các kỹ thuật như feature engineering.

Thu thập thêm dữ liệu huấn luyện: Tập dữ liệu huấn luyện lớn hơn sẽ giúp mô hình học được các mối quan hệ quan trọng hơn.

Sử dụng các kỹ thuật cải thiện khả năng học của mô hình: Có một số kỹ thuật cải thiện khả năng học của mô hình, chẳng hạn như boosting và bagging.

**Có một số cách để phân biệt Overfitting và Underfitting, bao gồm:**

Kiểm tra độ chính xác trên dữ liệu huấn luyện và dữ liệu thử nghiệm: Nếu độ chính xác trên dữ liệu huấn luyện cao hơn đáng kể so với độ chính xác trên dữ liệu thử nghiệm, thì có thể mô hình đang bị Overfitting.

Kiểm tra độ lệch và phương sai: Nếu độ lệch cao, thì có thể mô hình đang bị Overfitting. Nếu phương sai cao, thì có thể mô hình đang bị Underfitting.

Kiểm tra các đặc điểm của mô hình: Nếu mô hình có nhiều tham số, thì có thể mô hình đang bị Overfitting.

**ứng dụng (SLIDE)**