***Nhóm đồ án 3***

***Phạm Trường Giang - 15110036***

***Nguyễn Ngọc Hoàng Phúc***

**Robust Logistic Regression and Classification**

*Một phần của: Advances in Neural Information Processing Systems 27 (NIPS 2014)*

***Authors:***

**Jiashi Feng**

**Huan Xu**

**Shie Mannor**

**Shuicheng Yan**

# Tóm tắt:

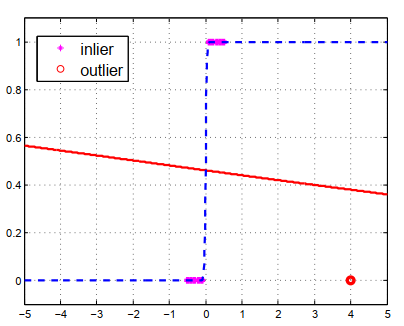
Xem xét hồi quy logistic với các ngoại lệ tùy ý (arbitrary outliers) trong ma trận hiệp phương sai (covariate matrix). Chúng tôi đề xuất một thuật toán hồi quy logistic mạnh mẽ mới, tiếng Anh là Robust Logistic Regression and Classification, gọi tắt là RoLR, ước tính tham số thông qua một thủ tục lập trình tuyến tính đơn giản. Chúng tôi chứng minh rằng RoLR mạnh mẽ cho một phần không đổi của ngoại lệ đối nghịch. Theo chúng tôi, đây là kết quả đầu tiên về ước tính mô hình hồi quy logistic khi ma trận hiệp phương sai bị sai với bất kỳ đảm bảo hiệu năng nào. Bên cạnh hồi quy, chúng tôi áp dụng RoLR để giải quyết các bài toán phân loại nhị phân trong đó một phần nhỏ các mẫu training bị sai.

# 1. Giới Thiệu

Hồi quy logistic (LR) là một mô hình phân lớp xác suất thống kê tiêu chuẩn đã được sử dụng rộng rãi trong các lĩnh vực: thị giác máy tính, tiếp thị, khoa học xã hội, … Khác với hồi quy tuyến tính, kết quả của LR trên mỗi mẫu là *xác suất* positive hoặc negative, trong đó xác suất phụ thuộc vào độ đo tuyến tính của mẫu. Do đó, LR được sử dụng rộng rãi để phân lớp. Chính thức hơn, với mẫu  có nhãn được ký hiệu là , xác suất của positive được dự đoán là cho bởi tham số β của mô hình LR . Để có được một tham số chạy tốt, thường là một tập các mẫu được gán nhãn được thu thập để học tham số β giúp tối đa hóa hàm khả năng gây ra trên các mẫu trainning.

Tuy nhiên, trong thực tế, các mẫu training thường nhiễu và một số có thể thậm chí có chứa adversarial corruptions. Ở đây bởi "adversarial", chúng tôi dự định rằng corruption có thể là tùy ý, không bị chặn và không phải từ bất kỳ phân phối cụ thể nào. Ví dụ: trong nhiệm vụ phân loại hình ảnh/video, một số hình ảnh hoặc video có thể bị hỏng bất ngờ do lỗi của cảm biến hoặc sự vướng mắc nghiêm trọng trên các đối tượng chứa. Những mẫu bị hỏng, được gọi là các outlier, có thể làm sai lệch các ước lượng tham số nghiêm trọng và do đó phá hủy hiệu suất của LR.

Để thấy sự nhạy cảm của LR với các outlier trực quan hơn, hãy xem xét một ví dụ đơn giản sau, nơi tất cả các mẫu của là từ không gian một chiều , như trong Hình 1. Chỉ sử dụng mẫu inlier cung cấp một tham số LR chính xác (ở đây chúng tôi chỉ ra đường cong hàm cảm sinh) giải thích các inlier tốt. Tuy nhiên, khi chỉ có một mẫu bị hỏng (ban đầu là negative nhưng bây giờ gần hơn với các mẫu positive), đường cong hồi quy kết quả bị phân tán xa khỏi mặt đất sự thật một và các dự đoán nhãn trên các nội dung liên quan là hoàn toàn sai. Điều này chứng tỏ LR đó thực sự mong manh để lấy mẫu tham nhũng. Một cách nghiêm ngặt hơn, sự không mạnh mẽ của LR có thể thể hiện qua tính toán hàm ảnh hưởng của nó [7] (chi tiết trong tài liệu bổ sung).



Hình 1: Đường cong hồi quy logistic ước tính (đường màu đỏ đậm) nằm cách xa đường cong chính xác (nét đứt xanh dương) do sự tồn tại của chỉ một ngoại lệ (vòng tròn màu đỏ).

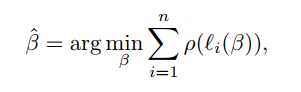
Như Hình 1 cho thấy, ước lượng tối đa khả năng LR là cực kỳ nhạy cảm với sự có mặt của dữ liệu bất thường trong mẫu. Pregibon cũng quan sát thấy sự không mạnh mẽ của LR trong [14]. Để giải quyết bài toán quan trọng này của LR, Pregibon [14], Cook và Weisberg [4] và Johnson [9] đề xuất các thủ tục để xác định các quan sát có ảnh hưởng đến ước tính β dựa trên một số đo lường ngoại lệ nhất định. Stefanski et al. [16, 10] và Bianco et al. [2] cũng đề xuất các nhà ước tính mạnh mẽ, tuy nhiên, yêu cầu phải ước tính mạnh mẽ ma trận covariate hoặc ranh giới trên các ngoại lệ. Hơn nữa, điểm phân tích1 của các phương pháp này thường tỷ lệ nghịch với chiều kích mẫu và giảm nhanh chóng đối với các mẫu có chiều cao.

Chúng tôi đề xuất một thuật toán hồi quy logistic mạnh mẽ mới, được gọi là RoLR, tối ưu hóa một mối tương quan tuyến tính mạnh mẽ giữa đáp ứng y và phép đo tuyến tính hβ; xi qua một thủ tục lập trình tuyến tính hiệu quả. Chúng tôi chứng minh rằng RoLR được đề xuất đạt được sự mạnh mẽ để tùy tiện biến đổi tham nhũng. Ngay cả khi một phần không đổi của các mẫu training bị hỏng, RoLR vẫn có thể tìm hiểu tham số LR với một giới hạn trên không tầm thường về lỗi. Bên cạnh sự bảo đảm lý thuyết này của RoLR về ước lượng tham số, chúng tôi cũng cung cấp các giới hạn rủi ro thực nghiệm và dân số cho RoLR. Hơn nữa, RoLR chỉ cần giải quyết một bài toán lập trình tuyến tính và do đó có thể mở rộng với các tập dữ liệu quy mô lớn, tương phản rõ ràng với các thuật toán tối ưu hóa LR trước đây thường sử dụng phương pháp lặp lại được tính toán (tốn kém tính toán) [11]. RoLR được đề xuất có thể dễ dàng thích nghi để giải quyết các bài toán phân loại nhị phân, nơi có các mẫu training bị hỏng. Chúng tôi cũng cung cấp bảo đảm hiệu suất phân loại lý thuyết cho RoLR. Do giới hạn về không gian, chúng tôi trì hoãn tất cả các bằng chứng cho tài liệu bổ sung.

Chúng tôi đề xuất một thuật toán hồi quy logistic mạnh mẽ mới, được gọi là RoLR, tối ưu hóa một mối tương quan tuyến tính mạnh mẽ giữa đáp ứng y và phép đo tuyến tính hβ; xi qua một thủ tục lập trình tuyến tính hiệu quả. Chúng tôi chứng minh rằng RoLR được đề xuất đạt được sự mạnh mẽ để tùy tiện biến đổi tham nhũng. Ngay cả khi một phần không đổi của các mẫu training bị hỏng, RoLR vẫn có thể tìm hiểu tham số LR với một giới hạn trên không tầm thường về lỗi. Bên cạnh sự bảo đảm lý thuyết này của RoLR về ước lượng tham số, chúng tôi cũng cung cấp các giới hạn rủi ro thực nghiệm và dân số cho RoLR. Hơn nữa, RoLR chỉ cần giải quyết một bài toán lập trình tuyến tính và do đó có thể mở rộng với các tập dữ liệu quy mô lớn, tương phản rõ ràng với các thuật toán tối ưu hóa LR trước đây thường sử dụng phương pháp lặp lại được tính toán (tốn kém tính toán) [11]. RoLR được đề xuất có thể dễ dàng thích nghi để giải quyết các bài toán phân loại nhị phân, nơi có các mẫu training bị hỏng. Chúng tôi cũng cung cấp bảo đảm hiệu suất phân loại lý thuyết cho RoLR. Do giới hạn về không gian, chúng tôi trì hoãn tất cả các bằng chứng cho tài liệu bổ sung.

# 2. Các công việc liên quan

Một số công việc trước đây đã điều tra nhiều cách tiếp cận để củng cố hồi quy logistic(LR) [15, 13, 17, 16, 10]. Phần lớn trong số họ là M-estimator dựa trên: giảm thiểu một phức tạp và hàm mất mát mạnh mẽ hơn so với hàm mất tiêu chuẩn (khả năng ghi log âm) của LR. Dành choví dụ, Pregiobon [15] đã đề xuất M-estimator sau:

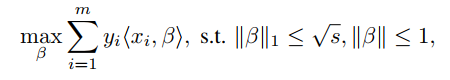
  

trong đó là negative log-likelihood của mẫu thứ i và là hàm kiểu Huber [8] chẳng hạn như

với c một tham số dương. Tuy nhiên, kết quả từ ước lượng như vậy không mạnh mẽ với các ngoại lệ với các biến số tác dụng đòn bẩy cao như được thể hiện trong [5].

Gần đây, Ding và cộng sự đã giới thiệu hồi quy T -logistic như là một thay thế mạnh mẽ cho LR tiêu chuẩn, thay thế phân phối theo cấp số nhân trong LR theo họ phân phối theo cấp số nhân. Tuy nhiên, hồi quy T -logistic chỉ đảm bảo rằng tham số đầu ra hội tụ tới một hàm tối ưu cục bộ của hàm mất thay vì hội tụ với tham số chân lý mặt đất.

Công việc của chúng tôi phần lớn được lấy cảm hứng từ việc theo dõi hai tác phẩm gần đây [3, 13] về hồi quy thưa thớt mạnh mẽ. Trong [3], Chen et al. đề xuất thay thế sản phẩm vector bên trong tiêu chuẩn bằng một sản phẩm được cắt tỉa, và thu được thuật toán hồi quy tuyến tính mới, mạnh mẽ đối với các tham nhũng biến đổi không liên kết. Trong công việc này, chúng tôi cũng sử dụng hoạt động đơn giản nhưng mạnh mẽ này để đạt được sự vững mạnh. [13], một lồi phương pháp lập trình để ước tính các thông số thưa thớt của mô hình hồi quy logistic được đề xuất:



trong đó s là tham số trước thưa thớt trên β. Tuy nhiên, phương pháp này không phải là mạnh mẽ để ma trận covariate bị hỏng. Một vài hoặc thậm chí một mẫu bị hỏng có thể thống trị mối tương quan trong hàm mục tiêu và tạo ra các ước tính bất thường xấu. Trong công việc này, chúng tôi đề xuất một thuật toán mạnh mẽ để khắc phục bài toán này.

# 3. Hồi quy logistic mạnh mẽ

**3.1. Thiết lập bài toán**

Chúng tôi xét bài toán của hồi quy logistic (LR). Cho biết ký hiệu hình cầu đơn vị và  ký hiệu đơn vị bóng Euclidean trong . Gọi  là tham số groundtruth của mô hình LR. Giả sử các mẫu training là các cặp covariate-response , nếu không bị hỏng, sẽ tuân theo mô hình LR sau:



trong đó hàm τ (·) được định nghĩa là:. Độ nhiễu là 1 i.i.d. Biến ngẫu nhiên Gaussian với số không trung bình và phương sai của. Đặc biệt, khi chúng ta xem xét trường hợp không ồn, chúng ta giả định  = 0. Vì LR chỉ phụ thuộc vào, chúng ta luôn có thể mở rộng các mẫu xi để làm cho độ lớn của  nhỏ hơn 1. Do đó, không mất tính tổng quát, chúng ta giả định rằng.

Trong số các mẫu , một số không đổi  của các mẫu có thể bị hỏng một cách bất lợi và chúng tôi không đưa ra giả định nào về các ngoại lệ này. Trong suốt bài báo, chúng tôi sử dụng  để biểu thị phần ngoại lệ. Chúng tôi gọi các mẫu “xác thực” còn lại không bị hỏng, tuân theo thiết kế chuẩn Gaussian chuẩn sau đây [12, 3].

**Definition 1** (Sub-Gaussian design).

Chúng ta nói rằng một ma trận ngẫu nhiên X = [x1; :::; xn] 2 Rp × n là phụ Gaussian với tham số  nếu: (1) mỗi cột được lấy mẫu độc lập với phân bố zero-mean với hiệp phương sai và (2) với bất kỳ đơn vị vector , biến ngẫu nhiên  là phụ Gaussian với tham số2 .

Các biến ngẫu nhiên phụ Gaussian ở trên có một số thuộc tính tập trung tốt, một trong số đó được ghi trong Lemma [12] sau.

**Lemma 1** (Sub-Gaussian Concentration [12])

Cho là *n* i.i.d. các biến ngẫu nhiên sub-Gaussian zero-mean với tham số  và phương sai lớn nhất . Ta có:

, với xác suất ít nhất đối với một số hằng số tuyệt đối c1.

Dựa trên thuộc tính concentration trên, chúng ta có thể đạt được giới hạn sau về độ lớn của một tập hợp các biến ngẫu nhiên phụ Gaussian [3].

**Lemma 2**

Giả sử  là n biến ngẫu nhiên sub-Gaussian độc lập với tham số. Và ta có  với xác suất ít nhất là .

Ngoài ra, Lemma này cung cấp một giới hạn thô về độ lớn của các mẫu inlier, và ràng buộc này đóng vai trò như một ngưỡng để xử lý trước các mẫu trong thuật toán RoLR sau đây.

**3.2 Thuật toán RoLR**

Bây giờ chúng tôi tiến hành giới thiệu các chi tiết của thuật toán hồi quy mạnh mẽ (RoLR) được đề xuất. Về cơ bản, RoLR trước tiên loại bỏ các mẫu có độ lớn quá lớn và sau đó tối đa hóa mối tương quan đã cắt của các mẫu còn lại với mô hình LR ước tính. Trực giác đằng sau RoLR tối đa hóa mối tương quan được cắt tỉa là: nếu các ngoại lệ có độ lớn quá lớn, chúng sẽ không đóng góp vào sự tương quan và do đó không ảnh hưởng đến việc học tham số LR. Nếu không, họ có giới hạn ảnh hưởng đến việc học LR (mà thực sự có thể bị ràng buộc bởi các mẫu inlier do chúng tôi áp dụng thống kê tỉa). Thuật toán 1 cung cấp các chi tiết thực hiện của RoLR.

**Algorithm 1** RoLR

**Input:** Các mẫu đào tạo bị ô nhiễm, một giới hạn trên về số lượng các ngoại lệ , số lượng các tham số n và số chiều mẫu p.

**Initialization:** đặt 

**Preprocessing:** Lấy mẫu  có độ lớn thỏa mãn .

Giải quyết bài toán lập trình tuyến tính sau đây (xem Eqn. (3)):



**Ouput:** .

Lưu ý rằng, trong thuật toán RoLR, chúng tôi cần tối ưu hóa thống kê được sắp xếp sau:

 (1)

trong đó là một thống kê được sắp xếp sao cho và z ký hiệu biến có liên quan. Bài toán trong Eqn. (2) tương đương với việc giảm thiểu tổng kết của các biến n hàng đầu, đó là một lồi và có thể được giải quyết bởi một giải pháp off-the-shelf (chẳng hạn như CVX). Ở đây, chúng ta lưu ý rằng nó cũng có thể được chuyển đổi thành bài toán lập trình tuyến tính sau đây (với một ràng buộc bậc hai), có hiệu quả tính toán cao hơn. Để thấy điều này, trước tiên chúng tôi giới thiệu các biến phụ trợ là chỉ số về các thuật ngữ tương ứng  rơi vào n nhỏ nhất. Sau đó, chúng tôi viết bài toán trong Eqn. (2) là



Ở đây các ràng buộc của là từ việc tái cấu trúc chuẩn của  Bây giờ, bài toán trên trở thành một chương trình tuyến tính max-min. Để tách các biến β và , chúng ta chuyển sang giải quyết dạng kép của bài toán giảm thiểu bên trong. Gọi ν, và ξi là bội số Lagrange cho các ràng buộc và tương ứng. Sau đó, biểu mẫu kép w.r.t. ti của bài toán trên là:

 (3)

Cải tiến hồi quy logistic thành một bài toán lập trình tuyến tính như trên làm tăng đáng kể khả năng mở rộng của LR trong việc xử lý các tập dữ liệu quy mô lớn, một thuộc tính rất hấp dẫn trong thực tế, vì lập trình tuyến tính được biết là hiệu quả tính toán và không có vấn đề gì với 1 × 106 biến trong một PC tiêu chuẩn.

**3.3 Đảm bảo hiệu suất cho RoLE**

Ngược lại với các thuật toán LR truyền thống, RoLR không thực hiện ước tính khả năng tối đa. Thay vào đó, RoLR tối đa hóa tương quan . Chiến lược này làm giảm độ phức tạp tính toán của LR, và quan trọng hơn là tăng cường độ mạnh của việc ước lượng tham số, sử dụng thực tế là các mẫu xác thực thường có mối tương quan dương giữa và , như được mô tả trong lemma sau đây.

**Lemma 3.** Cho . Giả sử mẫu được tạo ra bởi mô hình được mô tả ở mục (1). Kỳ vọng của được tính như sau:

trong đó, là một biến ngẫu nhiên Gaussian và là mức ồn trong (1), hơn nữa kỳ vọng trên có thể bị ràng buộc như sau:

trong đó và là 2 cực trị. Chúng còn có thể được biểu diễn và

Lemma sau đây cho thấy sự khác biệt của các mối tương quan là một đại diện hiệu quả cho sự khác biệt các tham số LR. Vì vậy, chúng ta luôn có thể giảm thiểu sự khác biệt của thông qua việc tối đa hóa .

**Lemma 4.** Cho như một tham số groundtruth trong (1) và . Biểu thị . Vậy

Và

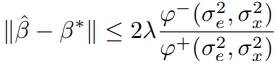
Dựa trên hai bổ đề này, cùng với một số tính chất tập trung của các mẫu bên trong (được hiển thị trong tài liệu bổ sung), chúng tôi bảo đảm việc thực hiện của RoLR trên mô hình phục hồi tham số LR.

**Định lý 1** (RoLR để khôi phục tham số LR). Cho là phân số ngoại lệ, là đầu ra của **thuật toán 1**, và là tham số ground truth. Giả sử có n mẫu đã được xác thực được tạo ra từ mô hình được mô tả ở (1). Với xác suất lớn hơn ta có



Ở đây lúc nào cũng là hằng số.

**Ghi chú 1.** Để làm cho các kết quả trên rõ ràng hơn, chúng ta xem xét trường hợp tiệm cận nơi . Vì vậy, các giới hạn trên trở thành



Với xác suất lớn hơn . Trong trường hợp không nhiễu, , và giả định , chúng ta có và . Tỉ lệ là . Vì vậy ràng buộc được đơn giản hóa

Nhớ lại rằng và giá trị cực đại của là 2. Như vậy so với kết quả trước đó là không đáng kể, chúng ta cần , cụ thể là . Nói cách khác, trong trường hợp không nhiễu, RoLR có thể ước tính tham số LR với một sai số không đáng kể (còn được gọi là “breakdown point”) với tối đa các mẫu bị ngoại lệ.

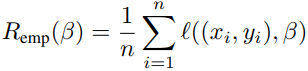
# 4. Empirical and Population Risk Bounds of RoLR

Bên cạnh việc phục hồi tham số, chúng tôi cũng lo ngại về hiệu suất dự đoán của mô hình LR ước tính trong thực tế. Hàm dự đoán tiêu chuẩn hao tổn của LR là không âm và hàm bị chặn được định nghĩa là

 (4)

Sự tin cậy của một yếu tố dự báo LR, được đo bằng rủi ro phổ biến của nó

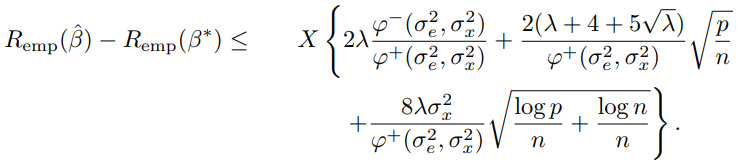
Trong đó P(X,Y) mô tả sự phân bố chung của sự biến đổi X và đáp ứng Y. Tuy nhiên, rủi ro phân bố hiếm khi có thể được tính toán trực tiếp như việc phân phối P(X,Y) thường không được biết. Trong thực hành, chúng ta thường xem xét rủi ro thực nghiệm được tính toán qua các mẫu cung cấp đã được training như sau



Lưu ý rằng rủi ro thực nghiệm chỉ được tính toán trên các mẫu xác thực, do đó không thể trực tiếp và được tối ưu hóa khi các ngoại lệ tồn tại.

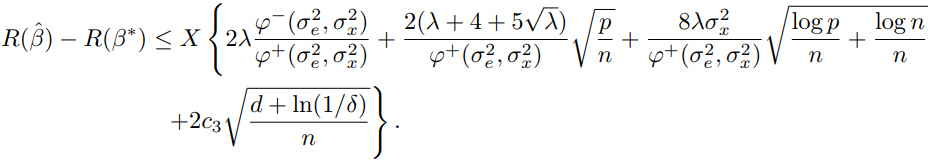
Dựa trên giới hạn của cung cấp trong **định lý 1**, chúng ta có thể dễ dàng có được ràng buộc về rủi ro thực nghiệm đối với RoLR là hàm hao tổn được định nghĩa ở **(4)** là Lipschitz liên tục.

**Hệ quả 1** (Ràng buộc về rủi ro thực nghiệm). Gọi là đầu ra của **Thuật toán 1** và là tham số tối ưu giảm thiểu rủi ro thực nghiệm. Giả sử rằng có n mẫu xác thực được tạo bởi mô hình được mô tả trong **(1)**. Định nghĩa . Chúng ta có với xác suất lớn hơn rủi ro thực nghiệm của bị giới hạn bởi



Do rủi ro thực nghiệm bị ràng buộc, chúng ta có thể dễ dàng có được sự ràng buộc về rủi ro phân bố bằng cách so sánh với các kết quả tổng quát tiêu chuẩn về các mức độ phức tạp của lớp hàm khác nhau. Một số được sử dụng rộng rãi các phép đo phức tạp bao gồm kích thước VC-dimension [18] và độ phức tạp của Rademacher và Gaussian. So với độ phức tạp của Rademacher phụ thuộc vào dữ liệu, VC-dimension thì tổng quát hơn mặc dù tổng quát hóa kết quả có thể hơi lỏng lẻo. Ở đây, chúng tôi áp dụng VC-dimension để đo lường sự phức tạp của hàm và thu được các rủi ro phân bố sau đây bị ràng buộc.

**Hệ quả 2** (Ràng buộc về rủi ro phân bố). Gọi là đầu ra của **Thuật toán 1** và là tham số tối ưu. Giả sử không giam tham số có kích thước VC-dimension hữu hạn là d. Có n mẫu đã xác thực được tạo ra bởi mô hình được mô tả ở (1). Định nghĩa . Chúng ta có với xác suất lớn hơn rủi ro phân bố của bị giới hạn bởi



Ở đây c2 và c3 luôn là hằng số.

# 5 Robust Binary Classification (Phân lớp nhị phân tăng cường)

5.1 Problem Setup

Khác với mô hình tạo mẫu cho LR, trong thiết lập phân loại nhị phân chuẩn, nhãn của một mẫu được xác định chính xác bằng dấu của phép đo tuyến tính của mẫu . Cụ thể, các mẫu được tạo bởi mô hình sau:

(5)

Ở đây là một biến Gaussian như ở phần (1). Vì được xác định quan hệ với , độ tương quan đạt được giá trị cực đại trong bước này (Lemma 5), mà điều này đảm bảo rằng RoLR vẫn hoạt động tốt để phân lớp. Chúng tôi một lần nữa giả định rằng các mẫu đã được training chứa n mẫu xác thực và ở hầu hết n1 các ngoại lệ.

**5.2 Performance Guarantee for Robust Classification (Đảm bảo hiệu suất)**

**Lemma 5**. Cho . Giả sử mẫu (x,y) được tạo bởi mô hình được mô tả ở **(5)**. Kỳ vọng của kết quả được tính như sau:

So sánh kết quả ở trên với kết quả trong **Lemma 3**, ở đây để phân loại nhị phân, chúng ta có thể tính toán chính xác kỳ vọng của mối tương quan và kỳ vọng này luôn lớn hơn kỳ vọng của LR đã được cài đặt. Sự tương quan phụ thuộc vào tỷ số tín hiệu bị nhiễu. Trong trường hợp không bị nhiễu, và tương quan dự kiến là , và được biết đến như sự phân chia nửa bình thường. Tương tự như vậy để phân tích RoLR cho LR, dựa trên **Lemma 5**, chúng ta có thể bảo đảm được hiệu suất cho RoLR trong việc giải quyết các bài toán phân loại.

**Định lý 2**. Gọi là đầu ra của **Thuật toán 1** và là tham số tối ưu giảm thiểu rủi ro thực nghiệm. Giả sử có n mẫu xác thực được tạo ra bởi mô hình được mô tả bởi (5). Chúng ta có với xác suất lớn hơn ,

Thí nghiệm của **Định lý 2** tương tự như **Định lý 1**. Ngoài ra, tương tự như trường hợp LR, dựa trên tham số bị ràng buộc lỗi, nó thì dễ dàng để có được giới hạn rủi ro thực nghiệm và phân bố của RoLR để phân loại. Do giới hạn về không gian, ở đây chỉ phác thảo cách để có được giới hạn rủi ro.

Đối với bài toán phân loại, hàm số hao tổn tự nhiên nhất là hao tổn 0-1. Tuy nhiên hàm tổn thất 0-1 là không lồi, không trơn tru, và ta không thể nhận một giá trị hàm số đáng kể bị ràng buộc trong khoản như ta thực nghiệm cho hàm logistic hao tổn. May mắn thay, một số hàm lồi hao tổn cho tổn thất 0-1 đã được đề xuất và đạt được hiệu suất phân loại tốt, bao gồm sự hao tổn căn bản, hao tổn lũy thừa và hao tổn logistic. Các hàm tổn thất này là tất cả các Lipschitz liên tục và do đó ta có thể ràng buộc các rủi ro thực nghiệm và rủi ro phân bố của chúng đối với hồi quy logistic.

# Tài liệu Kham khảo

1. Peter L Bartlett and Shahar Mendelson. Rademacher and gaussian complexities: Risk bounds and structural results. The Journal of Machine Learning Research, 3:463–482, 2003.
2. Ana M Bianco and V´ıctor J Yohai. Robust estimation in the logistic regression model. Springer, 1996.
3. Yudong Chen, Constantine Caramanis, and Shie Mannor. Robust sparse regression under adversarial corruption. In ICML, 2013.
4. R Dennis Cook and Sanford Weisberg. Residuals and influence in regression. 1982.
5. JB Copas. Binary regression models for contaminated data. Journal of the Royal Statistical Society. Series B (Methodological), pages 225–265, 1988.
6. Nan Ding, SVN Vishwanathan, Manfred Warmuth, and Vasil S Denchev. T-logistic regression for binary and multiclass classification. Journal of Machine Learning Research, 5:1–55, 2013.
7. Frank R Hampel. The influence curve and its role in robust estimation. Journal of the American Statistical Association, 69(346):383–393, 1974.
8. Peter J Huber. Robust statistics. Springer, 2011.
9. Wesley Johnson. Influence measures for logistic regression: Another point of view. Biometrika, 72(1):59–65, 1985.
10. Hans R Kunsch, Leonard A Stefanski, and Raymond J Carroll. Conditionally unbiased bounded-influence estimation in general regression models, with applications to generalized linear models. Journal of the American Statistical Association, 84(406):460–466, 1989.
11. Su-In Lee, Honglak Lee, Pieter Abbeel, and Andrew Y Ng. Efficient L1 regularized logistic regression. In AAAI, 2006.
12. Po-Ling Loh and Martin J Wainwright. High-dimensional regression with noisy and missing data: Provable guarantees with nonconvexity. Annals of Statistics, 40(3):1637, 2012.
13. Yaniv Plan and Roman Vershynin. Robust 1-bit compressed sensing and sparse logistic regression: A convex programming approach. Information Theory, IEEE Transactions on, 59(1):482–494, 2013.
14. Daryl Pregibon. Logistic regression diagnostics. The Annals of Statistics, pages 705–724, 1981.
15. Daryl Pregibon. Resistant fits for some commonly used logistic models with medical applications. Biometrics, pages 485–498, 1982.
16. Leonard A Stefanski, Raymond J Carroll, and David Ruppert. Optimally hounded score functions for generalized linear models with applications to logistic regression. Biometrika, 73(2):413–424, 1986.
17. Julie Tibshirani and Christopher D Manning. Robust logistic regression using shift parameters. arXiv preprint arXiv:1305.4987, 2013.
18. Vladimir N Vapnik and A Ya Chervonenkis. On the uniform convergence of relative frequencies of events to their probabilities. Theory of Probability & Its Applications, 16(2):264–280, 1971.