**Phạm Trường Giang – 15110036**

**Nguyễn Ngọc Hoàng Phúc – 15110099**

**Báo cáo Đồ án 3**

**Robust Logistic Regression and Classification**

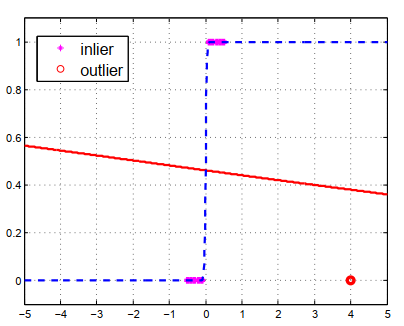
# I. Giới thiệu

Hồi quy logistic (LR) là một mô hình phân lớp xác suất thống kê tiêu chuẩn đã được ứng 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.

# II. Nhược điểm của Logistic Regression

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 dự đoán (đường màu đỏ đậm) lệch xa đường cong chính xác (nét đứt xanh dương) do sự tồn tại của chỉ một ngoại lai (outlier) (vòng tròn màu đỏ).

# III. Thuật toán Robust Logistic Regression

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 lai 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 lai , 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.

# IV. Hiệu suất của RoLR

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

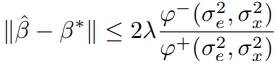
Dựa trên hai bổ đề này, cùng với một số tính chất tập trung của các mẫu bên trong (được hiển thị trong tài liệu bổ sung), chúng tôi bảo đảm việc thực hiện của RoLR trên mô hình phục hồi tham số LR.

**Định lý 1** (RoLR để khôi phục tham số LR). Cho là phân số ngoại lai, 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 lai.

# V. Tài liệu kham khảo

1. Paper “Robust Logistic Regression and Classification”, NIPS 2014
2. <https://machinelearningcoban.com/>
3. Course “Machine learning”, Andrew Ng, <https://www.coursera.org/learn/machine-learning>
4. <https://www.wikipedia.org/>