**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

# III. Giải thuật 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 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.

# IV. Điểm khác biệt giữa RoLR và logistic regression truyền thống

# V. Tài liệu kham khảo