LOGISTIC REGRESSION AND RIDGE REGRESSION USING BREAST CANCER WISCONSIN

Nguyễn Việt Đức, Nguyễn Kim Toàn

ABSTRACT

Hồi quy Logistic và chính quy hóa hồi quy Logistic trên bộ dữ liệu Breast Cancer Wisconsin.

***Index Terms*—** L2-Regularization, Logistic Regression, Ridge Regression, Breast Cancer Wisconsin.

1. INTRODUCTION

Hồi quy Logistic là một mô hình phân loại tuyến tính dự đoán kết quả nhị phân dựa trên một tập hợp các tính năng. Trong hồi quy Logistic, chúng ta quan tâm đến việc xác định xác suất mà một quan sát thuộc về một lớp nhất định. Chúng ta có thể ánh xạ một kết hợp tuyến tính giữa trọng số và tính năng sau đó biến đổi chúng thành các xác suất có giá trị là 0 và 1 thông qua hàm Sigmoid hay còn gọi là hàm Logistic.

Như vậy, dữ liệu đầu vào của mô hình là một tập các mẫu quan sát với một hoặc nhiều tính năng, và kết quả trả về là 0 hoặc 1. Ngoài ra quá trình chính quy hoá được bổ sung nhằm tránh việc trang bị quá khớp trong quá trình đào tạo mô hình.

Việc áp dụng thuật toán trên bộ dữ liệu Breast Cancer Wisconsin có thể giúp ích trong việc dự đoán khối u là dương tính hay âm tính với độ chính xác cao. Các bước đã thực hiện là:

Bước 1: Thay những giá trị còn thiếu trong tập dữ liệu của thuộc tính đang xét bằng các giá trị trung bình.

Bước 2: Phân chia bộ dữ liệu theo tỷ lệ huấn luyện và thử nghiệm.

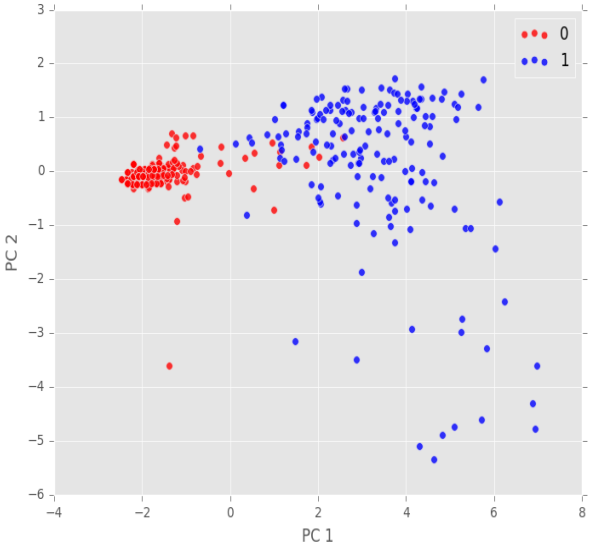
Bước 3: Chuẩn hóa dữ liệu điểm z: chuẩn hóa mọi giá trị trong tập dữ liệu sao cho giá trị trung bình của tất cả các giá trị là 0 và độ lệch chuẩn là 1.

Bước 4: Áp dụng thuật toán PCA để giảm chiều dữ liệu.

Bước 5: Huấn luyện mô hình và so sánh độ hiệu quả khi áp dụng chính quy hóa trên tập dữ liệu thử nghiệm.

2. METHOD

Thực hiện phân tích thành phần chính (PCA) để giảm chiều dữ liệu xuống còn 2 chiều và cải thiện hiệu quả tính toán của mô hình hồi quy Logistic. Như đã thấy bên dưới, dữ liệu gần như có thể phân tách tuyến tính theo lớp.

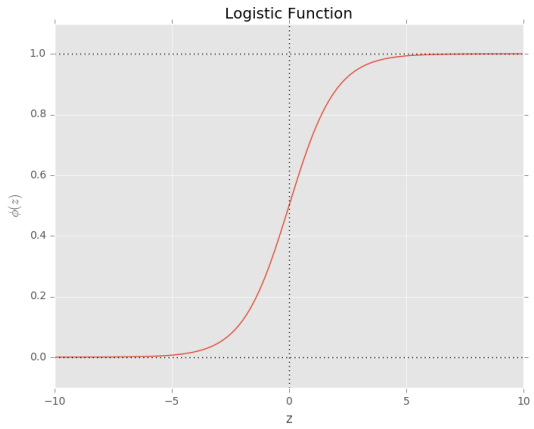


Hình 1: Phân tích thành phần chính(PCA)

Xây dựng Hàm Logistic hay còn gọi là hàm Sigmoid:

$$g(z) = \frac{1}{1 + e^{-z}}$$

$$z = w^T x = w_0 x_0 + w_1 x_1 + ... + w_m x_m$$



Hình 2: Hàm Logistic

Nếu g(z) có giá trị ở nửa trên, chúng ta sẽ phân loại mẫu là lớp 1, nếu không là lớp 0. Trọng số của hàm Logistic được học bằng tối thiểu hoá hàm log-likelihood J (hàm chi phí) thông qua thuật toán giảm dốc:

$$J(w) = -\sum_{i}^{m} y^i \log(g(z^i)) + (1 - y^i) log(1 - g(z^i))$$

Nếu mô hình hồi quy Logistic có phương sai cao (quá khớp với dữ liệu huấn luyện), có thể thực hiện chính quy hóa để xử phạt các hệ số có trọng số lớn. Bài viết này sử dụng chính quy hoá L2 được thêm vào chi phí:

$$\frac{\lambda}{2}||w||^2 = \frac{\lambda}{2}\sum_{j=1}^{m} w_j^2$$

Giả sử tham số chính quy hoá $C = \frac{1}{\lambda}$, hàm chi phí hồi quy Logistic trở thành:

$$J(w) = C[-\sum_{i}^{m} y^i \log(g(z^i)) + (1 - y^i) log(1 - g(z^i))] + \frac{1}{2}||w||^2$$

3. EXPERIMENTAL RESULTS AND ANALYSIS

Để kiểm tra mô hình Hồi quy Logistic, sử dụng trên bộ dữ liệu Breast Cancer Wisconsin UCI. Bộ dữ liệu bao gồm chín đặc điểm có giá trị thực được tính toán từ hình ảnh số hóa của một lần hút kim cuối cùng (FNA) của một khối vú với 699 lần quan sát. Các đặc điểm khác nhau của nhân tế bào được tính toán và mô tả hiện diện trong hình ảnh sinh thiết ở cả khối u vú lành tính và ác tính.

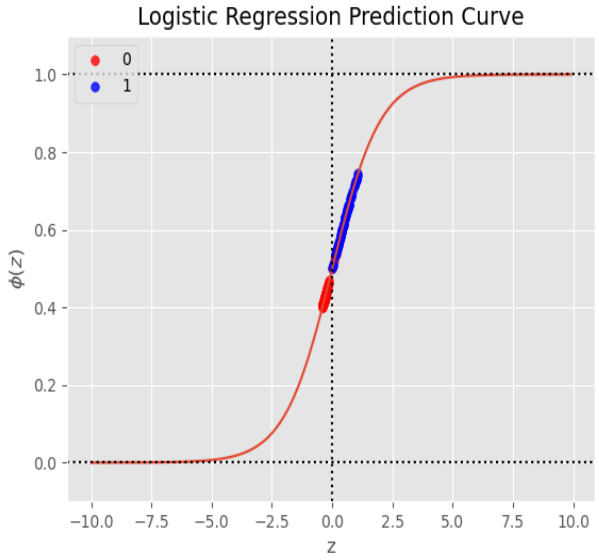
STT	Feature	Value Range
1	Mã số mẫu	ID number
2	Độ dày khối	1->10
3	Kích thước tế bào đồng nhất	1->10
4	Hình dạng tế bào đồng nhất	1->10
5	Kết dính biên	1->10
6	Kích thước tế bào biểu mô đơn	1->10
7	Bar Nuclei	1->10
8	Chất nhiễm sắc nhạt	1->10
9	Normal Nucleoli	1->10
10	Nguyên phân	1->10
11	Chẩn đoán	2 = u lành; 4 = u ác

Bảng 1: Mô tả bộ dữ liệu

Kết quả đánh giá hiệu suất trên tập dữ liệu thử nghiệm. Trong trường hợp không sử dụng chính quy hoá, kết quả thu được là: Accuracy : 0.9762; Recall : 0.9589; F1-score : 0.9655; Precision : 0.9722

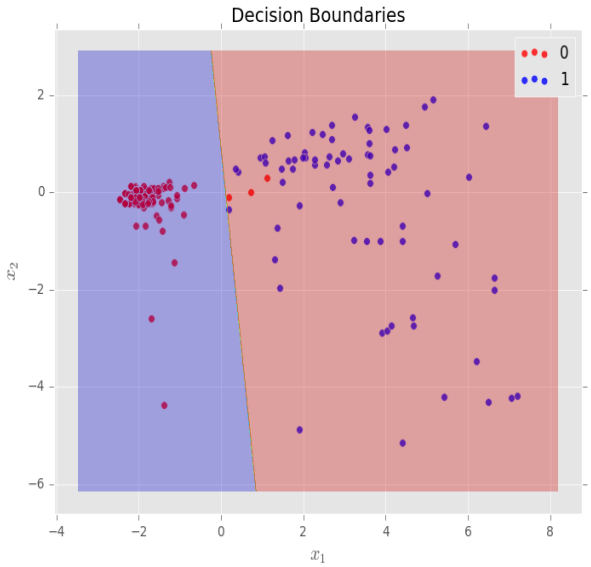
Sau khi sử dụng chính quy hoá L2: Accuracy : 0.9857; Recall : 1.0000; F1-score : 0.9799; Precision : 0.9605

Biểu đồ dự đoán sau cho thấy các mẫu được ánh xạ vào hàm Logistic một cách trơn tru và không bị dồn về phía 2 đầu mút:



Hình 3: Biểu đồ dự đoán các mẫu

Trong khi biểu đồ vùng quyết định cho thấy cách mô hình hồi quy Logistic phân chia không gian con của đối tượng theo lớp được dự đoán:



Hình 4: Ranh giới quyết định

4. CONCLUSION

Việc áp dụng chính quy hóa trên mô hình Hồi quy Logistic làm tránh việc trang bị quá khớp, đồng thời làm tăng độ chính

xác của bộ phân lớp. Thành phần chính quy hóa làm giảm độ lớn của các thông số trên mô hình, giảm độ phức tạp cũng như khối lượng tính toán. Nhược điểm của chính quy hóa là phải chọn một giá trị cho siêu tham số để áp dụng hồi quy này. Chính quy hóa L2 không thường được sử dụng trong trường hợp dữ liệu thưa, mà sử dụng chính quy hóa L1 (Lasso Regression).

5. REFERENCES

[1] Andrew Ng. Machine Learning: Logistic Regression. [Online]. Available from: https://www.coursera.org/learn/machine-learning/home/week/3

[2] Logistic Regression [Online]. Available from: https://en.wikipedia.org/wiki/Logistic_regression

[3] Regularization (mathematics) [Online]. Available from: https://en.wikipedia.org/wiki/Regularization_(mathematics)

[4] Breast Cancer Wisconsin dataset [Online]. Available from: https://archive.ics.uci.edu/ml/datasets/breast+cancer+wisconsin+(diagnostic)

[5] Logistic-Regression-Classifier-with-L2-Regularization[Online]. Available from:https://github.com/pickus91/Logistic-Regression-Classifier-with-L2-Regularizat