**3. Bài toán dự đoán chất lượng rượu vang "Vinho Verde" của Bồ Đào Nha với biến thể đỏ**

**3.1. Chi tiết dữ liệu sử dụng :**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Đặc điểm của Tập dữ liệu:** | Đa biến | **Số phiên bản:** | 4898 | **Khu vực:** | Kinh doanh |
| **Đặc điểm thuộc tính:** | Thực tế | **Số thuộc tính:** | 12 | **Ngày được tặng** | 2009-10-07 |
| **Nhiệm vụ liên quan:** | Phân loại, hồi quy | **Giá trị bị mất?** | N / A | **Số lượt truy cập web:** | 1402897 |

**Nguồn:**

Paulo Cortez, Đại học Minho, Guimarães, Bồ Đào Nha, <http://www3.dsi.uminho.pt/pcortez>

A. Cerdeira, F. Almeida, T. Matos và J. Reis, Ủy ban trồng trọt vùng Vinho Verde (CVRVV) , Porto, Bồ Đào Nha

@ 2009

link :<https://archive.ics.uci.edu/ml/datasets/Wine+Quality>

Thông tin dữ liệu : Các bộ dữ liệu này có thể được xem như các nhiệm vụ phân loại hoặc hồi quy. Các lớp được sắp xếp theo thứ tự và không cân bằng (ví dụ: có nhiều loại rượu bình thường hơn loại xuất sắc hoặc kém). Các thuật toán phát hiện ngoại lệ có thể được sử dụng để phát hiện một số loại rượu ngon hoặc kém.

Bộ dữ liệu có 12 thuộc tính. 1600 dòng và 12 cột.

Dữ liệu có 12 cột gồm :

+ fixed acidity ( độ axit cố định )

+ volatile acidity (độ axit bay hơi )

+ citric acid (axit xitric)

+ residual sugar ( đường dư )

+ chlorides ( clorua )

+ free sulfur dioxide ( lưu huỳnh dioxit )

+ total sulfur dioxide ( tổng lưu huỳnh dioxit )

+ density ( tỷ trọng )

+ pH

+ sulphates ( sunfat )

+ alcohol ( rượu )

+ quality ( chất lượng )

**3.2. Mô tả bài toán học máy :**

Bài toán sử dụng muốn dự đoán chất lượng của rượu

Sử dụng bài toán : phân loại

      Kỹ thuật LogisticRegression

**3.3. Chi tiết về cách xử lý dữ liệu**

- Dữ liệu thiếu, bị mất : Không

- Bài toán sử dụng kĩ thuật học máy : Logistic Regression

- Giới thiệu

   Logistic Regression là một thuật toán phân loại được dùng để gán các   đối tượng cho một tập giá trị rời rạc như 0,1,2…. Ví dụ điển hình là phân loại Email. Gồm có email công việc, email gia đình, email spam,…

Mô hình Logistic Regression

Đầu ra dự đoán của:

* Linear Regression:f(x)=wTxf(x)=wTx
* PLA:f(x)=sgn(wTx)f(x)=sgn(wTx)

Đầu ra dự đoán của logistic regression thường được viết chung dưới dạng:

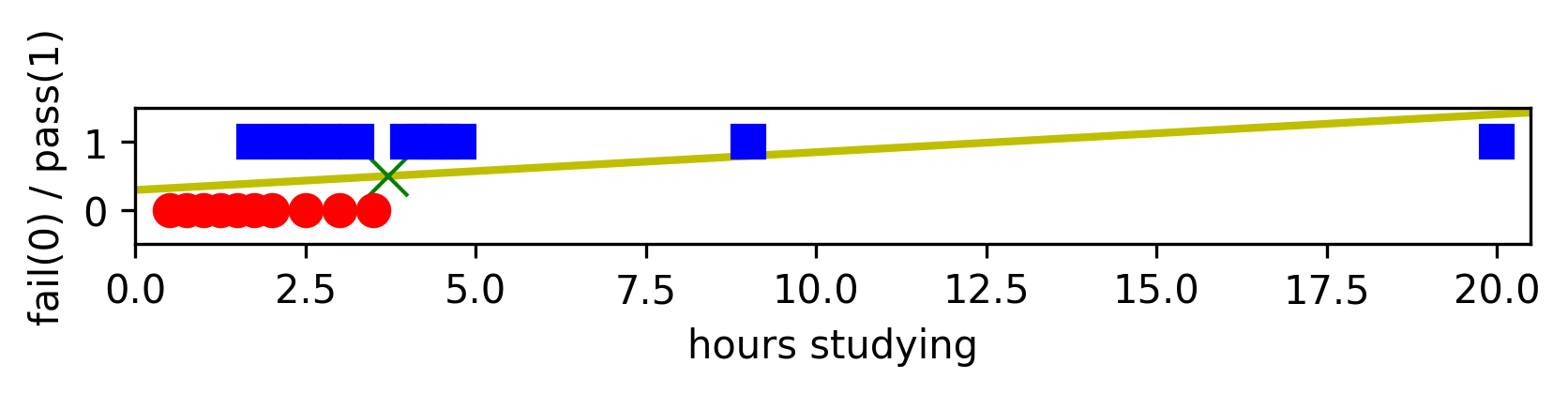
f(x)=θ(wTx)f(x)=θ(wTx)

Trong đó θθ được gọi là logistic function. Một số activation cho mô hình tuyến tính được cho trong hình dưới đây:



Hình : Các activation function khác nhau.

* Đường màu vàng biểu diễn linear regression. Đường này không bị chặn nên không phù hợp cho bài toán này. Có một *trick* nhỏ để đưa nó về dạng bị chặn: *cắt* phần nhỏ hơn 0 bằng cách cho chúng bằng 0, *cắt* các phần lớn hơn 1 bằng cách cho chúng bằng 1. Sau đó lấy điểm trên đường thẳng này có tung độ bằng 0.5 làm điểm phân chia hai *class*, đây cũng không phải là một lựa chọn tốt. Giả sử có thêm vài bạn *sinh viên tiêu biểu* ôn tập đến 20 giờ và, tất nhiên, thi đỗ. Khi áp dụng mô hình linear regression như hình dưới đây và lấy mốc 0.5 để phân lớp, toàn bộ sinh viên thi trượt vẫn được dự đoán là trượt, nhưng rất nhiều sinh viên thi đỗ cũng được dự đoán là trượt (nếu ta coi điểm x màu xanh lục là *ngưỡng cứng* để đưa ra kết luận). Rõ ràng đây là một mô hình không tốt. Anh chàng sinh viên tiêu biểu này đã *kéo theo* rất nhiều bạn khác bị trượt.



Hình 3: Tại sao Linear Regression không phù hợp?

* Đường màu đỏ (chỉ khác với activation function của PLA ở chỗ hai class là 0 và 1 thay vì -1 và 1) cũng thuộc dạng *ngưỡng cứng* (hard threshold). PLA không hoạt động trong bài toán này vì dữ liệu đã cho không *linearly separable*.

·         Các đường màu xanh lam và xanh lục phù hợp với bài toán của chúng ta hơn. Chúng có một vài tính chất quan trọng sau:

* Là hàm số liên tục nhận giá trị thực, bị chặn trong khoảng (0,1)(0,1).
* Nếu coi điểm có tung độ là 1/2 làm điểm phân chia thì các điểm càng xa điểm này về phía bên trái có giá trị càng gần 0. Ngược lại, các điểm càng xa điểm này về phía phải có giá trị càng gần 1. Điều này *khớp* với nhận xét rằng học càng nhiều thì xác suất đỗ càng cao và ngược lại.
* *Mượt* (smooth) nên có đạo hàm mọi nơi, có thể được lợi trong việc tối ưu.

Một vài tính chất của Logistic Regression

Logistic Regression thực ra được sử dụng nhiều trong các bài toán Classification.

Mặc dù có tên là Regression, tức một mô hình cho fitting, Logistic Regression lại được sử dụng nhiều trong các bài toán Classification. Sau khi tìm được mô hình, việc xác định class yy cho một điểm dữ liệu xx được xác định bằng việc so sánh hai biểu thức xác suất:P(y=1|x;w);  P(y=0|x;w)P(y=1|x;w);  P(y=0|x;w)Nếu biểu thức thứ nhất lớn hơn thì ta kết luận điểm dữ liệu thuộc class 1, ngược lại thì nó thuộc class 0. Vì tổng hai biểu thức này luôn bằng 1 nên một cách gọn hơn, ta chỉ cần xác định xem P(y=1|x;w)P(y=1|x;w) lớn hơn 0.5 hay không. Nếu có, class 1. Nếu không, class 0.

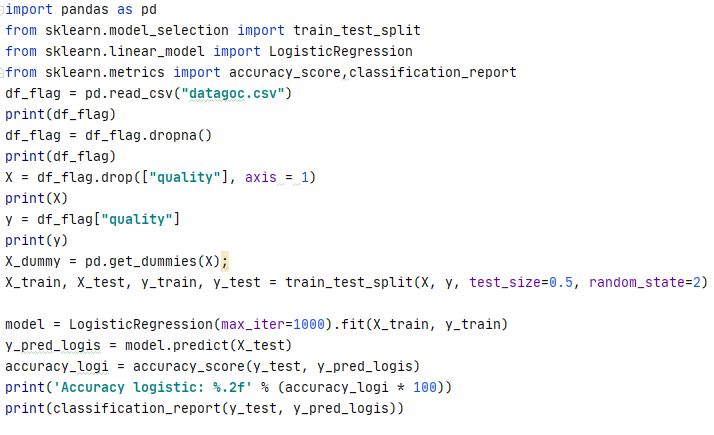
Boundary tạo bởi Logistic Regression có dạng tuyến tính

Thật vậy, theo lập luận ở phần trên thì chúng ta cần kiểm tra:

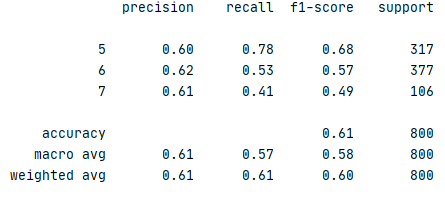
P(y=1|x;w)>0.5 ⇔11+e−wTx>0.5 ⇔e−wTx<1 ⇔wTx>0P(y=1|x;w)>0.5 ⇔11+e−wTx>0.5 ⇔e−wTx<1 ⇔wTx>0

Nói cách khác, boundary giữa hai class là đường có phương trình wTxwTx. Đây chính là phương trình của một siêu mặt phẳng. Vậy Logistic Regression tạo ra boundary có dạng tuyến tính.

-         Code thực hiện bài toán :



-         Kết quả đạt được



Giải thích :

Accuracy : Là độ chính xác của mô hình học máy

Recall : Bao nhiêu cái được lấy ra là đúng. Chỉ số này còn được gọi là độ bao phủ tức là xét xem mô hình tìm được có khả năng tổng quát hóa thể nào.

F1-Score : Từ 2 yếu tố độ chính xác và độ bao phủ phía trên người ta đưa ra 1 chỉ số khác là F1-score. Đây được gọi là một trung bình điều hòa của các tiêu chí Precision và Recall và đồng thời nó có giá trị lớn nếu cả 2 giá trị Precision và Recall đều lớn.

Với kết quả đạt được như trên thì chúng ta thu được độ chính xác của mô hình học máy là 61%

**3.4.Phương pháp đánh giá**

Cách đơn giản và hay được sử dụng nhất là *accuracy* (độ chính xác). Cách đánh giá này đơn giản tính tỉ lệ giữa số điểm được dự đoán đúng và tổng số điểm trong tập dữ liệu kiểm thử.

Đối với phân loại nhị phân, độ chính xác cũng có thể được tính theo mặt tích cực và tiêu cực như sau:

Trong đó *TP* = True Positives, *TN* = True Negatives, *FP* = False Positives, and *FN* = False Negatives.