Hồi quy logistic
(Logistic Regression)

Trường Đại học Công nghệ Thông tin, Đại học Quốc gia Thành phố Hồ Chí Minh

Tài liệu nội bộ

Tháng 2 năm 2020

Tổng quan

1. Ví dụ mở đầu
2. Hồi quy logistic
3. Hàm mất mát

Nội dung trình bày

1. Ví dụ mở đầu

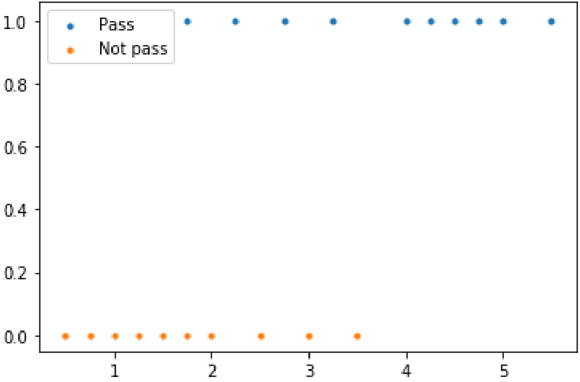
Ví dụ mở đầu

Quan sát 20 sinh viên dành thời gian (x) cho việc ôn thi và kết quả thi của (y) của các sinh viên này.

Hours	Pass	Hours	Pass
.5	0	2.75	1
.75	0	3	0
1	0	3.25	1
1.25	0	3.5	0
1.5	0	4	1
1.75	0	4.25	1
1.75	1	4.5	1
2	0	4.75	1
2.25	1	5	1
2.5	0	5.5	1

Nếu một sinh viên có thời gian ôn thi là $x^* = 4.1$ giờ thì có thi đạt không?

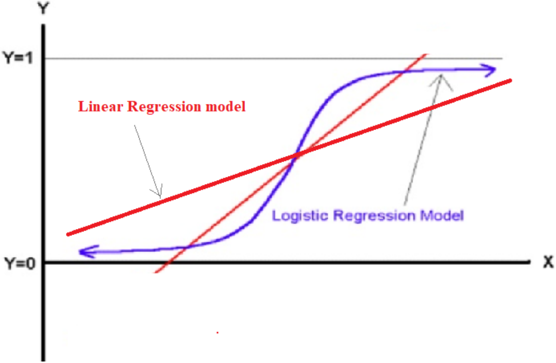
Đồ thị phân tán của tập dữ liệu



Hình 1: Đồ thị phân tán

$\rightarrow$ Không dùng được mô hình hồi qui tuyến tính.
Thời gian ôn thi này ảnh hưởng đến khả năng sinh viên vượt qua kỳ thi như thế nào?

Chọn đường hồi quy nào?



Hình 2: So sánh giữa đường tuyến tính và đường cong sigmoid

Nội dung trình bày

2. Hồi quy logistic

Hồi qui Logistic

Với giá trị x^*, theo xác suất có điều kiện, xác suất sinh viên này thuộc nhóm y ($y = \in \{0, 1\}$, y=1: thi đạt) là:

$$P(y/x^*) = \frac{P(x^*y)}{P(x^*)} = \frac{P(y)P(x^*/y)}{P(y)P(y/x^*) + P(\bar{y})P(x^*/\bar{y})} = \frac{1}{1 + \frac{P(\bar{y})P(x^*/\bar{y})}{P(y)P(x^*/y)}} \qquad (1)$$

Đặt

$$\alpha = ln\frac{P(y)P(x^*/y)}{P(\bar{y})P(x^*/\bar{y})} \rightarrow P(y/x^*) = \frac{1}{1 + e^{-\alpha}} = \sigma(\alpha) \qquad (2)$$

$\sigma(\alpha)$ được gọi là hàm sigmoid
Mô hình hồi qui logistic:

$$P(y) = \frac{1}{1 + e^{-(\theta_0 + \theta_1 x_1 + \ldots \theta_{n-1} x_{n-1})}} \qquad (3)$$

Mô hình logistic

$$\ln\left(\frac{p}{1-p}\right) = \theta_0 + \theta_1 x_1 + ...\theta_{n-1} x_{n-1} + \varepsilon = \boldsymbol{\theta}^T \mathbf{x} + \varepsilon \tag{4}$$

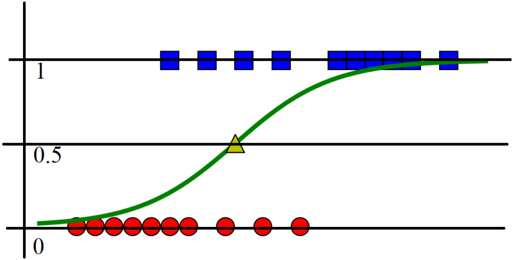
- p là xác suất sự kiện Y xảy ra
- $p/(1-p)$ được gọi là tỉ lệ **odds** , đó là tỉ số giữa xác suất xảy ra và xác suất không xảy ra của cùng một sự kiện
- $ln[p/(1-p)]$ là logarit của tỉ lệ odds, hay "logit"
- Công thức ước lượng xác suất

$$\hat{p} = \frac{1}{1 + e^{-\boldsymbol{\theta}^T \mathbf{x}}} \tag{5}$$

- Nếu $\boldsymbol{\theta}^T\mathbf{x} = 0$ thì $p = 0.5$. Phương trình $\boldsymbol{\theta}^T\mathbf{x} = 0$ được xem như biên quyết định (decision boundary) khi dùng hồi quy logistic trong bài toán phân loại (classification)

$$\hat{y} = \begin{cases} 0 \text{ nếu } \hat{p} < 0.5, \text{ hay } \boldsymbol{\theta}^T\mathbf{x} < 0 \\ 1 \text{ nếu } \hat{p} \geqslant 0.5, \text{ hay } \boldsymbol{\theta}^T\mathbf{x} \geqslant 0 \end{cases} \tag{6}$$

- Nếu $\boldsymbol{\theta}^T\mathbf{x} \to +\infty$ thì $p \to 1$
- Nếu $\boldsymbol{\theta}^T\mathbf{x} \to -\infty$ thì $p \to 0$



Hình 3: Hàm sigmoid và dữ liệu

Dự báo dựa vào mô hình hồi quy logistic
1 Thuộc tính

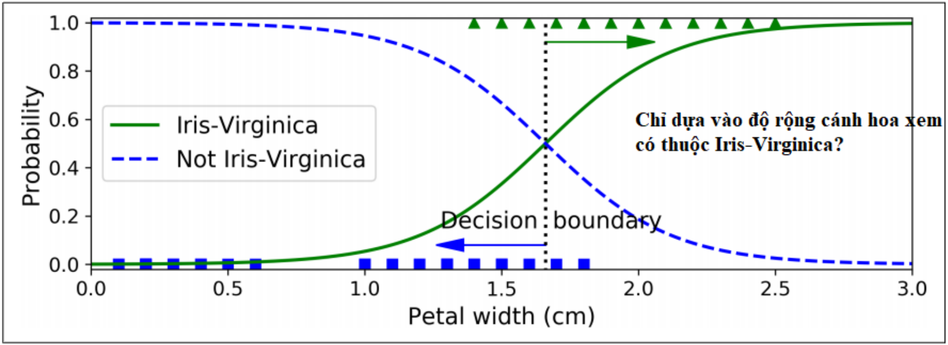


Figure 4-23. Estimated probabilities and decision boundary

Hình 4

Dự báo dựa vào mô hình hồi quy logistic
2 Thuộc tính

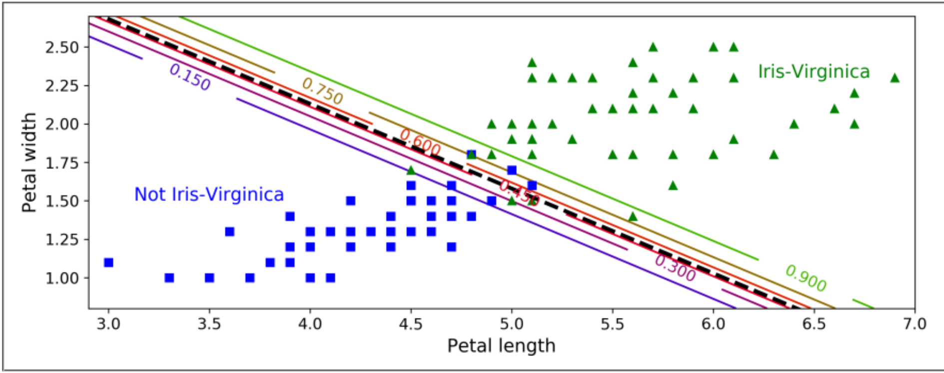


Figure 4-24. Linear decision boundary

Dựa vào 2 thuộc tính, hoa thuộc Iris-Virginica?

Hình 5

Nội dung trình bày

3. Hàm mất mát

Hàm mất mát trong hồi quy Logistic

Tại một điểm dữ liệu, đặt $\hat{p}^{(i)} = \sigma(\boldsymbol{\theta}^T \mathbf{x}^{(i)})$, ta có xác suất $y^{(i)}$ nhận giá trị 0 hoặc 1 là:

$$P(y_i/\boldsymbol{\theta}, \mathbf{x}^{(\mathbf{i})}) = (\hat{p}^{(i)})^{y^{(i)}}(1-\hat{p}^{(i)})^{1-y^{(i)}} \tag{7}$$

Với toàn bộ dữ liệu, ta cần chọn $\boldsymbol{\theta}$ sao cho cực đại biểu thức (phương pháp ước lượng hợp lý cực đại) $P(\mathbf{y}/\boldsymbol{\theta}, \mathbf{X}) = \prod_{i=1}^{m}(\hat{p}^{(i)})^{y^{(i)}})(1-\hat{p}^{(i)})^{1-y^{(i)})}$ (m là kích thước tập huấn luyện), tức là:

$$\boldsymbol{\theta} = \arg\max_{\boldsymbol{\theta}} P(\mathbf{y}|\boldsymbol{\theta}, \mathbf{X}) = \arg\max_{\boldsymbol{\theta}} \prod_{i=1}^{m}(\hat{p}^{(i)})^{y^{(i)}}(1-\hat{p}^{(i)})^{1-y^{(i)}} \tag{8}$$

Lấy logarit hai vế để chuyển tích thành tổng giúp đơn giản trong quá trình tính hàm mất mát, đổi dấu để bài toán chuyển thành cực tiểu hàm mất mát:

$$J(\boldsymbol{\theta}) = -\frac{1}{m}\sum_{i=1}^{m}\left[y^{(i)}\log\hat{p}^{(i)}) + (1-y^{(i)})\log(1-\hat{p}^{(i)})\right] \tag{9}$$

(Công thức 4-17, tr146)

Tối ưu hàm mất mát trong hồi quy Logistic

$$J(\boldsymbol{\theta}) = -\frac{1}{m}\sum_{i=1}^{m}\left[y^{(i)}log\hat{p}^{(i)}) + (1-y^{(i)})log(1-\hat{p}^{(i)})\right]$$

Để ý công thức đạo hàm hàm sigmoid:

$$\begin{aligned}\sigma'(s) &= \frac{e^{-s}}{(1+e^{-s})^2} = \frac{1}{1+e^{-s}}\frac{e^{-s}}{1+e^{-s}} && (10)\\ &= \sigma(s)(1-\sigma(s)) && (11)\end{aligned}$$

Tính đạo hàm hàm J theo θ với chỉ một điểm dữ liệu(dùng log cơ số e cho đơn giản)

$$J(\boldsymbol{\theta}) = -\left[y^{(i)}ln\hat{p}^{(i)}) + (1-y^{(i)})ln(1-\hat{p}^{(i)})\right]$$

$$J'(\boldsymbol{\theta}_j) = \left[\hat{p}^{(i)} - y^{(i)}\right]x_j^{(i)}$$

Biểu thức đạo hàm riêng hàm mất mát cho toàn bộ dữ liệu huấn luyện:

$$\frac{\partial J(\boldsymbol{\theta}; \mathbf{x}, y)}{\partial \boldsymbol{\theta}_j} = \frac{1}{m} \sum_{i=1}^{m} \left[(\sigma(\boldsymbol{\theta}^T \mathbf{x}^{(i)}) - y^{(i)} \right] x_j^{(i)} \tag{12}$$

Nếu dùng thuật toán Stochastic GD thì ta dùng từng điểm dữ liệu khi tính một epoch. Công thức gradient tại một điểm dữ liệu là:

$$\frac{\partial J(\boldsymbol{\theta}; \mathbf{x}^{(i)}, y^{(i)}}{\partial \boldsymbol{\theta}}) = \frac{1}{m} \left[\sigma(\boldsymbol{\theta}^T \mathbf{x}^{(i)}) - y^{(i)} \right] \mathbf{x}^{(i)} \tag{13}$$

1 Thử chạy các đoạn CT ở trang 147 và tìm hiểu ý nghĩa các câu lệnh.

enter

Các câu sau dùng dữ liệu cho bài toán được nêu ở ví dụ mở đầu. Dữ liệu được cho trong file data-vd-logit.xlsx

2 Hãy viết CT tìm mô hình hồi quy logistic

3 Sử dụng statsmodels tính các hệ số của hồi quy logistic

4 SV có thời gian học là 4.1 giờ thì có thi đạt không?

1 Khác biệt cơ bản của mô hình hồi quy tuyến tính và mô hình hôi quy logistic là gì?

2 Cho hai đại lượng X (kg)=trọng lượng của SV, Y= thích môn ML (no: không thích, yes: thích) có dữ liệu như sau (X,Y)={(60,yes), (55,no), (61,no), (70,yes),(59,yes), (65,yes), (80,yes), (63,no), (50,no), (75,yes), (73,yes), (51,no)}
Hãy xây dựng mô hình hồi quy và dựa vào đó dự báo xem SV có trọng lượng là 62kg có thích môn máy học không?

3 Tìm mô hình hồi quy trên tập dữ liệu huấn luyện dưới đây để dự báo ”play tennis”

Day	Outlook	Temp	Humidity	Wind	PlayTennis
1	Sunny	Hot	High	Weak	No
2	Sunny	Hot	High	Strong	No
3	Overcast	Hot	High	Weak	Yes
4	Rain	Mild	High	Weak	Yes
5	Rain	Cool	Normal	Weak	Yes
6	Rain	Cool	Normal	Strong	No
7	Overcast	Cool	Normal	Strong	Yes
8	Sunny	Mild	High	Weak	No
9	Sunny	Cool	Normal	Weak	Yes
10	Rain	Mild	Normal	Weak	Yes
11	Sunny	Mild	Normal	Strong	Yes
12	Overcast	Mild	High	Strong	Yes
13	Overcast	Hot	Normal	Weak	Yes
14	Rain	Mild	High	Strong	No

Hình 6

Tài liệu tham khảo

- Hands-on Machine Learning with Scikit-Learn, Keras & TensorFlow, 2nd Edition của tác giả Aurélien Géron.