MÔ HÌNH TUYẾN TÍNH
PHÂN LOẠI HAI LỚP

Lê Thành Sách

✉ ltsach@hcmut.edu.vn

Khoa Khoa học & Kỹ thuật Máy tính
Trường Đại học Bách Khoa - ĐHQG Tp.HCM

Tp.HCM. Ngày 9 tháng 9 năm 2019

Mục lục

ltsach@hcmut.edu.vn
Lê Thành Sách

Giới thiệu bài toán

ltsach@hcmut.edu.vn
Lê Thành Sách

Giới thiệu bài toán

Đầu vào

❶ Dữ liệu đầu vào:

$$\boldsymbol{X} = \begin{bmatrix} 1 & x_{1,1} & x_{1,2} & \cdots & x_{1,(M-1)} \\ 1 & x_{2,1} & x_{2,2} & \cdots & x_{2,(M-1)} \\ 1 & x_{3,1} & x_{3,2} & \cdots & x_{3,(M-1)} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_{N,1} & x_{N,2} & \cdots & x_{N,(M-1)} \end{bmatrix}$$

- Giả sử: dữ liệu gốc đã được rút trích đặc trưng và mỗi điểm dữ liệu là một hàng trong ma trận trên, có M đặc trưng.
- $\boldsymbol{X}$ có kích thước $N \times M$

ltsach@hcmut.edu.vn
Lê Thành Sách

Giới thiệu bài toán

Đầu vào

❷ **Tập nhãn:**

- **Bài toán này tập nhãn chỉ có 2 nhãn**
- Mỗi nhãn **thường** là tên của lớp; Ví dụ, phân loại trái cây vào "quả táo" và "không phải quả táo"
- $\Rightarrow$ **cần mã hóa nhãn** theo cách dễ xử lý và thuận tiện cho xây dựng mô hình dự báo
- $\Rightarrow$ **Dùng chỉ số**: một nhãn có chỉ số là 0 nhãn kia là 1[1]

[1] cách quy ước này giúp thuận tiện cho việc xây dựng mô hình ở sau

ltsach@hcmut.edu.vn
Lê Thành Sách

Giới thiệu bài toán

Đầu vào

❸ **Nhãn của dữ liệu:**

- Là một véctơ có N nhãn tương ứng với N điểm dữ liệu.

ltsach@hcmut.edu.vn
Lê Thành Sách

Giới thiệu bài toán

Đầu vào

Quy ước

1. $\boldsymbol{X}$: ma trận dữ liệu; kích thước $N \times M$
2. $\boldsymbol{t}$: véctơ chứa nhãn cho dữ liệu; kích thước $N \times 1$. Véctơ này chứa tên nhãn (chuỗi) hay chỉ số.
3. $\boldsymbol{y}$: là biểu diễn dạng chỉ số của nhãn t; kích thước: $N \times 1 \Rightarrow$ có thể dùng $\boldsymbol{y}$ và $\boldsymbol{t}$ thay thế nhau trong một số trường hợp
4. N: số điểm dữ liệu
5. M: số đặc trưng của mỗi điểm dữ liệu
6. $\hat{y}$ là giá trị dự báo từ mô hình

ltsach@hcmut.edu.vn
Lê Thành Sách

Giới thiệu bài toán

Mục tiêu

Mục tiêu của bài toán là xây dựng mô hình dự đoán từ tập huấn luyện, để sau đó, khi nhận mẫu dữ liệu mới $\boldsymbol{x}$ nó dự đoán nhãn của mẫu này, nghĩa là xuất ra một giá trị thuộc tập nhãn

Bài giảng này hướng dẫn xây dựng mô hình tuyến tính[1] cho bài toán phân lớp.

[1] Tuyến tính: chỉ ra rằng đường biên giữa các phân lớp là tuyến tính

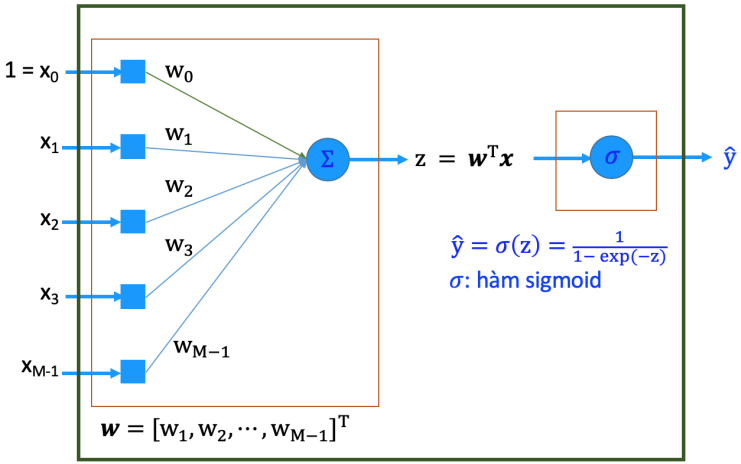
ltsach@hcmut.edu.vn
Lê Thành Sách

Phương pháp xây dựng mô hình

ltsach@hcmut.edu.vn
Lê Thành Sách

Phương pháp xây dựng mô hình

Ý tưởng



Hình 2.1: Mô hình hồi quy logistic 2 lớp

ltsach@hcmut.edu.vn
Lê Thành Sách

Phương pháp xây dựng mô hình

Ý tưởng

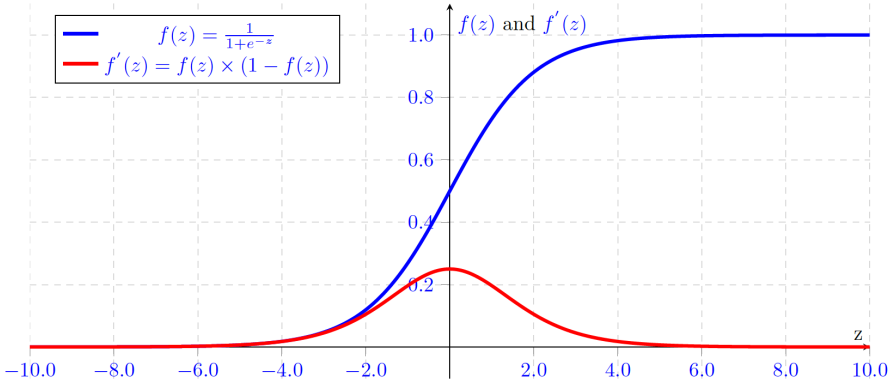
- Hai lớp là C_0 và C_1
- Mô hình nhận vào véctơ $\boldsymbol{x}$, tính toán và xuất ra $\hat{y}$, là xác suất của $\boldsymbol{x}$ thuộc lớp C_1, xem Hình 2.1
  - Ngầm hiểu: $1 - \hat{y}$ là xác suất của $\boldsymbol{x}$ thuộc lớp C_0
- Bên trong
  1. Sử dụng một mô hình tuyến tính để dự báo ra điểm $z = \boldsymbol{w}^T \boldsymbol{x}$[1]
  2. Cho z qua hàm **sigmoid** để xuất ra xác suất

[1] $\boldsymbol{w}$: tham số của mô hình

ltsach@hcmut.edu.vn
Lê Thành Sách

Phương pháp xây dựng mô hình

Ý tưởng



Hình 2.2: Hàm sigmoid và đạo hàm

ltsach@hcmut.edu.vn
Lê Thành Sách

Phương pháp xây dựng mô hình

Mô hình dự báo

Công thức của mô hình

$$\begin{aligned}\hat{y} &\stackrel{\triangle}{=} p(C_1|\boldsymbol{x}, \boldsymbol{w}) \\ &= \sigma(\boldsymbol{w}^T\boldsymbol{x}) \\ &= \frac{1}{1+\exp(-\boldsymbol{w}^T\boldsymbol{x})}\end{aligned} \tag{2.1}$$

ltsach@hcmut.edu.vn
Lê Thành Sách

Phương pháp xây dựng mô hình

Mô hình dự báo

1. Vào: Dữ liệu $\boldsymbol{x}$
2. Vào: Tham số của mô hình $\boldsymbol{w}$ và ngưỡng λ
3. Tính xác suất dự báo $\hat{y}$ theo C.T (2.1)
4. Nếu $\hat{y} \geq \lambda$, $\boldsymbol{x}$ thuộc lớp C_1
   Ngược lại, $\boldsymbol{x}$ thuộc lớp C_0

ltsach@hcmut.edu.vn
Lê Thành Sách

Ước lượng tham số của mô hình

ltsach@hcmut.edu.vn
Lê Thành Sách

Ước lượng tham số của mô hình

Nguyên tắc chung

Nguyên tắc chung

- Đưa về dạng bài toán tối ưu
  - Sử dụng hàm likelihood
  - Cực đại hóa likelihood cho tập dữ liệu đầu vào
- Giải bài toán tối ưu
  - Không thể phân tích toán học → sử dụng phương pháp lặp
    - Gradient Descent
    - Iterative Re-Weighted Least Squares

ltsach@hcmut.edu.vn
Lê Thành Sách

Ước lượng tham số của mô hình

Xây dựng hàm mục tiêu

Với một điểm dữ liệu: $< \boldsymbol{x}, y >$[1]
Xác suất dự báo được là:

$$p(y|\boldsymbol{x}, \boldsymbol{w}) = \begin{cases} \hat{y} & \text{if } y = 1 \\ 1 - \hat{y} & \text{if } y = 0 \end{cases}$$

Viết gọn:

$$p(y|\boldsymbol{x}, \boldsymbol{w}) = \hat{y}^y (1 - \hat{y})^{1-y} \tag{3.1}$$

[1] **Lưu ý:** y chỉ nhận giá trị thuộc tập :$\{0, 1\}$

ltsach@hcmut.edu.vn
Lê Thành Sách

Ước lượng tham số của mô hình

Xây dựng hàm mục tiêu

Với một điểm dữ liệu: $< \boldsymbol{x}_n, y_n >$[1]
Xác suất dự báo được là:

$$p(y_n|\boldsymbol{x}_n, \boldsymbol{w}) = \begin{cases} \hat{y}_n & \text{if } y_n = 1 \\ 1 - \hat{y}_n & \text{if } y_n = 0 \end{cases}$$

Viết gọn:

$$p(y_n|\boldsymbol{x}_n, \boldsymbol{w}) = \hat{y}_n^{y_n}(1 - \hat{y}_n)^{1-y_n} \qquad (3.2)$$

[1] **Lưu ý:** y_n chỉ nhận giá trị thuộc tập $:\{0, 1\}$

ltsach@hcmut.edu.vn
Lê Thành Sách

Ước lượng tham số của mô hình

Xây dựng hàm mục tiêu

Xác suất xảy ra N nhãn[1] trong tập huấn luyện:

$$p(\boldsymbol{t}|\boldsymbol{X}, \boldsymbol{w}) = \prod_{n=1}^{N} \hat{y}_n^{y_n}(1-\hat{y}_n)^{1-y_n} \tag{3.3}$$

Mục tiêu: nguyên lý cực đại hóa hàm hợp lý

- Tìm $\boldsymbol{w}$ sao cho $p(\boldsymbol{t}|\boldsymbol{X}, \boldsymbol{w})$ đạt cực đại
- Lưu ý: $\hat{y}_n = p(y_n|\boldsymbol{X}, \boldsymbol{w})$, phụ thuộc vào $\boldsymbol{w}$ trong C.T. (2.1)

[1] lấy mẫu theo nguyên tắc **i.i.d**

ltsach@hcmut.edu.vn
Lê Thành Sách

Ước lượng tham số của mô hình

Xây dựng hàm mục tiêu

Sử dụng **negative log-likelihood**:

$$\begin{aligned}\mathcal{L}(\boldsymbol{w}) &\stackrel{\triangle}{=} -p(\boldsymbol{t}|\boldsymbol{X}, \boldsymbol{w}) \\ &= \sum_{n=1}^{N} y_n \log \hat{y}_n + (1 - y_n) \log(1 - \hat{y}_n) \end{aligned} \tag{3.4}$$

Mục tiêu: nguyên lý cực đại hóa hàm hợp lý

- Tìm $\boldsymbol{w}$ sao cho $\mathcal{L}(\boldsymbol{w})$[1] đạt cực tiểu

[1] hàm: **cross-entropy**

ltsach@hcmut.edu.vn
Lê Thành Sách

Ước lượng tham số của mô hình

Tìm hệ số của mô hình

Nguyên tắc

- Khó dùng phân tích toán học để giải tìm nghiệm cho bài toán tối ưu có hàm mục tiêu trong C.T (3.4).
- Bài toán cực tiểu hóa $\mathcal{L}(\boldsymbol{w})$ là tối ưu không ràng buộc, có thể dùng phương pháp lặp.
  - Dựa vào đạo hàm bậc 1: **Gradient Descent**
  - Dựa vào đạo hàm bậc 2: **Newton-Raphson**

ltsach@hcmut.edu.vn
Lê Thành Sách

Ước lượng tham số của mô hình

Tìm hệ số của mô hình

Cần thiết

Tìm đạo hàm của hàm tổn thất $\mathcal{L}$ so với các tham số $\boldsymbol{w}$

$$\Delta \boldsymbol{w} \triangleq \frac{\partial \mathcal{L}(\boldsymbol{w}; \boldsymbol{x}, \boldsymbol{y})}{\partial \boldsymbol{w}}$$

ltsach@hcmut.edu.vn
Lê Thành Sách

Phương pháp

Sử dụng **chain-rule**, xem Hình 3.1

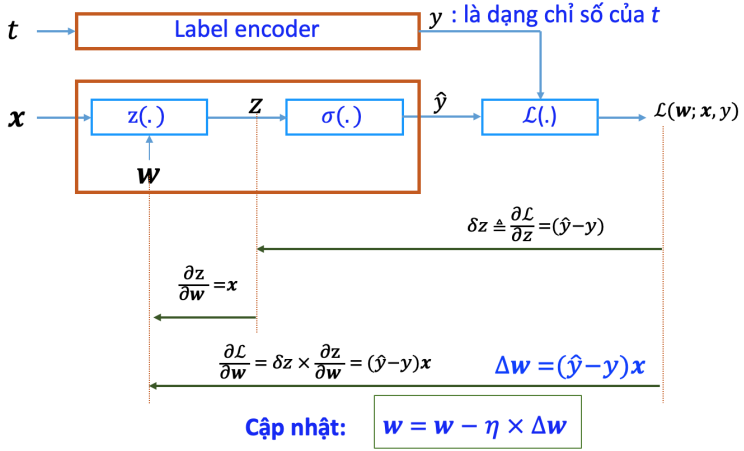
$$\frac{\partial \mathcal{L}(w; x, y)}{\partial w} = \frac{\partial \mathcal{L}}{\partial z} \bullet \frac{\partial z}{\partial w}$$

Dấu $\bullet$ là phép **"dot"**

ltsach@hcmut.edu.vn
Lê Thành Sách

Ước lượng tham số của mô hình

Tìm hệ số của mô hình



Hình 3.1: Quy trình tính và cập nhật tham số

ltsach@hcmut.edu.vn
Lê Thành Sách

Ước lượng tham số của mô hình

Tìm hệ số của mô hình

Đạo hàm của một số hàm trong sơ đồ tính toán, Hình 3.1:

$$\frac{\mathrm{d}\mathcal{L}(\boldsymbol{w};\boldsymbol{x},y)}{\mathrm{d}\hat{y}} = \frac{\hat{y}-y}{\hat{y}(1-\hat{y})} \tag{3.5}$$

$$\frac{\mathrm{d}\hat{y}}{\mathrm{d}z} = \hat{y}(1-\hat{y}) \tag{3.6}$$

$$\frac{\partial z}{\partial \boldsymbol{w}} = \boldsymbol{x} \tag{3.7}$$

Lưu ý: theo Jacobian, $\frac{\partial z}{\partial \boldsymbol{w}} = \boldsymbol{x}^T$; ở đây, phải chuyển vị để phù hợp với kích thước của $\boldsymbol{w}$, $M \times 1$

ltsach@hcmut.edu.vn
Lê Thành Sách

Ước lượng tham số của mô hình

Tìm hệ số của mô hình

Từ các C.T (3.5), (3.6), và (3.7),
Đạo hàm của hàm $\mathcal{L}(\boldsymbol{w}; \boldsymbol{x}, y)$ tính trên một điểm dữ liệu $< \boldsymbol{x}, y >$ là:

$$\begin{aligned}\Delta \boldsymbol{w} &= \frac{\partial \mathcal{L}(\boldsymbol{w}; \boldsymbol{x}, y)}{\partial \boldsymbol{w}} \\ &= \frac{\mathrm{d}\mathcal{L}(\boldsymbol{w}; \boldsymbol{x}, y)}{\mathrm{d}\hat{y}} \times \frac{\mathrm{d}\hat{y}}{\mathrm{d}z} \times \frac{\partial z}{\partial \boldsymbol{w}} \\ &= (\hat{y} - y)\boldsymbol{x} \end{aligned} \tag{3.8}$$

ltsach@hcmut.edu.vn
Lê Thành Sách

Ước lượng tham số của mô hình

Tìm hệ số của mô hình

Đạo hàm của hàm $\mathcal{L}(\boldsymbol{w}; \boldsymbol{X}, \boldsymbol{y})$ tính trên một tập của N điểm dữ liệu $< \boldsymbol{X}, \boldsymbol{y} >$ là:

$$\begin{aligned}\Delta \boldsymbol{w} &= \frac{\partial \mathcal{L}(\boldsymbol{w}; \boldsymbol{X}, \boldsymbol{y})}{\partial \boldsymbol{w}} \\ &= \sum_{n=1}^{N} (\hat{\boldsymbol{y}}_n - \boldsymbol{y}_n)\boldsymbol{x}_n \\ &= \boldsymbol{X}^T(\hat{\boldsymbol{y}} - \boldsymbol{y}) \qquad (3.9)\end{aligned}$$

ltsach@hcmut.edu.vn
Lê Thành Sách

Ước lượng tham số của mô hình

Giải thuật lặp với đạo hàm bậc 2

Đạo hàm bậc 2 của $\mathcal{L}(\boldsymbol{w})$:

$$\begin{aligned}\boldsymbol{H} &\triangleq \nabla\nabla\mathcal{L}(\boldsymbol{w}) \\ &= \sum_{n=1}^{N} \hat{y}_n(1-\hat{y}_n)\boldsymbol{x}\boldsymbol{x}^T \\ &= \boldsymbol{X}^T\boldsymbol{R}\boldsymbol{X}\end{aligned} \tag{3.10}$$

R là ma trận đường chéo có phần tử $R_{nn} = \hat{y}_n(1-\hat{y}_n)$

ltsach@hcmut.edu.vn
Lê Thành Sách