Giải thuật Perceptron

Breast cancer dataset

Đào Quang Dinh
Dương Lữ Điện
Lê Nguyễn Kha

Nội dung

- Mô tả tập dữ liệu
- Giải thuật Perceptron
- Học từ tập dữ liệu
- Đánh giá giải thuật

Mô tả tập dữ liệu

- Thu được từ bệnh viện của trường Đại học Wisconsin
- Thời gian: 09/01/1991
- Lấy từ UCI Machine Learning
- Số dòng dữ liệu: **699 dòng**

Mô tả tập dữ liệu

- Các thuộc tính và miền giá trị: Các thuộc tính có miền giá trị từ **1-10**
  - Sample code number
  - Clump Thickness (Độ dày khối u)
  - Uniformity of Cell Size (Sự đồng nhất kích thước TB)
  - Uniformity of Cell Shape (Sự đồng nhất hình dạng TB)
  - Marginal Adhesion (Độ bám dính)
  - Single Epithelial Cell Size (Kích thước biểu mô TB)
  - Bare Nuclei (Hạt nhân trần)
  - Bland Chromatin (Bromatin huyết thanh)
  - Normal Nucleoli (Hạt nhân bình thường)
  - Mitoses (Sự phân chia)
- Lớp: **2 lớp**

Mô tả tập dữ liệu

Trích dẫn một số dòng từ tập dữ liệu

Clump Thickness	Uniformity of Cell Size	Uniformity of Cell Shape	Marginal Adhesion	…	Class
3	1	2	1	…	0
2	1	1	1	…	0
10	10	10	8	…	1
6	2	1	1	…	0
5	4	4	9	…	1
2	5	3	3	…	1
6	6	6	9	…	0
10	4	3	1	…	1
6	10	10	2	…	1
5	6	5	6	…	1
10	10	10	4	…	1

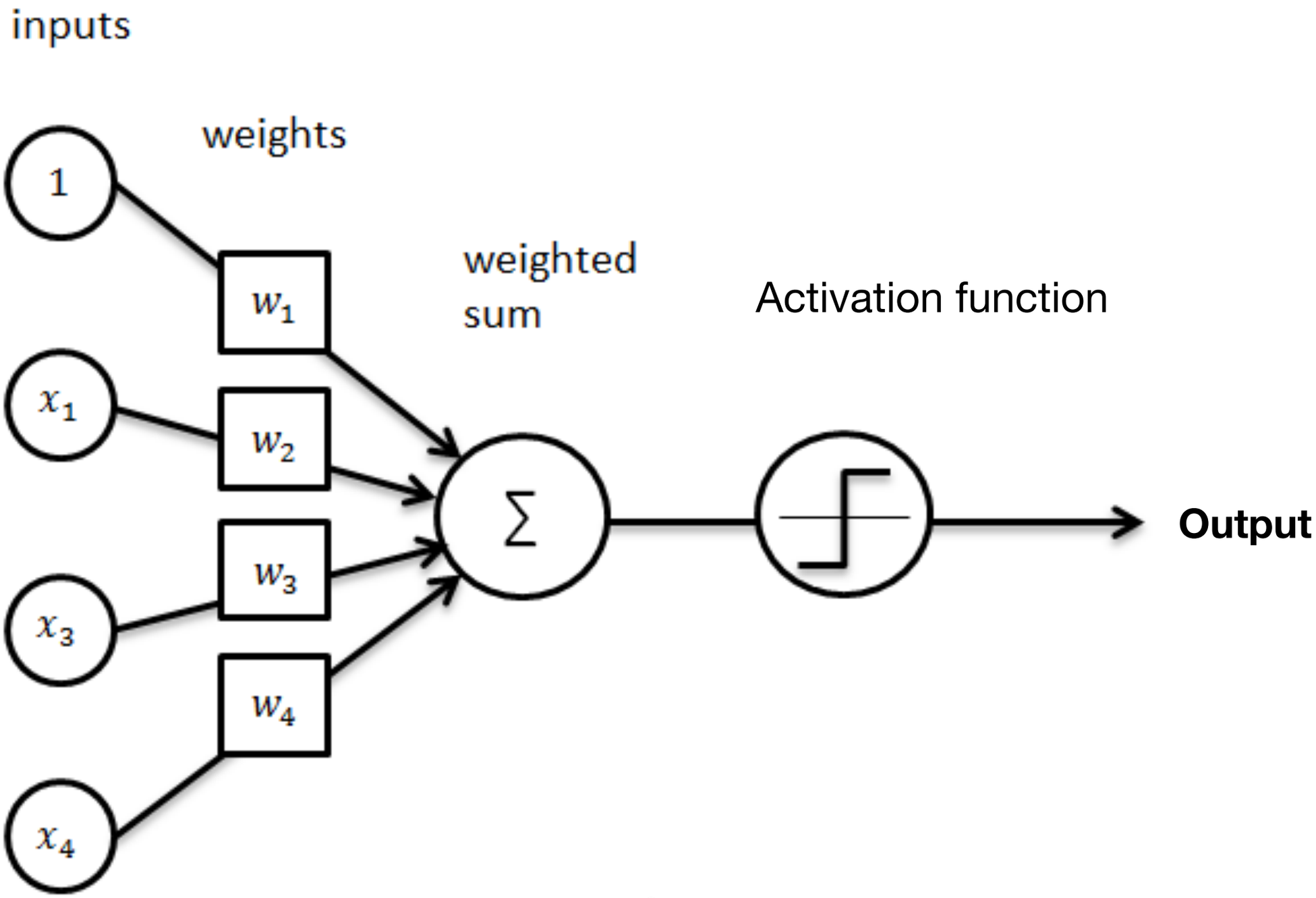
Mô tả tập dữ liệu

- Giá trị thuộc tính bị thiếu: 16 giá trị (dấu **?**)
- Tỷ lệ phân bố các lớp
  - Lớp dương (ác tính): 241 dòng (34.5%)
  - Lớp âm (lành tính): 458 dòng (65.5%)

Mô tả tập dữ liệu

- Làm sạch dữ liệu:
  - Loại bỏ cột đầu tiên (Sample ID)
  - Chuyển các giá trị bị thiếu (**?**) thành **NaN** để thực hiện tiền xử lý
- Đổi giá trị của lớp:
  - 2 thành **0** (lành tính)
  - 4 thành **1** (ác tính)
- Thêm header

Giải thuật Perceptron



Giải thuật Perceptron

- Hàm mạng:

$$u = g(x) = \sum_{i=0}^{n} w_i x_i$$

- Hàm kích hoạt: hàm **signum**

$$f(x) = \begin{cases} 1 & \text{if } x \geq 0 \\ 0 & \text{otherwise} \end{cases}$$

Giải thuật Perceptron

- Hàm cập nhật trọng số:

$$w_j = w_j + \eta(y_i - o_i)x_{ij}$$

Giải thuật Perceptron trên tập dữ liệu Breast cancer

Thứ tự mẫu (i)	Đầu vào giả (x_o)	Clump Thickness (x_1)	Uniformity of Cell Size (x_2)	Uniformity of Cell Shape (x_3)	Marginal Adhesion (x_4)	Class (y)
2	1	3	1	1	1	0
3	1	8	10	10	8	1

- Khởi tạo ngẫu nhiên trọng số:

  w0=0, w1=0.5, w2=1, w3=1.5, w4=2

- Tốc độ học: ŋ = 0.01

Giải thuật Perceptron trên tập dữ liệu Breast cancer

- Lần lặp 1:
  - i = 2 : u = 0.1 + 0,5.3 + 1.1 + 1,5.1 + 2.1 = 6 (output = 1)
  - y2 = 0 (khác với đầu ra)
  - Cập nhật:
    - w0 = w0 + ŋ (y0 – o0).x = 0 + 0,01.(0-1).1 = -0,01
    - w1 = w1 + ŋ (y1 – o1).x = 0,5 + 0,01.(0-1).3 = 0,47
    - w2 = w2 + ŋ (y2 – o2).x = 1 + 0,01.(0-1).1 = 0,99
    - w3 = w3 + ŋ (y3 – o3).x = 1,5 + 0,01.(0-1).1 = 1,49
    - w4 = w4 + ŋ (y4 – o4).x = 2 + 0,01.(0-1).1 = 1,99

Giải thuật Perceptron trên tập dữ liệu Breast cancer

- i = 3 :
- u = (-0,01).1 + 0,47.8 + 0,99.10 + 1,49.10 + 1,99.8 = 44,47

  (output = 1)
- y3 = 1 (giống với đầu ra thực tế)

Giải thuật Perceptron trên tập dữ liệu Breast cancer

- Lần lặp 2:
  - i = 2 :

    u = (-0,01).1 + 0,47.3 + 0,99.1 + 1,49.1 + 1,99.1 = 5,87 (output = 1)

    y2 = 0 (khác với đầu ra thực tế)

- Cập nhật:
  - w0 = -0,01 + 0,01.(0-1).1 = -0,02
  - w1 = 0,47 + 0,01.(0-1).3 = 0,44
  - w2 = 0,99 + 0,01.(0-1).1 = 0,98
  - w3 = 1,49 + 0,01.(0-1).1 = 1,48
  - w4 = 1,99 + 0,01.(0-1).1 = 1,98

Giải thuật Perceptron trên tập dữ liệu Breast cancer

- i = 3 :
  - u = -0,02.1 + 0,44.8 + 0,98.10 + 1,48.10 + 1,98.8 = 43,94

    output = 1
  - y3 = 1 (giống với đầu ra thực tế)
- Tạm dừng giải thuật, ta có:

  w0 = -0,02, w1 = 0,44, w2 = 0,98, w3 = 1,48 , w4 = 1,98

Học từ tập dữ liệu

- Tiền xử lý
- Chia tập dữ liệu
- Học từ tập dữ liệu
- Đánh giá mô hình

Tiền xử lý

Thay thế các thuộc tính có giá trị **NaN** bằng giá trị trung bình của thuộc tính đó (**mean**).

Chia tập dữ liệu

- Chia tập dữ liệu thành 3 phần:
  - **Training set** (dùng để học): 60%
  - **Test set** (dùng để chọn mô hình): 20%
  - **Cross validation set** (dùng để kiểm tra): 20%

Học từ tập dữ liệu

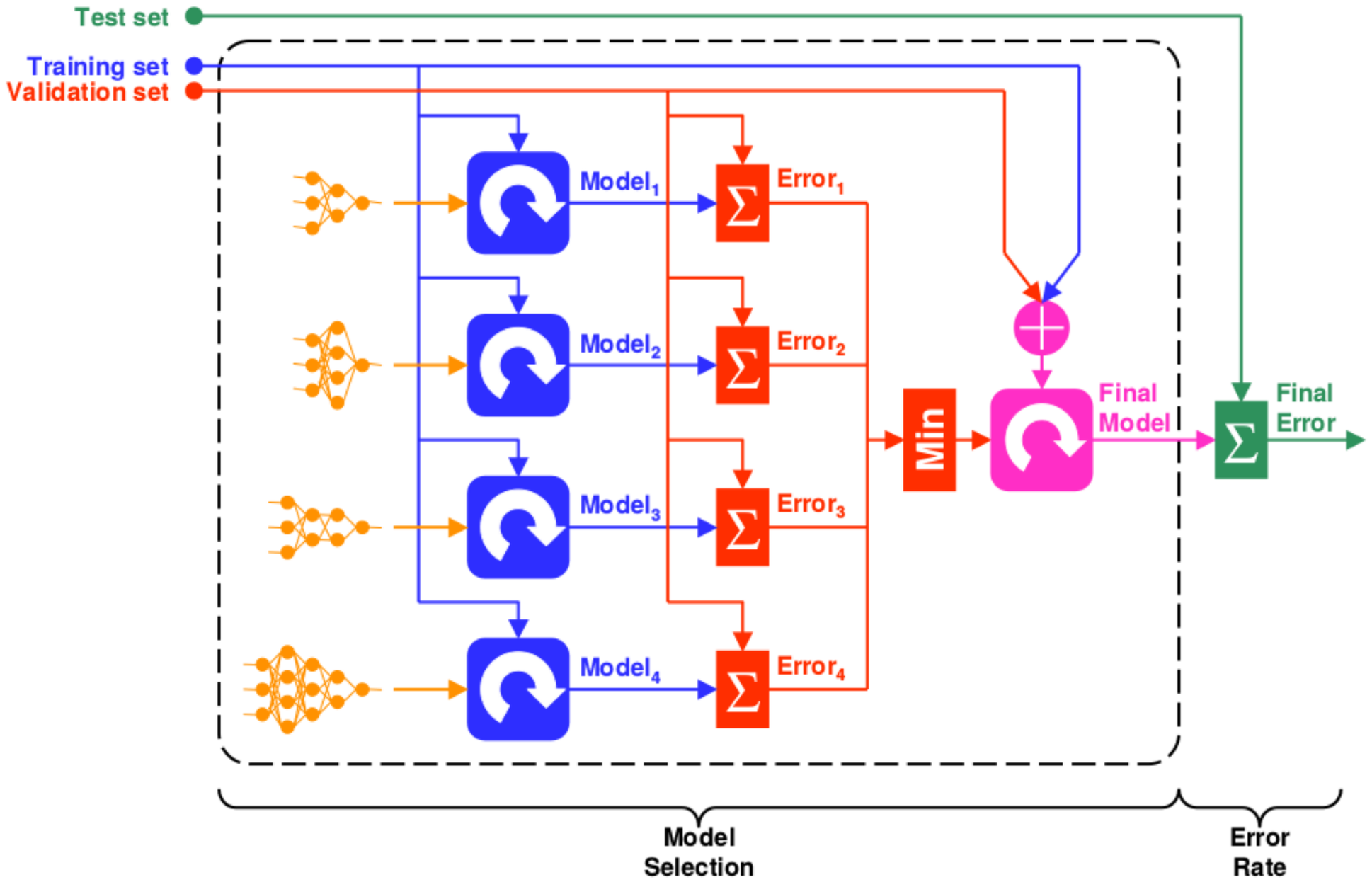
- Sử dụng thư việt Scikit Learn
- Tạo ra 9 mạng neural khác nhau
- Dùng **Training set** để cho máy học
- Dùng **Cross validation set** để chọn ra mô hình tốt nhất
- Dùng **Test set** để kiểm tra mô hình được chọn

Đánh giá mô hình

- Nghi thức **Hold-out cross-validation**

- Three-ways data splits

Đánh giá mô hình



Tham khảo

- http://research.cs.tamu.edu/prism/lectures/iss/iss_l13.pdf

- https://www.mff.cuni.cz/veda/konference/wds/proc/pdf10/WDS10_105_i1_Reitermanova.pdf

Lời cảm ơn

1. **O. L. Mangasarian** and **W. H. Wolberg**: "Cancer diagnosis via linear programming", SIAM News, Volume 23, Number 5, September 1990, pp 1 & 18.

2. **William H. Wolberg** and **O.L. Mangasarian**: "Multisurface method of pattern separation for medical diagnosis applied to breast cytology", Proceedings of the National Academy of Sciences, U.S.A., Volume 87, December 1990, pp 9193-9196.

3. **O. L. Mangasarian**, **R. Setiono**, and **W.H. Wolberg**: "Pattern recognition via linear programming: Theory and application to medical diagnosis", in: "Large-scale numerical optimization", Thomas F. Coleman and Yuying Li, editors, SIAM Publications, Philadelphia 1990, pp 22-30.

4. **K. P. Bennett** & **O. L. Mangasarian**: "Robust linear programming discrimination of two linearly inseparable sets", Optimization Methods and Software 1, 1992, 23-34 (Gordon & Breach Science Publishers).

Thank you for your attention!

Q&A