Multinomial Logistic Regression

Lê Hồng Phương
Data Science Laboratory
Vietnam National University, Hanoi
<*phuonglh@hus.edu.vn*>

November 8, 2021

Content

Mô hình hồi quy logistic đa lớp

- Khi bài toán phân loại có nhiều lớp, ta có thể mở rộng mô hình hồi quy logistic nhị phân ở trên cho trường hợp đa lớp.
- Mô hình hồi quy logistic đa lớp còn được gọi là **mô hình entropy cực đại** (maximum entropy–maxent), một dạng của mô hình log–tuyến tính.

Mô hình hồi quy logistic đa lớp

Mô hình entropy cực đại được phát minh nhiều lần, trong nhiều lĩnh vực khác nhau:

- Trong lí thuyết xác suất, dưới các tên *mô hình entropy cực đại*, *mô hình log–tuyến tính*, *trường ngẫu nhiên Markov* và *họ hàm mũ*;
- Trong thống kê toán học dưới tên *hồi quy logistic*;
- Trong cơ học thống kê và vật lí, dưới các tên *phân phối Gibbs*, *phân phối Boltzmann*;
- Trong các mạng nơ-ron dưới tên *máy Boltzmann* và *hàm kích hoạt softmax*.

Mô hình hồi quy logistic đa lớp

Xác suất để đối tượng $\mathbf{x}$ thuộc lớp $k \in \{1, 2, \dots, K\}$ được mô hình bởi:

$$P(y = k | \mathbf{x}; \theta_k) = \frac{1}{Z(\mathbf{x})} \exp(\theta_k^\top \mathbf{x}), \qquad (1)$$

trong đó Z là số hạng chuẩn hoá để đảm bảo phân phối xác suất:

$$1 = \sum_{k=1}^{K} P(y = k | \mathbf{x}; \theta_k) \Rightarrow Z(\mathbf{x}) = \sum_{k=1}^{K} \exp(\theta_k^\top \mathbf{x}). \qquad (2)$$

Mô hình hồi quy logistic đa lớp

- Tham số $\theta_k = (\theta_{k0}, \theta_{k1}, \dots, \theta_{kD})^\top$ là một véc-tơ tham số $D+1$ chiều ứng với lớp k.
- Mỗi lớp k có một véc-tơ tham số θ_k ứng với $D+1$ đặc trưng (đặc trưng thứ 0 được cố định là đơn vị).
- Ta có ma trận tham số của mô hình:

$$\begin{pmatrix} \theta_{10} & \theta_{11} & \cdots & \theta_{1D} \\ \theta_{20} & \theta_{21} & \cdots & \theta_{2D} \\ \dots & \dots & \dots & \dots \\ \theta_{K0} & \theta_{K1} & \cdots & \theta_{KD} \end{pmatrix}.$$

Mô hình hồi quy logistic đa lớp

- Vì điều kiện chuẩn hoá

$$\sum_{k=1}^{K} P(y = k | \mathbf{x}; \theta_k) = 1,$$

  nên ta chỉ cần ước lượng $(K - 1)$ véc-tơ tham số θ_k.
- Do đó, véc-tơ tham số θ của mô hình có $(K - 1) * (D + 1)$ chiều.

Content

Ước lượng tham số

Công thức tính xác suất để đối tượng $\mathbf{x}$ thuộc lớp y trong mô hình entropy cực đại:

$$P(y|\,\mathbf{x};\theta) = \frac{\exp\left(\sum_{j=0}^{D}\theta_{yj}x_j\right)}{\sum_{k=1}^{K}\exp\left(\sum_{j=0}^{D}\theta_{kj}x_j\right)}. \tag{3}$$

Trung bình của log-hợp lí của tập dữ liệu huấn luyện là:

$$\ell(\theta) = \frac{1}{N}\sum_{i=1}^{N}\log P(y_i|\,\mathbf{x}_i;\theta).$$

Ước lượng tham số

- Để ước lượng các tham số của mô hình, ta cần tìm θ^* cực tiểu hoá hàm mục tiêu sau:
$$J(\theta) = -\ell(\theta) + \lambda R(\theta), \tag{4}$$
trong đó $R(\theta)$ là số hạng hiệu chỉnh dùng để tránh hiện tượng quá khớp và tăng độ chính xác của mô hình.
- Mục tiêu của việc hiệu chỉnh là để làm trơn mô hình, phạt các tham số lớn.
- Tham số $\lambda \geq 0$ dùng để điều khiển tính cân bằng của mô hình trong việc phù hợp với dữ liệu quan sát và việc hiệu chỉnh.

Content

Hiệu chỉnh dạng L_1

Nếu sử dụng hiệu chỉnh dạng L_1 thì hàm mục tiêu là

$$J_1(\theta) = -\ell(\theta) + \lambda \sum_{k=1}^{K} \sum_{j=1}^{D} |\theta_{kj}|. \tag{5}$$

Chú ý rằng hàm mục tiêu J_1 không phải là hàm lồi nên nghiệm tối ưu cục bộ có thể không phải là nghiệm tối ưu toàn cục.

Hiệu chỉnh dạng L_2

Hiệu chỉnh dạng L_2 là một hàm toàn phương, hàm mục tiêu là:

$$J_2(\theta) = -\ell(\theta) + \frac{\lambda}{2}\sum_{k=1}^{K}\sum_{j=1}^{D}\theta_{kj}^2. \tag{6}$$

Dễ thấy hàm mục tiêu J_2 là hàm lồi nên ta có thể dùng các thuật toán tối ưu lồi để tìm tham số tối ưu θ^* của mô hình.

Hiệu chỉnh dạng L_2

- Kiểu hiệu chỉnh L_2 tương đương với việc giả định rằng các tham số θ_j tuân theo phân phối chuẩn với trung bình $\mu = 0$ và phương sai σ^2.
- Do đó, nếu một tham số θ_j càng xa giá trị trung bình 0 thì xác suất của nó càng nhỏ (tỉ lệ với độ lệch chuẩn σ).
- Ta có:

$$P(\theta_j) = \frac{1}{\sqrt{2\pi\sigma_j^2}} \exp\left(-\frac{\theta_j^2}{2\sigma_j^2}\right).$$

Hiệu chỉnh dạng L_2

- Theo công thức Bayes:

$$P(\theta|(\mathbf{x}_i, y_i)) \propto P((\mathbf{x}_i, y_i)|\theta)P(\theta),$$

  trong đó $P(\theta)$ là xác suất tiên nghiệm của tham số.
- Nếu giả định các tham số θ_j là độc lập thì ta có

$$P(\theta) = \prod_{j=1}^{D} P(\theta_j).$$

Hiệu chỉnh dạng L_2

- Nếu viết dưới dạng các log xác suất:

$$\log P(\theta|\{\mathbf{x}_i, y_i\}_{i=1}^N) = \log P(\{\mathbf{x}_i, y_i\}_{i=1}^N|\theta) + \sum_{j=1}^D \log\left\{\frac{1}{\sqrt{2\pi\sigma_j^2}}\exp\left(-\frac{\theta_j^2}{2\sigma_j^2}\right)\right\} + c,$$

  với c là một hằng số.

- Từ đó, hàm mục tiêu J sẽ có dạng

$$J(\theta) = -\ell(\theta) + \sum_{j=1}^D \frac{\theta_j^2}{2\sigma_j^2},$$

  và đây chính là dạng hiệu chỉnh L_2.

Ước lượng tham số

Ta có

$$\begin{aligned}\ell(\theta) &= \frac{1}{N}\sum_{i=1}^{N} \log P(y_i|\,\mathbf{x}_i;\theta) \\ &= \frac{1}{N}\sum_{i=1}^{N}\left\{\theta_{y_i}^\top \mathbf{x}_i - \log\left(\sum_{k=1}^{K}\exp(\theta_k^\top \mathbf{x}_i)\right)\right\}.\end{aligned}$$

Từ đó

$$\begin{aligned}\frac{\partial \ell(\theta)}{\partial \theta_{kj}} &= \frac{1}{N}\sum_{i=1}^{N}\left\{\frac{\partial}{\partial\theta_{kj}}(\theta_{y_i}^\top \mathbf{x}_i) - \frac{\partial}{\partial\theta_{kj}}\log\left(\sum_{k=1}^{K}\exp(\theta_k^\top \mathbf{x}_i)\right)\right\} \\ &= \frac{1}{N}\sum_{i=1}^{N}\left\{\delta(y_i = k)\,\mathbf{x}_{ij} - \frac{1}{\sum_{k=1}^{K}\exp(\theta_k^\top \mathbf{x}_i)}\frac{\partial}{\partial\theta_{kj}}\left(\sum_{k=1}^{K}\exp(\theta_k^\top \mathbf{x}_i)\right)\right\}.\end{aligned}$$

Ước lượng tham số

Do

$$\frac{\partial}{\partial \theta_{kj}} \left(\sum_{k=1}^{K} \exp(\theta_k^\top \mathbf{x}_i) \right) = \exp(\theta_k^\top \mathbf{x}_i)\, \mathbf{x}_{ij},$$

nên

$$\begin{aligned}
\frac{\partial}{\partial \theta_{kj}} \ell(\theta) &= \frac{1}{N} \sum_{i=1}^{N} \left\{ \delta(y_i = k)\, \mathbf{x}_{ij} - \frac{\exp(\theta_k^\top \mathbf{x}_i)}{\sum_{k=1}^{K} \exp(\theta_k^\top \mathbf{x}_i)} \mathbf{x}_{ij} \right\} \\
&= \frac{1}{N} \sum_{i=1}^{N} \delta(y_i = k)\, \mathbf{x}_{ij} - \frac{1}{N} \sum_{i=1}^{N} P(y = k \,|\, \mathbf{x}_i; \theta)\, \mathbf{x}_{ij} .
\end{aligned}$$

Ước lượng tham số

$$\frac{\partial}{\partial \theta_{kj}} \ell(\theta) = \frac{1}{N} \sum_{i=1}^{N} \delta(y_i = k)\, \mathbf{x}_{ij} - \frac{1}{N} \sum_{i=1}^{N} P(y = k|\, \mathbf{x}_i; \theta)\, \mathbf{x}_{ij}$$

Nhận xét:

- Đại lượng $\frac{1}{N}\sum_{i=1}^{N} \delta(y_i = k)\, \mathbf{x}_{ij}$ là kì vọng mẫu của đặc trưng thứ j trên mẫu huấn luyện ứng với lớp k;
- Đại lượng $\frac{1}{N}\sum_{i=1}^{N} P(y = k|\, \mathbf{x}_i; \theta)\, \mathbf{x}_{ij}$ là kì vọng của đặc trưng thứ j ứng với mô hình $P(y = k|\, \mathbf{x}; \theta)$.

Ước lượng tham số

- Với mô hình entropy cực đại hiệu chỉnh dạng L_2 thì đạo hàm riêng của hàm mục tiêu $J_2(\theta)$ ứng với tham số θ_{kj} là

$$\frac{\partial J_2(\theta)}{\partial \theta_{kj}} = -\left(\frac{1}{N} \sum_{i=1}^{N} \delta(y_i = k)\,\mathbf{x}_{ij} - \frac{1}{N} \sum_{i=1}^{N} P(y = k|\,\mathbf{x}_i;\theta)\,\mathbf{x}_{ij} \right) + \lambda\theta_{kj}.$$

- Để ước lượng véc-tơ tham số θ, ta cần giải hệ phương trình

$$\boxed{\frac{\partial}{\partial \theta_{kj}} J_2(\theta) = 0, \quad \forall k = 1, 2, \dots, K; \quad \forall j = 0, 1, \dots, D.}$$

Ước lượng tham số

- Có nhiều thuật toán được dùng để ước lượng tham số của mô hình entropy cực đại.
- Hai phương pháp chính:
  - **phương pháp thang lặp**
  - **phương pháp tối ưu**

Ước lượng tham số: Phương pháp thang lặp

- Thuật toán GIS (Generalized Iterative Scaling)
- Thuật toán IIS (Improved Iterative Scaling)
- Thuật toán SCGIS (Sequential Conditional Generalized Iterative Scaling)

Ước lượng tham số: Phương pháp tối ưu

- *Phương pháp gradient bậc một*: phương pháp giảm gradient, phương pháp gradient liên hợp;
- *Phương pháp gradient bậc hai*: phương pháp Newton và các phương pháp tựa-Newton:
    - thuật toán BFGS
    - thuật toán L-BFGS
    - thuật toán OWL-QN (Orthant-wise Limited-memory Quasi-Newton)
    - thuật toán Newton cụt

Ước lượng tham số

- Chú ý rằng các phương pháp tối ưu cũng sử dụng các thủ tục lặp để tìm chuỗi $\{\theta^{(n)}\}_{n=1}^{\infty}$ hội tụ tới giá trị tối ưu của tham số.

Ước lượng tham số: Một số nhận xét

- Các thuật toán tối ưu có tốc độ và hiệu quả cao hơn các thuật toán thang lặp.
- Phương pháp thang lặp cập nhật mỗi thành phần θ_j của θ tại một thời điểm, nên chi phí tại mỗi bước lặp là nhỏ nhưng số bước lặp là lớn.
- Ngược lại, phương pháp (tựa) Newton có chi phí cao tại mỗi bước lặp vì phải tính đúng (xấp xỉ) Hessian của hàm mục tiêu nhưng có tốc độ hội tụ nhanh.

Ước lượng tham số: Một số nhận xét

- Trong nhiều trường hợp, các phương pháp L-BFGS và gradient liên hợp là tốt hơn giảm gradient ngẫu nhiên trong nhiều trường hợp.
- Nếu số lượng tham số là tương đối nhỏ thì L-BFGS cho kết quả tốt, còn với các bài toán có số chiều lớn thì phương pháp gradient liên hợp thường cho kết quả tốt.
- Các phương pháp gradient liên hợp và L-BFGS cũng có thể tận dụng được các thuật toán tính toán song song tốt hơn.

Ước lượng tham số: Một số nhận xét

- Mô hình entropy cực đại hiệu chỉnh dạng L_2 thường cho kết quả cao hơn một chút mô hình entropy cực đại hiệu chỉnh dạng L_1.
- Tuy nhiên, dạng chuẩn hoá L_1 có độ hiệu quả gần tương tự mà lại có tốc độ huấn luyện nhanh hơn nhiều so với dạng hiệu chỉnh L_2.

Ước lượng tham số: Một số nhận xét

- Với dạng hiệu chỉnh L_2, đạo hàm của các số hạng hiệu chỉnh $\lambda\theta_j \to 0$ khi $\theta_j \to 0$.
- Tác động của số hạng hiệu chỉnh giảm dần nếu θ_j nhỏ.
- Từ đó, dạng hiệu chỉnh L_2 làm các tham số thường là nhỏ, xấp xỉ 0, nhưng không bằng 0.

Ước lượng tham số: Một số nhận xét

- Với dạng hiệu chỉnh L_1, đạo hàm của các số hạng hiệu chỉnh là $\lambda\mathtt{sign}(\theta_j) \in \{-\lambda, \lambda\}$ trừ khi $\theta_j = 0$.
- Tác động của các số hạng hiệu chỉnh là không đổi, không phụ thuộc vào mức độ lớn nhỏ của θ_j.
- Do đó, dạng chuẩn hoá L_1 sinh mô hình thưa, theo nghĩa sẽ cho kết quả ước lượng trong đó có nhiều tham số $\theta_j = 0$.
- Vì vậy, dạng hiệu chỉnh L_1 còn được dùng làm phương pháp chọn các đặc trưng.

Content

Phân loại ảnh

- Tập ảnh được cung cấp bởi nhóm nghiên cứu thị giác máy tính của Đại học Massachusetts, Hoa Kỳ.
- Tập dữ liệu gồm 210 ảnh dùng để huấn luyện mô hình và 2100 ảnh dùng để kiểm tra độ chính xác của mô hình.
- Mỗi ảnh được phân vào một trong 7 lớp sau: mặt gạch (brickface), bầu trời (sky), lá cây (foliage), xi-măng (cement), cửa sổ (window), đường đi (path) và cỏ (grass).
- Mỗi lớp có 30 mẫu huấn luyện và 300 mẫu kiểm tra.
- Các mẫu ảnh được trích ra từ 7 bức ảnh ngoài trời và được phân đoạn bằng tay để tạo phân loại cho từng điểm ảnh. Mỗi mẫu ảnh là một vùng điểm ảnh kích thước 3×3.

Phân loại ảnh

Mỗi mẫu có 19 đặc trưng là các số thực:

1. *region-centroid-col*: chỉ số cột của điểm ảnh trung tâm của vùng;
2. *region-centroid-row*: chỉ số hàng của điểm ảnh trung tâm của vùng;
3. *region-pixel-count*: số điểm ảnh của vùng, ở đây bằng 9;
4. *short-line-density-5*: kết quả của một thuật toán trích đoạn thẳng, là số đoạn thẳng độ dài 5 (hướng bất kì) với độ tương phản thấp, nhỏ hơn hoặc bằng 5, đi qua vùng ảnh;
5. *short-line-density-2*: giống như short-line-density-5 nhưng đếm số đoạn thẳng có độ tương phản cao, lớn hơn hoặc bằng 5;
6. *vedge-mean*: đo độ tương phản của các điểm ảnh nằm kề nhau theo chiều ngang trong vùng. Có 6 điểm ảnh, giá trị trung bình và độ lệch chuẩn cho trước. Đặc trưng này được sử dụng để phát hiện cạnh dọc.
7. *vegde-sd*: xem đặc trưng 6;

Phân loại ảnh

8. *hedge-mean*: đo độ tương phản của các điểm ảnh kề nhau theo chiều dọc. Được sử dụng để phát hiện đoạn nằm ngang;
9. *hedge-sd*: xem đặc trưng 8;
10. *intensity-mean*: giá trị trung bình trong vùng của $(R + G + B)/3$;
11. *rawred-mean*: giá trị trung bình trong vùng của giá trị R;
12. *rawblue-mean*: giá trị trung bình trong vùng của giá trị G;
13. *rawgreen-mean*: giá trị trung bình trong vùng của giá trị G;
14. *exred-mean*: đo màu đỏ thừa: $(2R - (G + B))$;
15. *exblue-mean*: đo màu xanh da trời thừa : $(2B - (G + R))$;
16. *exgreen-mean*: đo màu xanh lá cây thừa: $(2G - (R + B))$;
17. *value-mean*: biến đổi phi tuyến $3 - d$ của RGB.
18. *saturation-mean*: xem đặc trưng 17;
19. *hue-mean*: xem đặc trưng 17.

Phân loại ảnh

θ_j	brickface	sky	foliage	cement	window	path	grass
θ_0	0.0282	−0.0635	0.1531	−0.0025	0.1958	−0.2794	−0.0318
θ_1	−0.0197	−0.0125	0.0068	0.0137	0.0154	0.0037	−0.0073
θ_2	−0.0007	−0.039	−0.0613	−0.0045	−0.0718	0.1286	0.0488
θ_3	0.2541	−0.5715	1.3779	−0.0226	1.7625	−2.5145	−0.2858
θ_4	−0.0047	−0.0007	−0.0076	0.029	0.0179	0.0013	0.0006
θ_5	−0.0002	−0.0015	0.0037	−0.0062	−0.0051	0.0095	0
θ_6	−0.1484	−0.3321	0.3887	0.1627	0.1284	−0.234	0.0347
θ_7	−0.0497	−0.327	0.2637	0.0627	0.1015	−0.1441	0.0928
θ_8	0.0118	−0.2038	0.3646	0.0244	−0.3072	0.0579	0.0522
θ_9	−0.1309	−0.1823	0.0351	0.0579	0.122	0.0833	0.0149

Phân loại ảnh

θ_j	brickface	sky	foliage	cement	window	path	grass
θ_{10}	-0.0905	0.0828	-0.1482	0.03	-0.0013	0.0564	0.0709
θ_{11}	0.3326	-0.0878	-0.445	0.1565	-0.0414	0.1107	-0.0257
θ_{12}	-0.0122	0.0536	-0.1626	0.1011	0.0288	0.0924	-0.1011
θ_{13}	-0.592	0.2826	0.163	-0.1677	0.0087	-0.034	0.3394
θ_{14}	1.2693	-0.5118	-0.8904	0.3797	-0.1203	0.163	-0.2896
θ_{15}	0.2351	-0.0876	-0.0431	0.2133	0.0903	0.108	-0.516
θ_{16}	-1.5044	0.5994	0.9336	-0.593	0.03	-0.271	0.8055
θ_{17}	-0.0026	0.0501	0.1919	-0.1542	-0.1771	-0.0146	0.1066
θ_{18}	0.0036	-0.0181	0.6181	-0.0386	-0.4281	-0.1235	-0.0134
θ_{19}	0.0435	0.1154	-1.1184	-0.0238	0.3079	0.3985	0.277

Độ chính xác của mô hình:

- Trên tập huấn luyện: 96.66%
- Trên tập kiểm tra: 93.09%

Content

Iris

- Tập dữ liệu về hoa Iris[1] nổi tiếng trong lĩnh vực nhận dạng.
- Xuất hiện trong bài báo của Ronald Fisher năm 1936, ngày nay vẫn được dùng thường xuyên.
- Tập huấn luyện: 130 mẫu, tập kiểm tra: 20 mẫu

Đặc trưng	Lớp
độ dài của lá đài	Setosa
độ rộng của lá đài	Versicolour
độ dài của cánh hoa	Virginica
độ rộng của cánh hoa	

[1] http://archive.ics.uci.edu/ml/datasets/Iris

Phân loại hoa Iris

- L-BFGS, không sử dụng hiệu chỉnh tham số, độ chính xác của mô hình trên tập kiểm tra là 100% và trên tập huấn luyện là 98.46%
- Các tham số được ước lượng như sau:

Lớp	θ_0	θ_1	θ_2	θ_3	θ_4
`setosa`	3.281	5.633	16.834	−26.748	−12.396
`versicolor`	17.056	−1.582	−5.558	9.068	−1.777
`virginica`	−20.336	−4.051	−11.275	17.680	14.172

Phân loại hoa Iris

- Khi sử dụng hiệu chỉnh tham số thì số bước lặp và độ chính xác của mô hình ứng với các tham số λ được cho trong bảng sau:

λ	**Số bước**	**KT**	**HL**
1.0	30	100%	97.69%
2.0	17	100%	96.92%
3.0	19	100%	97.69%
4.0	16	100%	96.92%
5.0	14	100%	96.15%
6.0	20	100%	96.15%

Phân loại hoa Iris

- Ta thấy mô hình có hiệu chỉnh đạt độ chính xác cao nhất là 97.69% khi $\lambda = 3.0$.
- Các tham số của mô hình khi đó là:

Lớp	θ_0	θ_1	θ_2	θ_3	θ_4
`setosa`	0.236	0.546	1.279	−1.813	−0.847
`versicolor`	0.244	0.560	1.279	−1.813	−0.847
`virginica`	−0.507	−0.954	−1.083	1.810	1.337

Phân loại hoa Iris

- Ta thấy khi dùng phương pháp hiệu chỉnh L_2, các tham số của mô hình có giá trị tuyệt đối bé hơn nhiều giá trị tuyệt đối của các tham số trong mô hình không hiệu chỉnh; đồng thời các tham số phân bố xung quanh giá trị 0, phù hợp với khảo sát lí thuyết.[2].
- Số bước lặp của thuật toán tối ưu L-BFGS cũng phụ thuộc vào số hạng hiệu chỉnh λ.

[2] Với các tham số θ nhỏ, việc hiệu chỉnh cũng giúp giảm thiểu khả năng tràn số khi cài đặt mô hình entropy cực đại.

So sánh độ chính xác

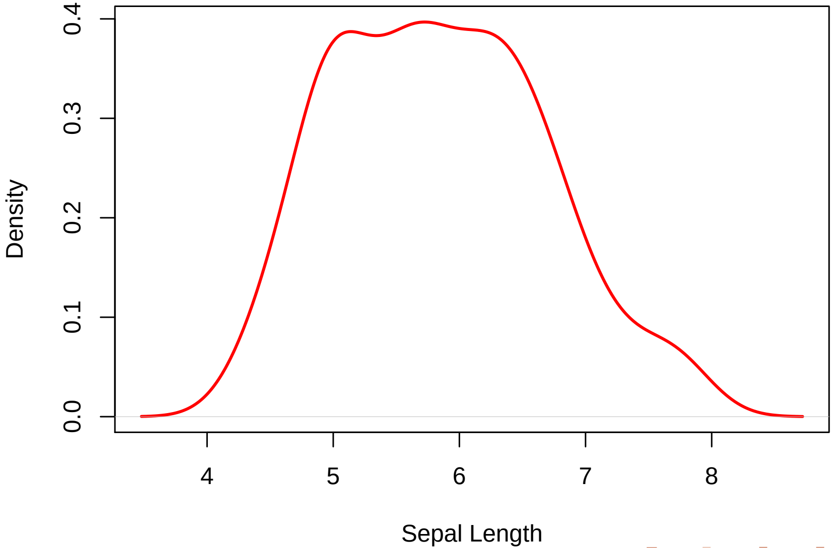
Độ chính xác của một số mô hình phân loại trên tập dữ liệu Iris:

Mô hình	**KT**	**HL**
Chuẩn một chiều (dùng riêng σ_j)	**100.00%**	95.38%
Chuẩn một chiều (dùng riêng σ_j)	85.00%	87.69%
GDA	**100.00%**	97.69%
MLR L-BFGS	**100.00%**	**98.46%**
MLR L-BFGS, L_2	**100.00%**	97.69%

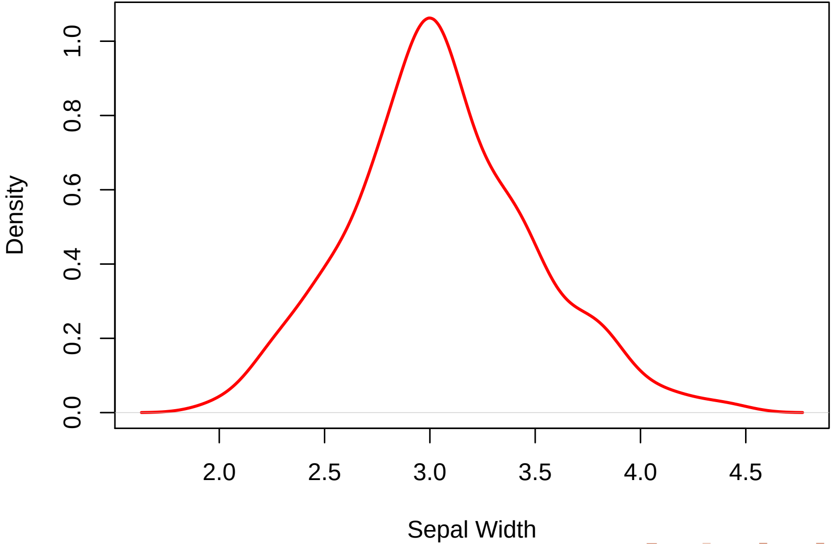
So sánh độ chính xác

- Ta thấy mô hình MLR cho kết quả tốt nhất trên cả tập kiểm tra và tập huấn luyện.
- Các đặc trưng trong tập dữ liệu là kích thước của các phần tử tự nhiên (lá và cánh hoa) nên thường tuân theo phân phối chuẩn.
- Mô hình GDA với giả định dữ liệu phân phối chuẩn tỏ ra mô hình hoá dữ liệu tốt.

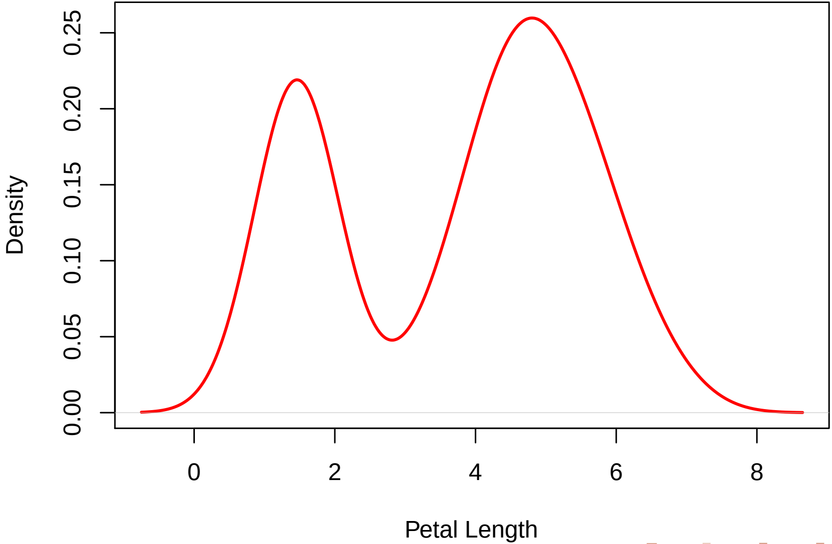
Lá đài, cánh hoa và phân phối chuẩn



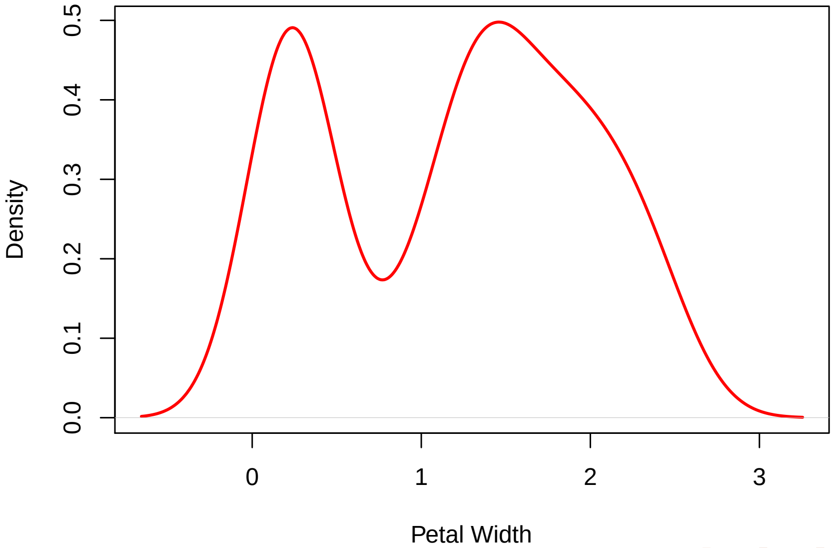
Lá đài, cánh hoa và phân phối chuẩn



Lá đài, cánh hoa và phân phối chuẩn



Lá đài, cánh hoa và phân phối chuẩn

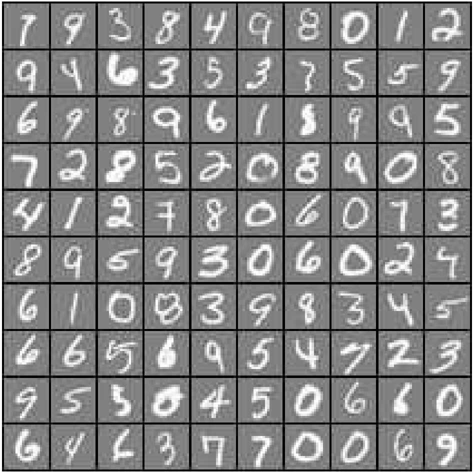


Content

Handwritten Digit Recognition

- Application of MLR on a dataset of handwritten digits. The set contains 5000 training examples.
- This is a subset of the MNIST handwritten digit dataset:
  - *http://yann.lecun.com/exdb/mnist/*
- Each training example is a 20x20 pixels grayscale image of the digit.
- Each pixel is represented by a floating point number indicating the grayscale intensity at that location.
- The 20x20 grid of pixel is unrolled into a 400-dimensional vector.

Handwritten Digit Recognition



Handwritten Digit Recognition

- L_2 regularization with $\lambda = 0.1$
- Logistic regression using one-versus-all classification:
    - Train 10 binary logistic regression classifiers
    - Training accuracy: 94.98%
- MLR: 96.72%
    - $J(0) = 2.302585, J(\theta^*) = 0.147061$

Content

Wine Origin

- Data of a chemical analysis of wines grown in the same region in Italy but derived from three different cultivars.
- 13 constituents found in each of the three types of wines:
- *http://archive.ics.uci.edu/ml/datasets/Wine*

- Training accuracy: 100%.
  - No regularization
  - $J(0) = 1.098612, J(\theta^*) = 0.007770$

Content

Wine Quality

- Two datasets related to red and white vinho verde wine samples, from the north of Portugal.
- The goal is to model wine quality based on physicochemical tests.
- *http://archive.ics.uci.edu/ml/datasets/Wine+Quality*

- Number of instances: red wine $= 1599$; white wine $= 4898$; $D = 11$.

Wine Quality

No regularization:

	Red Wine	**White Wine**
Training accuracy	60.85%	54.35%
$J(0)$	1.791759	1.945910
$J(\theta^*)$	0.921343	1.093946

Content

Hàm đặc trưng

- Ta tổng quát hoá mô hình entropy cực đại ở mục trước với việc sử dụng các **hàm đặc trưng**.
- Việc sử dụng hàm đặc trưng cho phép biểu diễn ngắn gọn tập dữ liệu quan sát (các đối tượng $\mathbf{x}_i$ và lớp y_i) và tổng quát hoá mô hình.

Hàm đặc trưng

- Giả sử X và Y là các biến ngẫu nhiên xác định tương ứng trên các tập $\mathcal{X}$ và $\mathcal{Y}$.
- Để ngắn gọn kí hiệu, với $(\mathbf{x}, y) \in (\mathcal{X}, \mathcal{Y})$, ta viết $P(X = \mathbf{x} \,|\, Y = y)$ đơn giản là $P(\mathbf{x} \,|\, y)$.
- Ta định nghĩa hàm đặc trưng f như sau:

$$f : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}^D.$$

Hàm đặc trưng

- Với mỗi $(\mathbf{x}, y) \in (\mathcal{X}, \mathcal{Y})$, $f(\mathbf{x}, y)$ là một véc-tơ D chiều ứng với D đặc trưng:
$$f(\mathbf{x}, y) = (f_1(\mathbf{x}, y), f_2(\mathbf{x}, y), \dots, f_D(\mathbf{x}, y)).$$
- Các hàm đặc trưng thành phần $f_j(\mathbf{x}, y) \in \mathbb{R}$, tuy nhiên trong mô hình entropy cực đại chúng thường nhận giá trị nhị phân thông qua một hàm chỉ số nào đó của $\mathbf{x}$ và y.

Hàm đặc trưng

- Tổng quát, mỗi hàm đặc trưng được định nghĩa bởi:

$$f_j(\mathbf{x}, y) = A_a(\mathbf{x})B_b(y),$$

  trong đó chỉ số dưới a đánh số một tập hàm xác định trên $\mathbf{x}$, chỉ số dưới b đánh số một tập hàm xác định trên y.
- Nếu các hàm này là hàm nhị phân xác định việc có hay không có một tính chất nào đó của $\mathbf{x}$ và y thì tích $A_a(\mathbf{x})B_b(y)$ là một dạng hội logic.

Hàm đặc trưng

- Ta có thể xác định mọi thông tin hữu ích cho việc phân loại bằng các hàm đặc trưng tương ứng xác định trên các lớp y và các thuộc tính x_i của $\mathbf{x}$.
- Các hàm này không nhất thiết phải độc lập nhau.
- Ví dụ, nếu $\mathbf{x}$ là một từ, ta có thể xây dựng các hàm đặc trưng khai thác các thông tin của $\mathbf{x}$ như:

$$
\begin{aligned}
A_1(\mathbf{x}) &= \delta(\mathbf{x} \text{ bắt đầu bằng một chữ cái in hoa}) \\
A_2(\mathbf{x}) &= \delta(\mathbf{x} \text{ bắt đầu bằng T}) \\
A_3(\mathbf{x}) &= \delta(\mathbf{x} \text{ là Thomson}) \\
A_4(\mathbf{x}) &= \delta(\mathbf{x} \text{ có 7 chữ cái})
\end{aligned}
$$

Hàm đặc trưng

- Các giá trị của hàm đặc trưng này thường được trích rút tự động từ các *mẫu đặc trưng* tương ứng.
- Trong các mô hình entropy cực đại ứng dụng trong học máy, số chiều D của mỗi véc-tơ đặc trưng là lớn, có thể từ hàng trăm ngàn tới hàng triệu đặc trưng.
- Ta kí hiệu $\theta \in \mathbb{R}^D$ là véc-tơ tham số của mô hình.

Content

Mô hình hồi quy logistic với hàm đặc trưng

Xác suất của mỗi lớp được xác định bởi

$$P(y|\,\mathbf{x};\theta) = \frac{\exp(\theta^\top f(\mathbf{x}, y))}{\sum_{y' \in \mathcal{Y}} \exp(\theta^\top f(\mathbf{x}, y'))}. \tag{7}$$

Mẫu số của xác suất này chính là số hạng chuẩn hoá

$$Z(\theta) = \sum_{y' \in \mathcal{Y}} \exp(\theta^\top f(\mathbf{x}, y')),$$

để đảm bảo phân phối xác suất:

$$\sum_{y \in \mathcal{Y}} P(y|\,\mathbf{x};\theta) = 1, \forall\, \mathbf{x} \in \mathcal{X}.$$

Mô hình hồi quy logistic với hàm đặc trưng

Cho trước mẫu huấn luyện $(\mathbf{x}_i, y_i), i = 1, 2, \dots, N$, trung bình log-hợp lí của dữ liệu là:

$$\begin{aligned}\ell(\theta) &= \frac{1}{N}\sum_{i=1}^{N} \log P(y_i | \mathbf{x}_i; \theta) \\ &= \frac{1}{N}\sum_{i=1}^{N} \left[\theta^\top f(\mathbf{x}_i, y_i) - \log\left(\sum_{y\in\mathcal{Y}} \exp(\theta^\top f(\mathbf{x}_i, y))\right)\right].\end{aligned}$$

Mô hình hồi quy logistic với hàm đặc trưng

Ta có, $\forall j = 1, 2, \ldots, D$:

$$
\begin{aligned}
\frac{\partial}{\partial \theta_j} \ell(\theta) &= \frac{1}{N} \sum_{i=1}^{N} \left[f_j(\mathbf{x}_i, y_i) - \frac{1}{\sum_{y \in \mathcal{Y}} \exp(\theta^\top f(\mathbf{x}_i, y))} \frac{\partial}{\partial \theta_j} \left(\sum_{y \in \mathcal{Y}} \exp(\theta^\top f(\mathbf{x}_i, y)) \right) \right] \\
&= \frac{1}{N} \sum_{i=1}^{N} \left[f_j(\mathbf{x}_i, y_i) - \frac{1}{\sum_{y \in \mathcal{Y}} \exp(\theta^\top f(\mathbf{x}_i, y))} \sum_{y \in \mathcal{Y}} \exp(\theta^\top f(\mathbf{x}_i, y)) f_j(\mathbf{x}_i, y) \right] \\
&= \frac{1}{N} \sum_{i=1}^{N} \left[f_j(\mathbf{x}_i, y_i) - \sum_{y \in \mathcal{Y}} \frac{\exp(\theta^\top f(\mathbf{x}_i, y))}{\sum_{y \in \mathcal{Y}} \exp(\theta^\top f(\mathbf{x}_i, y))} f_j(\mathbf{x}_i, y) \right] \\
&= \frac{1}{N} \sum_{i=1}^{N} \left[f_j(\mathbf{x}_i, y_i) - \sum_{y \in \mathcal{Y}} P(y | \mathbf{x}_i; \theta) f_j(\mathbf{x}_i, y) \right]
\end{aligned}
$$

Mô hình hồi quy logistic với hàm đặc trưng

Ta thấy:

- $\frac{1}{N}\sum_{i=1}^{N} f_j(\mathbf{x}_i, y_i) = \hat{\mathbb{E}}[f_j(\mathbf{x}, y)]$ là kì vọng mẫu của đặc trưng thứ j trên tập huấn luyện;
- $\frac{1}{N}\sum_{i=1}^{N}\sum_{y\in\mathcal{Y}} P(y|\,\mathbf{x}_i;\theta) f_j(\mathbf{x}_i, y) = \mathbb{E}[f_j(\mathbf{x}, y)]$ là *giá trị kì vọng* của đặc trưng thứ j theo phân phối xác suất của mô hình.

Như vậy, việc tìm θ gắn với việc giải hệ phương trình:

$$\hat{\mathbb{E}}[f_j(\mathbf{x}, y)] = \mathbb{E}[f_j(\mathbf{x}, y)], \forall j = 1, 2, \ldots, D.$$

Nói cách khác, ta cần tìm mô hình trong đó kì vọng của mỗi đặc trưng j khớp với giá trị của nó trên tập huấn luyện.

Mô hình hồi quy logistic với hàm đặc trưng

Khi áp dụng mô hình entropy cực đại với dạng hiệu chỉnh L_2, ta có hàm mục tiêu cần cực tiểu hoá và các đạo hàm riêng của nó là:

$$J_2(\theta) = -\frac{1}{N}\sum_{i=1}^{N}\left[\theta^\top f(\mathbf{x}_i, y_i) - \log\left(\sum_{y\in\mathcal{Y}} \exp(\theta^\top f(\mathbf{x}_i, y))\right)\right] + \frac{\lambda}{2}\theta^\top\theta$$

$$\frac{\partial}{\partial\theta_j} J_2(\theta) = -\frac{1}{N}\sum_{i=1}^{N}\left[f_j(\mathbf{x}_i, y_i) - \sum_{y\in\mathcal{Y}} P(y|\,\mathbf{x}_i;\theta) f_j(\mathbf{x}_i, y)\right] + \lambda\theta_j.$$

Bài tập

Cài đặt các thuật toán ước lượng tham số của mô hình hồi quy logistic đa lớp:

1. Thuật toán giảm gradient theo loạt
2. Thuật toán giảm gradient ngẫu nhiên
3. Thuật toán Newton

Chạy các thuật toán trên các dữ liệu thử nghiệm và thông báo kết quả.