

4. Regresia logistică

Reprezentarea funcției ipoteză
Marginea de decizie
Funcția logistică de cost
Algoritmul de minimizare al gradientului

Așa cum s-a menționat în lucrarea de laborator *Regresia Liniară*, avem de-a face cu o problemă de clasificare atunci când valoarea de ieșire a algoritmului este discretă, $y \in \{0, 1, \dots, k\}$. O primă încercare simplă de a clasifica în valori discrete caracteristicile de intrare este de a rula metoda de regresie liniară și de a mapa toate predicțiile mai mari de 0.5 la valoarea de ieșire $y = 1$, iar pe cele mai mici decât 0.5 la valoarea $y = 0$. Această abordare nu ar funcționa corect, deoarece operația de clasificare nu este guvernată de o funcție liniară.

În această lucrare ne vom concentra asupra problemei de **clasificare binară**, în care y poate lua doar două valori, 0 sau 1. De exemplu, în cazul în care dorim să construim un clasificator ce ne poate spune dacă un email este spam sau nu, atunci $x^{(i)}$ ar fi un set de caracteristici ce descrie email-ul, iar y ar putea fi 1 dacă email-ul este spam și 0 în caz contrar. Altfel spus, $y \in \{0, 1\}$. De obicei clasa 0 se numește clasă negativă ("−"), iar 1 clasa pozitivă ("+"). Valoarea $y^{(i)}$ este denumită *eticheta* (eng. label) exemplului de antrenare $x^{(i)}$.

4.1 Reprezentarea ipotezei

Pentru a putea construi metoda de regresie logistică, vom modifica forma ipotezei $h_\theta(x)$ în așa fel încât să ia valori doar în intervalul $[0, 1]$:

$$0 \leq h_\theta(x) \leq 1. \quad (4.1)$$

Reducerea valorilor pentru $h_\theta(x)$ la intervalul $[0, 1]$ se face prin integrarea lui $\theta^T x$ în așa numita *funcție logistică*, denumită și *sigmoid*:

$$h_\theta(x) = g(\theta^T x), \quad (4.2)$$

$$z = \theta^T x, \quad (4.3)$$

$$g(z) = \frac{1}{1 + e^{-z}}. \quad (4.4)$$

Forma funcției sigmoid $g(z)$ este redată în figura 4.1.

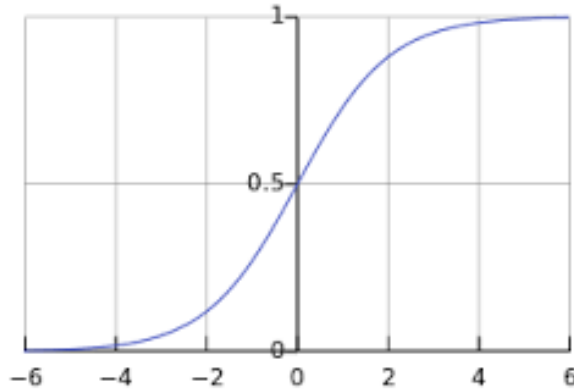


Fig. 4.1 Forma funcției sigmoid.

Funcția $g(z)$ mapează orice număr real la intervalul $[0, 1]$, fiind astfel utilă la transformarea funcției într-o funcție potrivită pentru clasificare.

$h_\theta(x)$ ne va indica probabilitatea ca ieșirea y să fie 1. Spre exemplu, $h_\theta(x) = 0.7$ ne spune ca există o probabilitate de 70% ca ieșirea să fie 1. Probabilitatea ca predicția să fie 0 este complementul probabilității ieșirii de a fi 1 (de exemplu, atunci când probabilitatea să fie 1 este de 70%, probabilitatea ieșirii să fie 0 este de 30%):

$$h_\theta(x) = P(y = 1|x; \theta) = 1 - P(y = 0|x; \theta), \quad (4.5)$$

$$P(y = 0|x; \theta) + P(y = 1|x; \theta) = 1. \quad (4.6)$$

4.2 Marginea de decizie

Pentru a putea obține valori discrete de 0 și 1 pentru clasificare, putem converti ieșirea ipotezei astfel:

$$h_\theta(x) \geq 0.5 \rightarrow y = 1, \quad (4.7)$$

$$h_\theta(x) < 0.5 \rightarrow y = 0. \quad (4.8)$$

Ieșirea funcției logistice g este egală sau mai mare de 0.5 atunci când valoarea de intrare este mai mare sau egală cu zero:

$$g(z) \geq 0.5 \text{ dacă } z \geq 0. \quad (4.9)$$

Generalizat, valoarea de ieșire a lui $g(z)$ pentru $z \in (-\infty, \infty)$ este:

$$z = 0, e^0 = 1 \Rightarrow g(z) = \frac{1}{2}, \quad (4.10)$$

$$z \rightarrow \infty, e^{-\infty} \rightarrow 0 \Rightarrow g(z) = 1, \quad (4.11)$$

$$z \rightarrow -\infty, e^{\infty} \rightarrow \infty \Rightarrow g(z) = 0. \quad (4.12)$$

Utilizând funcția logistică, sau sigmoid, valoarea ipotezei $h_{\theta}(x)$ pentru intrarea $\theta^T x$ este:

$$h_{\theta}(x) = g(\theta^T x) \geq 0.5 \text{ dacă } \theta^T x \geq 0. \quad (4.13)$$

Astfel, putem deduce pentru y următoarele valori:

$$\theta^T x \geq 0 \Rightarrow y = 1, \quad (4.14)$$

$$\theta^T x < 0 \Rightarrow y = 0. \quad (4.15)$$

Marginea de decizie este linia ce separă zona în care $y = 0$ de cea în care $y = 1$. Această margine este creată de funcția ipoteză $h_{\theta}(x)$.

Exemplu

Luând în considerare o problemă de clasificare binară, în care utilizăm două caracteristici, x_1 și x_2 , și următorul set de parametri θ ai funcției $h_{\theta}(x)$:

$$\theta = \begin{bmatrix} 5 \\ -1 \\ 0 \end{bmatrix}, \quad (4.16)$$

valoarea de ieșire y va fi 1 dacă:

$$y = 1 \text{ dacă } 5 + (-1)x_1 + 0x_2 \geq 0. \quad (4.17)$$

Astfel:

$$\begin{aligned} 5 - x_1 &\geq 0 \\ -x_1 &\geq -5 \\ x_1 &\leq 5 \end{aligned}$$

În acest caz, marginea de decizie este o linie verticală poziționată pe axa x la valoarea $x_1 = 5$. Orice punct aflat la stânga marginii de decizie va avea eticheta $y = 1$, în timp ce orice punct aflat la dreapta marginii va avea eticheta $y = 0$, așa cum este ilustrat în figura 4.2.

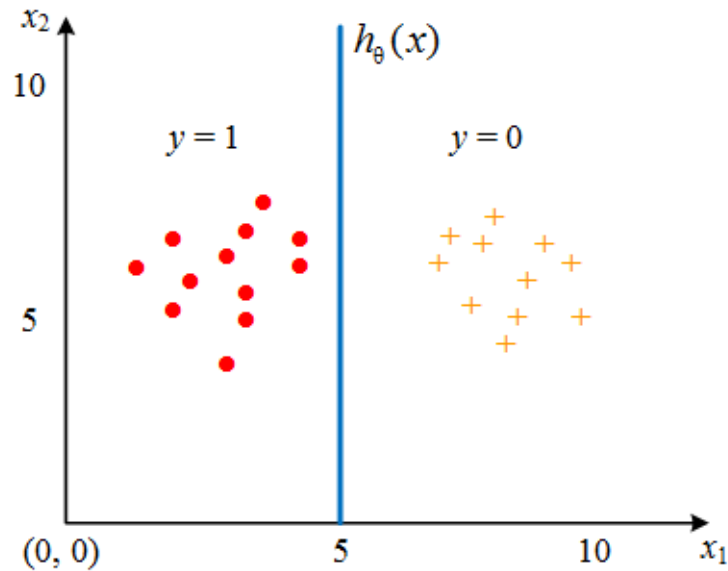


Fig. 4.2 Marginea de decizie pentru o posibilă ipoteză $h_\theta(x)$.

Valoarea de intrare pentru funcția sigmoid $g(z)$ (e.g. $\theta^T x$) nu trebuie să fie neapărat liniară. Această valoare poate fi, spre exemplu, o funcție ce descrie un cerc (e.g. $z = \theta_0 + \theta_1 x_1^2 + \theta_2 x_2^2$), sau orice altă funcție ce poate descrie datele de intrare.

4.3 Funcția de cost

În cazul în care am utiliza pentru regresia logistică funcția de cost utilizată în regresia liniară, am obține o ieșire perturbată (sub formă "ondulată") ce ar produce o mulțime de minime locale. Cu alte cuvinte, funcția $J(\theta)$ nu ar fi convexă.

Funcția de cost pentru regresia logistică are următoarea formă:

$$J(\theta) = \frac{1}{m} \sum_{i=1}^m \text{cost}(h_\theta(x^{(i)}), y^{(i)}), \quad (4.18)$$

$$\text{cost}(h_\theta(x), y) = -\log(h_\theta(x)) \text{ pentru } y = 1, \quad (4.19)$$

$$\text{cost}(h_\theta(x), y) = -\log(1 - h_\theta(x)) \text{ pentru } y = 0. \quad (4.20)$$

Forma funcției $J(\theta)$ față de ipoteza $h_\theta(x)$ este ilustrată pentru cele două cazuri, $y = 0$ și $y = 1$ în figura 4.3(a), respectiv în figura 4.3(b).

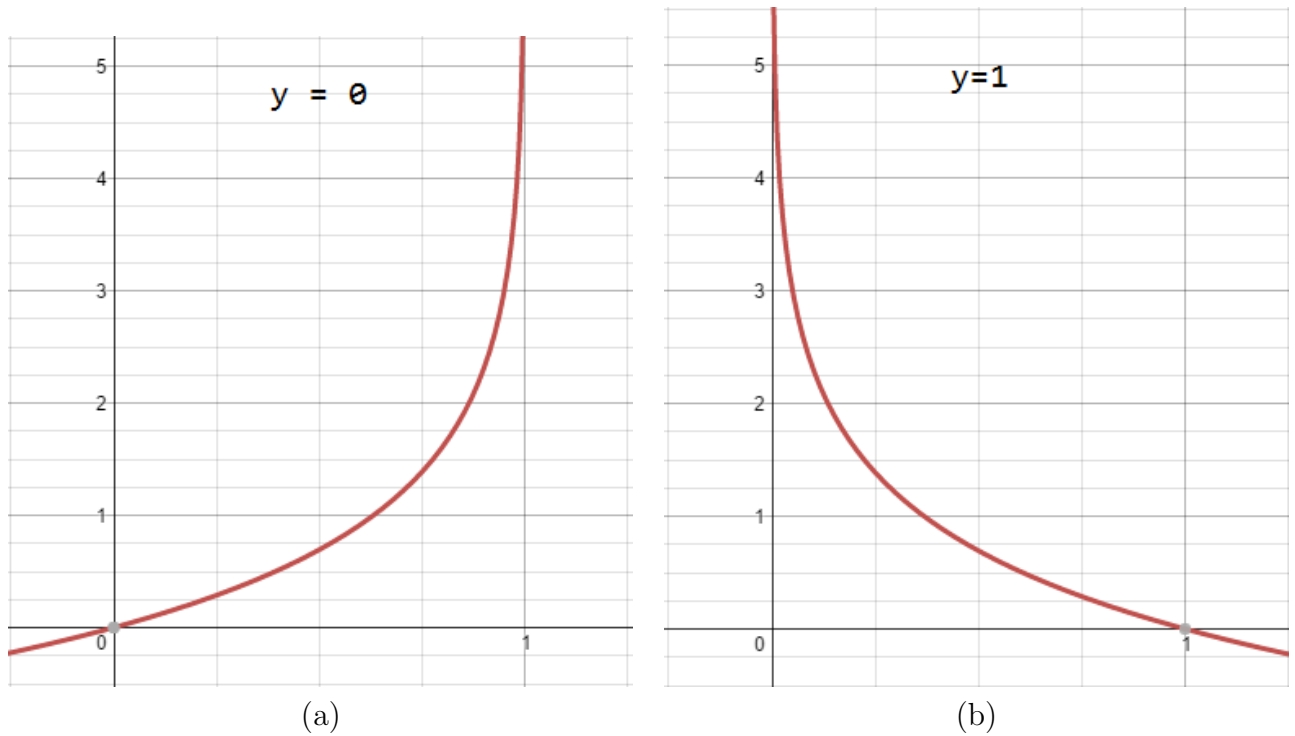


Fig. 4.3 Forma funcției de cost $J(\theta)$ pentru regresia logistică: (a) $y = 0$; (b) $y = 1$.

Astfel:

$$\text{cost}(h_{\theta}(x), y) = 0 \text{ dacă } h_{\theta}(x) = y, \quad (4.21)$$

$$\text{cost}(h_{\theta}(x), y) \rightarrow \infty \text{ dacă } y = 0 \text{ și } h_{\theta}(x) \rightarrow 1, \quad (4.22)$$

$$\text{cost}(h_{\theta}(x), y) \rightarrow \infty \text{ dacă } y = 1 \text{ și } h_{\theta}(x) \rightarrow 0. \quad (4.23)$$

În cazul în care eticheta y este 0, atunci funcția de cost va fi zero dacă și ieșirea ipotezei este 0. Când ipoteza se apropie de valoarea 1, atunci funcția de cost va tinde către o valoare infinită.

Analog, în cazul în care eticheta y este 1, atunci funcția de cost va fi zero dacă ieșirea ipotezei este 1. Când ipoteza se apropie de valoarea 0, atunci funcția de cost va tinde către o valoare infinită.

Această formă a funcției de cost garantează că $J(\theta)$ este convexă pentru regresia logistică.

4.4 Forma compactă a funcției de cost

Cele două forme ale funcției de cost pot fi scrise în aceeași formulă astfel:

$$\text{cost}(h_{\theta}(x), y) = -y \log(h_{\theta}(x)) - (1 - y) \log(1 - h_{\theta}(x)). \quad (4.24)$$

Se poate observa că al doilea termen $(1 - y) \log(1 - h_\theta(x))$ va fi zero atunci când y este 1, neafectând astfel rezultatul. Dacă y este 0, atunci primul termen $-y \log(h_\theta(x))$ va fi zero și nu va afecta rezultatul.

Pentru setul de antrenare compus din m exemple, $J(\theta)$ va avea următoarea expresie:

$$J(\theta) = -\frac{1}{m} \sum_{i=1}^m [y^{(i)} \log(h_\theta(x^{(i)})) + (1 - y^{(i)}) \log(1 - h_\theta(x^{(i)}))]. \quad (4.25)$$

Forma vectorială a ecuației 4.25 este:

$$h = g(x\theta), \quad (4.26)$$

$$J(\theta) = \frac{1}{m} \cdot (-y^T \log(h) - (1 - y)^T \log(1 - h)). \quad (4.27)$$

4.5 Algoritm de minimizare al gradientului

Pentru determinarea parametrilor optimi θ , vom utiliza algoritmul de minimizare a gradientului descris în lucrarea de laborator *Regresia Liniară*.

Forma generală a algoritmului este:

$$\begin{array}{l} \text{Repetă } \{ \\ \quad \theta_j := \theta_j - \alpha \frac{\partial}{\partial \theta_j} J(\theta) \\ \} \end{array}$$

Luând în considerare derivata parțială a lui $J(\theta)$:

$$\begin{array}{l} \text{Repetă până la convergență: } \{ \\ \quad \theta_j := \theta_j - \alpha \frac{1}{m} \sum_{i=1}^m (h_\theta(x^{(i)}) - y^{(i)}) x_j^{(i)} \\ \} \end{array}$$

Varianta vectorizată a algoritmului este:

$$\theta := \theta - \alpha \frac{1}{m} x^T (g(x\theta) - y). \quad (4.28)$$

4.6 Cerințe

Cerința din această lucrare de laborator este de a construi un model de clasificare binară ce va estima, pe bazele rezultatelor a două teste, șansele unui elev de a fi admis la facultate.

4.6.1 Baza de date de antrenare

Pentru a implementa algoritmul, descărcați datele din arhiva "lab4Data.zip".

În acest exercițiu vom presupune că un liceu are o bază de date reprezentând 40 de elevi ce sunt admiși la facultate și 40 ce nu sunt admiși. Fiecare exemplu de antrenare $(x^{(i)}, y^{(i)})$ conține scorul unui elev la două teste standard și o etichetă ce indică dacă elevul a fost admis sau nu la facultate.

Structura bazei de date de antrenare are următoarea formă:

- (a) fișierul "lab4x.dat" conține caracteristicile x (rezultatele la cele două examene); prima coloană conține rezultatele pentru examenul 1, iar a doua coloană rezultatele pentru examenul 2;
- (b) fișierul "lab4y.dat" conține etichetele y , unde $y = 1$ reprezintă admiterea elevului la facultate; elevul nu a fost admis atunci când $y = 0$.