3. Regresia logistică

Reprezentarea funcţiei ipoteză
Marginea de decizie
Funcţia logistică de cost
Algoritmul de minimizare al gradientului

Aşa cum s-a menţionat în cursul *Regresia Liniară*, avem de-a face cu o problemă de clasificare atunci când valoarea de ieşire a algoritmului este discretă, $y \in \{0, 1, ..., k\}$. O primă încercare simplă de a clasifica în valori discrete caracteristicile de intrare este de a rula metoda de regresie liniară şi de a mapa toate predicţiile mai mari de 0.5 la valoarea de ieşire $y = 1$, iar pe cele mai mici decât 0.5 la valoarea $y = 0$. Această abordare nu ar funcţiona corect, deoarece operaţia de clasificare nu este guvernată de o funcţie liniară.

În acest curs 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 şi 0 în caz contrar. Altfel spus, $y \in \{0, 1\}$. De obicei clasa 0 se numeşte clasă negativă ("-"), iar 1 clasa pozitivă ("+"). Valoarea $y^{(i)}$ este denumită *eticheta* (eng. label) exemplului de antrenare $x^{(i)}$.

3.1 Reprezentarea ipotezei

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

$$0 \leq h_\theta(x) \leq 1. \tag{3.1}$$

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

$$h_\theta(x) = g(\theta^T x), \tag{3.2}$$

$$z = \theta^T x, \tag{3.3}$$

$$g(z) = \frac{1}{1 + e^{-z}}. \tag{3.4}$$

Forma funcţiei sigmoid $g(z)$ este redată în figura 3.1.

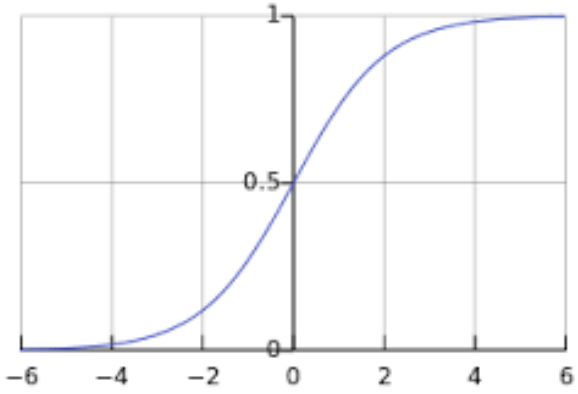


Fig. 3.1 Forma funcţiei sigmoid.

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

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

$$P(y = 1|x; \theta) = h_\theta(x), \tag{3.5}$$

$$P(y = 0|x; \theta) = 1 - h_\theta(x), \tag{3.6}$$

astfel ca:

$$P(y = 0|x; \theta) + P(y = 1|x; \theta) = 1. \tag{3.7}$$

Ecuațiile de mai sus pot fi scrise compact:

$$P(y|x; \theta) = (h_\theta(x))^y (1 - h_\theta(x))^{1-y}, \tag{3.8}$$

Luând în considerare m exemple de antrenare, putem scrie probabilitatea $L(\theta)$ a parametrilor modelului, demunită și likelihood:

$$\begin{aligned} L(\theta) &= P(y|x; \theta) \\ &= \prod_{i=1}^{m} P(y^{(i)}|x^{(i)}; \theta) \\ &= \prod_{i=1}^{m} (h_\theta(x)^{(i)})^{y^{(i)}} (1 - h_\theta(x^{(i)}))^{1-y^{(i)}} \end{aligned} \tag{3.9}$$

Parametrii optimi ai modelului sunt obținuți prin maximizarea lui $L(\theta)$. Înmulțirile introduse de produs pot fi simplificate prin înlocuirea lor cu suma logaritmilor probabilității $L(\theta)$:

$$\begin{aligned} l(\theta) &= \log L(\theta) \\ &= \sum_{i=1}^{m} y^{(i)} \log h_\theta(x^{(i)}) + (1 - y^{(i)}) \log(1 - h_\theta(x^{(i)})) \end{aligned} \tag{3.10}$$

unde $l(\theta)$ poartă denumirea de log-likelihood.

3.2 Marginea de decizie

Pentru a putea obţine valori discrete de 0 şi 1 pentru clasificare, putem converti ieşirea ipotezei astfel:

$$h_\theta(x) \geq 0.5 \rightarrow y = 1, \tag{3.11}$$

$$h_\theta(x) < 0.5 \rightarrow y = 0. \tag{3.12}$$

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

$$g(z) \geq 0.5 \text{ dacă } z \geq 0. \tag{3.13}$$

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

$$z = 0, e^0 = 1 \Rightarrow g(z) = \frac{1}{2}, \tag{3.14}$$

$$z \rightarrow \infty, e^{-\infty} \rightarrow 0 \Rightarrow g(z) = 1, \tag{3.15}$$

$$z \rightarrow -\infty, e^{\infty} \rightarrow \infty \Rightarrow g(z) = 0. \tag{3.16}$$

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

$$h_\theta(x) = g(\theta^T x) \geq 0.5 \text{ dacă } \theta^T x \geq 0. \tag{3.17}$$

Astfel, putem deduce pentru y următoarele valori:

$$\theta^T x \geq 0 \Rightarrow y = 1, \tag{3.18}$$

$$\theta^T x < 0 \Rightarrow y = 0. \tag{3.19}$$

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

Exemplu

Luând în considerare o problemă de clasificare binară, în care utilizăm două caracteristici, x_1 şi x_2, şi următorul set de parametrii θ ai funcţiei $h_\theta(x)$:

$$\theta = \begin{bmatrix} 5 \\ -1 \\ 0 \end{bmatrix}, \tag{3.20}$$

valoarea de ieşire y va fi 1 dacă:

$$y = 1 \text{ dacă } 5 + (-1)x_1 + 0x_2 \geq 0. \tag{3.21}$$

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 poziţionată pe axa x la valoarea $x_1 = 5$. Orice punct aflat la stânga marginei de decizie va avea eticheta $y = 1$, în timp ce orice punct aflat la dreapta marginii va avea eticheta $y = 0$, aşa cum este ilustrat în figura 3.2.

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

3.3 Funcţia de cost

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

Funcţia de cost pentru regresia logistică are următoarea formă:

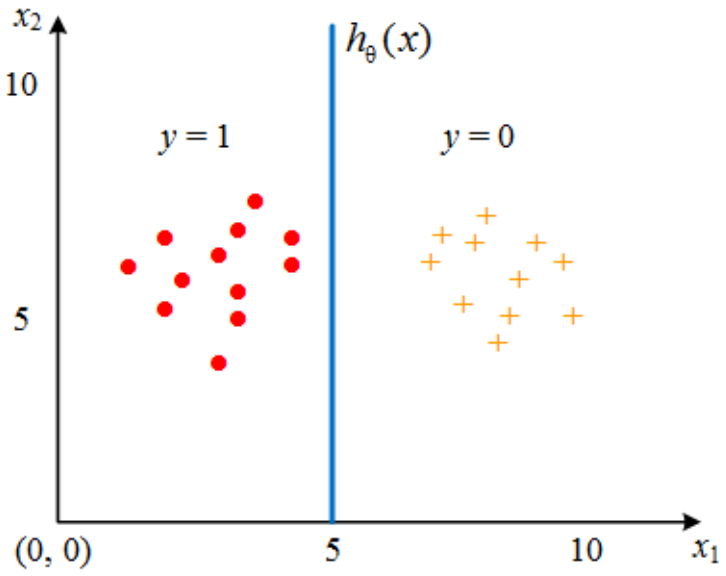


Fig. 3.2 Marginea de decizie pentru o posibilă ipoteză $h_\theta(x)$.

$$J(\theta) = \frac{1}{m}\sum_{i=1}^{m} \text{cost}(h_\theta(x^{(i)}), y^{(i)}), \tag{3.22}$$

$$\text{cost}(h_\theta(x), y) = -\log(h_\theta(x)) \text{ pentru } y = 1, \tag{3.23}$$

$$\text{cost}(h_\theta(x), y) = -\log(1 - h_\theta(x)) \text{ pentru } y = 0. \tag{3.24}$$

Forma funcţiei $J(\theta)$ faţă de ipoteza $h_\theta(x)$ este ilustrată pentru cele două cazuri, $y = 0$ şi $y = 1$ în figura 3.3(a), respectiv în figura 3.3(b).

Astfel:

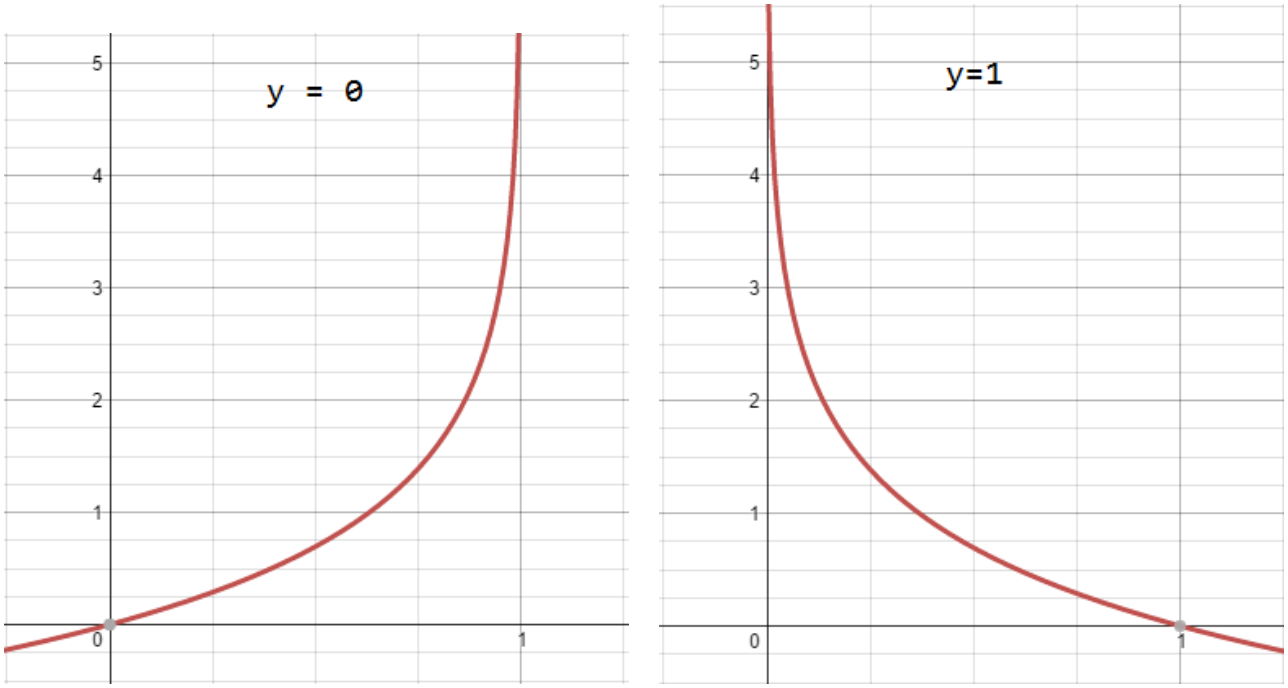
$$\text{cost}(h_\theta(x), y) = 0 \text{ dacă } h_\theta(x) = y, \tag{3.25}$$

$$\text{cost}(h_\theta(x), y) \to \infty \text{ dacă } y = 0 \text{ şi } h_\theta(x) \to 1, \tag{3.26}$$

$$\text{cost}(h_\theta(x), y) \to \infty \text{ dacă } y = 1 \text{ şi } h_\theta(x) \to 0. \tag{3.27}$$

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

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



(a) (b)

Fig. 3.3 Forma funcţiei de cost $J(\theta)$ pentru regresia logistică: (a) $y = 0$; (b) $y = 1$.

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

3.4 Forma compactă a funcţiei de cost

Cele două forme ale funcţiei de cost pot fi scrise în aceeaşi formulă astfel:

$$\text{cost}(h_\theta(x), y) = -y \ \log(h_\theta(x)) - (1 - y) \log(1 - h_\theta(x)). \tag{3.28}$$

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 şi nu va afecta rezultatul.

Deoarece dorim sa minimizam valoarea funcției de cost, probabilitatea exprimată în ecuația 3.10 va fi evaluată prin adăugarea semnului negativ. Funcția de cost devine astfel negativul probabilității $l(\theta)$, denumită și *negative log-likelihood*.

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

$$\begin{aligned} J(\theta) &= -\log L(\theta) \\ &= -l(\theta) \\ &= -\frac{1}{m} \sum_{i=1}^{m} [y^{(i)} \log(h_\theta(x^{(i)})) + (1 - y^{(i)}) \log(1 - h_\theta(x^{(i)}))]. \end{aligned} \tag{3.29}$$

Forma vectorială a ecuaţiei 3.29 este:

$$h = g(\theta^T x), \tag{3.30}$$

$$J(\theta) = -\frac{1}{m} \cdot \left(y^T \log(h) + (1-y)^T \log(1-h) \right). \tag{3.31}$$

3.5 Algoritmul de minimizare al gradientului

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

Forma generală a algoritmului este:

$$\begin{aligned} &\text{Repetă } \{ \\ &\qquad \theta_j := \theta_j - \alpha \frac{\partial}{\partial \theta_j} J(\theta) \\ &\} \end{aligned}$$

Derivata funcției sigmoid are forma:

$$\begin{aligned} g'(z) &= \frac{d}{dz} \frac{1}{1+e^{-z}} \\ &= \frac{1}{(1+e^{-z})^2} \left(e^{-z} \right) \\ &= \frac{1}{(1+e^{-z})} \left(1 - \frac{1}{(1+e^{-z})} \right) \\ &= g(z)(1-g(z)). \end{aligned} \tag{3.32}$$

Folosind formula $g'(z) = g(z)(1-g(z))$, vom calcula derivata funcției de cost pentru un singur exemplu de antrenare:

$$\begin{aligned} \frac{\partial}{\partial \theta_j} J(\theta) &= -\frac{\partial}{\partial \theta_j} l(\theta) \\ &= -\left(y \frac{1}{g(\theta^T x)} - (1-y) \frac{1}{1-g(\theta^T x)} \right) \frac{\partial}{\partial \theta_j} g(\theta^T x) \\ &= -\left(y \frac{1}{g(\theta^T x)} - (1-y) \frac{1}{1-g(\theta^T x)} \right) g(\theta^T x)(1-g(\theta^T x)) \frac{\partial}{\partial \theta_j} \theta^T x \\ &= -(y(1-g(\theta^T x)) - (1-y) g(\theta^T x)) x_j \\ &= -(y - h_\theta(x)) x_j \end{aligned} \tag{3.33}$$

Se poate observa că derivata parţială a lui $J(\theta)$ pentru regresia logistică are aceeași formă ca și în cazul regresiei liniare, putând fi astfel aplicati același algoritm de minimizare

a gradientului:

Repetă până la convergenţă: {

$$\theta_j := \theta_j - \alpha \frac{1}{m} \sum_{i=1}^{m} (h_\theta(x^{(i)}) - y^{(i)}) x_j^{(i)}$$

}

Varianta vectorizată a algoritmului este:

$$\theta := \theta - \alpha \frac{1}{m} x^T (g(x\theta) - y). \tag{3.34}$$

Bibliografie

[1] C. M. Bishop, *Pattern Recognition and Machine Learning (Information Science and Statistics)*. Secaucus, NJ, USA: Springer-Verlag New York, Inc., 2006.

[2] I. Goodfellow, Y. Bengio and A. Courville, *Deep Learning*. MIT Press, 2016, http://www.deeplearningbook.org.

[3] M. Lutz, *Learning Python*, 2nd Ed. Sebastopol, CA, USA: O'Reilly & Associates, Inc., 2003.

[4] A. Ng, "Stanford cs229 - machine learning," 2008.

[5] S. J. Russell and P. Norvig, *Artificial Intelligence: A Modern Approach*, 2nd Ed. Pearson Education, 2003.

[6] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. u. Kaiser and I. Polosukhin, "Attention is all you need," *Advances in Neural Information Processing Systems*, Vol. 30. Curran Associates, Inc., 2017.