Perceptron Logístico e Regra Delta Generalizada

Prof. Dr. Guilherme de Alencar Barreto

Julho/2021

Departmento de Engenharia de Teleinformática
Programa de Pós-Graduação em Engenharia de TEleinformática (PPGETI)
Universidade Federal do Ceará (UFC), Fortaleza-CE

gbarreto@ufc.br

1 Introdução

Esta seção discute o efeito de se utilizar uma não-linearidade suave na saída do modelo de neurônio da rede perceptron simples, em vez de uma não-linearidade abrupta, tal como a função sinal. Esta mudança implicará em alterações na regra de aprendizagem que passará a ser chamada de regra delta generalizada.

Esta pequena mudança promove um grande impacto na aplicabilidade da rede perceptron simples, uma vez que permite sua aplicação também em problemas de regressão. Recorde que antes a rede perceptron simples era usada apenas em problemas de classificação de padrões. Isto é possível porque a regra de aprendizado será baseado no gradiente do erro quadrático instantâneo, como no modelo adaline. Assim, o modelo de neurônio logístico a ser discutido nesta nota de aula, unifica os modelos adaline e perceptron simples, que possuem origens distintas [1, 2].

O perceptron simples com função de ativação sigmoidal (i.e. em forma de S) passa a ser chamado doravante de *perceptron simples Logístico* ou simplesmente de perceptron logístico. Aqui iremos discutir duas funções sigmoidais muito utilizadas nos modelos atuais de redes neurais artificiais, a saber, a função sigmóide logística e a função tangente hiperbólica. Porém, pode-se usar qualquer função não-linear suave, que seja diferenciável até segunda ordem, pelo menos.

Manteremos a mesma notação para o vetor de entrada ($\mathbf{x}$), ativação do i-ésimo neurônio (u_i), saída do i-ésimo neurônio (y_i), peso sináptico conectando a j-ésima entrada ao i-ésimo neurônio ($\mathbf{w}_i$), vetor de pesos do i-ésimo neurônio ($\mathbf{w}_i$) e número de neurônios (q).

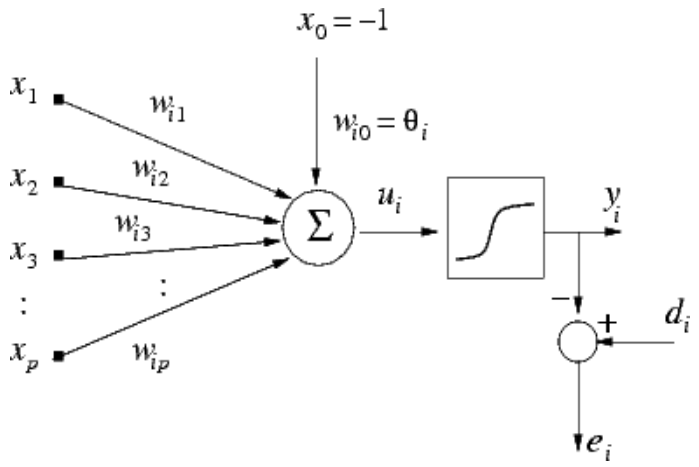


Figura 1: Arquitetura do modelo perceptron logístico.

2 Regra Delta Generalizada

A ativação do i-ésimo neurônio do perceptron logístico é calculada como segue

$$\begin{aligned} u_i(t) &= \sum_{j=1}^{p} w_{ij}(t)x_j(t) - \theta \\ &= \sum_{j=1}^{p} w_{ij}(t)x_j(t) + w_{i0}(t)x_0(t) \\ &= \sum_{j=0}^{p} w_{ij}(t)x_j(t) \\ &= \mathbf{w}_i^T(t)\mathbf{x}(t) = \mathbf{x}^T(t)\mathbf{w}_i(t) \end{aligned} \tag{1}$$

em que foi feito $x_0 = -1$ e $w_{i0} = \theta_i$. O superscrito T indica a operação de transposição dos vetores.

Como agora existe uma não-linearidade suave na saída do neurônio, conforme ilustrado na Figura 1, então a sua saída passa a ser representada como

$$y_i(t) = \phi(u_i(t)) = \phi\left(\mathbf{w}_i^T(t)\mathbf{x}(t)\right) = \phi_i(t), \tag{2}$$

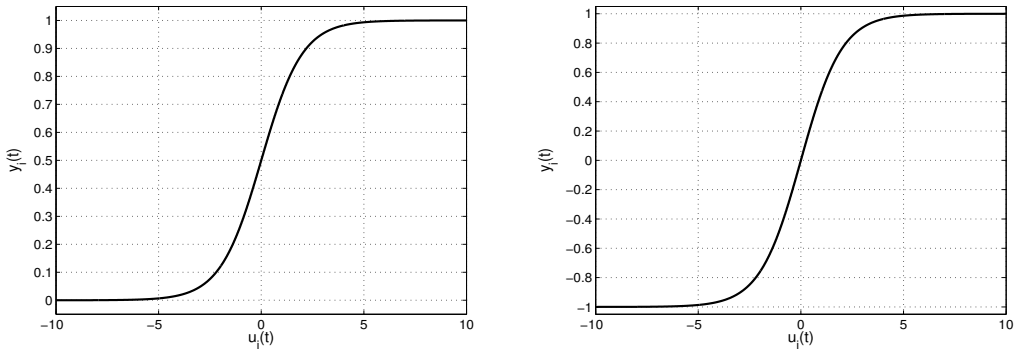
em que $\phi(\cdot)$ é uma não-linearidade do tipo sigmoidal, possuindo a forma da letra S esticada. Dois tipos são comumente usados, a sigmóide logística

$$y_i(t) = \phi_i(t) = \frac{1}{1 + \exp\{-u_i(t)\}}, \tag{3}$$

e a tangente hiperbólica

$$y_i(t) = \phi_i(t) = \frac{1 - \exp\{-u_i(t)\}}{1 + \exp\{-u_i(t)\}}. \tag{4}$$

O domínio tanto da função sigmóide logística, quanto da tangente hiperbólica é a reta dos números reais. Contudo, a imagem da função sigmóide logística está restrita ao intervalo $(0, 1)$, enquanto a da tangente hiperbólica está restrita ao intervalo $(-1, +1)$. De um extremo a outro a função é sempre crescente. O gráfico dessas duas funções estão mostrados na Figura 2.



(a) (b)

Figura 2: (a) Função sigmóide logística e (b) tangente hiperbólica.

A acurácia do i-ésimo neurônio do perceptron logístico no instante t, como aproximador do mapeamento $\mathbf{F}(\cdot)$, é medida com base no erro quadrático instantâneo definido como

$$J_i(t) = \frac{1}{2}e_i^2(t), \tag{5}$$

$$= \frac{1}{2}\left(d_i(t) - y_i(t)\right)^2, \tag{6}$$

$$= \frac{1}{2}\left(d_i(t) - \phi(u_i(t))\right)^2, \tag{7}$$

$$= \frac{1}{2}\left(d_i(t) - \phi(\mathbf{w}_i^T(t)\mathbf{x}(t))\right)^2, \tag{8}$$

em que $e_i(t) = d_i(t) - y_i(t)$ é o erro do i-ésimo neurônio em resposta à apresentação do par entrada-saída atual $(\mathbf{x}(t), \mathbf{d}(t))$.

Assim, define-se a seguir uma medida global do desempenho do perceptron logístico, chamada de **erro quadrático médio** (EQM), com base nos erros produzidos para todos os pares entrada-saída $\{\mathbf{x}(t), \mathbf{d}(t)\}$:

$$J[\mathbf{W}] = \frac{1}{2N}\sum_{t=1}^{N}\sum_{i=1}^{q} J_i(t) = \frac{1}{2N}\sum_{t=1}^{N}\sum_{i=1}^{q} e_i^2(t) = \frac{1}{2N}\sum_{i=1}^{q}\sum_{t=1}^{N}\left[d_i(t) - y_i(t)\right]^2 \tag{9}$$

em que $\mathbf{W}$ denota o conjunto de todos os parâmetros ajustáveis do modelo (pesos e limiares).

Os parâmetros do perceptron logístico devem ser especificados a fim de que os neurônios produzam sempre uma saída $y_i(t)$ bem próxima da saída esperada $d_i(t)$ para um vetor de entrada $\mathbf{x}(t)$ qualquer. Em outras palavras, o funcional $J_i(t)$ deve ter o menor valor possível para aquele conjunto de dados específico. Dá-se o nome de parâmetros ótimos aos valores de $\mathbf{w}_i$ que alcançam este objetivo.

Um procedimento iterativo de se chegar aos parâmetros ótimos envolve o uso do método de otimização baseada no gradiente descendente:

$$\mathbf{w}_i(t+1) = \mathbf{w}_i(t) - \alpha\frac{\partial J_i(t)}{\partial \mathbf{w}_i(t)} \tag{10}$$

tal que a variável $0 < \alpha < 1$ é a taxa de aprendizagem. Utilizando a regra da cadeia, a derivada presente na Eq. (10) pode ser escrita da seguinte forma:

$$\frac{\partial J_i(t)}{\partial \mathbf{w}_i(t)} = \frac{dJ_i(t)}{de_i(t)}\frac{de_i(t)}{dy_i(t)}\frac{dy_i(t)}{du_i(t)}\frac{\partial u_i(t)}{\partial \mathbf{w}_i(t)} \tag{11}$$

em que os resultados das quatro derivadas para o perceptron logístico são mostrados a seguir:

$$\frac{dJ(t)}{de_i(t)} = e_i(t) \tag{12}$$

$$\frac{de_i(t)}{dy_i(t)} = -1 \tag{13}$$

$$\frac{dy_i(t)}{du_i(t)} = \frac{d\phi(u_i(t))}{du_i(t)} = \phi_i'(t) \tag{14}$$

$$\frac{\partial u_i(t)}{\partial \mathbf{w}_i(t)} = \mathbf{x}(t) \tag{15}$$

Assim, temos que a regra recursiva de ajuste dos pesos, $w_j(t)$, é dada por

$$\begin{aligned} \mathbf{w}_i(t+1) &= \mathbf{w}_i(t) - \alpha \frac{\partial J_i(t)}{\partial \mathbf{w}_i(t)}, \\ &= \mathbf{w}_i(t) + \alpha e_i(t) \phi_i'(t) \mathbf{x}(t), \\ &= \mathbf{w}_i(t) + \alpha \delta_i(t) \mathbf{x}(t), \end{aligned} \tag{16}$$

em que $\delta_i(t)$ é chamado de gradiente local do i-ésimo neurônio. A regra de aprendizagem mostrada na Eq. (16) é conhecida como *regra delta generalizada* ou *regra LMS generalizada*, para diferenciar da regra LMS original que é também conhecida como regra delta. Em todo caso, se a função de ativação $\phi(u_i)(t)$ for a função identidade, ou seja, $\phi(u_i)(t) = u_i(t)$, teremos a regra LMS usual.

A derivada instantânea[1] da função de ativação $\phi_i(t)$, denotada por $\phi_i'(t)$, para a função logística é dada por

$$\phi_i'(t) = \frac{d\phi_i(t)}{du_i(t)} = \frac{d(1 + \exp\{-u_i(t)\})^{-1}}{du_i(t)}, \tag{17}$$

$$= (-1) \cdot (1 + \exp\{-u_i(t)\})^{-2} \cdot \frac{d \exp\{-u_i(t)\}}{du_i(t)}, \tag{18}$$

$$= (-1) \cdot (-1) \cdot (1 + \exp\{-u_i(t)\})^{-2} \cdot \exp\{-u_i(t)\}, \tag{19}$$

$$= \frac{\exp\{-u_i(t)\}}{(1 + \exp\{-u_i(t)\})^2} = \frac{1}{1 + \exp\{-u_i(t)\}} \cdot \frac{\exp\{-u_i(t)\}}{1 + \exp\{-u_i(t)\}}, \tag{20}$$

$$= \phi_i(t)[1 - \phi_i(t)] = y_i(t)[1 - y_i(t)], \tag{21}$$

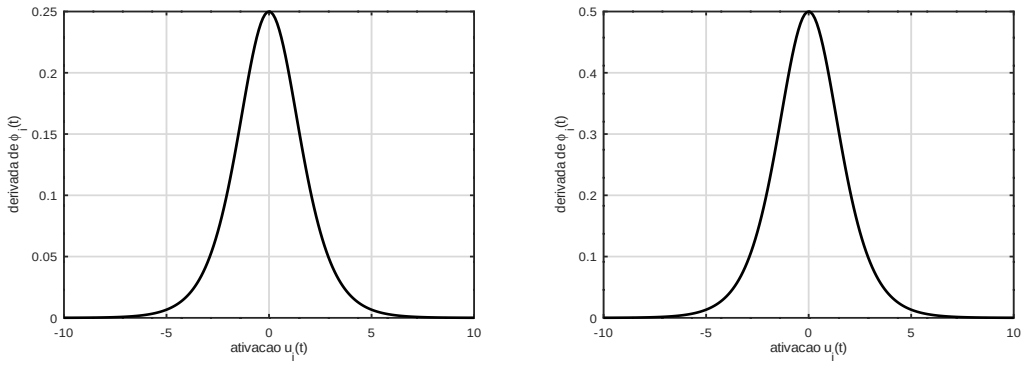
e para a função tangente hiperbólica (exercício!) temos que

$$\phi_i'(t) = \frac{1}{2}(1 - \phi_i^2(t)) = \frac{1}{2}(1 - y_i^2(t)). \tag{22}$$

Uma das vantagens em se utilizar as funções sigmóide logística e tangente hiperbólica reside no fato de suas derivadas instantâneas não precisam ser computadas via métodos numéricos, pois estão em função do valor da saída no instante t, $y_i(t)$. Os gráficos das derivadas das funções de ativação sigmóide logística e tangente hiperbólica estão mostrados na Figura 3.

Além disso, a derivada da função de ativação é usada no cálculo do gradiente local do erro do i-ésimo neurônio, $\delta_i(t)$, que pode ser entendido como elemento modulador do erro $e_i(t)$. Assim, valores altos do erro (positivos ou negativos) gerarão valores pequenos da derivada $\phi_i'(t)$, o que

[1] Derivada no instante atual t.



(a) (b)

Figura 3: (a) Função sigmóide logística e (b) tangente hiperbólica.

confere certa estabilidade ao aprendizado do modelo perceptron logístico ao longo do tempo. Tal estabilidade é verificada por uma curva de aprendizagem (convergência do EQM) mais suave que a da regra LMS convencional. Porém, essa modulação do erro pode também causar o problema conhecido como *paralisia da rede*, em que os pesos não são modificados mesmo para erros grandes, principalmente no começo do processo de aprendizado. Para dirimir a ocorrência de paralisia, recomenda-se iniciar os pesos com valores bem pequenos.

O perceptron logístico pode ser utilizado tanto para aproximação de funções, quanto para classificação de padrões. Em ambos os casos, o treinamento, a convergência e a generalização seguem os mesmos procedimentos que os descritos para o modelo adaline.

3 Exercícios Computacionais

Exercício 1 - Qual é a implicação do aparecimento do fator $\phi_i'(t)$ na regra delta generalizada? Dica: Pesquise por *paralisia da rede* em sítios da internet.

Exercício 2 - Escolher um conjunto de dados no Repositório da Universidade da Califórnia em Irvine (`http://www.ics.uci.edu/~mlearn/MLSummary.html`) e utilizar os pares de vetores entrada-saída correspondentes para avaliar o perceptron simples Logístico em um problema de classificação de padrões e em um problema de regressão. Determinar a curva de aprendizagem (EQM em função da época de treinamento) e os índices de desempenhos correspondentes a cada tarefa. Classificação: taxas de acerto médio, desvio padrão, mediana, mínimo e máximo. Regressão: R2 médio, desvio padrão, mediana, mínimo e máximo.

Referências

[1] B. Widrow and M. A. Lehr. Perceptrons, adalines, and backpropagation. In M. A. Arbib, editor, *The Handbook of Brain Theory and Neural Networks*, pages 871–877. MIT Press, 2nd edition, 2002.

[2] B. Widrow and M. A. Lehr. 30 years of adaptive neural networks: Perceptron, madaline and backpropagation. *Proceedings of the IEEE*, 78(9):1415–1442, 1990.