Proceeding Series of the Brazilian Society of Computational and Applied Mathematics, v. 10, n. 1, 2023.

Trabalho apresentado no XLII CNMAC, Universidade Federal de Mato Grosso do Sul - Bonito - MS, 2023

Proceeding Series of the Brazilian Society of Computational and Applied Mathematics

Regularização de uma Máquina de Aprendizado Extremo Usando a Norma $\ell_{r,p}$

Angélica M. Narváez Vivas[1], IMECC, Campinas, SP
João Florindo [2], IMECC, Campinas, SP
Thadeu Senne[3], UNIFESP, São José dos Campos, SP

O processo de aprendizado das redes neurais pode ser altamente custoso computacionalmente em certas situações, já que todos os parâmetros dessas redes são ajustados iterativamente. Uma solução proposta para resolver esta questão são as máquinas de aprendizado extremo (sigla ELM, do inglês). Estas são redes neurais do tipo *feedforward* de uma única camada escondida, em que os pesos entre a entrada e a camada escondida são aleatórios, enquanto os pesos entre a camada escondida e a saída são determinados analiticamente usando mínimos quadrados. Em teoria, esse algoritmo tende a fornecer um bom desempenho de generalização em uma velocidade de aprendizado extremamente rápida.

Considere um problema de aprendizado supervisado com N amostras de treino $\left\{(\mathbf{x_i}, \mathbf{t_i}) \,|\mathbf{x_i} \in \mathbb{R}^d, \mathbf{t_i} \in \mathbb{R}^m\right\}_{i=1}^N$ e m classes. Onde $\mathbf{x_i} = [x_{i1}, x_{i2}, ..., x_{id}]^T$ é a $i-$ésima amostra de treino, d é a dimensão do dado e $\mathbf{t_i} = [t_{i1}, t_{i2}, ..., t_{im}]^T$ é o alvo da i-ésima amostra. A matriz de saída da camada oculta da ELM com L nós pode ser expressa como

$$\mathbf{H} = \begin{bmatrix} \mathbf{h}(\mathbf{x}_1) \\ \vdots \\ \mathbf{h}(\mathbf{x}_N) \end{bmatrix} = \begin{bmatrix} g\left(\mathbf{w}_1^T\mathbf{x}_1 + \mathbf{b}_1\right) \cdots g\left(\mathbf{w}_L^T\mathbf{x}_1 + \mathbf{b}_L\right) \\ \vdots \\ g\left(\mathbf{w}_1^T\mathbf{x}_N + \mathbf{b}_1\right) \cdots g\left(\mathbf{w}_L^T\mathbf{x}_N + \mathbf{b}_L\right) \end{bmatrix} \tag{1}$$

em que $(\mathbf{w}_i, b_i)$ são os pesos e *biases* entre a camada de entrada e os nós da camada oculta, que compõem a matriz $\mathbf{W} = (\mathbf{w}_1, \mathbf{w}_2, \cdots, \mathbf{w}_L)$. Estes parâmetros podem ser criados aleatoriamente de acordo com qualquer distribuição de probabilidade contínua. $g(\cdot)$ é a função de ativação Sigmoide. A ELM Regularizada tenta aproximar as N amostras com com o menor erro possível, resolvendo $\mathbf{H}\beta = \mathbf{T}$, em que $\beta = [\beta_1, \beta_2, \cdots, \beta_L]^T \in \mathbb{R}^{L\times M}$ é a matriz de pesos de saída e $\mathbf{T} \in \mathbb{R}^{n\times m}$ é a matriz dos alvos das amostras [1].

O objetivo da ELM regularizada é minimizar simultaneamente o erro de treinamento e a norma dos pesos de saída para melhorar a estabilidade e a capacidade de generalização. Usualmente é empregada a norma Euclideana ou a norma l_1 [4]. Visando diminuir o problema de robustez ocasionados por *outliers* nos dados, este estudo propõe uma ELM regularizada, robusta e esparsa e que explora o uso da minimização da norma $\ell_{r,p}$ na parcela de regularização e a minimização da norma $\ell_{2,1}$ na parcela da função de perda. O problema irrestrito com a função objetivo $F(\beta) = f(\beta) + ch(\beta)$ é o seguinte:

$$\min_{\beta\in\mathbb{R}^{L\times m}} \|\beta\|_{r,p} + c\,\|\mathbf{H}\beta - \mathbf{T}\|_{2,1}\,, \tag{2}$$

[1] anarvis93@gmail.com
[2] florindo@unicamp.br
[3] senne@unifesp.br

em que c é o parâmetro de penalidade. A norma $\ell_{2,1}$ no caso leva a um processo de equilíbrio interessante, já que a norma ℓ_2 interna promove uma solução densa enquanto que a norma ℓ_1 externa promove esparsidade. Deste modo, a regularização baseada na norma $\ell_{2,1}$ tende reduzir os pesos do modelo como qualquer regularização, porém além disso levando os pesos menos significantes a zero. Isso é interesante em uma rede neural artificial porque zerar pesos equivale a remover neurônios e isso reduz a complexidade intrínseca do modelo, fazendo com que a ELM resultate seja mais compacta e ainda menos propensa a *overfitting*, ou seja, generalize melhor para dados não vistos no treinamento. [2]

Uma vez que (2) é um problema irrestrito, a função objetivo $F(\beta)$ é uma função convexa e o conjunto viável $\mathbb{R}^{L\times m}$ é um conjunto convexo, logo, temos a existência dos minimizadores globais e as condições necessárias de otimalidade de primeira ordem também são suficientes. Do mesmo modo, podemos encontrar uma direção de máxima descida para $F(\beta)$ em $\beta' \in \mathbb{R}^n$ empregando o método do gradiente descendente [3]. Os testes computacionais foram feitos em *Python* empregando o pacote *autograd* para diferenciação automática, com taxa de aprendizagem fixa de 10^{-7}. A Figura 1 mostra o gráfico da função custo e a norma do erro com o valor do parâmetro de penalidade $c = 10$, o número das camadas ocultas $L = 10$. A curva decai nos dois casos, como o esperado, confirmando a correto funcionamento da regularização proposta. A base de dados utilizada é a *transfusion.data*.

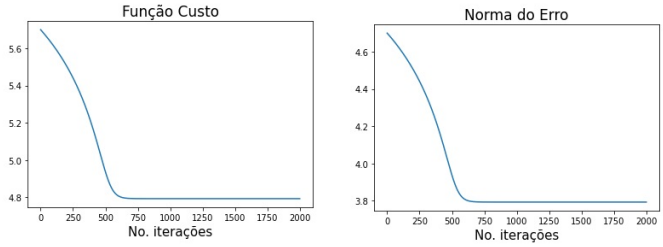

Figura 1: função de custo $F(\beta)$ e norma do erro *versus* iterações.

Os resultados obtidos sugerem que seja possível otimizar a função de custo ELM regularizada na norma $\ell_{r,p}$, com $r, p = 1, 2, 3$, pois a norma do erro tem um decaimento ótimo para garantir a generalização do modelo. Como passo futuro, pretende-se continuar empregando o método do gradiente acrescentando-se a ténica de busca linear para obter o melhor tamanho do passo e conseguir que sua convergência seja mais rápida do que quando o tamanho do passo é fixo.

Referências

[1] Guang-Bin Huang, Qin-Yu Zhu e Chee-Kheong Siew. "Extreme learning machine: theory and applications". Em: **Neurocomputing** 70.1-3 (2006), pp. 489–501.

[2] Rui Li et al. "$\ell_{2,1}$ Norm Based Loss Function and Regularization Extreme Learning Machine". Em: **IEEE Access** 7 (2018), pp. 6575–6586.

[3] José Mario Martınez e Sandra Augusta Santos. "Métodos computacionais de otimização". Em: **Colóquio Brasileiro de Matemática, Apostilas** 20 (1995).

[4] Feiping Nie et al. "Efficient and robust feature selection via joint $\ell_{2,1}$ norms minimization". Em: **Advances in neural information processing systems** 23 (2010).