Máquinas de Aprendizado Extremo
(*Extreme Learning Machines* – ELMs)

SUMÁRIO

1. Introdução e Motivação

- Todas as propostas de redes neurais não-recorrentes (*feedfoward*) apresentadas no curso, como o perceptron de múltiplas camadas (MLP) e a rede neural com funções de ativação de base radial (RBF), produzem a sua saída (podendo ser múltiplas saídas) como uma combinação linear das ativações dos neurônios da camada anterior.
- Tomando uma única camada intermediária, pode-se afirmar, portanto, que redes neurais MLP e RBF sintetizam mapeamentos multidimensionais de entrada-saída por meio de uma composição aditiva de funções-base, na forma:

$$\hat{s}_{kl} = \sum_{j=1}^{n} w_{kj} f\left(\mathbf{v}_j, b_j, \mathbf{x}_l\right) + w_{k0}$$

onde

- $\hat{s}_{kl}$ é a k-ésima saída da rede neural para o l-ésimo padrão de entrada $\mathbf{x}_l$;
- $f\left(\mathbf{v}_j, b_j, \bullet\right)$ é a j-ésima função da base de funções-base.

- No caso da rede neural MLP, as funções-base são funções de expansão ortogonal (*ridge functions*), enquanto que, no caso da rede neural RBF, as funções-base têm um comportamento radial em relação a um centro de ativação máxima.
- Nos dois casos, como em outros casos de composição aditiva de funções-base, há demonstração teórica da capacidade de aproximação universal. A capacidade de aproximação universal é uma **propriedade existencial**. Ela afirma que existe um número *n* finito de neurônios e uma certa configuração de pesos sinápticos que permitem obter um erro de aproximação arbitrariamente baixo para os dados de treinamento, supondo que se considera uma região compacta do espaço de entrada e que o mapeamento original, que é amostrado para produzir os dados de treinamento, é contínuo.
- É intuitivo concluir, também, que quanto maior o número *n* de neurônios na camada intermediária, maior é a flexibilidade do modelo matemático resultante, ou seja, maiores são as "possibilidades de contorção" do mapeamento a ser

sintetizado. Na Parte 3 deste tópico, já definimos espaço de hipóteses e associamos a este espaço o número *n* de neurônios da camada intermediária.

- Por outro lado, é sabido também que há o risco de sobre-ajuste aos dados, produzindo modelos que generalizam mal frente a novos dados de entrada-saída.
- A máxima capacidade de generalização está associada a modelos otimamente regularizados, ou seja, que se contorcem na medida certa, de acordo com as demandas de cada aplicação.
- Com isso, uma definição adequada do número de neurônios e dos pesos sinápticos é fundamental para garantir uma boa capacidade de generalização.
- Um resultado fundamental da literatura, restrito a problemas de classificação de padrões, foi apresentado por BARTLETT (1997; 1998). Nesses trabalhos, como o próprio título indica, conclui-se que controlar a norma dos pesos sinápticos é mais relevante para a capacidade de generalização do que controlar o tamanho da rede neural, ou seja, o número *n* de neurônios na camada intermediária.

- De fato, pode-se introduzir o conceito de ⟨número efetivo de neurônios na camada intermediária⟩, o qual é determinado pela configuração dos pesos da camada de saída da rede neural.
- As máquinas de aprendizado extremo exploram este resultado "de forma extrema", ou seja, jogam toda a responsabilidade por garantir uma boa capacidade de generalização aos pesos da camada de saída, permitindo que os pesos da camada intermediária, responsáveis por definir as funções-base, sejam definidos de modo aleatório.
- Por serem definidos de modo aleatório, portanto desvinculados das demandas da aplicação, deve-se considerar um valor elevado para n, podendo inclusive ultrapassar o valor de N, que representa o número de amostras para treinamento.
- Por mais que pareça estranho trabalhar com valores de n elevados e até maiores que N, as máquinas de aprendizado extremo se sustentam em três argumentos muito poderosos:

- ✓ O problema de treinamento passa a ser linear nos parâmetros ajustáveis, o que representa uma enorme economia de recursos computacionais para se realizar o treinamento supervisionado;
- ✓ A capacidade de generalização pode ser maximizada controlando-se a norma dos pesos na camada de saída, não dependendo de forma significativa do número *n* de neurônios na camada intermediária;
- ✓ Há recursos computacionais disponíveis para implementar redes neurais sobredimensionadas.

- E já que as funções-base podem ser definidas aleatoriamente, então não há razão também para que elas tenham formas sigmoidais ou tenham base radial. Logo, o elenco de funções-base pode ser também arbitrário, embora as demonstrações de capacidade de aproximação universal para ELMs restrinjam ainda as alternativas de funções-base.
- Por outro lado, são incluídas funções trigonométricas e até a função sinal.

2. Exemplos de máquinas de aprendizado extremo

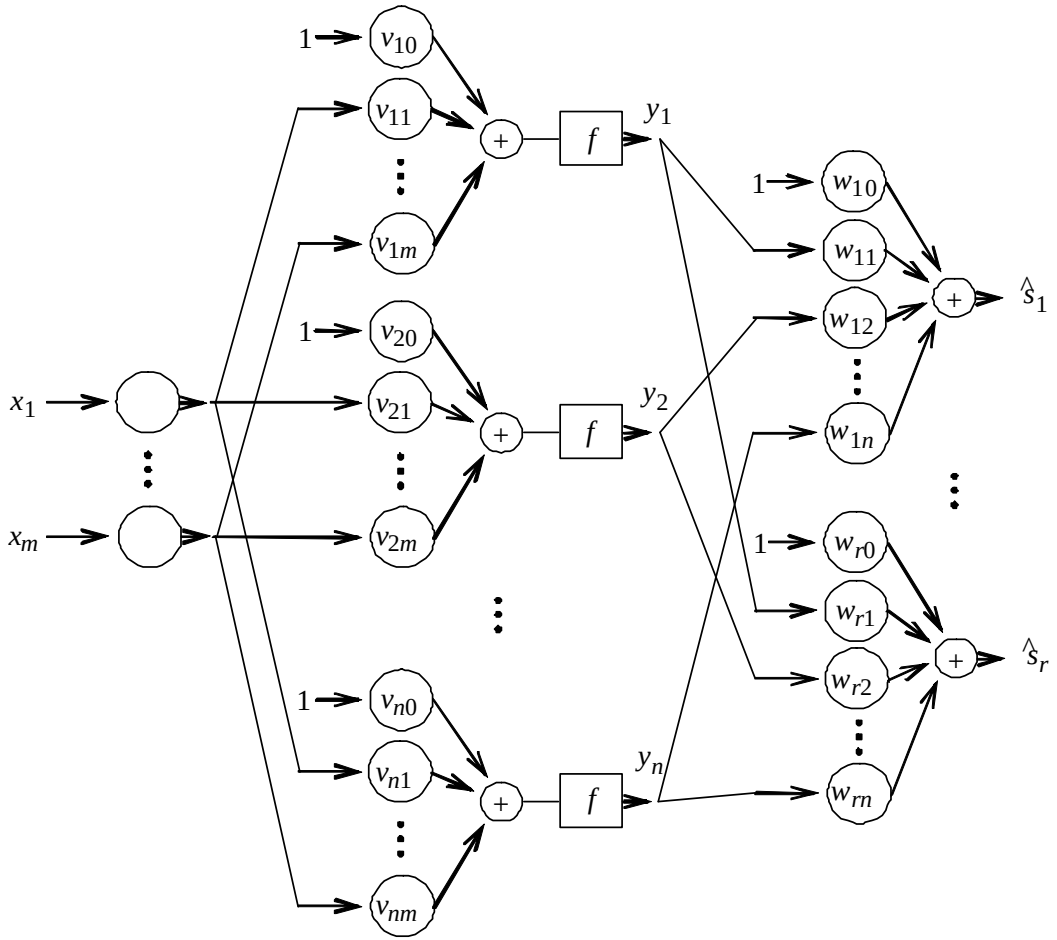

Figura 1 – Rede neural perceptron com uma camada intermediária

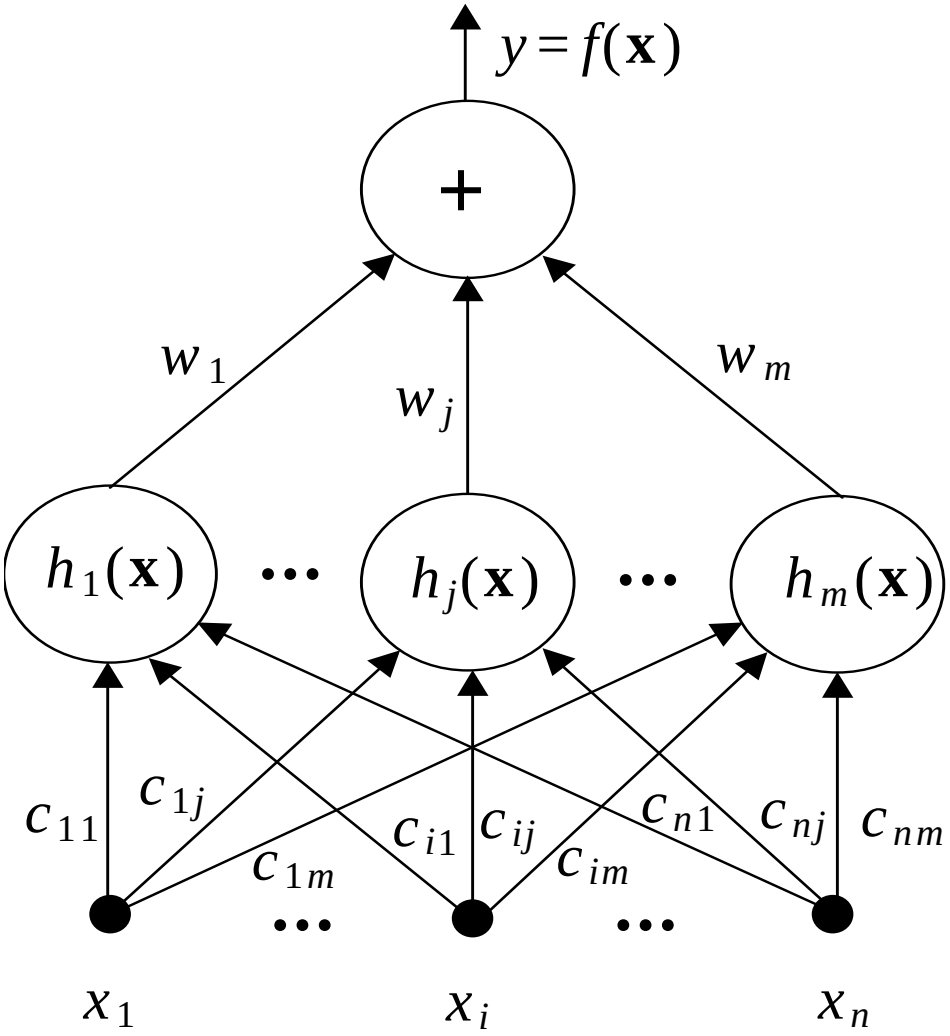

Figura 2 – Rede neural com funções de ativação de base radial (não estão indicados os pesos de polarização, associados às entradas constantes dos neurônios)

3. Treinamento das ELMs

- Treinar uma máquina de aprendizado extremo é equivalente a resolver o seguinte problema de otimização para cada uma das saídas da rede neural:

$$\mathbf{w}_k^* = \arg \min_{\mathbf{w}_k \in \Re^{n+1}} \left\| \mathbf{w}_k \right\|^2 + C_k \times J\left(\mathbf{w}_k\right)$$

onde

1. k é o índice da saída;
2. n é o número de neurônios na camada intermediária;
3. $\left\| \cdot \right\|^2$ é a norma euclidiana;
4. C_k é um coeficiente de ponderação, a ser determinado, por exemplo, por métodos de busca unidimensional;
5. $J\left(\mathbf{w}_k\right) = \frac{1}{2} \sum_{l=1}^{N} \left[\sum_{j=1}^{n} w_{kj} f\left(\mathbf{v}_j, b_j, \mathbf{x}_l\right) + w_{k0} - s_{kl} \right]^2$;
6. N é o número de amostras disponíveis para treinamento.

3.1. Como encontrar os pesos sinápticos

- Uma vez fornecido o coeficiente de ponderação C_k, para a k-ésima saída da rede neural, o vetor de pesos sinápticos é obtido como segue:

  1. Monta-se a matriz H_{inicial} de dimensão $N \times n$, com as ativações de todos os neurônios para todos os padrões de entrada, produzindo:

$$H_{\text{inicial}} = \begin{bmatrix} f(\mathbf{v}_1, b_1, \mathbf{x}_1) & f(\mathbf{v}_2, b_2, \mathbf{x}_1) & \cdots & f(\mathbf{v}_n, b_n, \mathbf{x}_1) \\ f(\mathbf{v}_1, b_1, \mathbf{x}_2) & \ddots & & \vdots \\ \vdots & & & \\ f(\mathbf{v}_1, b_1, \mathbf{x}_N) & \cdots & & f(\mathbf{v}_n, b_n, \mathbf{x}_N) \end{bmatrix}$$

  2. Acrescenta-se uma coluna de um's à matriz H_{inicial}, produzindo a matriz H:

$$H = \begin{bmatrix} f(\mathbf{v}_1, b_1, \mathbf{x}_1) & f(\mathbf{v}_2, b_2, \mathbf{x}_1) & \cdots & f(\mathbf{v}_n, b_n, \mathbf{x}_1) & 1 \\ f(\mathbf{v}_1, b_1, \mathbf{x}_2) & \ddots & & \vdots & 1 \\ \vdots & & & & \vdots \\ f(\mathbf{v}_1, b_1, \mathbf{x}_N) & \cdots & & f(\mathbf{v}_n, b_n, \mathbf{x}_N) & 1 \end{bmatrix}$$

  3. Monta-se o vetor $\mathbf{s}_k$, contendo todos os padrões de saída, na forma:

$$\mathbf{s}_k = \begin{bmatrix} s_{k1} & s_{k2} & \cdots & s_{kN} \end{bmatrix}^T$$

4. Considerando que a matriz H tenha posto completo, o vetor w_k é obtido como segue:

4.1. Se $n \leq N$, $\mathbf{w}_k = \left(\frac{I}{C_k} + H^T H \right)^{-1} H^T \mathbf{s}_k$;

4.2. Se $n > N$, $\mathbf{w}_k = H^T \left(\frac{I}{C_k} + HH^T \right)^{-1} \mathbf{s}_k$.

3.2. Como encontrar o coeficiente de ponderação

- A maximização da capacidade de generalização requer a definição de um valor adequado para o coeficiente de ponderação C_k, associado à saída k.
- Sugere-se aqui o uso de uma busca unidimensional (por exemplo, via seção áurea), empregando um conjunto de validação. O valor ótimo de C_k é aquele que minimiza o erro junto ao conjunto de validação.

4. Versão incremental para ELMs

Será apresentada no Tópico 7 do curso, com base em HUANG, CHEN & SIEW (2006).

5. Referências bibliográficas

BARTLETT, P.L. For valid generalization the size of the weights is more important than the size of the network. *Advances in Neural Information Processing Systems*, volume 9, pp. 134-140, 1997.

BARTLETT, P.L. The sample complexity of pattern classification with neural networks: the size of the weights is more important than the size of the network. *IEEE Transactions on Information Theory*, vol. 44, no. 2, pp. 525-536,1998.

HUANG, G.-B., CHEN, L., SIEW, C.-K. Universal Approximation Using Incremental Constructive Feedfoward Networks with Random Hidden Nodes. *IEEE Transactions on Neural Networks*, vol. 17, no. 4, pp. 879-892, 2006.

HUANG, G.-B., WANG, D.H., LAN, Y. Extreme learning machines: a survey. *International Journal of Machine Learning and Cybernetics*, vol. 2, pp. 107-122, 2011.

HUANG, G.-B., ZHOU, H., DING, X., ZHANG, R. Extreme Learning Machines for Regression and Multiclass Classification. *IEEE Transactions on Systems, Man, and Cybernetics – Part B: Cybernetics*, vol. 42, no. 2, pp. 513-529, 2012.

HUANG, G.-B., ZHU, Q.-Y., SIEW, C.-K. Extreme learning machine: a new learning scheme of feedforward neural networks. *Proceedings of the International Joint Conference on Neural Networks* (IJCNN'2004), vol. 2, pp. 985-990, 2004.

HUANG, G.-B., ZHU, Q.-Y., SIEW, C.-K. Extreme learning machine: theory and applications. Neurocomputing, vol. 70, pp. 489-501, 2006.