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

SUMÁRIO

1. Introdução e motivação

- Todas as propostas de redes neurais não-recorrentes (*feedfoward*) já apresentadas e a serem apresentadas no curso, como o perceptron de múltiplas camadas (MLP) e a rede neural com função 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 do conjunto 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 ou mínima.
- 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. Ainda neste Tópico 4, iremos definir espaço de hipóteses e associar a este espaço o número *n* de neurônios da camada intermediária.

- Por outro lado, é sabido também que há o risco de sobreajuste 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 (exibem grau adequado de flexibilidade), 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 determinados de modo aleatório, de acordo com uma certa distribuição de probabilidade.
- 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 contraintuitivo 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 expressiva 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 parâmetros das funções-base podem ser definidos aleatoriamente, então não há razão também para que as próprias funções-base sejam *ridge functions* de formato sigmoidal 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.

2. Exemplos de máquinas de aprendizado extremo

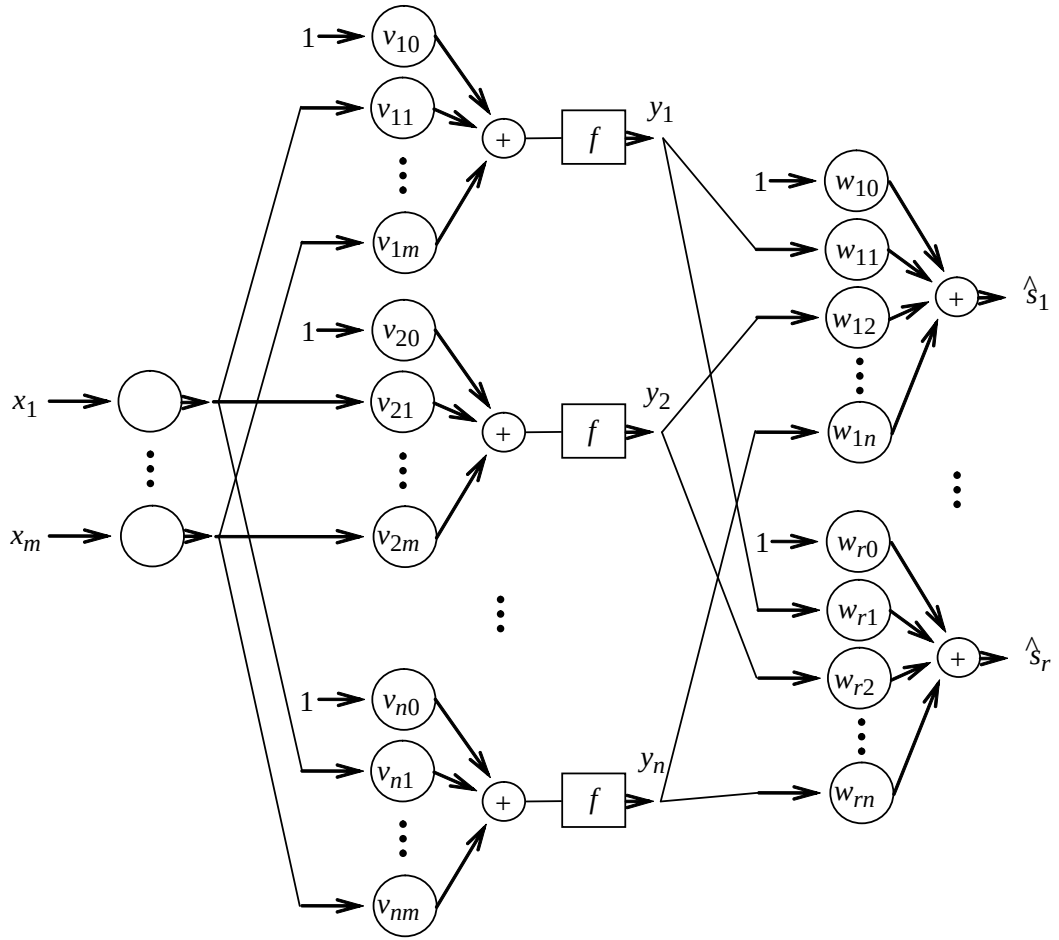


Figura 1 – Rede neural perceptron de múltiplas camadas (MLP) com uma camada intermediária

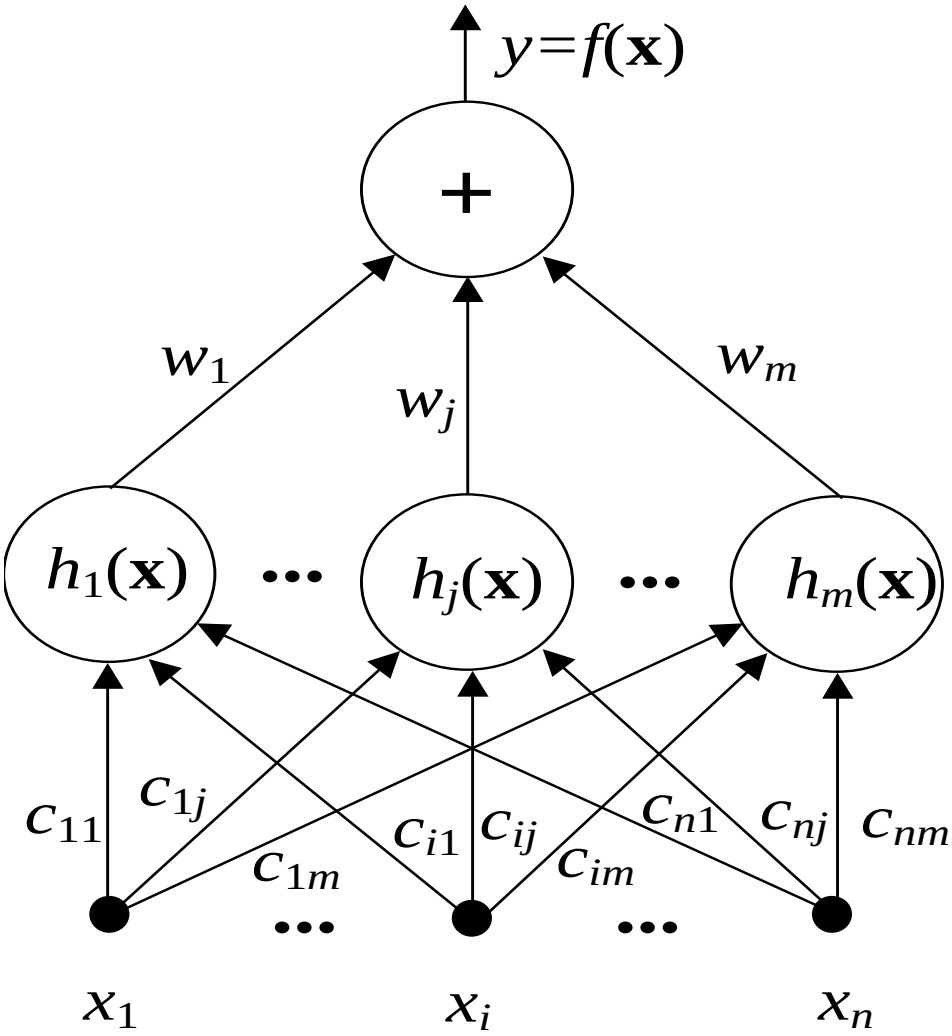
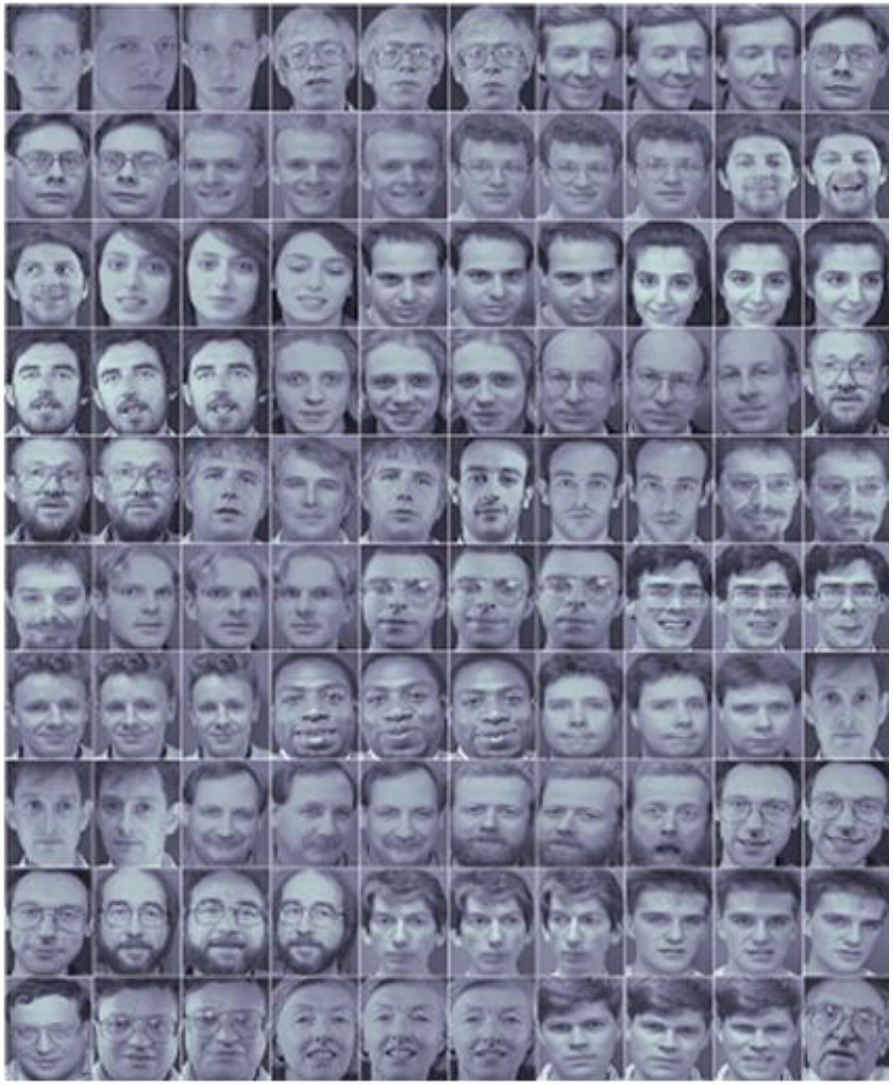


Figura 2 – Rede neural com funções de ativação de base radial (não está indicado o peso de polarização no neurônio de saída)

3. Síntese de mapeamentos multidimensionais

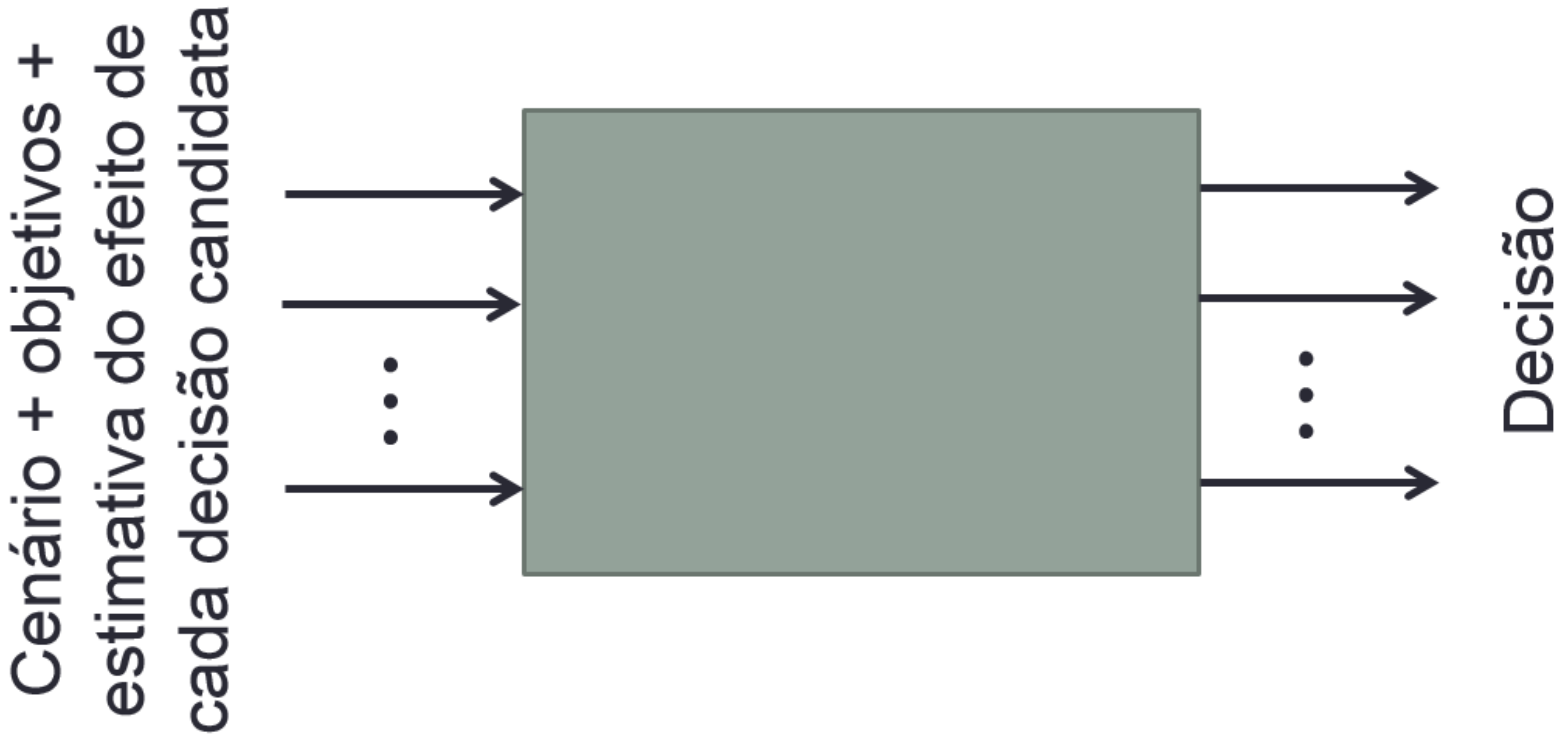
- São muitos os problemas da literatura que podem ser resolvidos a partir da síntese de mapeamentos multidimensionais. Seguem alguns exemplos.
- Reconhecimento de padrões / Classificação:




Atributos
Classes

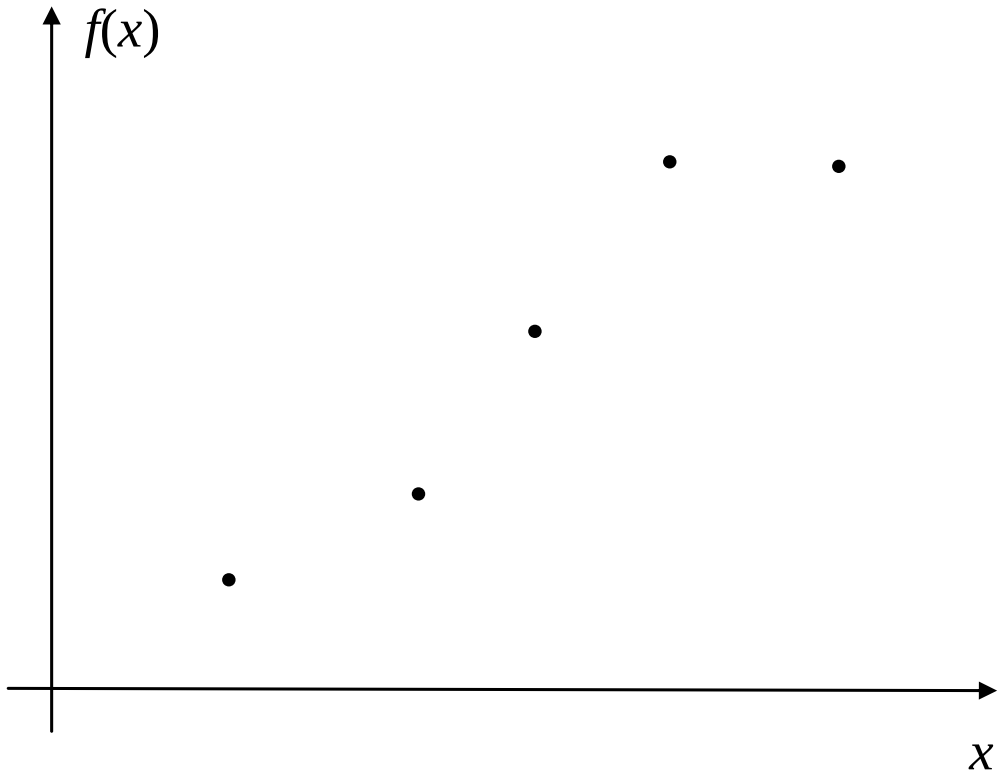
- Planejamento:

- Tomada de decisão:

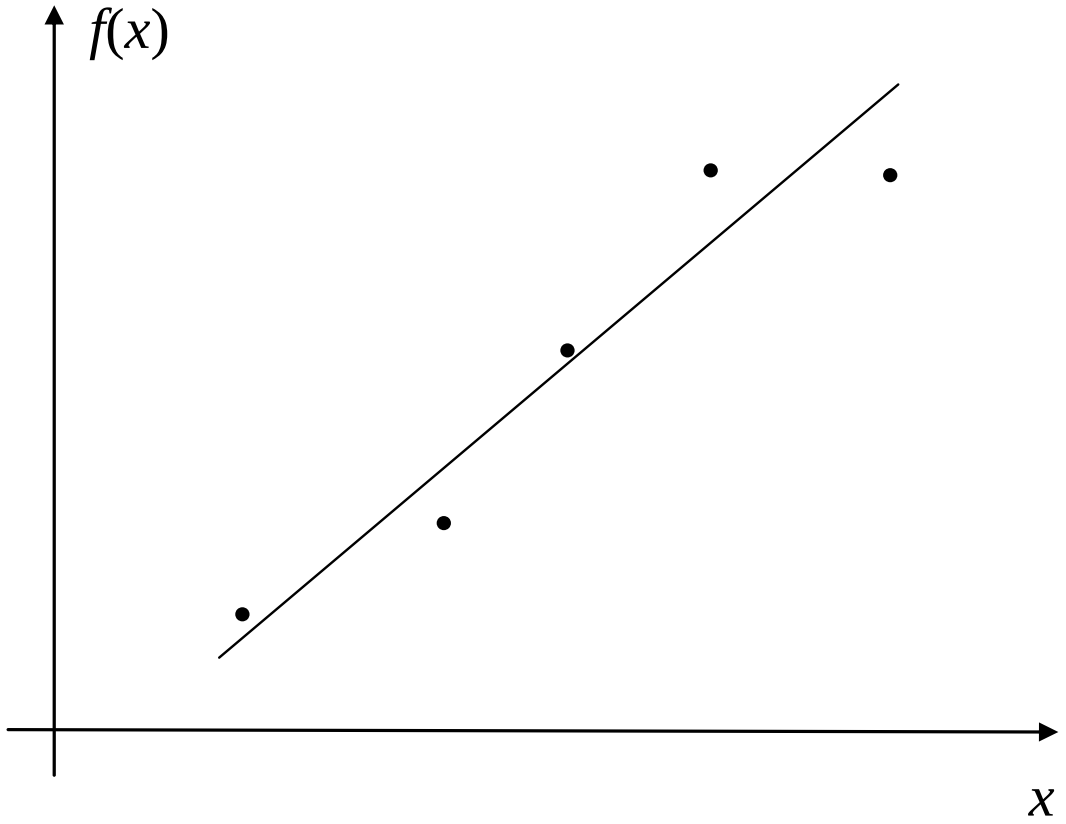


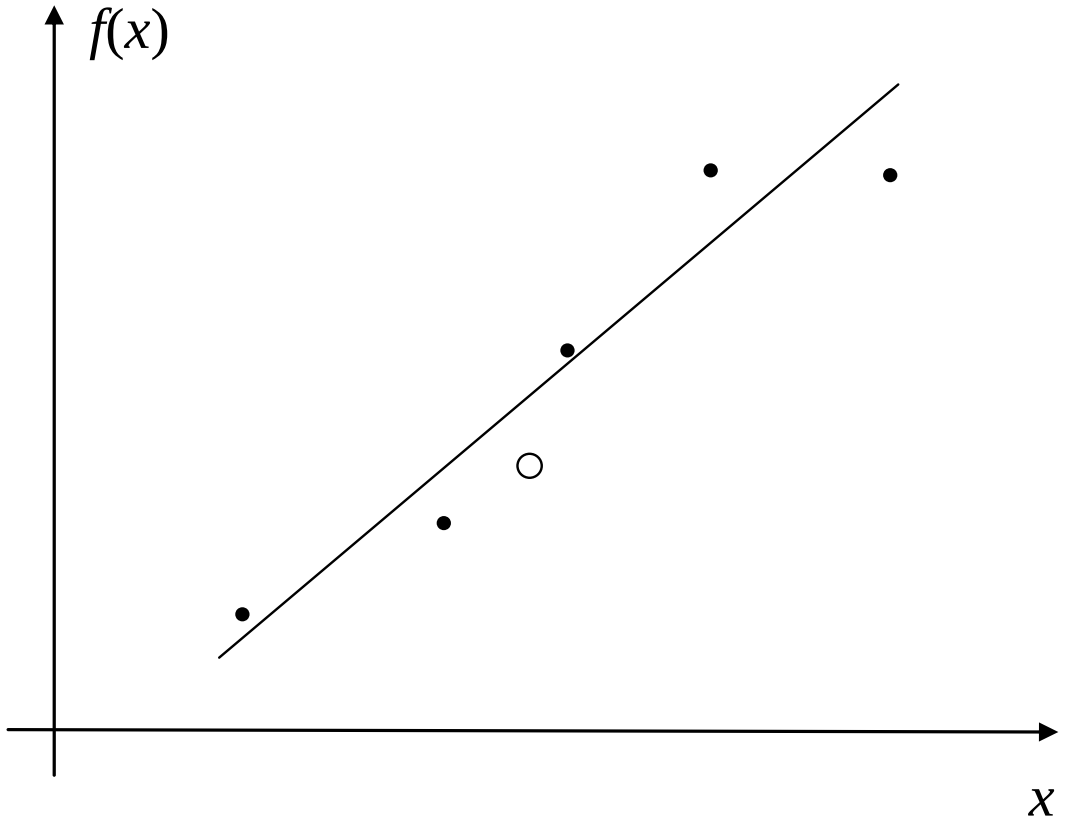
- Predição:

4. O dilema bias-variância

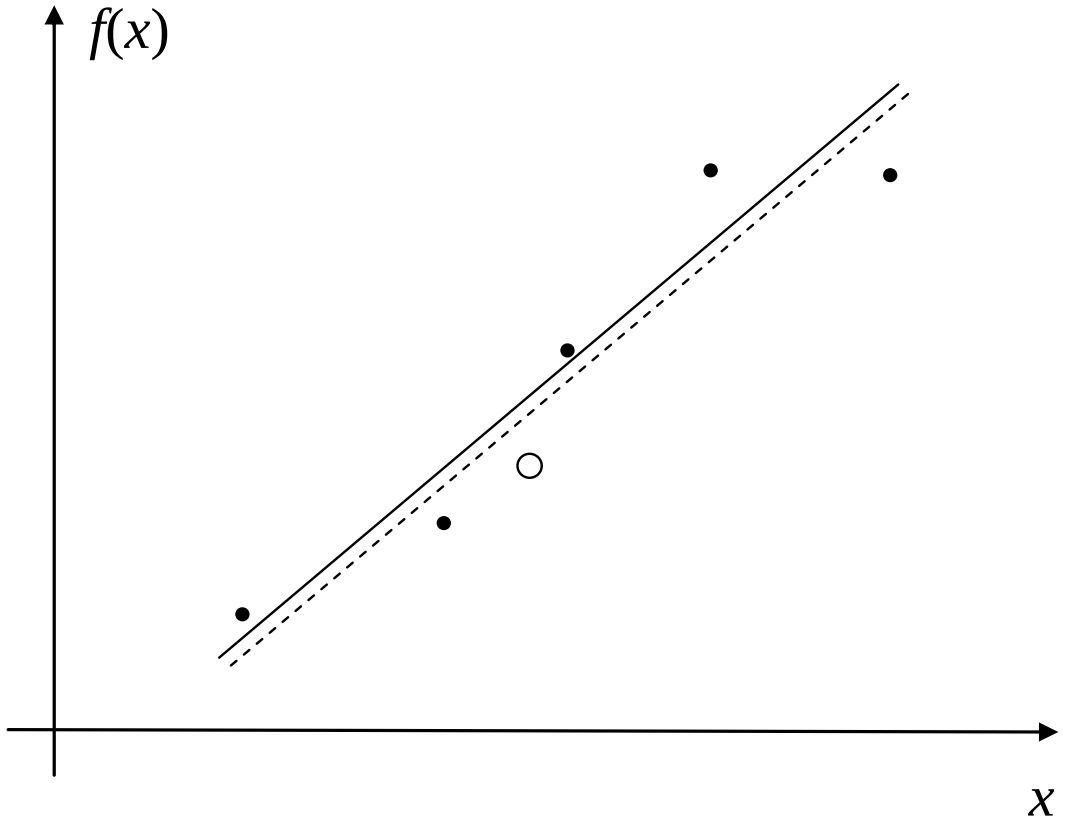


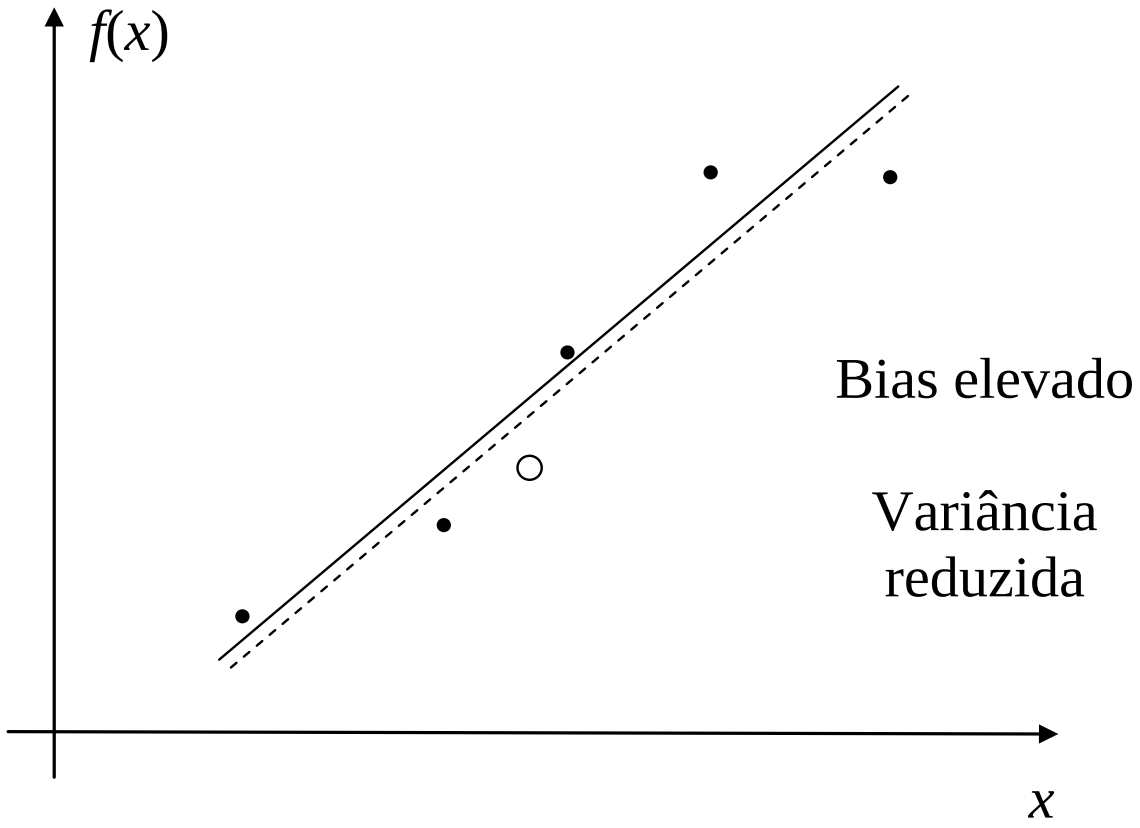
- Objetivo: Aproximar os pontos com um modelo de baixa flexibilidade.

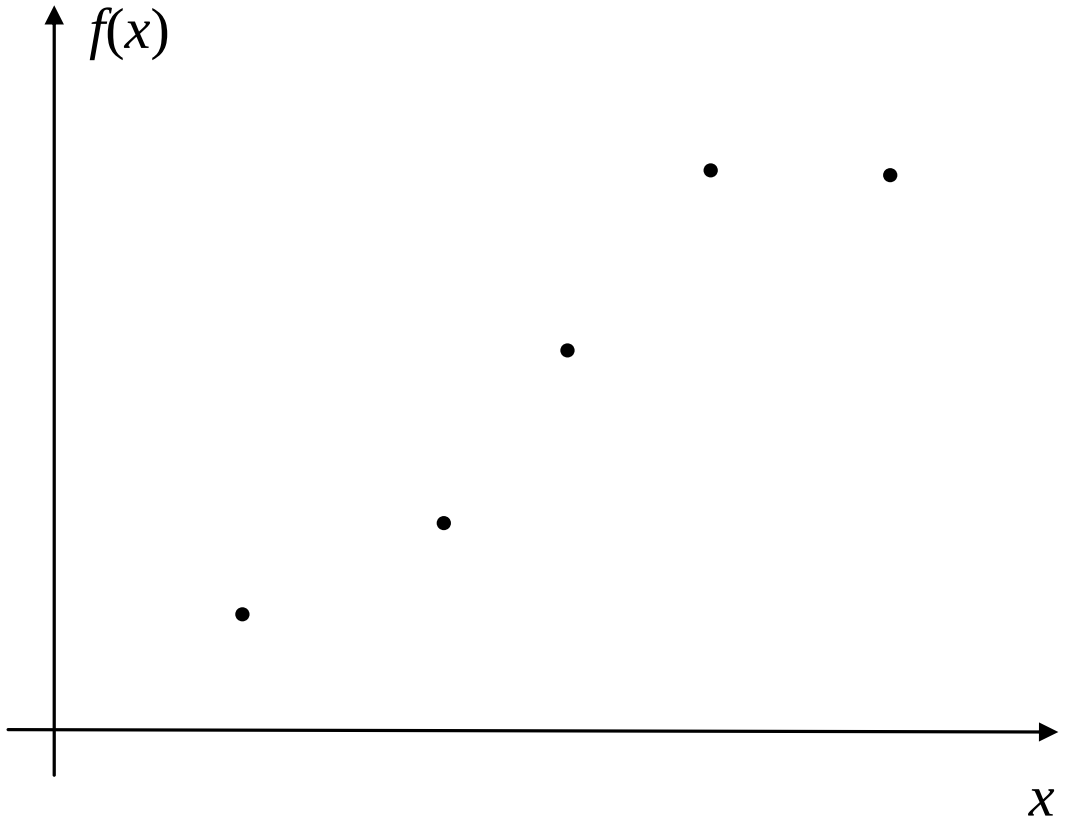

f(x)
x



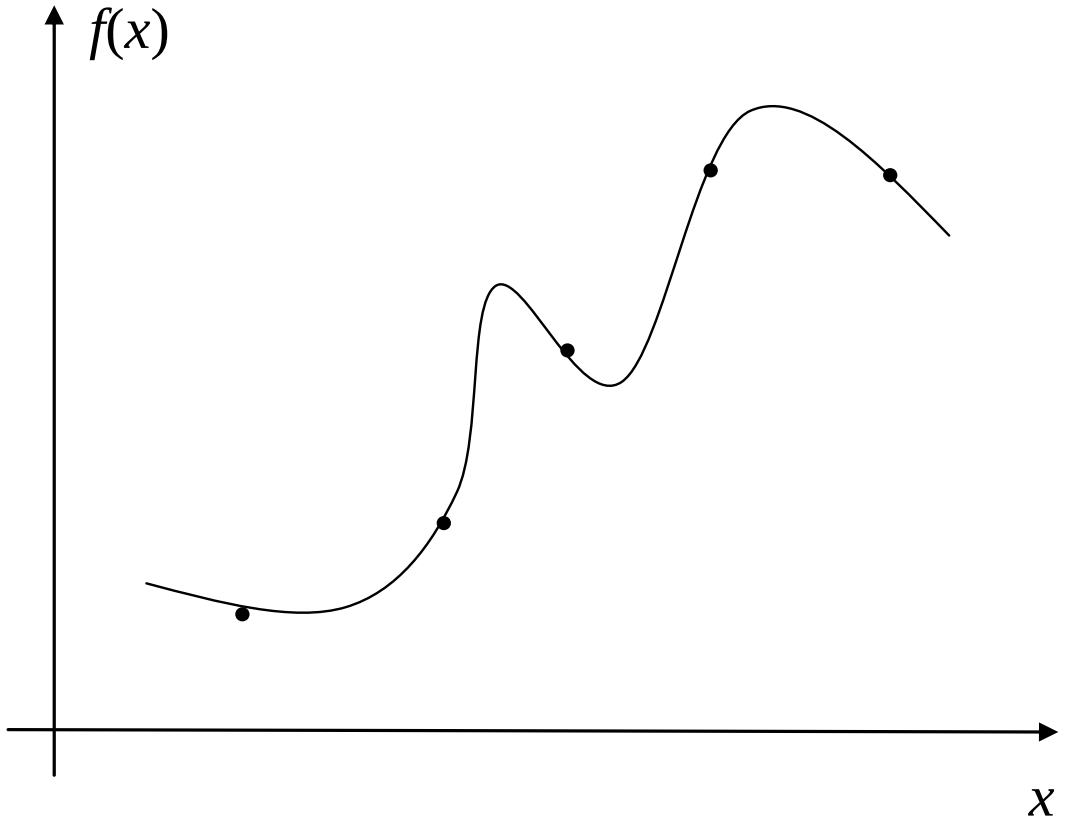
- O que ocorre quando chega uma nova “informação”?

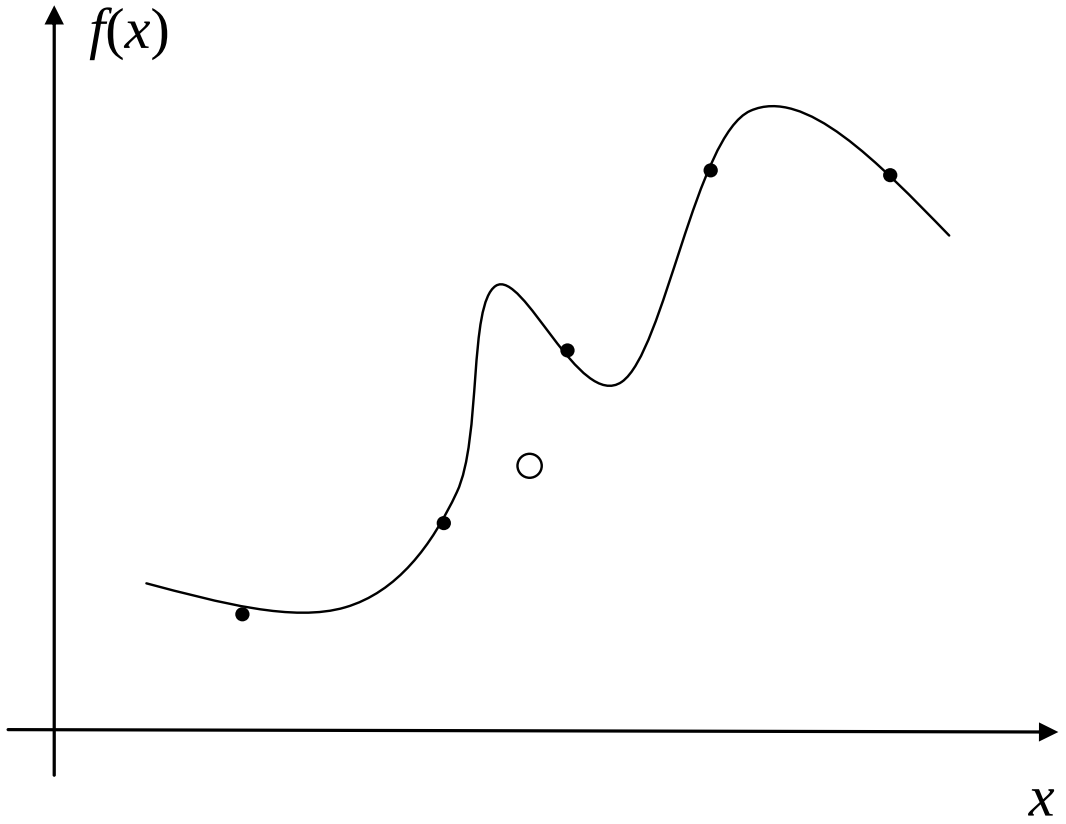

f(x)
x


f(x)
Bias elevado
Variância
reduzida
x

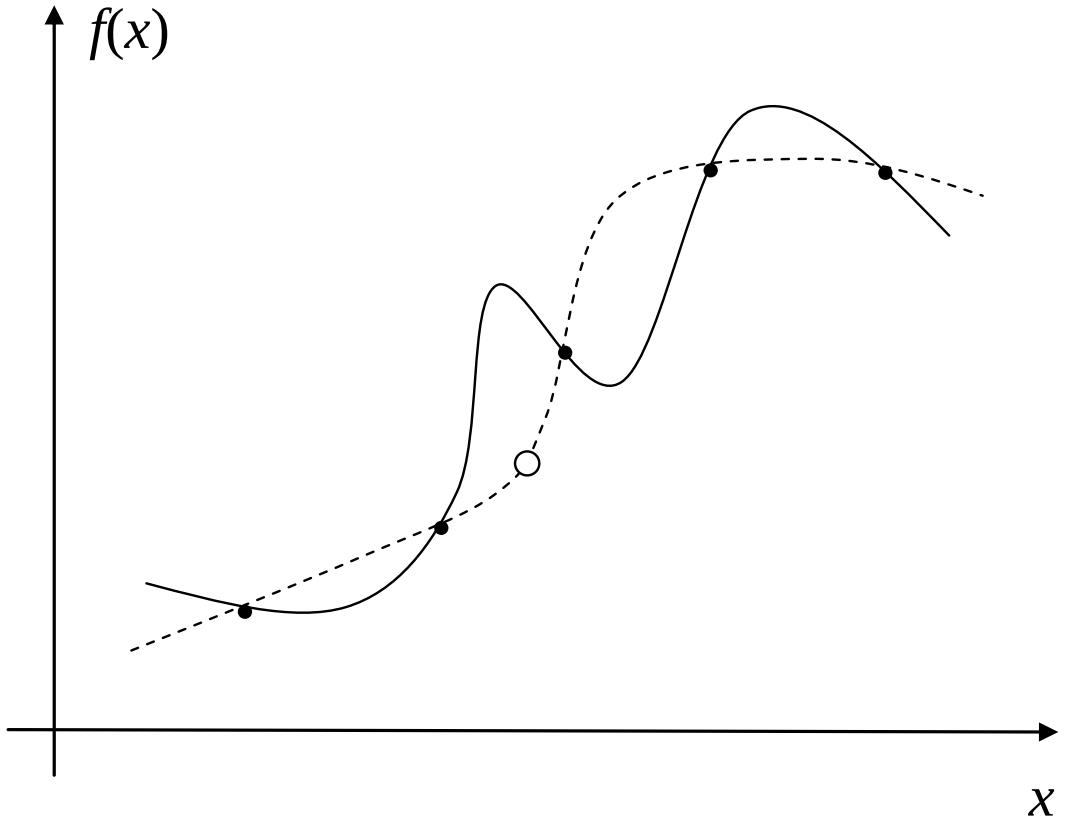


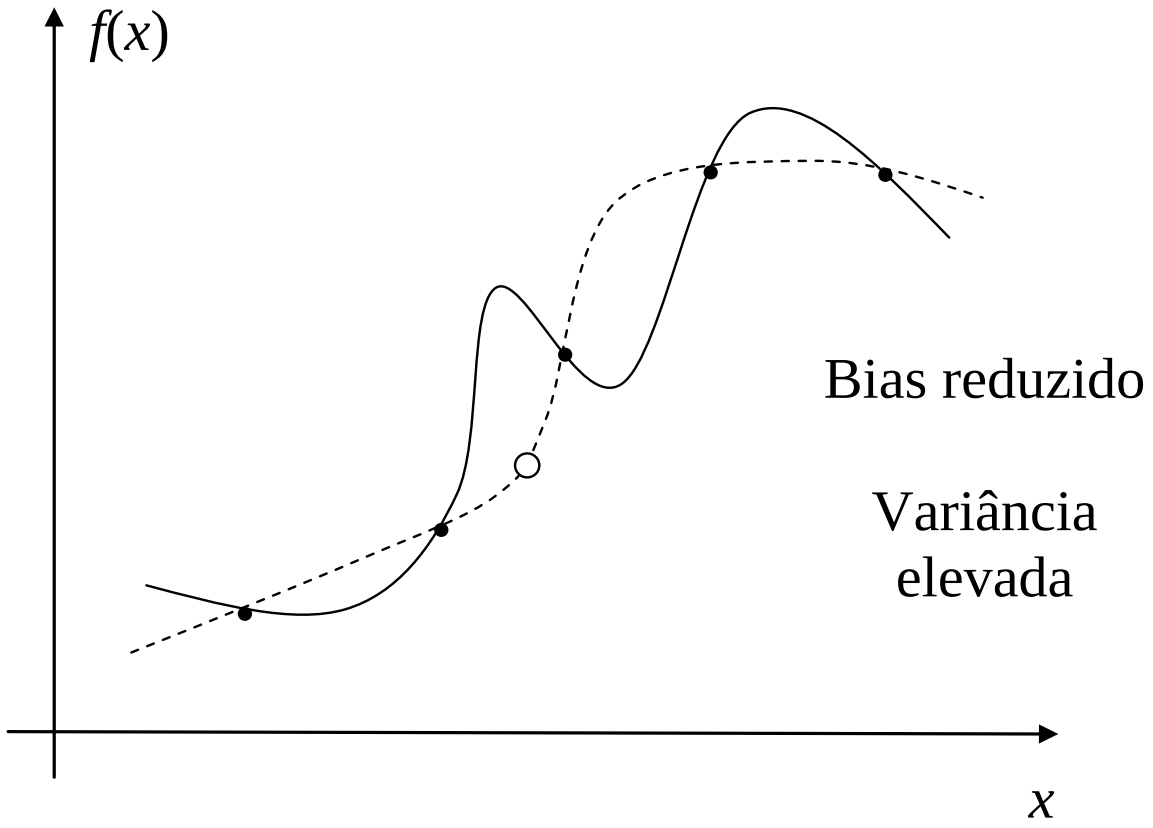
- Objetivo: Aproximar os pontos com um modelo de alta flexibilidade.

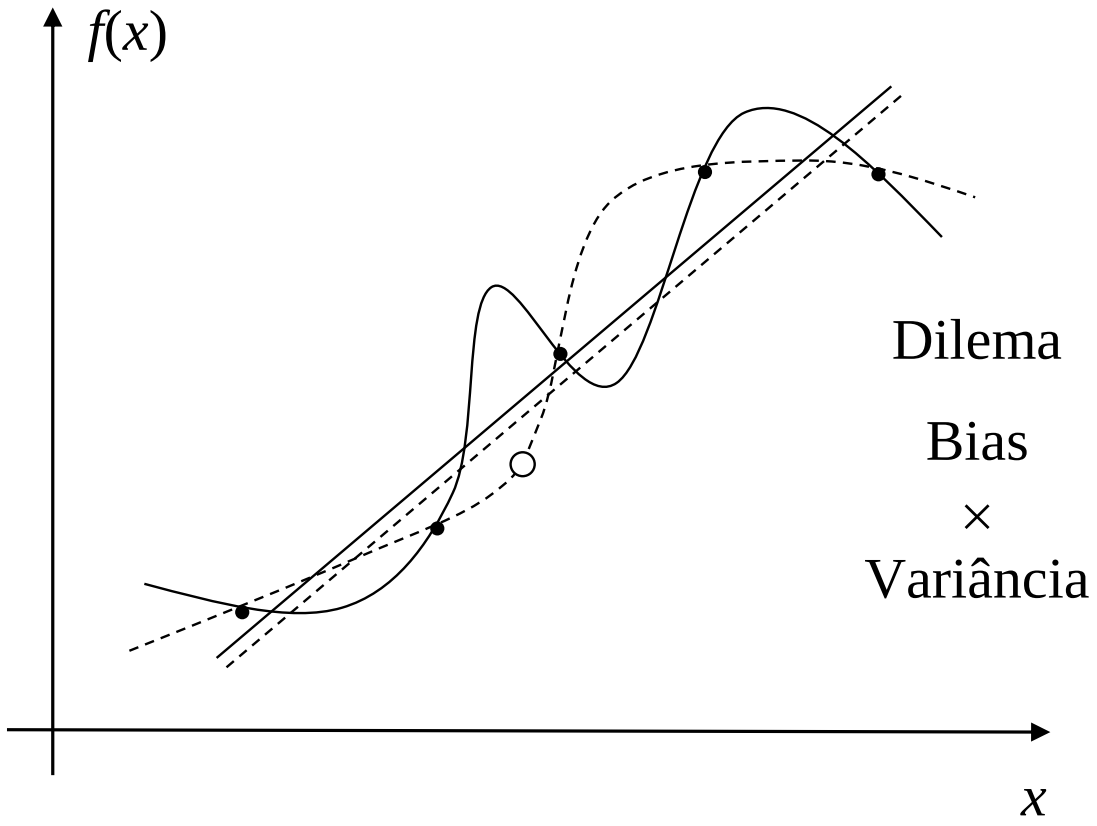

f(x)
x



- O que ocorre quando chega uma nova “informação”?


f(x)
x


f(x)
Bias reduzido
Variância
elevada
x


f(x)
Dilema
Bias
×
Variância
x

5. 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:

$$\mathbf{w}_k^* = \arg \min_{\mathbf{w}_k \in \Re^{n+1}} J(\mathbf{w}_k) + C_k \times \|\mathbf{w}_k\|_2^2$$

onde

- ✓ k é o índice da saída;
- ✓ n é o número de neurônios na camada intermediária;
- ✓ $\|\cdot\|_2$ é a norma euclidiana;
- ✓ C_k é um coeficiente de ponderação, a ser determinado, por exemplo, por métodos de busca unidimensional;
- ✓ $J(\mathbf{w}_k) = \frac{1}{2}\sum_{l=1}^{N}\left[\sum_{j=1}^{n} w_{kj} f(\mathbf{v}_j, b_j, \mathbf{x}_l) + w_{k0} - s_{kl}\right]^2$;
- ✓ N é o número de amostras disponíveis para treinamento.

- Este problema de otimização é conhecido na literatura como *ridge regression* (HASTIE, TIBSHIRANI & FRIEDMAN, 2009; HOERL & KENNARD, 1970).

5.1. 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 1'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 $n \leq N$, $\mathbf{w}_k = \left(H^T H + C_k I\right)^{-1} H^T \mathbf{s}_k$;

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

5.2. 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.

5.3. Versão incremental para ELMs

- Sugere-se consultar HUANG, CHEN & SIEW (2006).

6. Regularização para funções unidimensionais

- Há outras formas de impor suavidade quando aproximando funções a partir de dados amostrados, além de *ridge regression*.

- Para funções unidimensionais, são indicadas algumas técnicas, como (HASTIE, TIBSHIRANI & FRIEDMAN, 2009):
  - ✓ *k*-vizinhos mais próximos;
  - ✓ *Splines* polinomiais suavizantes;
  - ✓ Polinômios de Hermite com um número reduzido de funções-base;
  - ✓ Funções *kernel*.

- Nas próximas duas seções, serão abordadas técnicas alternativas / extensões ao *ridge regression*, retornando ao caso multidimensional. No entanto, essas técnicas alternativas não admitem solução na forma fechada, como em *ridge regression*.

7. LASSO

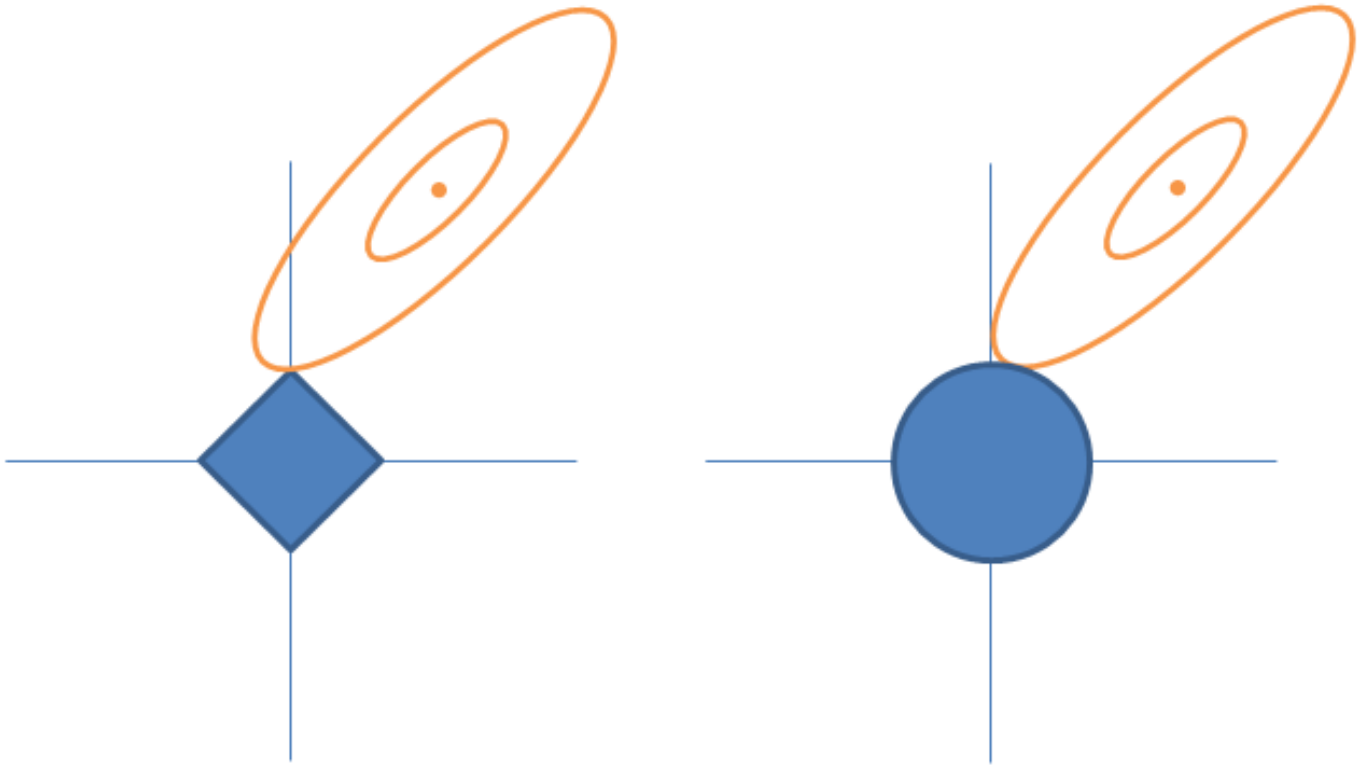
- Uma formulação alternativa considera a penalização com a norma l_1, produzindo:

$$\min_{\theta} \; J(\theta) + c\|\theta\|_1$$

onde $\|\theta\|_1 = \sum_{i=1}^{P} |\theta_i|$ e $c \geq 0$.

- Esta estratégia de regularização é conhecida como LASSO, do inglês *Least Absolute Shrinkage and Selection Operator*, e foi proposta por Tibshirani em 1996: *Tibshirani, R. "Regression shrinkage and selection via the LASSO", Journal of the Royal Statistical Society B, vol. 58, pp. 267–288, 1996*.
- A vantagem do LASSO está principalmente no fato de que múltiplos termos do vetor θ acabam sendo anulados, o que representa um processo de seleção de variáveis e leva a modelos de aproximação mais parcimoniosos.
- A motivação geométrica que permite justificar a anulação de um subconjunto de parâmetros ajustáveis é apresentada na figura a seguir. Nesta figura, apresentam-se

os lugares geométricos de valores constantes para $J(\theta)$ e $\|\theta\|_1$, à esquerda, e os lugares geométricos de valores constantes para $J(\theta)$ e $\|\theta\|_2$, à direita.



Inspirado na interpretação geométrica presente em HASTIE, TIBSHIRANI & FRIEDMAN (2009).

8. Elastic Net

- Uma solução intermediária entre *ridge regression* e LASSO é a Elastic Net, proposta em: *Zou, H., Hastie, T. "Regularization and variable selection via the elastic net", Journal of the Royal Statistical Society B, vol. 67, pp. 301–320, 2005*.
- No caso da Elastic Net, o problema de otimização regularizado assume a forma:

$$\min_{\theta} \; J(\theta) + c[\alpha\|\theta\|_1 + (1-\alpha)\|\theta\|_2^2]$$

  onde $\alpha \in [0, +1]$.
- O cálculo dos coeficientes c e α requer uma busca visando minimizar o erro junto a um conjunto de dados de validação.
- Repare que a Elastic Net contempla todas as demais propostas, para valores específicos de c e α:
  - ✓ Quadrados mínimos irrestrito: $c = 0$;
  - ✓ *Ridge Regression*: $c > 0$ e $\alpha = 0$;
  - ✓ LASSO: $c > 0$ e $\alpha = 1$;
  - ✓ *Elastic net*: $c > 0$ e $0 < \alpha < 1$.

9. Experimentos – ELM + *Ridge Regression*

- **Fonte**: *Kulaif, A.C.P. "Técnicas de regularização para máquinas de aprendizado extremo", Dissertação de Mestrado, Faculdade de Engenharia Elétrica e de Computação, Unicamp, 2014.*
- A seguir, empregando: (1) quadrados mínimos irrestrito, (2) *ridge regression* com *c* fixo e (3) *ridge regression* com *c* variável, tem-se curvas de erro de aproximação, conforme se varia o número de neurônios, para três problemas de regressão da literatura.

$$\min_{\theta} \; J(\theta) + c[\alpha\|\theta\|_1 + (1-\alpha)\|\theta\|_2^2]$$

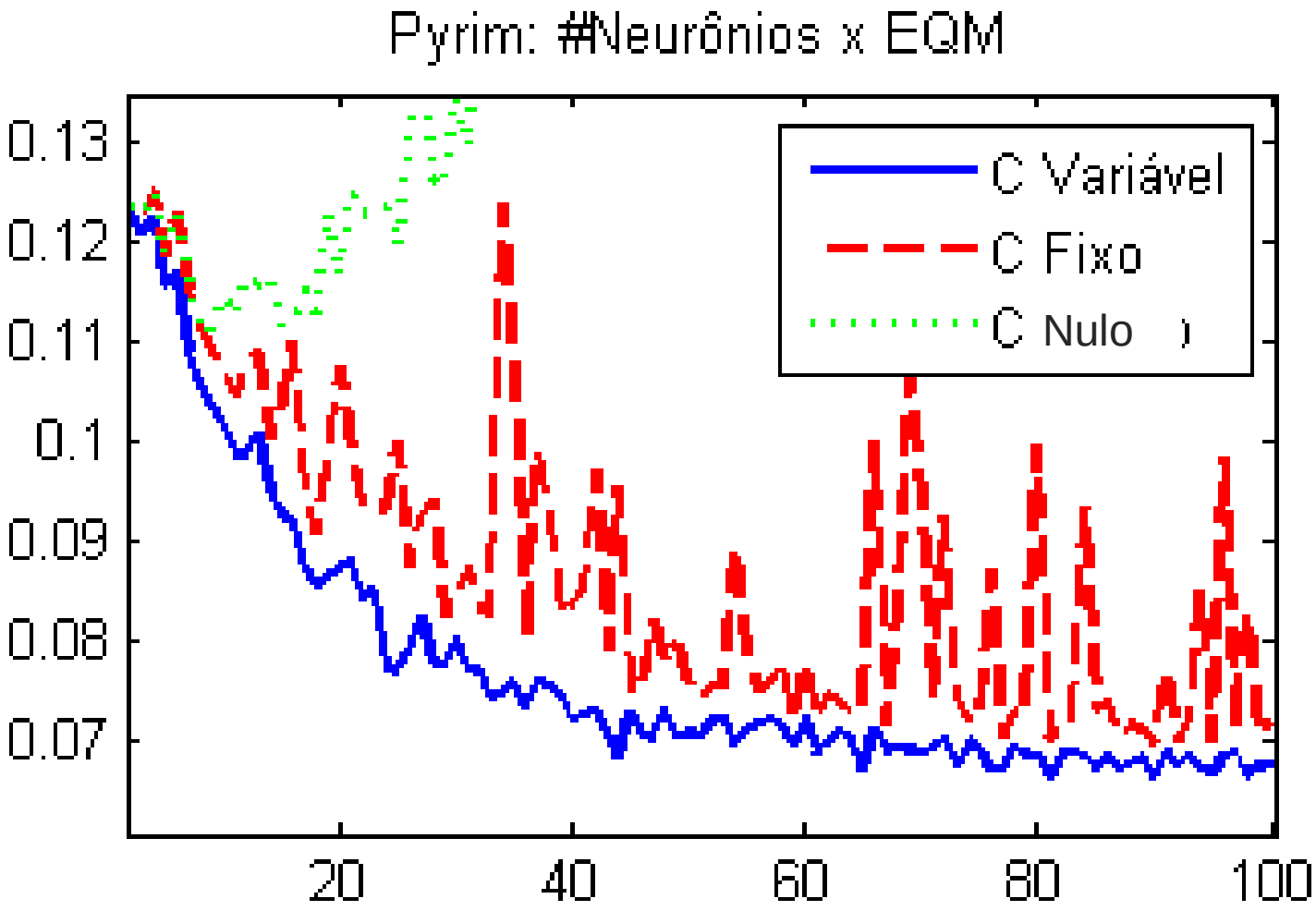
  - ✓ Quadrados mínimos irrestrito: $c = 0$;
  - ✓ *Ridge Regression*: $c > 0$ e $\alpha = 0$;

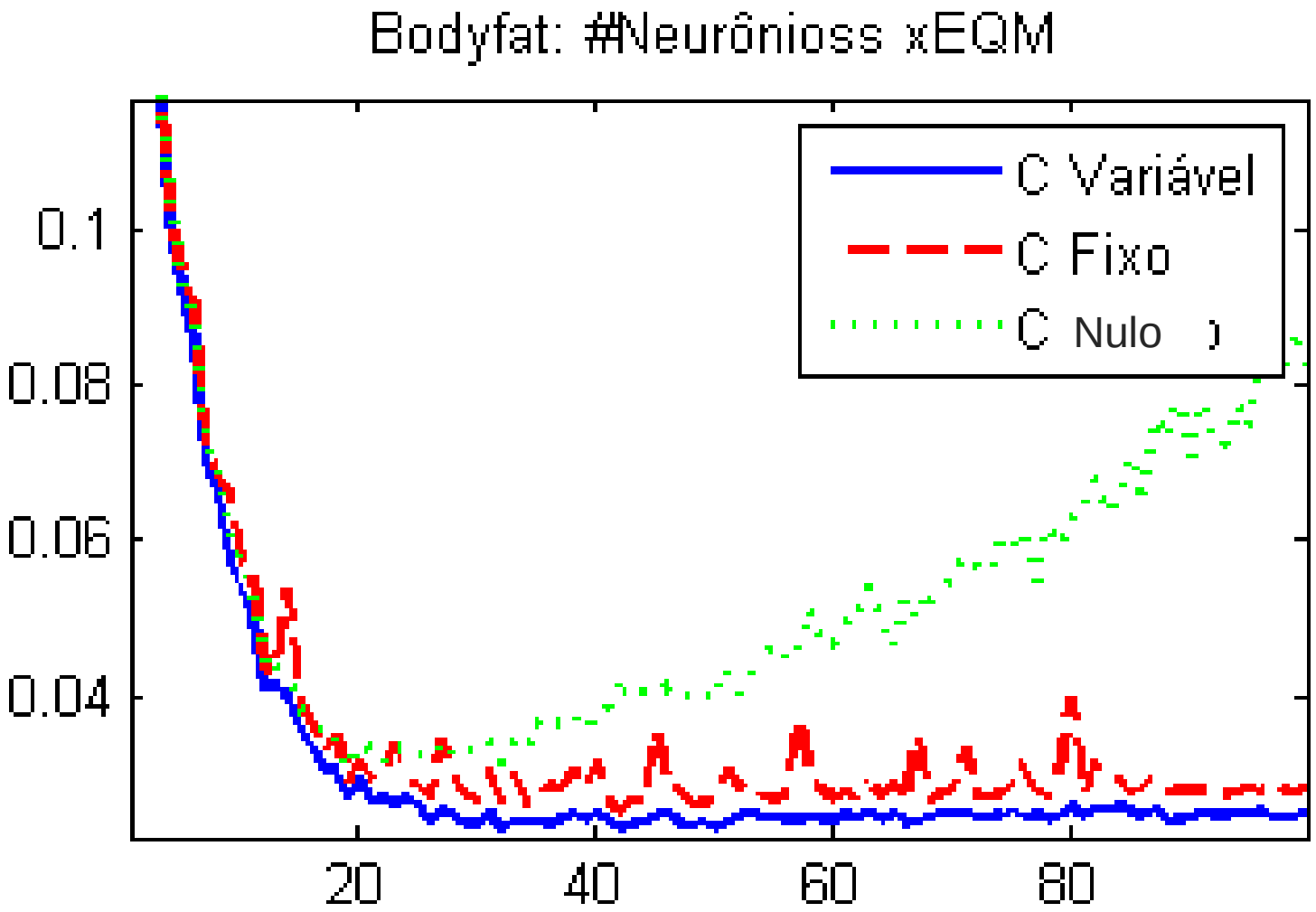
Tabela 1 – Especificações dos Conjuntos de Dados (UCI Machine Learning Repository – https://archive.ics.uci.edu/ml/about.html)

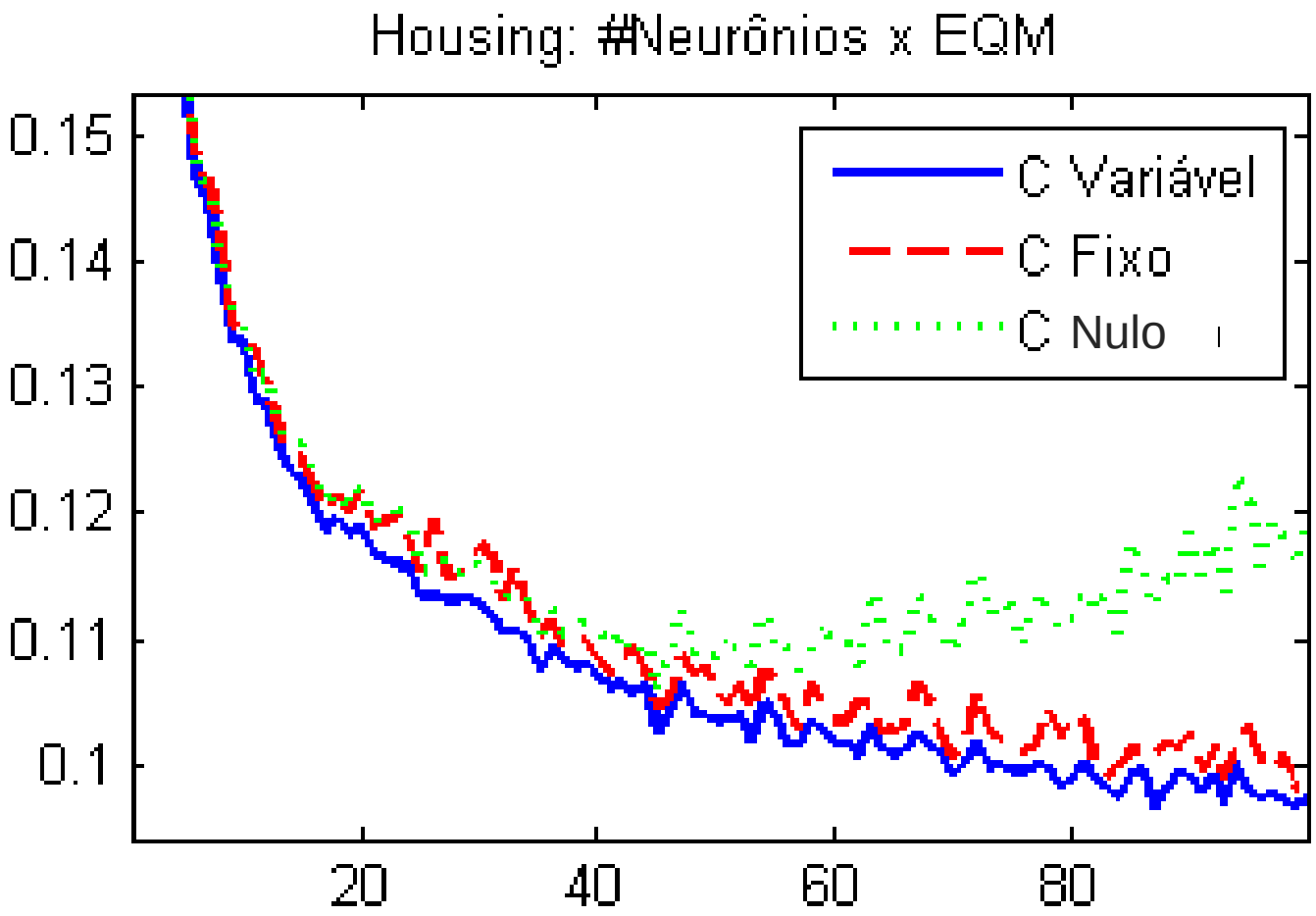
Conjuntos de Dados	Amostras	Atributos
Pyrim	74	27
Bodyfat	252	14
Housing	506	13

Tabela 2 – Valor fixo de *c* encontrado para cada conjunto de dados

Conjunto de Dados	*c*
Pyrim	2^0
Bodyfat	2^7
Housing	2^2


Pyrim: #Neurônios x EQM
0.13
0.12
0.11
0.1
0.09
0.08
0.07
20
40
60
80
100
C Variável
C Fixo
C Nulo


Bodyfat: #Neurônioss xEQM
C Variável
C Fixo
C Nulo
0.1
0.08
0.06
0.04
20
40
60
80


Housing: #Neurônios x EQM
0.15
0.14
0.13
0.12
0.11
0.1
20
40
60
80
C Variável
C Fixo
C Nulo

10.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.

HASTIE, T., TIBSHIRANI, R., FRIEDMAN, J. The Elements of Statistical Learning – Data Mining, Inference, and Prediction, Springer, 2nd edition, 2009.

HOERL, A.E., KENNARD, R. Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*, vol. 12, pp. 55-67, 1970.

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.