Um tema importante no curso:

<u>Sobreapredizado / Sobreajuste</u>

Conceito, entendimento da sua origem e formas de limitá-lo

35

... Passos preliminares ao tema: resgatemos alguns aspectos d prova de Cybenko referentes ao grau de ajuste entre modelo neural e função sendo aproximada por esse modelo

36

recordando

A aproximação universal com RNAs do tipo MLP, segundo Cybenko (& Kolmogorov)

37

Cybenko – a prova matemática, disponível para download na internet, é bastante complexa

recordando

Math. Control Signals Systems (1989) 2: 303–314

Mathematics of Control, Signals, and Systems

Approximation by Superpositions of a Sigmoidal Function*

G. Cybenko†

Abstract. In this paper we demonstrate that finite linear combinations of compositions of a fixed, univariate function and a set of affine functionals can uniformly approximate any continuous function of n real variables with support in the unit hypercube; only mild conditions are imposed on the univariate function. Our results settle an open question about representability in the class of single hidden layer neural networks. In particular, we show that arbitrary decision regions can be arbitrarily well approximated by continuous feedforward neural networks with only a single internal, hidden layer and any continuous sigmoidal nonlinearity. The paper discusses approximation properties of other possible types of nonlinearities that might be implemented by artificial neural networks.

Key words. Neural networks, Approximation, Completeness.

1. Introduction

A number of diverse application areas are concerned with the representation of general functions of an n-dimensional real variable, $x \in \mathbb{R}^n$, by finite linear combinations of the form

$$\sum_{j=1}^{N} \alpha_j \sigma(y_j^T x + \theta_j), \qquad (1)$$

where $y_j \in \mathbb{R}^n$ and $\alpha_j, \theta \in \mathbb{R}$ are fixed. (y^T is the transpose of y so that $y^T x$ is the inner product of y and x.) Here the univariate function σ depends heavily on the context of the application. Our major concern is with so-called sigmoidal σ's:

$$\sigma(t) \to \begin{cases} 1 & \text{as} \quad t \to +\infty, \\ 0 & \text{as} \quad t \to -\infty. \end{cases}$$

Such functions arise naturally in neural network theory as the activation function of a neural node (or unit as is becoming the preferred term) [L1], [RHM]. The main result of this paper is a demonstration of the fact that sums of the form (1) are dense in the space of continuous functions on the unit cube if σ is any continuous sigmoidal

* Date received: October 21, 1988. Date revised: February 17, 1989. This research was supported in part by NSF Grant DCR-8619103, ONR Contract N000-86-G-0202 and DOE Grant DE-FG02-85ER25001.

† Center for Supercomputing Research and Development and Department of Electrical and Computer Engineering, University of Illinois, Urbana, Illinois 61801, U.S.A.

4. Results for Other Activation Functions

In this section we discuss other classes of activation functions that have approximation properties similar to the ones enjoyed by continuous sigmoidals. Since these other examples are of somewhat less practical interest, we only sketch the corresponding proofs.

There is considerable interest in discontinuous sigmoidal functions such as hard limiters ($\sigma(x) = 1$ for $x \geq 0$ and $\sigma(x) = 0$ for $x < 0$). Discontinuous sigmoidal functions are not used as often as continuous ones (because of the lack of good training algorithms) but they are of theoretical interest because of their close relationship to classical perceptrons and Gamba networks [MP].

Assume that σ is a bounded, measurable sigmoidal function. We have an analog of Theorem 2 that goes as follows:

Theorem 4. *Let σ be bounded measurable sigmoidal function. Then finite sums of the form*

$$G(x) = \sum_{j=1}^{N} \alpha_j \sigma(y_j^T x + \theta_j)$$

are dense in $L^1(I_n)$. In other words, given any $f \in L^1(I_n)$ and $\varepsilon > 0$, there is a sum, $G(x)$, of the above form for which

$$\|G - f\|_{L^1} = \int_{I_n} |G(x) - f(x)|\, dx < \varepsilon.$$

The proof follows the proof of Theorems 1 and 2 with obvious changes such as replacing continuous functions by integrable functions and using the fact that $L^\infty(I_n)$ is the dual of $L^1(I_n)$. The notion of being discriminatory accordingly changes to the following: for $h \in L^\infty(I_n)$ the condition that

$$\int_{I_n} \sigma(y^T x + \theta) h(x)\, dx = 0$$

for all y and θ implies that $h(x) = 0$ almost everywhere. General sigmoidal functions are discriminatory in this sense as already seen in Lemma 1 because measures of the form $h(x)\, dx$ belong to $M(I_n)$.

Since convergence in L^1 implies convergence in measure [A], we have an analog of Theorem 3 that goes as follows:

Theorem 5. *Let σ be a general sigmoidal function. Let f be the decision function for any finite measurable partition of I_n. For any $\varepsilon > 0$, there is a finite sum of the form*

$$G(x) = \sum_{j=1}^{N} \alpha_j \sigma(y_j^T x + \theta_j)$$

and a set $D \subset I_n$, so that $m(D) \geq 1 - \varepsilon$ and

$$|G(x) - f(x)| < \varepsilon \qquad \text{for} \quad x \in D.$$

bed are quite powerful, we
that remain to be answered
immation (or equivalently,
imation of a given quality?
y a role in determining the
suspect quite strongly that
will require astronomical
dimensionality that plagues
Some recent progress con-
roximated and the number
ound in [MSJ] and [BH],
iseness of the results of this
more attention.

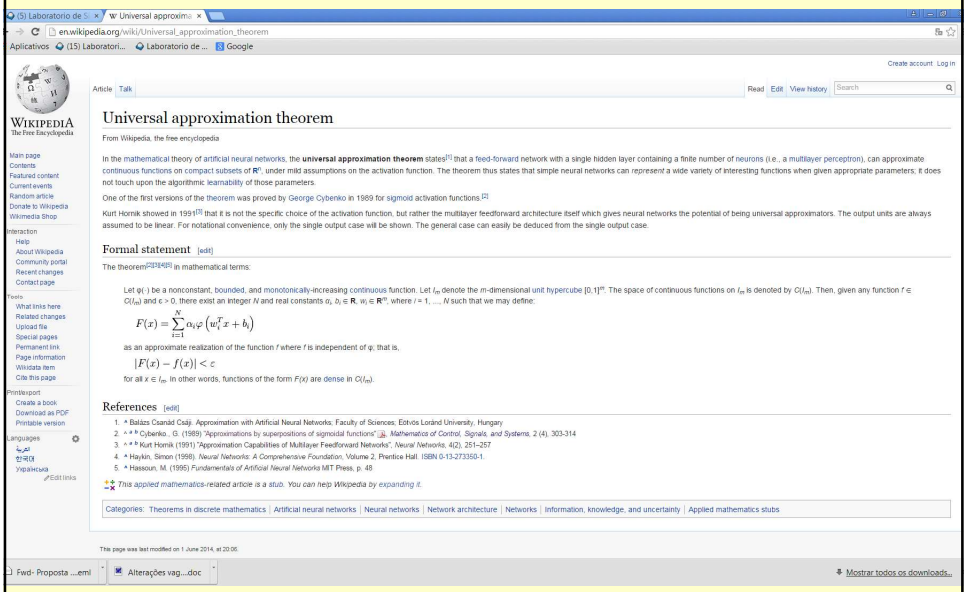
n, Christopher Chase, Lee
marov, Richard Lippmann,
trences, and improvements

New York, 1972.
eralization?, *Neural Comput.* (to
stems and control, *IEEE Control*
), Classifying learnable geometric
eedings of the 18th Annual ACM
p. 273–282.
and the Pompeiu problem, *Ann.*
nets using the Radon transform,
wo Hidden Layers are Sufficient,
University, 1988.
of linear combinations, *SIAM J.*
us mappings by neural networks,
IEE Trans. Acoust. Speech Signal
dforward networks are universal
n Neural Net and Conventional
87.
nal Classifiers, Technical Report,
–475.
tworks by sigmoidal functions,
s, University of Lowell, 1988.

Métodos Numéricos e Re

38

Cybenko – Enunciado da Prova ... (premissas + resultado)



Universal approximation theorem

From Wikipedia, the free encyclopedia

In the mathematical theory of artificial neural networks, the **universal approximation theorem** states[1] that a feed-forward network with a single hidden layer containing a finite number of neurons (i.e., a multilayer perceptron), can approximate continuous functions on compact subsets of $\mathbf{R}^n$, under mild assumptions on the activation function. The theorem thus states that simple neural networks can *represent* a wide variety of interesting functions when given appropriate parameters; it does not touch upon the algorithmic learnability of those parameters.

One of the first versions of the theorem was proved by George Cybenko in 1989 for sigmoid activation functions.[2]

Kurt Hornik showed in 1991[3] that it is not the specific choice of the activation function, but rather the multilayer feedforward architecture itself which gives neural networks the potential of being universal approximators. The output units are always assumed to be linear. For notational convenience, only the single output case will be shown. The general case can easily be deduced from the single output case.

Formal statement [edit]

The theorem[2][3][4][5] in mathematical terms:

Let $\varphi(\cdot)$ be a nonconstant, bounded, and monotonically-increasing continuous function. Let I_m denote the m-dimensional unit hypercube $[0,1]^m$. The space of continuous functions on I_m is denoted by $C(I_m)$. Then, given any function $f \in C(I_m)$ and $\varepsilon > 0$, there exist an integer N and real constants $\alpha_i, b_i \in \mathbf{R}$, $w_i \in \mathbf{R}^m$, where $i = 1, \ldots, N$ such that we may define:

$$F(x) = \sum_{i=1}^{N} \alpha_i \varphi\left(w_i^T x + b_i\right)$$

as an approximate realization of the function f where f is independent of φ; that is,

$$|F(x) - f(x)| < \varepsilon$$

for all $x \in I_m$. In other words, functions of the form $F(x)$ are dense in $C(I_m)$.

References [edit]

1. ^ Balázs Csanád Csáji. Approximation with Artificial Neural Networks; Faculty of Sciences; Eötvös Loránd University, Hungary
2. ^ a b Cybenko., G. (1989) "Approximations by superpositions of sigmoidal functions", *Mathematics of Control, Signals, and Systems*, 2 (4), 303-314
3. ^ a b Kurt Hornik (1991) "Approximation Capabilities of Multilayer Feedforward Networks", *Neural Networks*, 4(2), 251–257
4. ^ Haykin, Simon (1998). *Neural Networks: A Comprehensive Foundation*. Volume 2. Prentice Hall. ISBN 0-13-273350-1.
5. ^ Hassoun, M. (1995) *Fundamentals of Artificial Neural Networks* MIT Press, p. 48

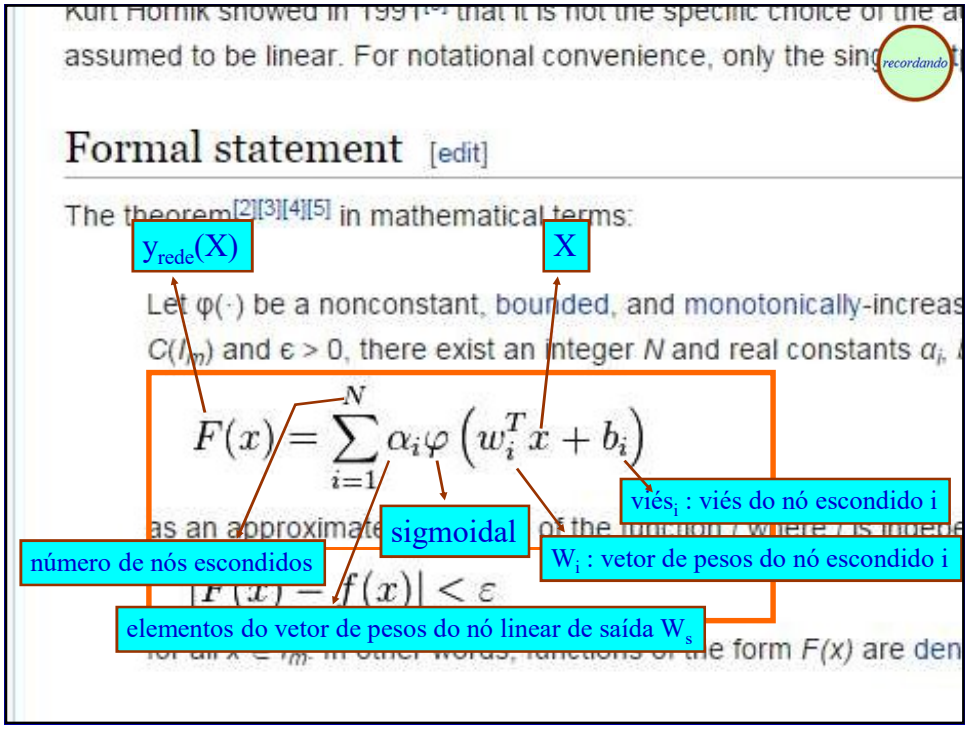
This applied mathematics-related article is a stub. You can help Wikipedia by expanding it.

Categories: Theorems in discrete mathematics | Artificial neural networks | Neural networks | Network architecture | Networks | Information, knowledge, and uncertainty | Applied mathematics stubs

This page was last modified on 1 June 2014, at 20:06.

Métodos Numéricos e Reconhecimento de Padrões – © Prof. Emilio Del Moral – EPUSP

39



40

Kurt Hornik showed in 1991[6] that it is not the specific choice of the a

assumed to be linear. For notational convenience, only the sing

recordando

Formal statement [edit]

The theorem[2][3][4][5] in mathematical terms:

Let φ(·) be a nonconstant, bounded, and monotonically-increas

$C(I_m)$ and ε > 0, there exist an integer N and real constants α_i,

$$F(x) = \sum_{i=1}^{N} \alpha_i \varphi\left(w_i^T x + b_i\right)$$

$y_{rede}(X)$ Fescondida_sistema(X)

as an approximate realization of the function f where f is indepe

Limite de erro

$$|F(x) - f(x)| < \varepsilon$$

for all $x \in I_m$. In other words, functions of the form $F(x)$ are den

41

Questões intrigantes, p/ esta aula e p/ pensar em casa ...

- *No que impacta escolhermos o "epslon" de Cybenko de alto valor? O que muda na estrutura de Cybenko com isso?*
- *No que impacta escolhermos o "epslon" de Cybenko de baixo valor?*
- *Como definimos o número de nós da primeira camada do MLP ? Isto pode ser definido a priori, antes de testar o seu desempenho? (por exemplo com base no número de entradas da rede e/ou com base no número de exemplares de treino M?)*
- *O que ganhamos e o que perdemos se escolhermos usar POUCOS nós na construção da rede neural?*
- *O que ganhamos e o que perdemos se escolhermos usar MUITOS nós na construção da rede neural?*

42

... Um parênteses para discutirmos um pouco a aproximação universal usando tangentes hiperbólicas, sigmoides, etc ... / funções em formato de "S"

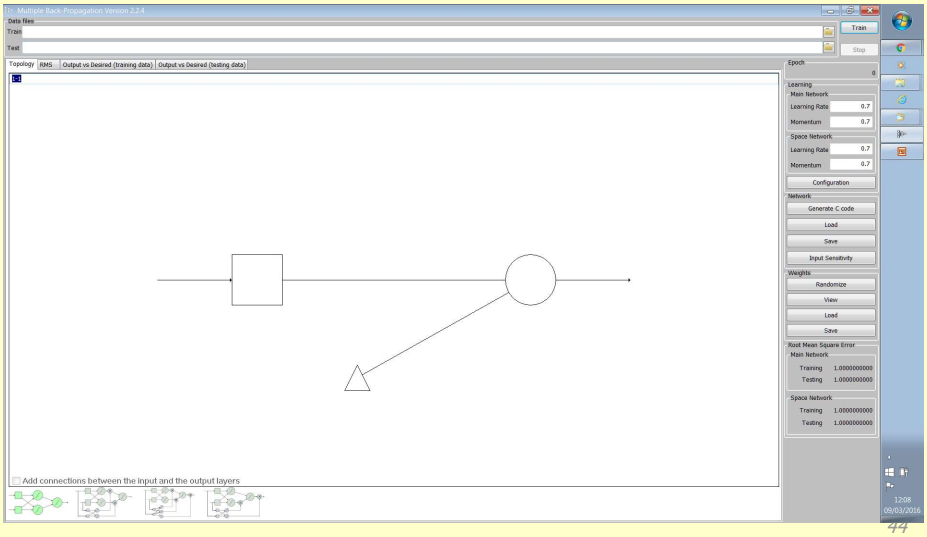
... ao menos no caso simples e bem particular de função escalar univariada ... aproximação de uma função y de uma única variável x_1:

$$y(x_1)$$

43

43

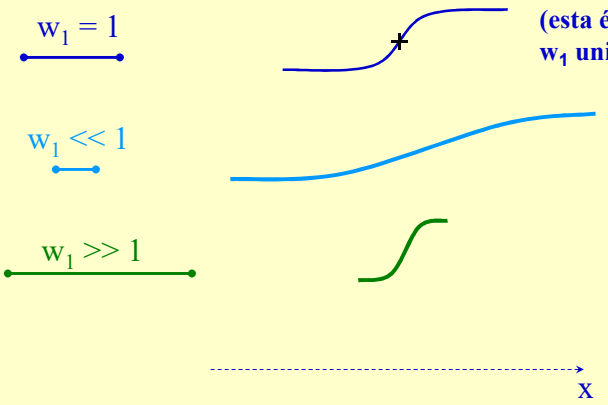
O que conseguimos fazer com um único neurônio sigmoidal, no caso de regressões ("y contínuo")?



44

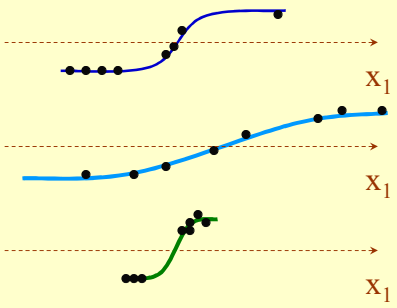
44

O que conseguimos fazer com um único neurônio sigmoidal $y(w_1.x_1)$ c/ escalamento de x_1 via w_1 e *VIÉS NULO*



(esta é a tgh "de fábrica": w_1 unitário e w_0 nulo)

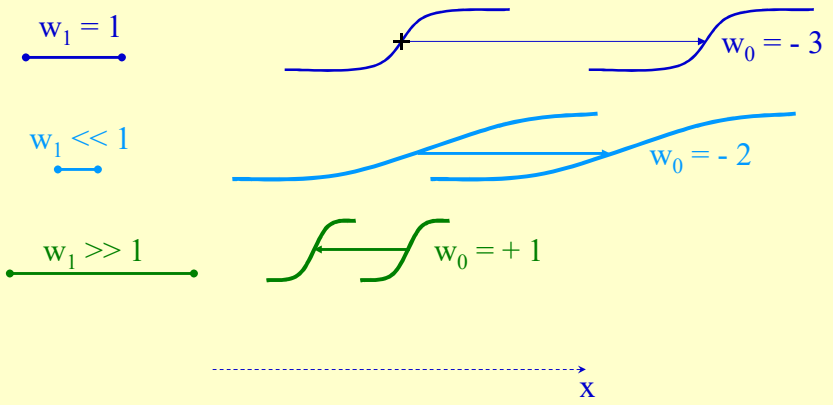
Que tipo de dados empíricos modelamos com um único neurônio sigmoidal em regressões ("$y(x_1)$ contínuo")?



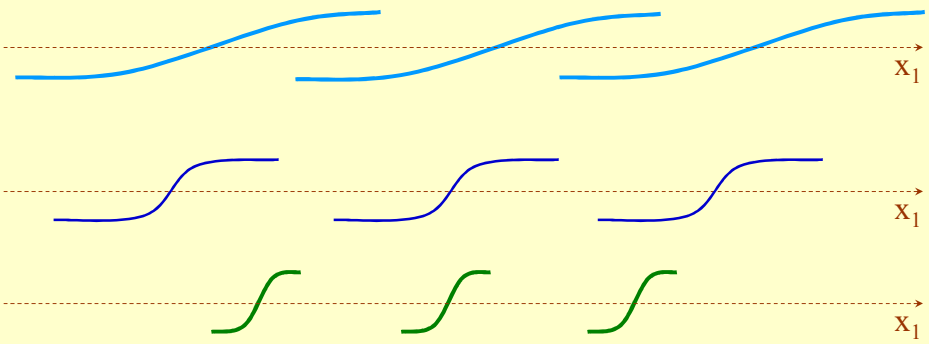
Os pontos pretos são pares empíricos (x^μ, y^μ); As curvas coloridas, são regressões sigmoidais aderentes a tais pares.

O que conseguimos fazer com um único neurônio sigmoidal $y(w_1.x_1 + w_0.1)$, c/ escalamento de x_1 via w_1

... e agora também com o viés, via viés w_0

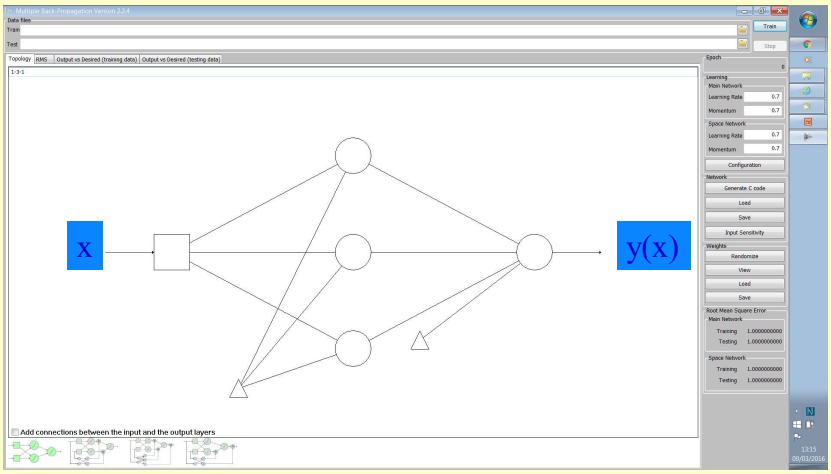


O que conseguimos fazer com um único neurônio sigmoidal, no caso de regressões ("$y(x_1)$ contínuo")?



$y = tgh\ (w_1.x_1 + viés)$

Regressão univariada com Cybenko "café com leite" de 3 nós na primeira camada ...



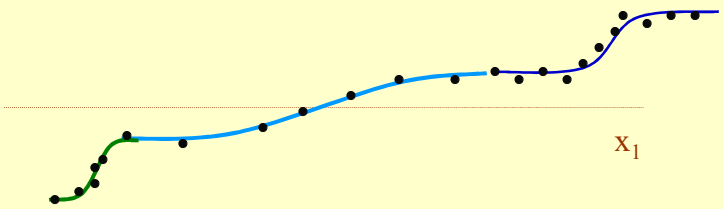
49

Métodos Numéricos e Reconhecimento de Padrões – © Prof. Emilio Del Moral – EPUSP

49

Cybenko "café com leite" (regressão genérica univariada), para aproximação universal de funções de 1 variável x_1 apenas?

... superposição de várias sigmóides deslocadas e escaladas



Vocês enxergam acima 3 nós "tgh" na primeira camada, com com 3 viéses distintos e 3 escaladores de x_1 distintos, e mais um 4o nó combinador (somatória simples de 3 entradas) na camada de saída?

50

Métodos Numéricos e Reconhecimento de Padrões – © Prof. Emilio Del Moral – EPUSP

50

Algumas discussões adicionais sobre o Cybenko "café com leite" da regressão univariada ...

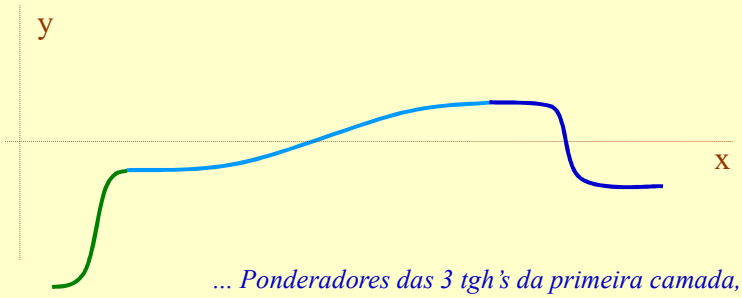
- *Vimos acima como se comporta o regressor univariado de Cybenko "café com leite" quando o nó de saída tem função de ativação identidade, seus pesos ponderadores das saídas da primeira camada são todos unitários positivos e o peso de viés é nulo.*

- *O que ocorre se os esses pesos ponderadores não forem mais unitários? (podem ser agora positivos, negativos, encolhedores (módulo < 1) ou amplificadores (módulo > 1))*

- *E se o seu peso de viés do 4o nó não for mais nulo?*

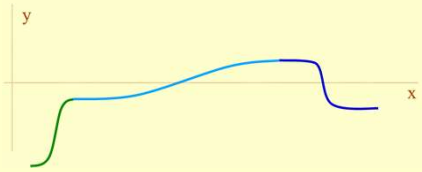
Cybenko "café com leite", para aproximação universal de funções de 1 variável x apenas?

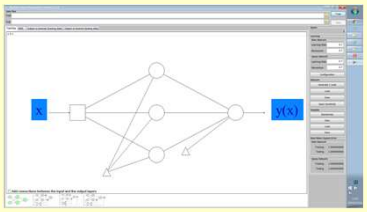
... superposição de várias sigmóides deslocadas e escaladas em x e em y ...



... Ponderadores das 3 tgh's da primeira camada, que são implemantados nos pesos sinápticos do 4o nó, não são mais unitários nem necessariamente positivos

Isto indica claramente que ao menos no caso de funções univariadas no domínio e na imagem (uma única variável x no argumento e uma única variável y na "saída" da função) uma RNA de duas camadas (com vários neurônios na segunda, não apenas 3 como ilustrado) pode aproximar qualquer função contínua univariada com erro bem pequeno se necessário: se desejado, basta usarmos mais e mais nós na segunda camada do MLP, aumentando assim arbitrariamente a precisão da aproximação da função alvo da modelagem.





53

Cybenko foi além de mostrar a viabilidade de aproximação em casos unidimensionais, ele fez a prova de *Aproximação Universal* no âmbito de funções de múltiplas variáveis!

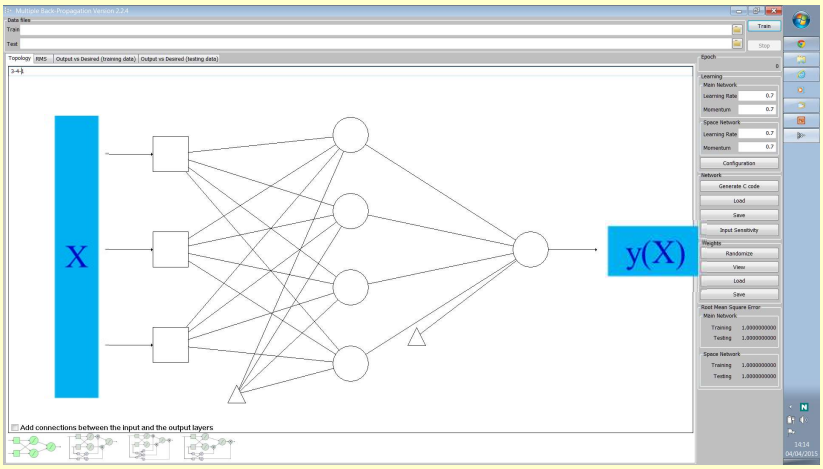
Qualquer Função(X) genérica pode ser aproximada por um MLP – O que é bom para Estimação / Regressão Contínua (um de alvos neste curso) !!!

E ...

É também bom para o Reconhecimento de Padrões (nosso segundo alvo de modelagem) com o MLP ... Trata-se de um caso específico de função binária na saída !!!

54

Cybenko foi para um terreno mais complexo: temos um vetor de entradas X em lugar de um x unidimensional



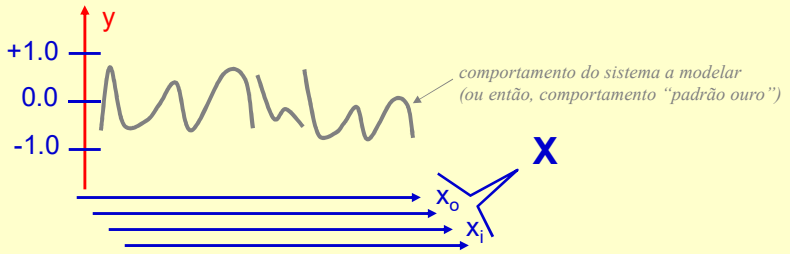
55

Métodos Numéricos e Reconhecimento de Padrões – © Prof. Emilio Del Moral – EPUSP

55

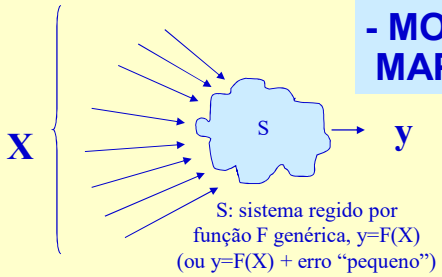
A função y(X) "a descobrir", num caso geral de função analógica y(X) ....

recordando



comportamento do sistema a modelar
(ou então, comportamento "padrão ouro")

- ESTIMADOR ;
- MODELAGEM POR MAPEAMENTO X→y



(.... VALORES CONTÍNUOS!!!)

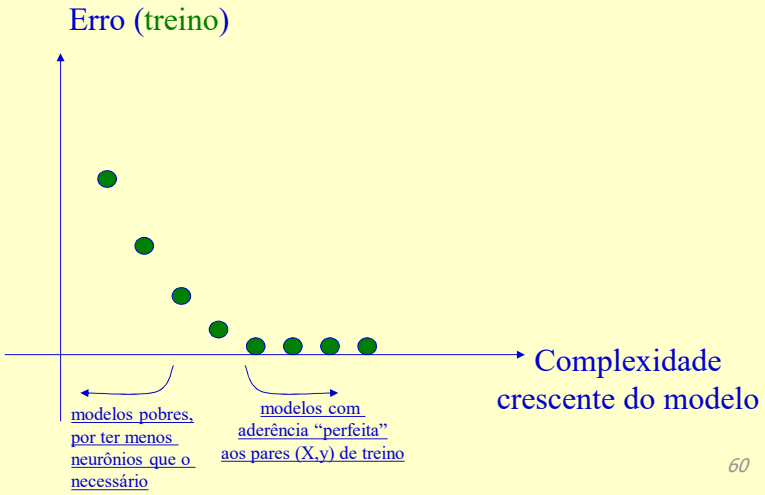
S: sistema regido por
função F genérica, y=F(X)
(ou y=F(X) + erro "pequeno")

56

© Prof. Emilio Del Moral – EPUSP

56

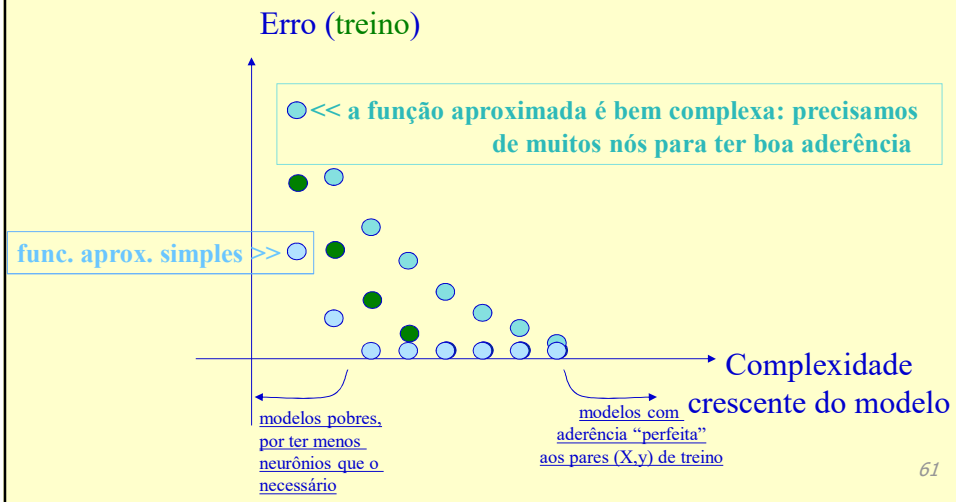
Aumento de aderência aos dados de treino com o aumento de nós da RNA ...



60

60

Aumento de aderência aos dados de treino com o aumento de nós da RNA ...



61

61

Isto quer dizer que sempre é melhor termos um modelo com mais nós neurais que um modelo com menos nós neurais?

Afinal, da mesma maneira que a computação de um regressor polinomial de grau seis engloba a computação dos regressores polinomiais de graus menores, os modelos com mais nós neurais englobam os mais simples (em termos de capacidades de computações possíveis) correto?

Sim, correto! Mas há um limite no "lucro" em tal estratégia, dado pelo fenômeno de

<u>Sobreaprendizado e perda de generalização</u> ...

63

63

O conceito de sobreaprendizado nos dá critérios adicionais para a definição do número de neurônios / grau de complexidade de uma rede neural, critérios esses que vão bem além de simples econômica computacional – Esses critérios vão mais na direção de aumento de precisão na generalização

64

64

Sobreaprendizado em polinômios:

Entendamos o conceito de sobreajuste / sobreaprendizado num universo mais familiar (e mais simples) que a modelagem com RNAs; o universo de modelagem por regressão polinomial univariada, usada para representar dados com comportamento linear ou não linear .

Depois, vocês mesmos podem pensar nos equivalentes dos nossos raciocínios feitos aqui para o universo de polinômios, mas para o universo de RNAs e mesmo de outros tipos de modelos com número de parâmetros variável (complexidade variável) que você conheça ... 85

85

Falemos em lousa um pouco sobre a reta média para um conjunto de pares (x,y), a parábola média, a cúbica média ... etc

$$y \sim ax+b \; ; \quad y \sim ax^2 + bx + c \; ; \quad y \sim ax^3 + bx^2 + cx + d$$

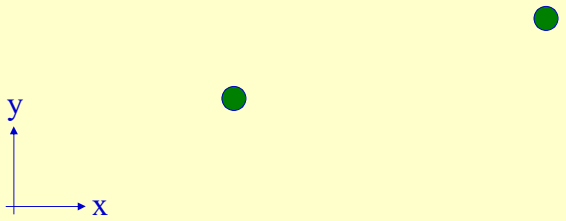
e mais além, falemos sobre regressão polinomial univariada, com o grau do polinômio aproximador podendo ser 1, 2, 3, ou mesmo graus bastante mais altos como 50, 51 etc.

$$y \sim ax^{51} + bx^{50} + cx^{49} + \ldots$$

86

86

Sobreaprendizado ilustrado em sua fenomenologia, na regressão polinomial (de vários graus) univariada



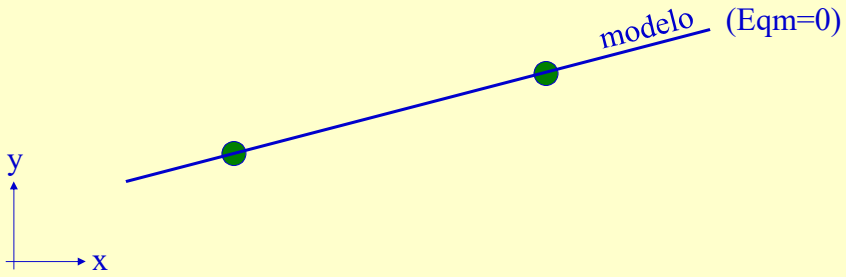
Os dados empíricos (x^{μ},y^{μ}) estão em verde;

87

87

Sobreaprendizado ilustrado em sua fenomenologia, na regressão polinomial (de vários graus) univariada

façamos uso de modelagem linear ...



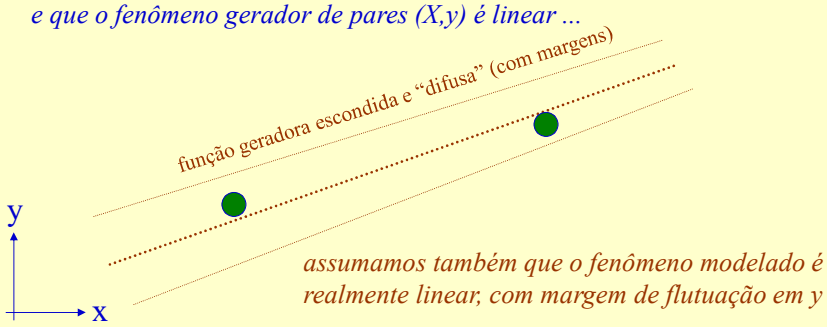
Os dados empíricos (x^{μ},y^{μ}) estão em verde;
O modelo linear gerado a partir dos dados, em azul.

88

88

Sobreaprendizado ilustrado em sua fenomenologia, na regressão polinomial (de vários graus) univariada

e que o fenômeno gerador de pares (X,y) é linear ...



assumamos também que o fenômeno modelado é realmente linear, com margem de flutuação em y

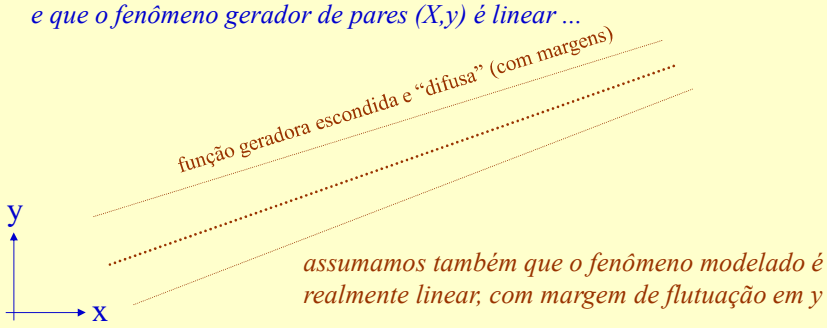
Os dados empíricos (x^{μ},y^{μ}) estão em verde;
O modelo linear gerado a partir dos dados, em azul.
O fenômeno gerador de pares (x,y) é linear em essência, mas tem alguma flutuação randômica em y. A tendência e os limites da flutuação estão representados em marrom

89

89

Sobreaprendizado ilustrado em sua fenomenologia, na regressão polinomial (de vários graus) univariada

e que o fenômeno gerador de pares (X,y) é linear ...



assumamos também que o fenômeno modelado é realmente linear, com margem de flutuação em y

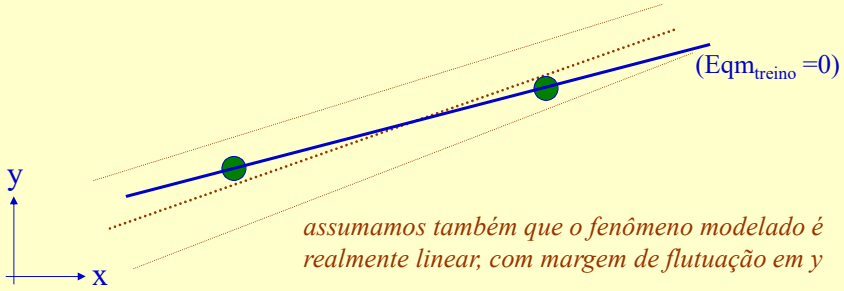
O fenômeno gerador de pares (x,y) é linear em essência, mas tem alguma flutuação randômica em y. A tendência e os limites da flutuação estão representados em marrom

90

90

Sobreaprendizado ilustrado em sua fenomenologia, na regressão polinomial (de vários graus) univariada

façamos uso de modelagem linear ...



assumamos também que o fenômeno modelado é realmente linear, com margem de flutuação em y

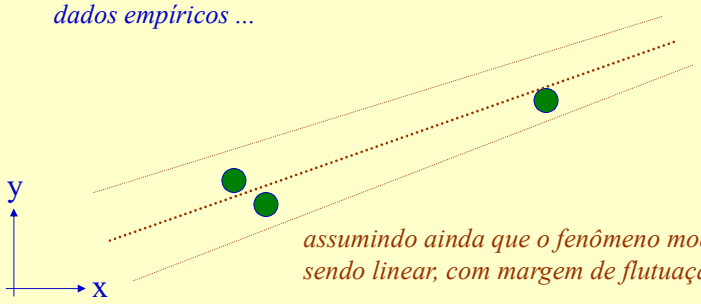
Os dados empíricos (x^{μ},y^{μ}) estão em verde;
O modelo linear gerado a partir dos dados, em azul.
O fenômeno gerador de pares (x,y) é linear em essência, mas tem alguma flutuação randômica em y. A tendência e os limites da flutuação estão representados em marrom

91

91

Sobreaprendizado ilustrado em sua fenomenologia, na regressão polinomial (de vários graus) univariada

Consideremos agora nova situação com mais dados empíricos ...



assumindo ainda que o fenômeno modelado segue sendo linear, com margem de flutuação em y

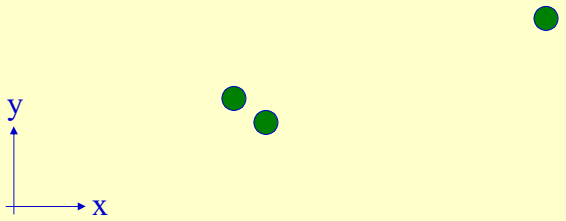
Os dados empíricos (x^{μ},y^{μ}) estão em verde;

92

92

Sobreaprendizado ilustrado em sua fenomenologia, na regressão polinomial (de vários graus) univariada

Consideremos agora nova situação com mais dados empíricos ... mas na modelagem não se sabe se é o fenômeno linear ou quadrático ...



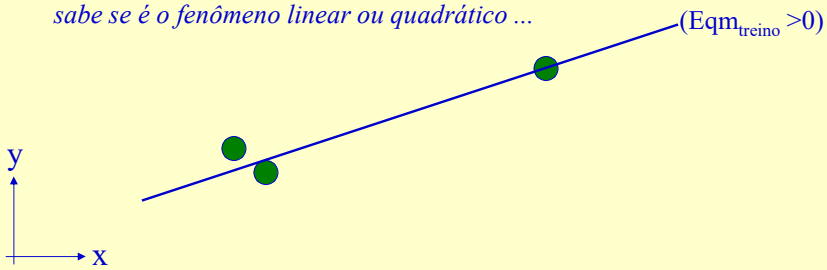
Os dados empíricos (x^{μ},y^{μ}) estão em verde;

93

93

Sobreaprendizado ilustrado em sua fenomenologia, na regressão polinomial (de vários graus) univariada

Consideremos agora nova situação com mais dados empíricos ... mas na modelagem não se sabe se é o fenômeno linear ou quadrático ...



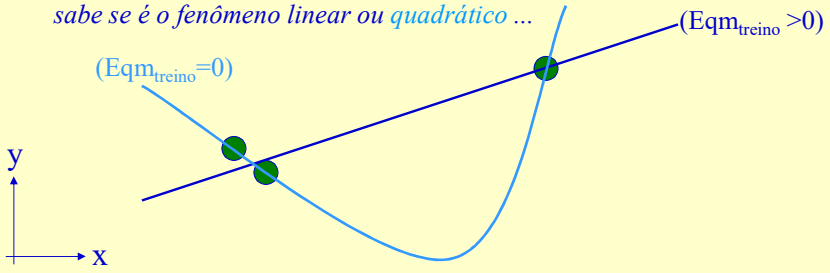
Os dados empíricos (x^{μ},y^{μ}) estão em verde;

94

94

Sobreaprendizado ilustrado em sua fenomenologia, na regressão polinomial (de vários graus) univariada

Consideremos agora nova situação com mais dados empíricos ... mas na modelagem não se sabe se é o fenômeno linear ou quadrático ...



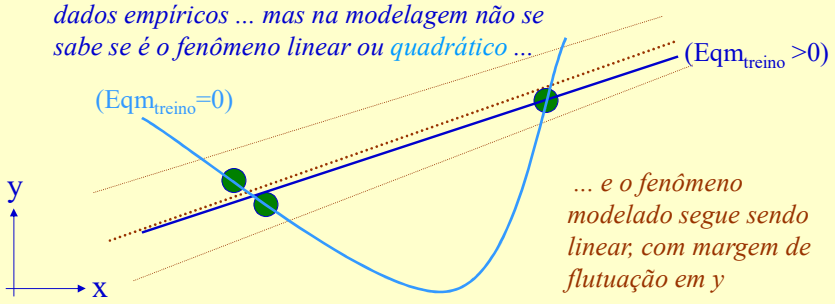
Os dados empíricos (x^{μ},y^{μ}) estão em verde;
Dois modelos polinomiais gerados a partir dos dados, em azuis.
O fenômeno gerador de pares (x,y) é linear em essência, mas tem alguma flutuação randômica em y. A tendência e os limites da flutuação estão representados em marrom

95

95

Sobreaprendizado ilustrado em sua fenomenologia, na regressão polinomial (de vários graus) univariada

Consideremos agora nova situação com mais dados empíricos ... mas na modelagem não se sabe se é o fenômeno linear ou quadrático ...

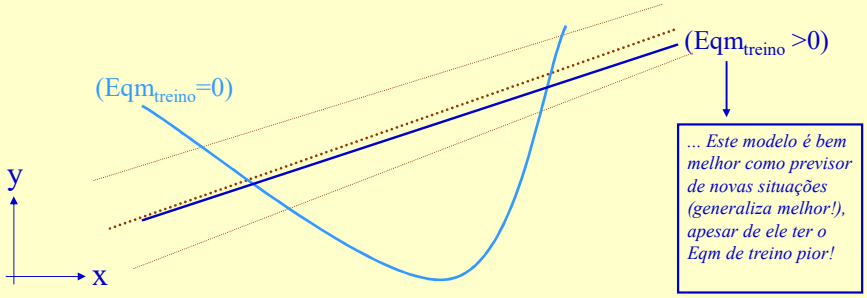


Os dados empíricos (x^{μ},y^{μ}) estão em verde;
Dois modelos polinomiais gerados a partir dos dados, em azuis.
O fenômeno gerador de pares (x,y) é linear em essência, mas tem alguma flutuação randômica em y. A tendência e os limites da flutuação estão representados em marrom

96

96

Sobreaprendizado ilustrado em sua fenomenologia, na regressão polinomial (de vários graus) univariada



... Este modelo é bem melhor como previsor de novas situações (generaliza melhor!), apesar de ele ter o Eqm de treino pior!

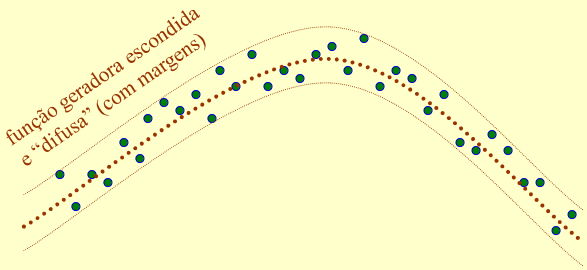
Os dados empíricos (x^{μ},y^{μ}) estão em verde;
Dois modelos polinomiais gerados a partir dos dados, em azuis. O fenômeno gerador de pares (x,y) é linear em essência, mas tem alguma flutuação randômica em y. A tendência e os limites da flutuação estão representados em marrom

97

97

Sobreaprendizado ilustrado em sua fenomenologia, na regressão polinomial (de vários graus) univariada

um novo exemplo



Os dados empíricos (x^{μ},y^{μ}) estão em verde;

O fenômeno gerador de pares (x,y) é quadrático em essência, mas tem alguma flutuação randômica em y. A tendência e os limites da flutuação estão representados em marrom

98

98

Sobreaprendizado ilustrado em sua fenomenologia, na regressão polinomial (de vários graus) univariada

um novo exemplo

Os dados empíricos (x^{μ},y^{μ}) estão em verde;

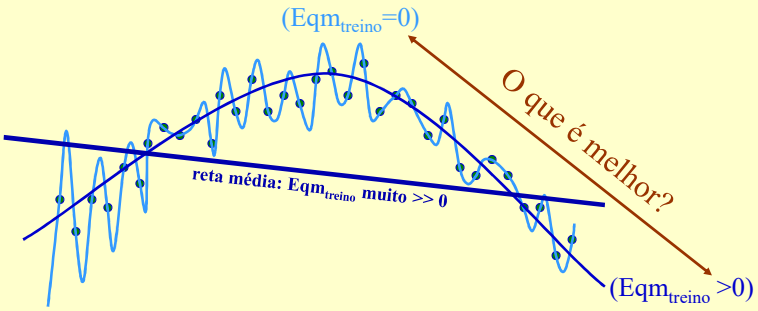
O fenômeno gerador de pares (x,y) é quadrático em essência, mas tem alguma flutuação randômica em y. A tendência e os limites da flutuação estão representados em marrom

99

99

Sobreaprendizado ilustrado em sua fenomenologia, na regressão polinomial (de vários graus) univariada

um novo exemplo



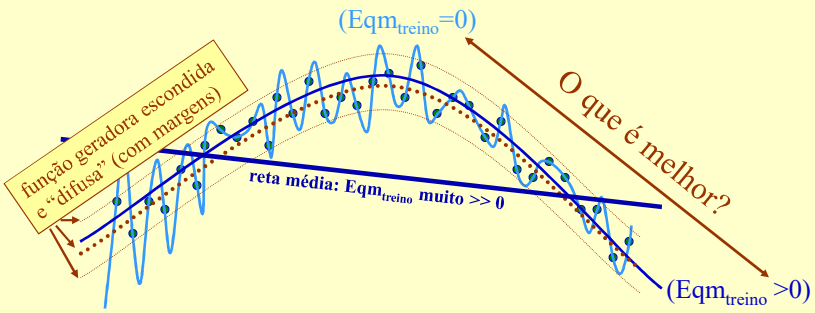
Os dados empíricos (x^{μ},y^{μ}) estão em verde;
Três modelos polinomiais gerados a partir dos dados, em azuis.

100

100

Sobreaprendizado ilustrado em sua fenomenologia, na regressão polinomial (de vários graus) univariada

um novo exemplo

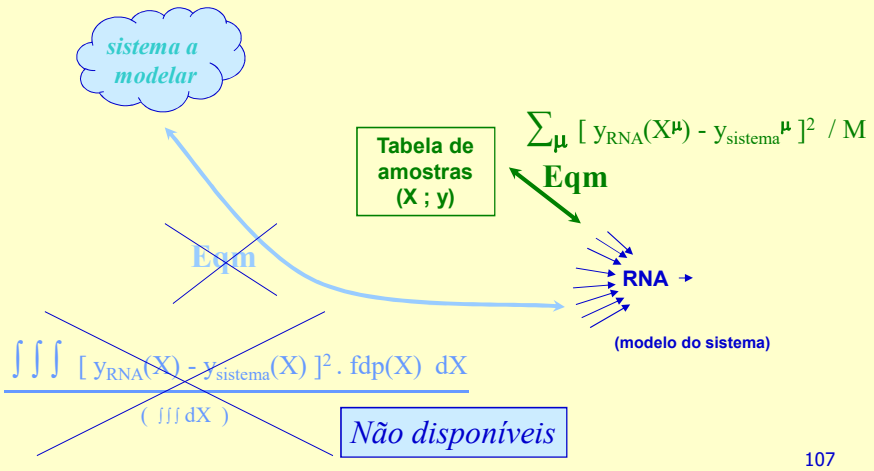


Os dados empíricos (x^{μ},y^{μ}) estão em verde;
Três modelos polinomiais gerados a partir dos dados, em azuis.

101

101

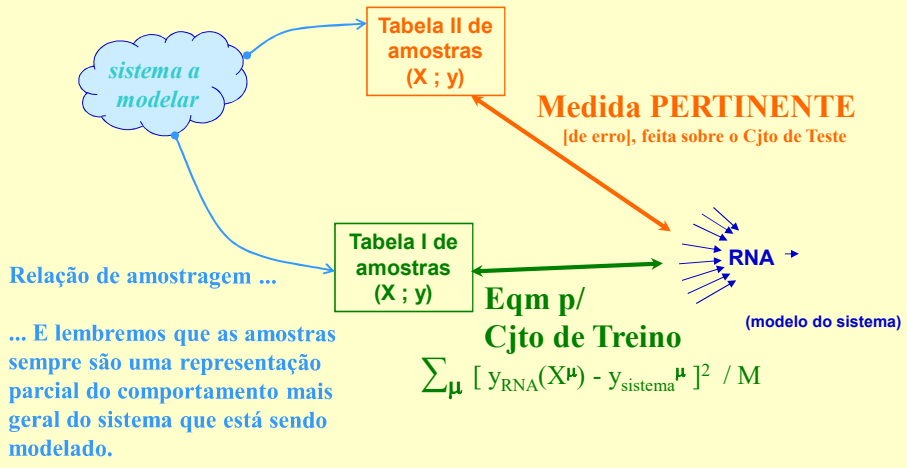
Sistema ... Amostras de treino ... RNA ...



107

107

O treinamento mira minimizar o Eqm das amostras (X ; y) de treino. A avaliação de qualidade de generalização é feita a partir de Eqm com novas amostras (X ; y) ... Estas são chamadas amostras de teste :

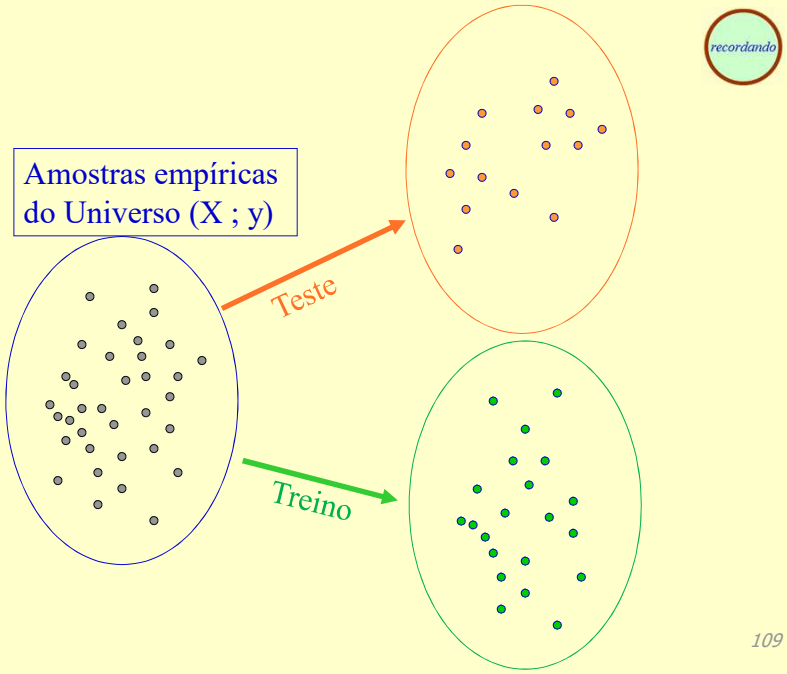


... E lembremos que as amostras sempre são uma representação parcial do comportamento mais geral do sistema que está sendo modelado.

108

© Prof. Emilio Del Moral – EPUSP

108



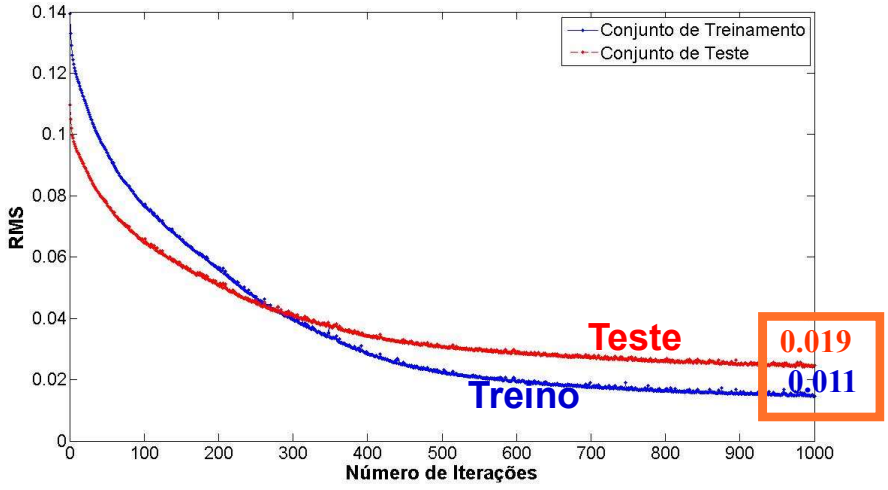
109

© Prof. Emilio Del Moral – EPUSP

109

Raiz do Erro Quadrático Médio (RMS) p/ conjuntos de treino e teste ... $RMS_{teste} > RMS_{treino}$

recordando

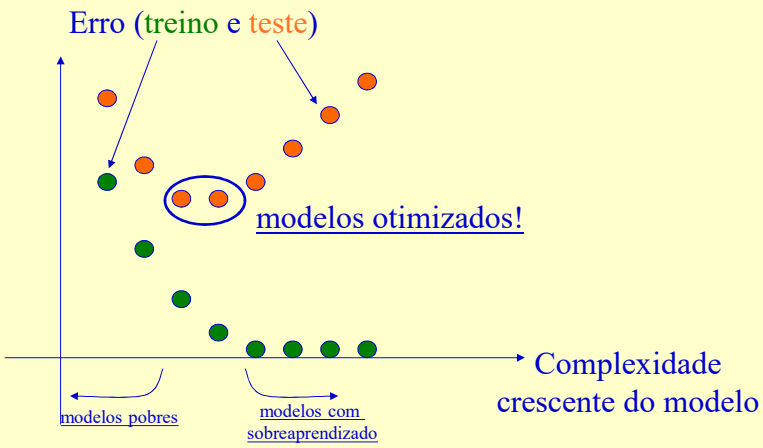


110

© Prof. Emilio Del Moral Hernandez

110

Sobreaprendizado em "sumário executivo"



124

© Prof. Emilio Del Moral – EPUSP

124