



1

Temas programados para a 5ª semana c/ Prof. Emilio

#9 (30/março – 2ªf) Foco da semana: Regressores - Casos simples de aproximação de funções univariadas. Teorema de Cybenko: o MLP como aproximador universal de funções multivariadas; implicações práticas do teorema para a implementação de regressores e reconhedores de padrões não lineares multivariados genéricos.

#10 (01/abril – 4ªf) ... Medidas de qualidade diversas para regressores multivariados (distintas do erro quadrático médio); Flutuação do desempenho do modelo com as particulares amostras de treino e de teste e técnicas de reamostragem; técnica de validação cruzada, k-fold cross validation e leave one out. Sobreajuste / sobreaprendizado / perda de generalização em regressão polinomial e em redes neurais; limitação do número de nós neurais para evitar o sobreajuste e otimizar a generalização da rede neural; partição do volume de observações em conjuntos de treino, validação e teste.

06 e 08 de abril: Semana Santa – não há aula

© Prof. Emilio Del Moral Hernandez

2

Emílio Del Moral Hernandez

Slides comentados
Prof. Emilio Del Moral Hernandez

Site do Grupo ICONE — <http://www.lsi.usp.br/ICONE/>
... e facebook: /www.facebook.com/ICONE.EPUSP/

Grupo de Inteligência Computacional, Modelagem e Neurocomputação - ICONE
Laboratório de Sistemas Inteligíveis - LSI
Escola Politécnica de USP - EPUSP
Universidade de São Paulo

18:22 Role para ver detalhes

© Prof. Emilio Del Moral Hernandez

12

Aulas remotas de PSI3471-2020
com temáticas programadas para
as semana de 30/03 e 01/04
Prof. Emilio Del Moral Hernandez/

Temas da quinta semana c/ Prof Emilio

#9 (30/março - 2#f) Foco da semana: Regressores - Casos simples de aproximação de funções univariadas. Teorema de Cybenko: o MLP como aproximador universal de funções multivariadas; implicações práticas do teorema para a implementação de regressores e reconhecedores de padrões não lineares multivariados genéricos.

#10 (01/abril - 4#f) ... Medidas de qualidade diversas para regressores multivariados (distintas do erro quadrático médio); Flutuação do desempenho do modelo com as particulares amostras de treino e de teste e técnicas de reamostragem; técnica de validação cruzada, k-fold cross validation e leave one out. Sobreajuste / sobreaprendizado / perda de generalização em regressão polinomial e em redes neurais; limitação do número de nós neurais para evitar o sobreajuste e otimizar a generalização da rede neural; partição do volume de observações em conjuntos de treino, validação e teste.

06 e 08 de abril: Semana Santa - não há aula

... nestes slides: primeira destas 2 aulas (#9 -30/03)

13

Temas da quinta semana c/ Prof. Emilio

14

Nesta 2ª feira dia 30-03-2020, abordaremos ...

#9 (30/março – 2ªf) Foco da semana: Regressores - Casos simples de aproximação de funções univariadas. Teorema de Cybenko: o MLP como aproximador universal de funções multivariadas; implicações práticas do teorema para a implementação de regressores e reconhecedores de padrões não lineares multivariados genéricos.

#10 (01/abril – 4ªf) ... Medidas de qualidade diversas para regressores multivariados (distintas do erro quadrático médio);

Flutuação do desempenho do modelo com as particulares amostras de treino e de teste e técnicas de reamostragem, técnica de validação cruzada, k-fold cross validation e leave one out. Sobreajuste / sobreaprendizado / perda de generalização em regressão polinomial e em redes neurais; limitação do número de nós neurais para evitar o sobreajuste e otimizar a generalização da rede neural; partição do volume de observações em conjuntos de treino, validação e teste.

06 e 08 de abril: Semana Santa – não há aula

© Prof. Emilio Del Moral Hernandez

14

14

O foco agora:

Regressores

$y(X)$ é analógico!

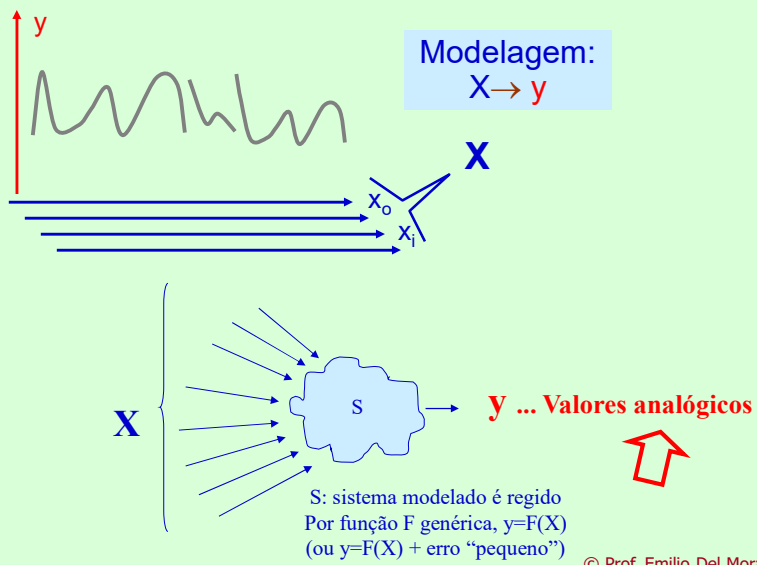
(y não é binário, nem inteiro, nem categórico)

© Prof. Emilio Del Moral – EPUSP

15

função com valores analógicos $y(X)$

16



© Prof. Emilio Del Moral Hernandez

16

16

Medidas de qualidade para regressores

17

Medindo qualidade de regressores ...

$RMS_{\text{treino/teste}}$ é uma possibilidade para refletir a distância média entre o modelo e dados empíricos ... E é já calculado facilmente em bibliotecas que operam otimização de parâmetros de modelo com base na minimização do Eqm sobre os dados empíricos
(note que o $RMS = \text{raiz quadrada do } Eqm$)

Além dessa comodidade, pode ser interpretado como o desvio padrão dos erros, ou seja tem um significado estatístico

PORÉM ... o RMS não é a única medida de qualidade para regressores e em geral NÃO é medida mais adequada para o usuário final do modelo!

18

© - Prof. Emilio Del Moral Hernandez

18

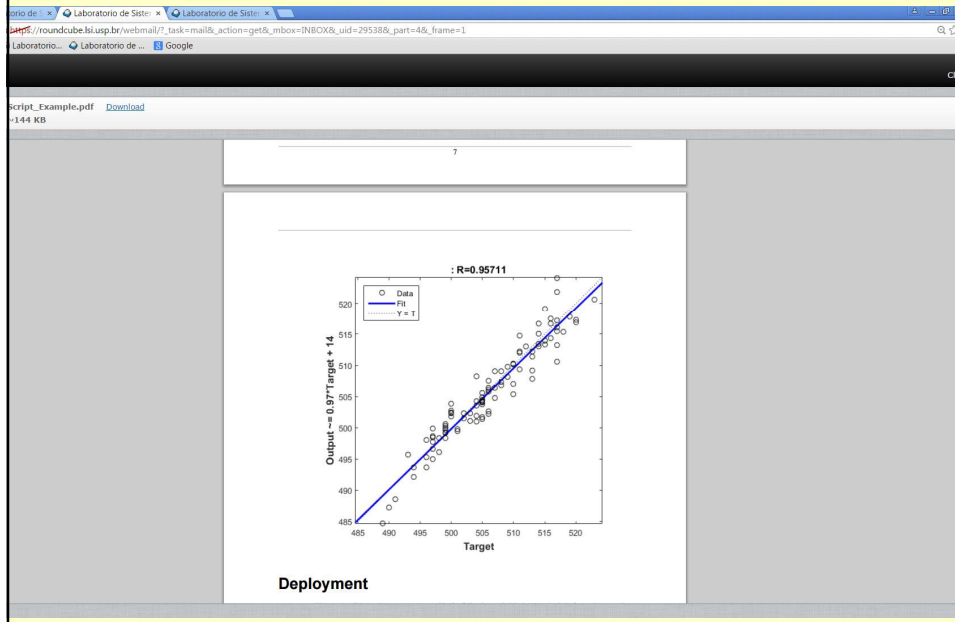
Algumas potenciais medidas de qualidade em regressores (cada aplicação pode preferir umas ou outras ...)

- ~~Eqm e RMS~~
- Módulo Médio do Erro
- Máximo Módulo do Erro
- Média dos Erros Positivos
- Máximo Módulo dos Erros Positivos
- Média dos Erros Negativos
- Máximo Módulo dos Erros Negativos
- Esses todos derivados dos erros acima, mas em suas versões normalizadas, com relação ao módulo de y
- Estes todos acima, mas sujeitos a conhecimento de X (local) – qualidade dependente do valor do argumento X da regressão
- Histograma de erros (ou seja, a “densidade de probabilidade de erros empírica”)
- Faixa de valores de erros que se enquadram num certo número de “deciles” – ou terciles, ou quartiles, etc etc – da distribuição de erros (seja sobre a distribuição o erro com sinal $+$ $-$, ou seja sobre o erro em módulo)
- Combinações específicas de vários acima ... Como bem percebido por colegas em sala, várias combinações fazem muito sentido, como informação de caracterização mais completa ao cliente / usuário do regressor!

11

19

Medindo qualidade de regressores ... Plots y corretos versus y estimados + valor de "R" é outra possibilidade



20

Coefficiente de correlação de Pearson ...

pt.wikipedia.org/wiki/Coefficiente_de_correla%C3%A7%C3%A3o_de_Pearson

Coefficiente de correlação de Pearson

Origem: Wikipédia, a enciclopédia livre.

Esta página ou secção cita fontes fiáveis e independentes, mas que **não cobrem** todo o conteúdo, o que **compromete a verificabilidade** (desde Agosto de 2011). Por favor, insira mais referências no texto. Material sem fontes poderá ser removido.
—Encontre fontes: Google (notícias, livros e acadêmico)

Em estatística descritiva, o **coeficiente de correlação de Pearson**, também chamado de "coeficiente de correlação produto-momento" ou simplesmente de "rho de Pearson" mede o grau da correlação (e a direção dessa correlação - se positiva ou negativa) entre duas variáveis de escala métrica (intervalar ou de rácio/razão).

Este coeficiente, normalmente representado por ρ assume apenas valores entre -1 e 1.

- $\rho = 1$ Significa uma correlação perfeita positiva entre as duas variáveis.
- $\rho = -1$ Significa uma correlação negativa perfeita entre as duas variáveis - Isto é, se uma aumenta, a outra sempre diminui.
- $\rho = 0$ Significa que as duas variáveis não dependem linearmente uma da outra. No entanto, pode existir uma dependência não linear. Assim, o resultado $\rho = 0$ deve ser investigado por outros meios.

Índice [esconder]

- Cálculo
- Interpretando ρ^n
- Interpretação geométrica
- Referências
- Ver também

Cálculo [editar] [editar código-fonte]

Calcula-se o coeficiente de correlação de Pearson segundo a seguinte fórmula:

$$\rho = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^n (x_i - \bar{x})^2} \cdot \sqrt{\sum_{i=1}^n (y_i - \bar{y})^2}} = \frac{\text{cov}(X, Y)}{\sqrt{\text{var}(X) \cdot \text{var}(Y)}}$$

onde x_1, x_2, \dots, x_n e y_1, y_2, \dots, y_n são os valores medidos de ambas as variáveis. Para além disso

$$\bar{x} = \frac{1}{n} \cdot \sum_{i=1}^n x_i$$

e

21

- 2 Interpretando ρ
- 3 Interpretação geométrica
- 4 Referências
- 5 Ver também

Cálculo [editar | editar código-fonte]

Calcula-se o coeficiente de correlação de Pearson segundo a seguinte fórmula:

$$\rho = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^n (x_i - \bar{x})^2} \cdot \sqrt{\sum_{i=1}^n (y_i - \bar{y})^2}} = \frac{\text{cov}(X, Y)}{\sqrt{\text{var}(X) \cdot \text{var}(Y)}}$$

onde x_1, x_2, \dots, x_n e y_1, y_2, \dots, y_n são os valores medidos de ambas as variáveis

$$\bar{x} = \frac{1}{n} \cdot \sum_{i=1}^n x_i$$

e

Coeficiente de correlação de Pearson ...

22

© - Prof. Emilio Del Moral Hernandez

22

Pergunta: O Eqm (ou RMS) indica a qualidade do modelo? *Onde usar Eqm ou onde não?*

Algumas potenciais medidas de qualidade em regressores (cada aplicação pode preferir umas ou outras ...)

- Eqm e RMS
- Módulo Médio do Erro
- Máximo Módulo do Erro
- Média dos Erros Positivos
- Máximo Módulo dos Erros Positivos
- Média dos Erros Negativos
- Máximo Módulo dos Erros Negativos
- Esses todos derivados dos erros acima, mas em suas versões normalizadas, com relação ao módulo de y
- Estes todos acima, mas sujeitos a conhecimento de X (local) – qualidade dependente do valor do argumento X da regressão
- Histograma de erros (ou seja, a “densidade de probabilidade de erros empírica”)
- Faixa de valores de erros que se enquadram num certo número de “deciles” – ou terciles, ou quartiles, etc etc – da distribuição de erros (seja sobre a distribuição o erro com sinal + -, ou seja sobre o erro em módulo)
- Combinações específicas de vários acima ... Como bem percebido por colegas em sala, várias combinações fazem muito sentido, como informação de caracterização mais completa ao cliente / usuário do regressor!

11

PSI2672 – Reconhec. de Padrões, Modelagem e Redes Neurais – Prof. Emilio Del Moral Hernandez – © 2015

mandez

23

Aproximação univariada

24

... Um parênteses para discutirmos um pouco a aproximação universal usando tangentes hiperbólicas, sigmóides, etc ... (funções em formato de “S”)

Abordemos o caso simples e bem particular de função escalar univariada, ou seja, pensemos sobre a aproximação de uma função de

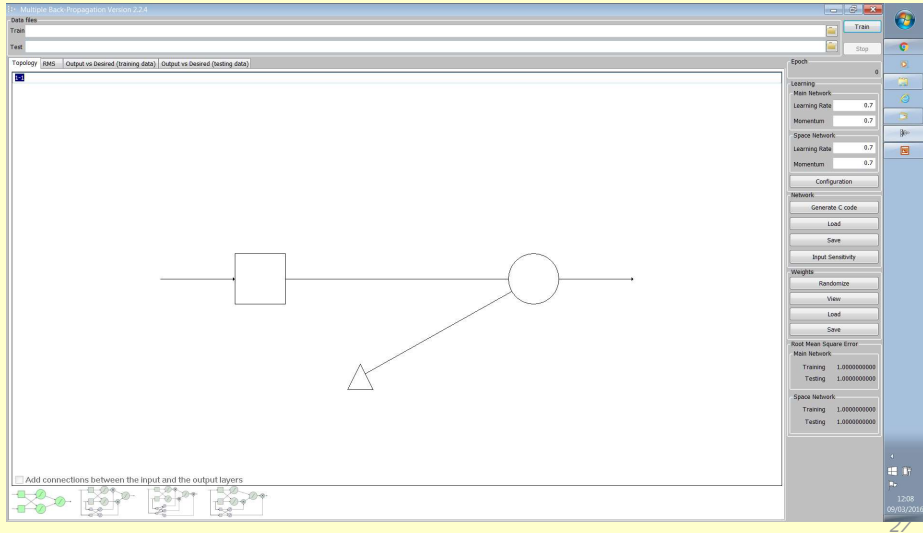
uma única variável x_1 : $y(x_1)$

25

© Prof. Emilio Del Moral – EPUSP

25

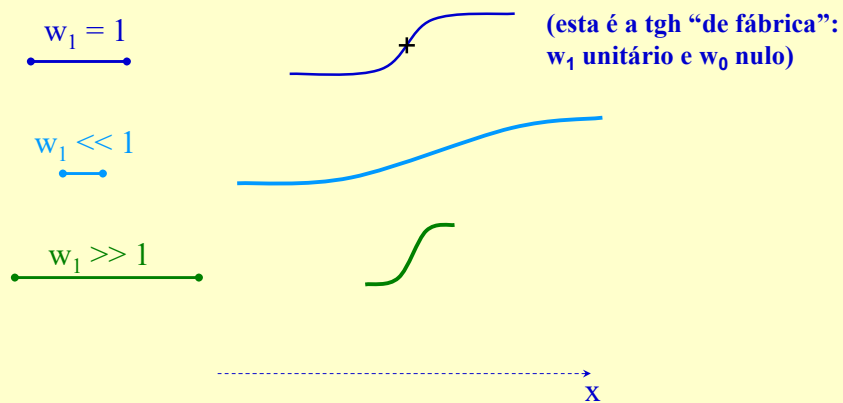
O que conseguimos fazer com **um único neurônio sigmoidal**, no caso de regressões (“y contínuo”)?



© Prof. Emilio Del Moral – EPUSP

27

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



28

© Prof. Emilio Del Moral – EPUSP

28

... e que tipo de dados empíricos conseguimos modelar com um único neurônio sigmoidal em regressões “ $y(x_1)$ ”?

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

29

© Prof. Emilio Del Moral – EPUSP

29

Em termos de Excel, teríamos ... recordando

Cliente (μ)	Idade (x_1)	Renda (x_2)	Clics (x_3)	Consumo do Produto B (x_4)	Consumo do Produto C (x_5)	Consumo do Produto A (y)
			302	958	136	9800
			186	985	196	8760
						520
						11640
						9640
						5320

*Equivalente em txt
Para uso do MBP*

```

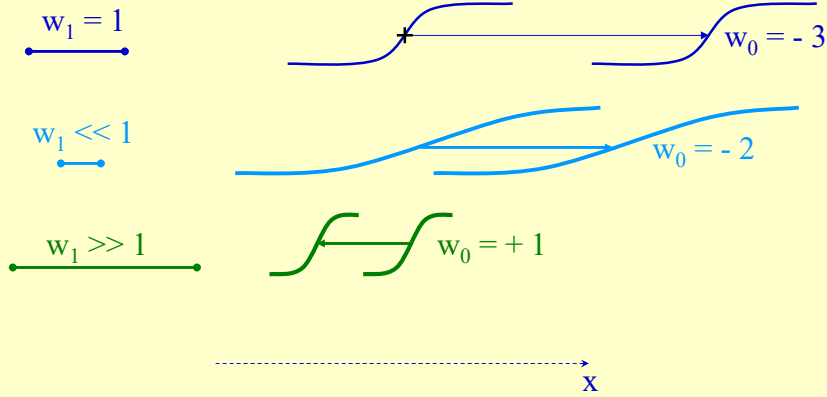
treino em txt para exemplo de consumo A e B - Bloco de notas
Arquivo Editar Formatar Exibir Ajuda
Idade Renda Clics ConsumoA ConsumoB ConsumoA
M-2 16 50 78 302 958 136 9800
M-1 30 65 128 186 985 196 8760
M 19 57 150 221 1093 35 520
(...)
16 19 51 707 131 11640
30 75 7 29 78 9640
19 47 116 285 124 5320
    
```

75

Moral – EPUSP

30

O que conseguimos fazer com **um único neurônio sigmoidal** $y(w_1 \cdot x_1 + w_0 \cdot 1)$, c/ escalamento de x_1 via w_1 ... e agora também com o viés, via viés w_0

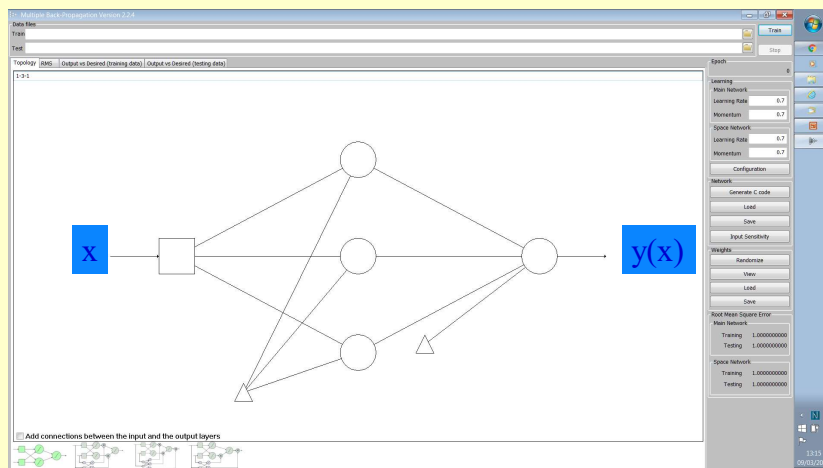


31

© Prof. Emilio Del Moral – EPUSP

31

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



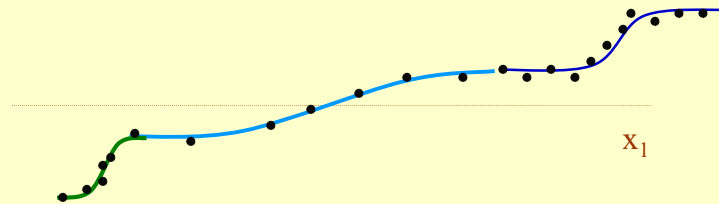
34

© Prof. Emilio Del Moral – EPUSP

34

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?

35

© Prof. Emilio Del Moral – EPUSP

35

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?

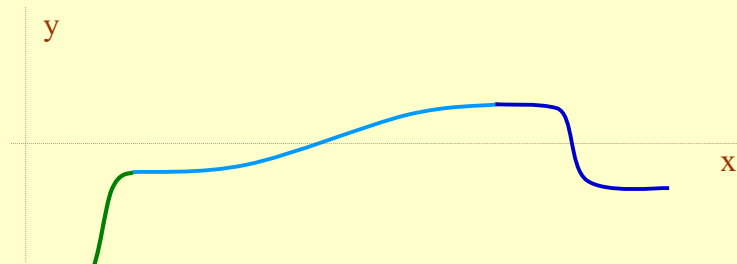
36

© Prof. Emilio Del Moral – EPUSP

36

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

... superposição de várias sigmóides deslocadas e escaladas:



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

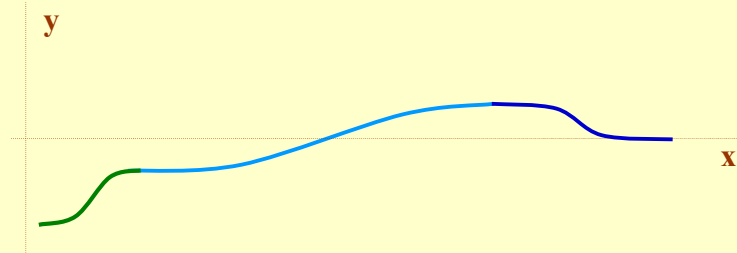
37

© Prof. Emilio Del Moral – EPUSP

37

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

... superposição de várias sigmóides deslocadas e escaladas:



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

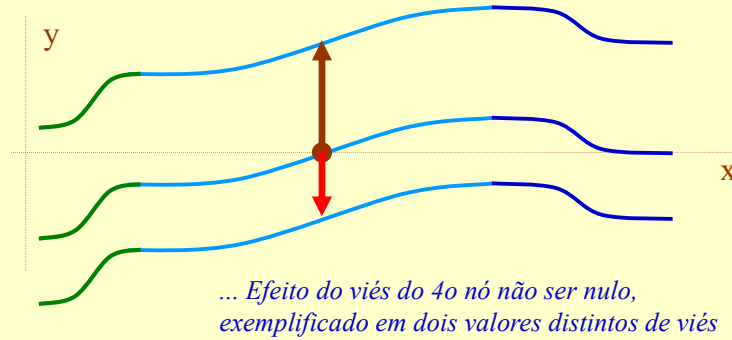
38

© Prof. Emilio Del Moral – EPUSP

38

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

... superposição de várias sigmóides deslocadas e escaladas:

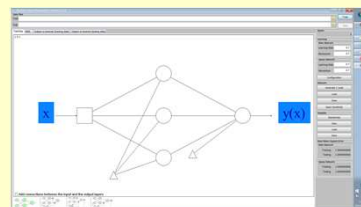
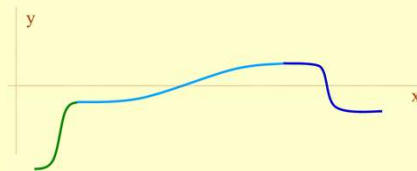


39

© Prof. Emilio Del Moral – EPUSP

39

A discussão anterior 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.



40

© Prof. Emilio Del Moral – EPUSP

40

A) 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 Estimacão / Regressão Contínua (um dos alvos de aplicacão deste curso) !!!

E ...

É também bom para o Reconhecimento de Padrões (outro alvo de modelagem)

41

© Prof. Emilio Del Moral – EPUSP

41

Cybenko – Teorema;

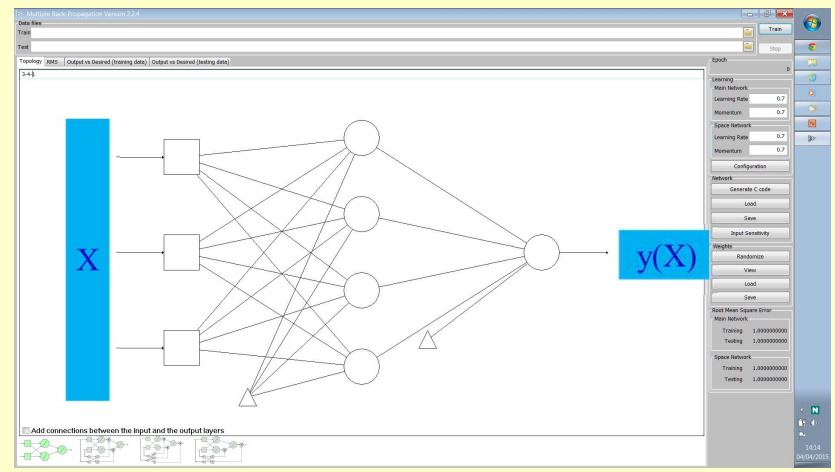
*Enunciado, Premissas,
“Sumário Executivo”*

O objetivo não é demonstrá-lo, mas entendê-lo

42

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

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



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

The screenshot shows the Wikipedia article for the 'Universal approximation theorem'. The title is 'Cybenko – Enunciado da Prova ... (premissas + resultado)'. The article text includes:

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 \mathbb{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 $\epsilon > 0$, there exist an integer N and real constants $\alpha_i, b_i \in \mathbb{R}, w_i \in \mathbb{R}^m$ where $i = 1, \dots, N$ such that we may define:

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

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

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

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

References [edit]

- ^[1] Balazs Csornai Csajj. Approximation with Artificial Neural Networks. Faculty of Sciences, Eötvös Loránd University, Hungary
- ^[2] G. Cybenko. (1989) "Approximation by superpositions of sigmoid functions" *Journal of Mathematical Control, Signals, and Systems*, 2 (4), 303-314
- ^[3] 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 0-13-27350-1
- ^[5] Hameed, M. (1995) *Fundamentals of Artificial Neural Networks* MIT Press, p. 43

⚠ 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 20:05.

Fwd: Proposta ...eml | Alterações vag...doc | Mostrar todos os downloads...

© Prof. Emilio Del Moral Hernandez

54

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.

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 $C(I_m)$ and $\epsilon > 0$, there exist an integer N and real constants $\alpha_i, b_i \in \mathbb{R}, w_i \in \mathbb{R}^m$ where $i = 1, \dots, N$ such that we may define:

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

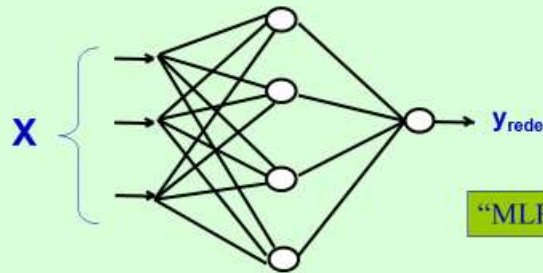
as an approximate realization of the function f where f is independent of φ , that is,

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

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

56

Rede de uma única camada de neurônios sigmoidais + um neurônio de saída



57

Kurt Hornik showed in 1991^{[2][3][4][5]} that it is not tr assumed to be linear. For notational conven

Formal statement [edit]

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

$y_{rede}(X)$

X

Let $\varphi(\cdot)$ be a nonconstant, bounded, and monotonically-increas $C(I_m)$ and $\epsilon > 0$, there exist an integer N and real constants α_i

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

número de nós escondidos

sigmoidal

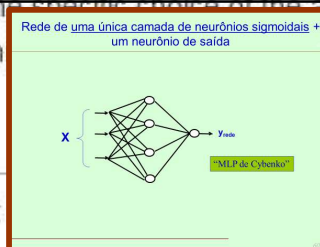
viés_i : viés do nó escondido i

W_i : vetor de pesos do nó escondido i

elementos do vetor de pesos do nó linear de saída W_s

as an approximate of the function f where f is indeed

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



59

Kurt Hornik showed in 1991^[2] that it is not the specific choice of the activation function assumed to be linear. For notational convenience, only the single output is shown.

Formal statement [\[edit\]](#)

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

Let $\phi(\cdot)$ be a nonconstant, bounded, and monotonically-increasing continuous function on $C(I_m)$ and $\epsilon > 0$, there exist an integer N and real constants a_i, b_i such that

$$F(x) = \sum_{i=1}^N a_i \phi(T_i x + b_i)$$

as an approximate realization of the function f where f is independent of x .

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

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

$Y_{\text{rede}}(X)$

Fescondida sistema(X)

Limite de erro

61

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

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

Mathematics of Control, Signals, and Systems
© 1989 Springer-Verlag New York Inc.

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 functions 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, Complexity.

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

$$\sum_{j=1}^m a_j \sigma_j^T x + \theta_j \quad (1)$$

where $y_j \in \mathbb{R}^n$ and $a_j, \theta_j \in \mathbb{R}$ are fixed, $(\sigma_j^T)^T$ is the transpose of y_j so that $y_j^T x$ is the inner product of y_j 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:

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

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-861903, ONR Contract N000146-86-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.

310

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 Gamma 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 a bounded measurable sigmoidal function. Then finite sums of the form

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

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

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

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 $L^1(I_n)$ is the dual of $L^\infty(I_n)$. 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 $\epsilon > 0$, there is a finite sum of the form

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

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

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

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

n, Christopher Chase, Lee narov, Richard Lippmann, 'tences, and improvements

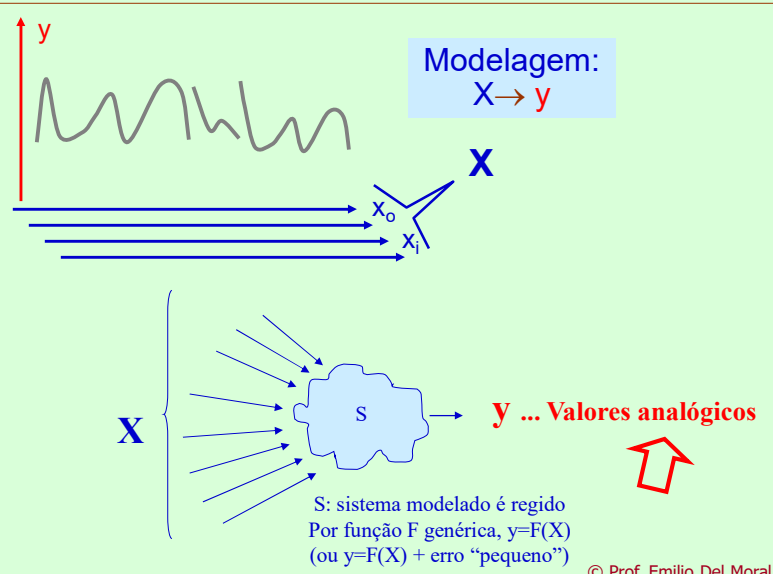
New York, 1972. uralization?, *Neural Comput.* (to stems and control, *IEEE Control* \. Classifying learnable geometric rdings of the 18th Annual ACM p. 273-282. 1 and the Pompeiu problem, *Ann.* ets using the Radon transform, EE Trans. *Acoust. Speech Signal* stward networks are universal a Neural Net and Conventional 87. mal Classifiers, Technical Report, -475. tworks by sigmoidal functions, a, University of Lowell, 1988.

63

De onde vem o grande poder do MLP em resolver tantos problemas diversos?

A prova de Cybenko nos respondeu essa dúvida!!

função com valores analógicos y(X)



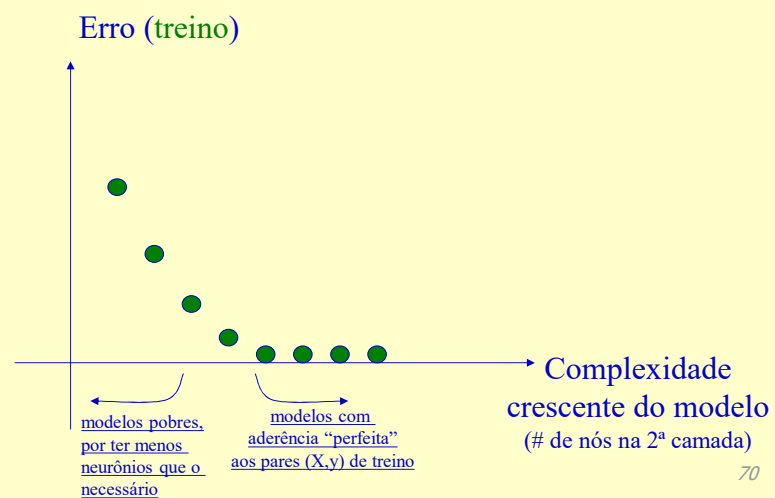
Questões intrigantes, p/ esta aula e p/ pensar em casa ...

- No que impacta escolhermos o “epsilon” de Cybenko de alto valor? O que muda na estrutura de Cybenko com isso?
- No que impacta escolhermos o “epsilon” 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?

© Prof. Emilio Del Moral Hernandez

69

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



© Prof. Emilio Del Moral – EPUSP

70

Resumindo, Cybenko foi além de mostrar a viabilidade de aproximação em casos unidimensionais; e foi além de mostrar a viabilidade de fronteiras de separação genéricas.

Qualquer Função(X) genérica pode ser aproximada por um MLP

– Isso é bom para Estimação / Regressão Contínua (um alvo neste curso) !!!

- É também bom para o Reconhecimento de Padrões (outro alvo de modelagem do curso)

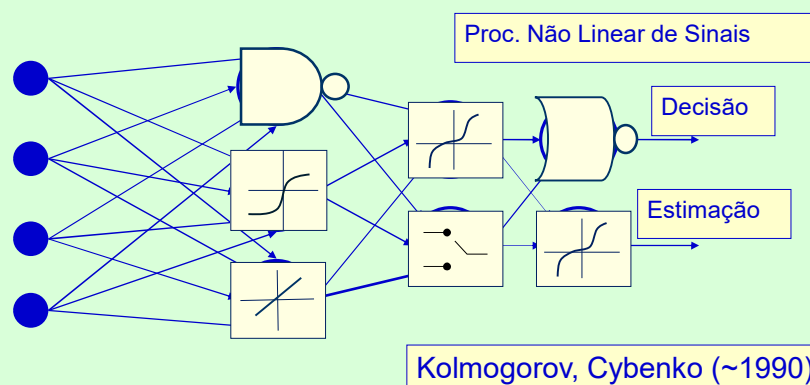
88

© Prof. Emilio Del Moral – EPUSP

88

O Multi Layer Perceptron (MLP)

- Múltiplas entradas / Múltiplas saídas / Múltiplas camadas
- Variáveis (internas e externas) analógicas ou digitais
- Relações lineares ou não lineares entre elas



Prof. Emilio Del Moral Hernandez

90

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

- Segundo Cybenko, quando queremos um aproximador universal de funções com um MLP que tenha os ingredientes da prova que ele demonstrou, que funções de ativação (função não linear do nó neural) podem ou não podem ser usadas na implementação das camadas internas?
- Podemos usar a Tgh, a Sigmóide, a Linear, a Linear por etapas (todas dessa classe ou só algumas?), a Relu, a função Sinal, a função Degrau, a Gaussiana?
(basta checar quais obedecem às premissas de Cybenko e quais não)

Emílio Del Moral Hernandez

*Nos falamos em breve ...
Prof. Emilio Del Moral Hernandez*

18:22

Role para ver detalhes

© Prof. Emilio Del Moral Hernandez