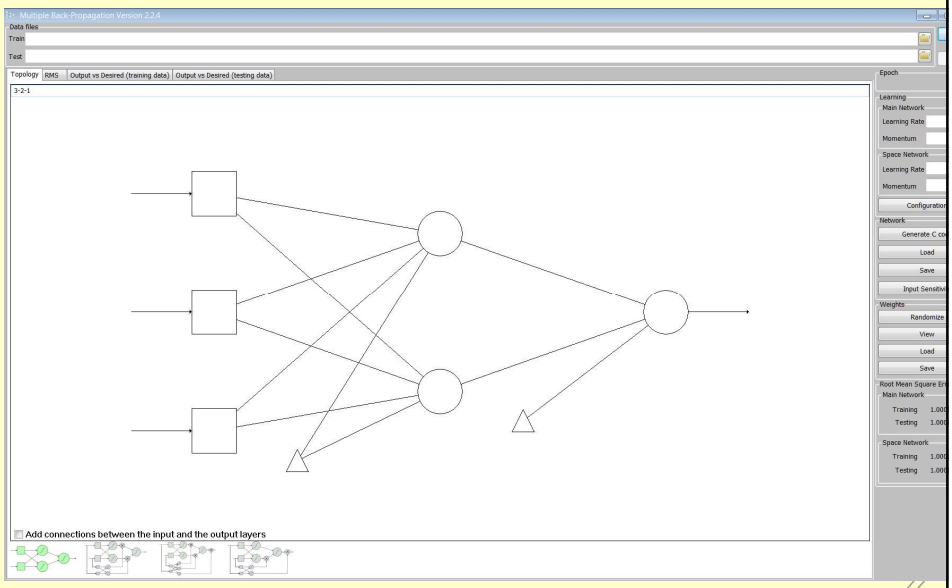
Aprendizado do MLP por Error Back Propagation ...

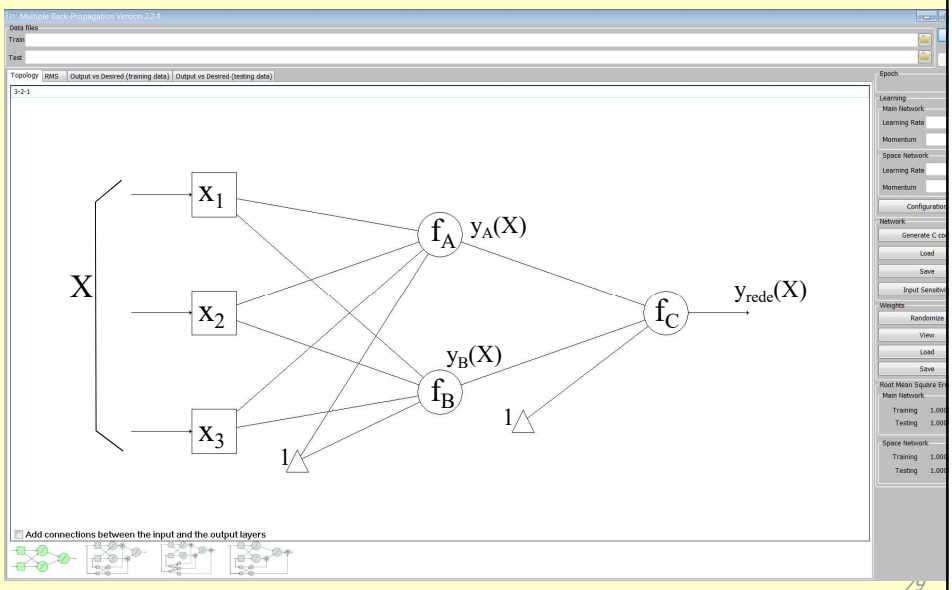
$$\Delta\vec{W} = -\eta.\vec{\nabla}Eqm\_$$

Gradiente de Eqm no espaço de pesos = ($\partial Eqm(W)/\partial w_1$, $\partial Eqm(W)/\partial w_2$, $\partial Eqm(W)/\partial w_3$, ...)

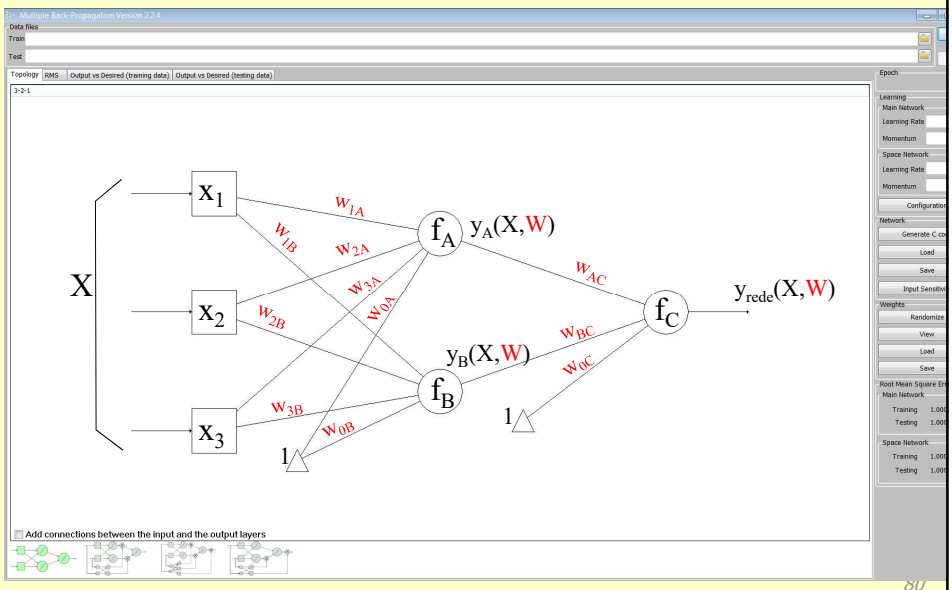
Chegando às fórmulas das derivadas parciais, necessárias à Bússola do Gradiente

Relembrando o que está por trás de um desenho como o que segue ...

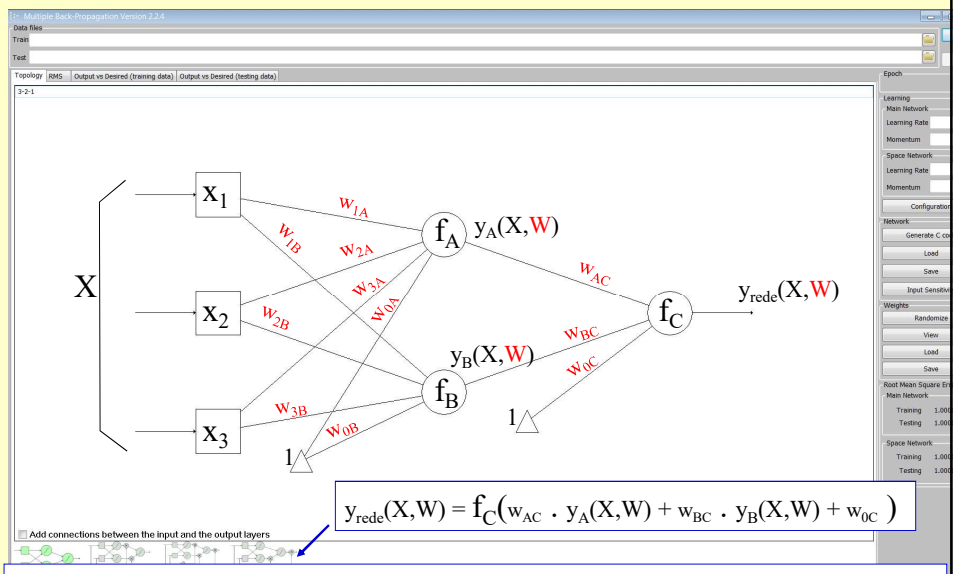




79

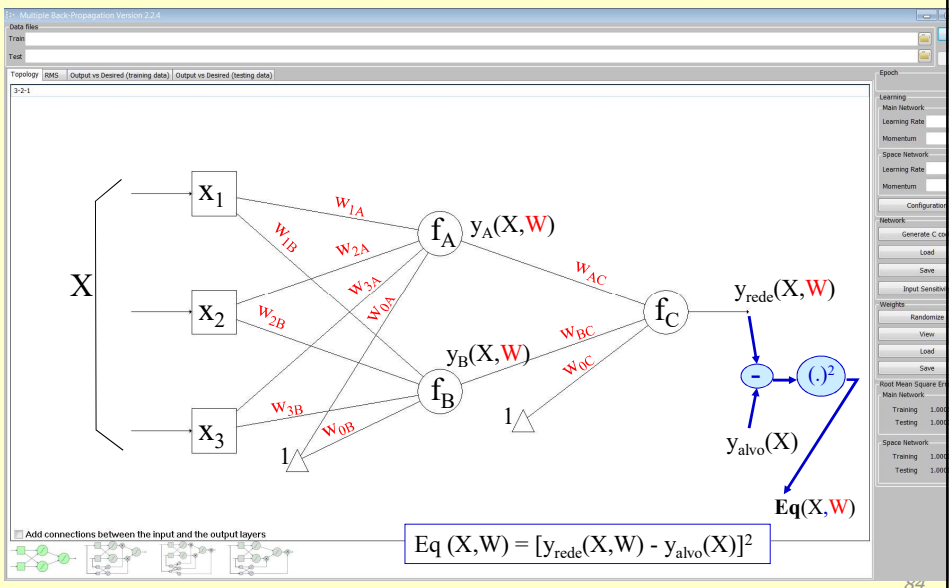


80



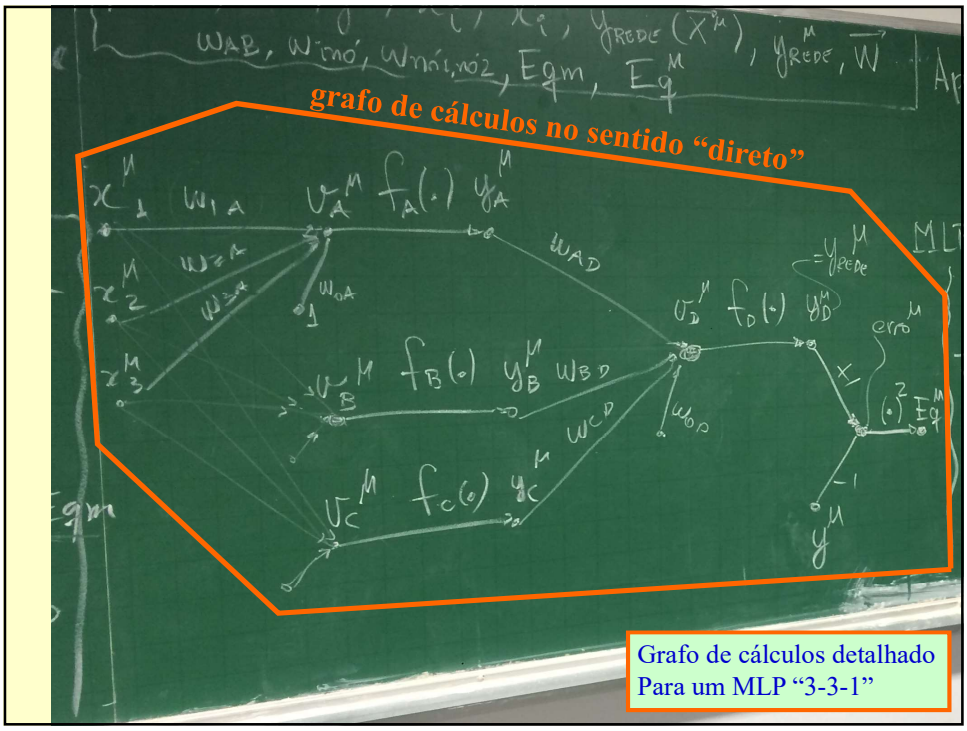
$$y_{rede}(X,W) = f_C\left(w_{AC} \cdot y_A(X,W) + w_{BC} \cdot y_B(X,W) + w_{0C}\right)$$

$$y_{rede}(X,W) = f_C\left(w_{AC} \cdot f_A(w_{1A}.x_1 + w_{2A}.x_2 + \ldots + w_{0A}) + w_{BC} \cdot f_B(w_{1B}.x_1 + w_{2B}.x_2 + \ldots + w_{0B}) + w_{0C}\right)$$



$$Eq(X,W) = [y_{rede}(X,W) - y_{alvo}(X)]^2$$

84



Grafo de cálculos detalhado
Para um MLP "3-3-1"

Chamada oral sobre a lição de casa: estudar / reestudar os conceitos e a parte operacional de derivadas parciais, do vetor Gradiente ...

- **Derivadas parciais (que são as componentes do gradiente):**

$\partial f(a,b,c)/\partial a \qquad \partial f(a,b,c)/\partial b \qquad \partial f(a,b,c)/\partial c$

- **Vetor Gradiente, útil ao método do máximo declive:**

$(\partial Eqm(W)/\partial w_1\ ,\ \partial Eqm(W)/\partial w_2,\ \partial Eqm(W)/\partial w_3,\ \ldots)$ $\qquad \Delta\vec{W} = -\eta.\vec{\nabla}Eqm\_$

Invertamos o operador gradiente e a somatória

.. afinal, gradiente é uma derivada, e a derivada de um soma de várias funções é igual à soma das derivadas individuais de cada componente da soma:

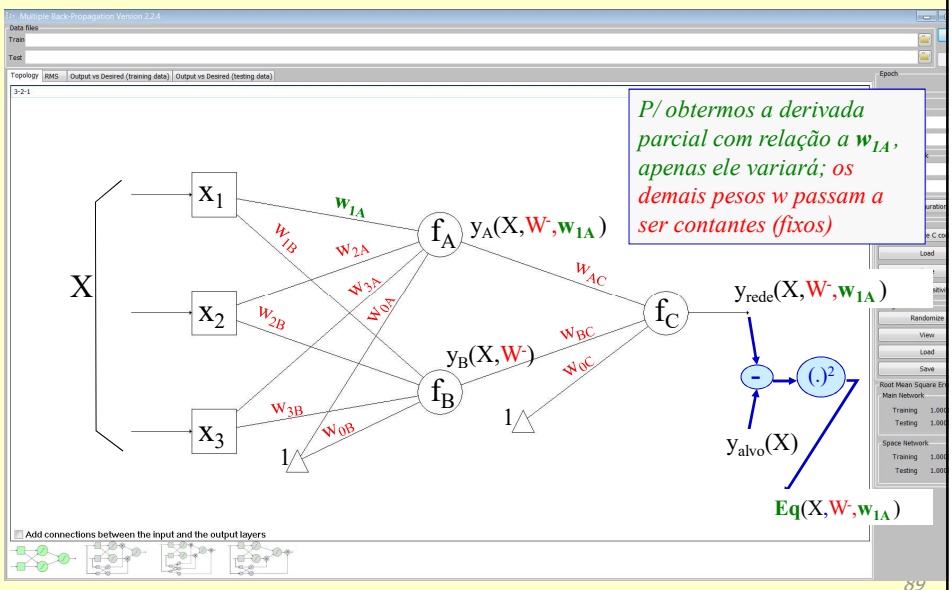
$$\mathbf{Grad(Eqm) =}$$

$$\mathbf{Grad(\Sigma_{\mu} Eq^{\mu}) / M}$$

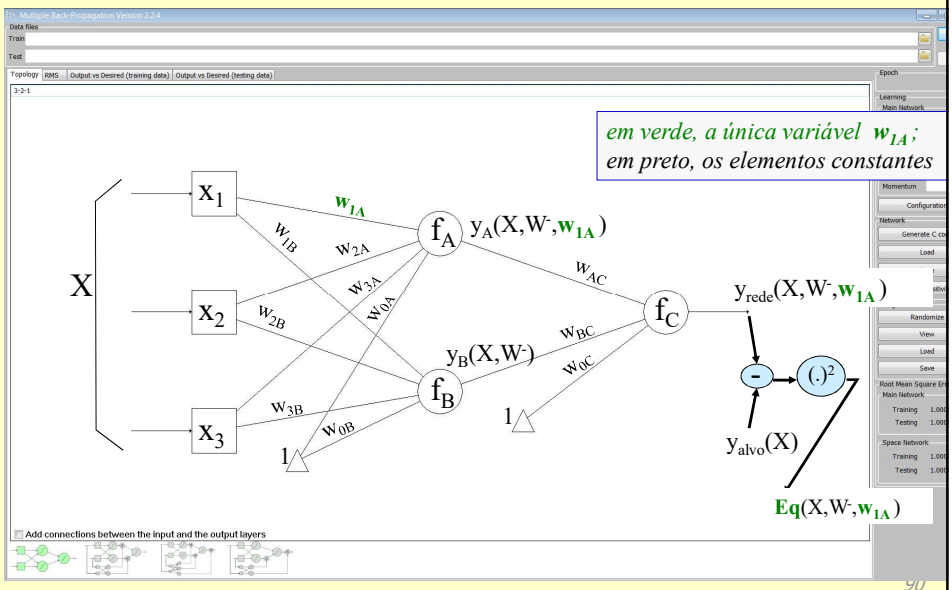
$$\mathbf{\Sigma_{\mu} Grad (Eq^{\mu}) / M}$$

Note que a inversão do gradiente com a somatória nada mais é que usar de forma repetida – e em separado para cada dimensão do vetor ***Grad(Σ_{μ} Eq$^{\mu}$)*** – a seguinte propriedade simples e sua velha conhecida ...

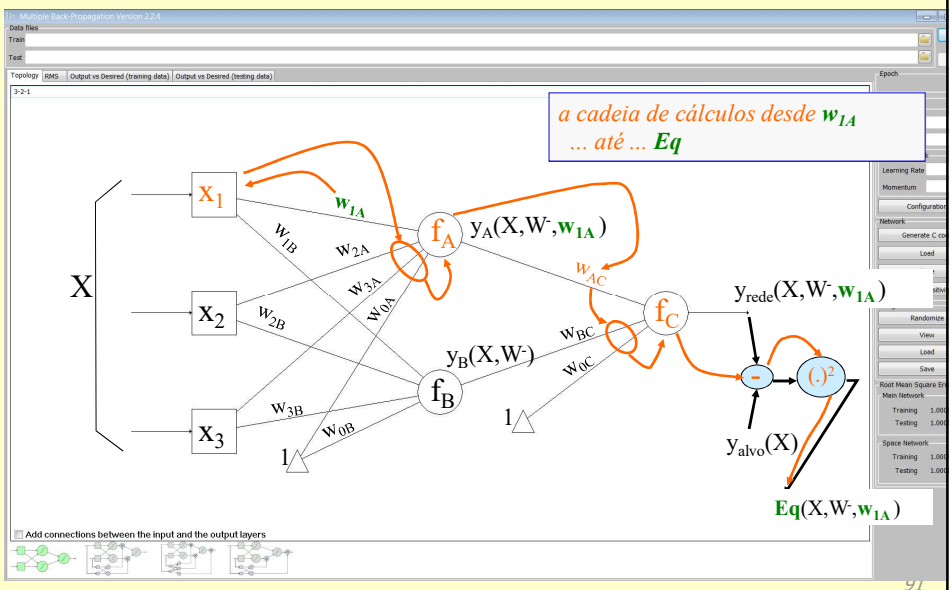
$$\mathbf{d(f_1(x)+f_2(x)) / dx = df_1(x)/dx + df_2(x)/dx}$$

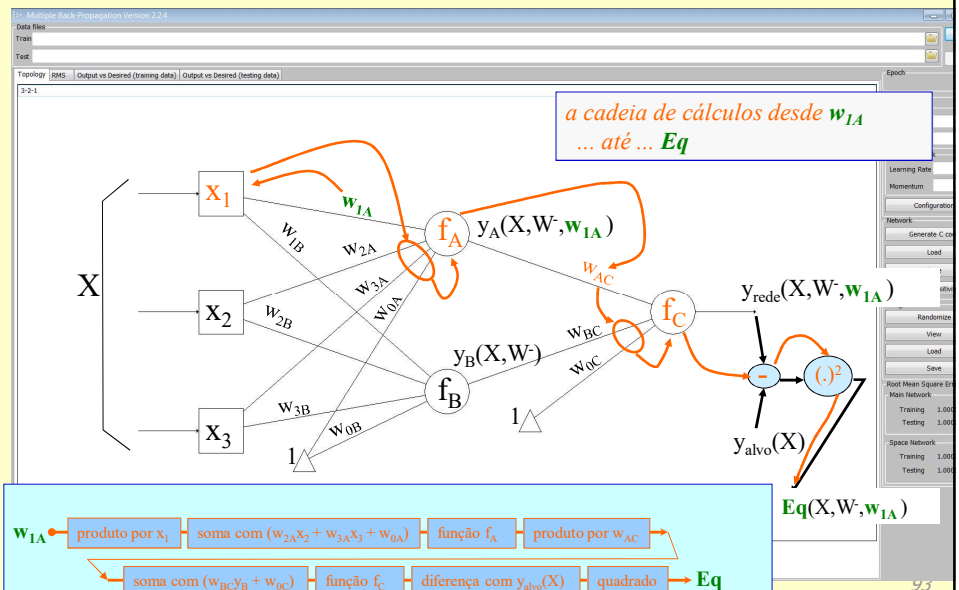


89



90

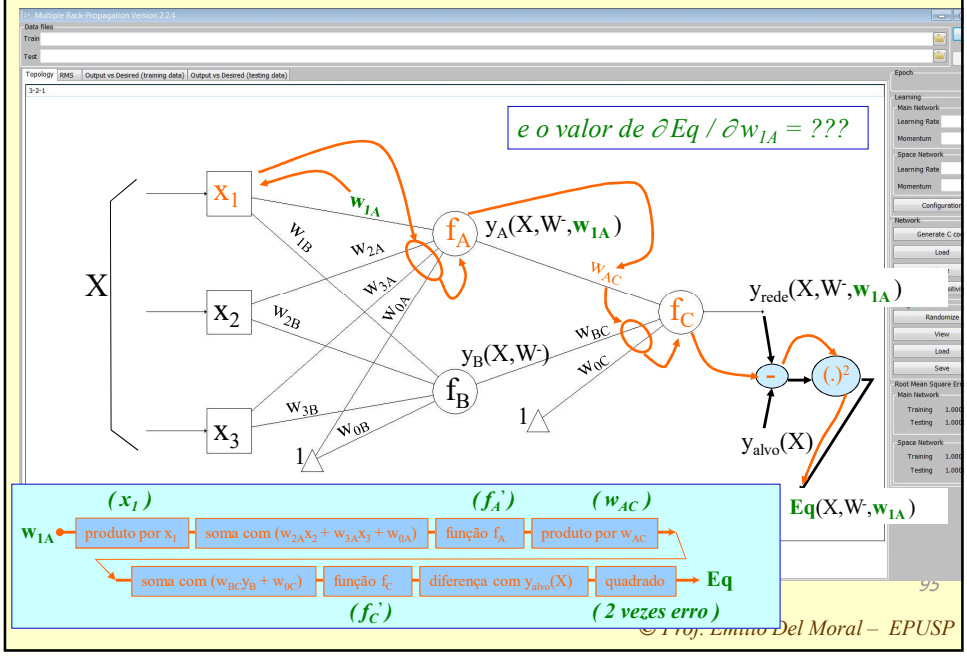


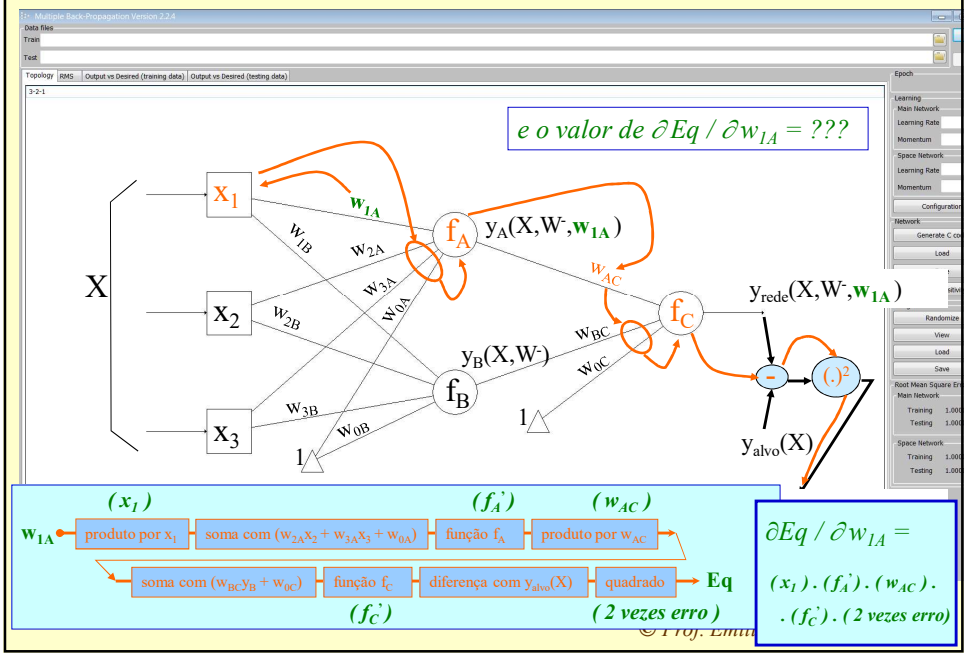


Note que aqui temos uma cadeia com muitos estágios que levam da váriável w_{1A}, à variável Eq^{μ}, e para a qual podemos calcular a derivada da saída (Eq^{μ}) com relação à entrada (w_{1A}) aplicando de forma repetida a seguinte propriedade simples e sua velha conhecida ...

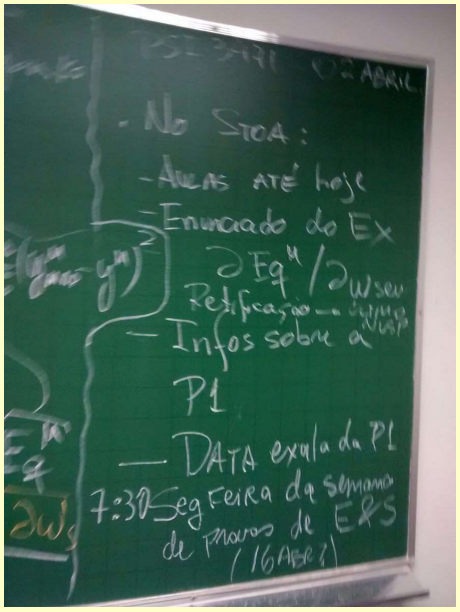
$$\mathbf{d(f_1(f_2(x)) / dx \;\; = \;\; df_1(x)/df_2 \, . \, df_2(x)/dx}$$

..., ou seja, calculando isoladamente o valor da derivada para cada estágio da cadeia, e finalizando o cálculo de derivada de ponta a ponta nessa cadeia toda através do produto dos diversos valores de cada estágio.



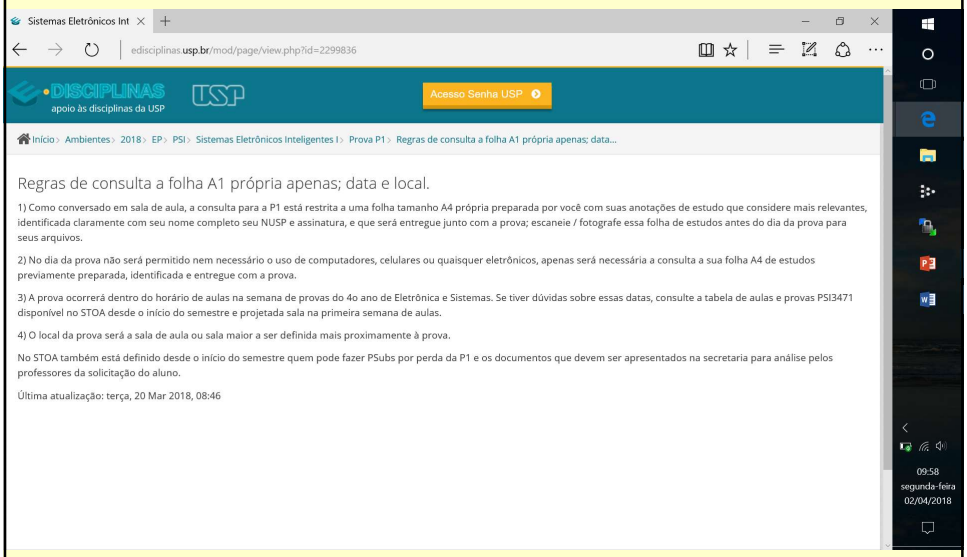


Avisos que detalhei na aula de 02 de abril



A nossa P1 ocorrerá no horário de aula de PSI3471 e na semana de P1 (de E&S), como no calendário divulgado no início do semestre; mais precisamente, ocorrerá no dia 16 de abril, 2ª feira às 7:30 (com duração prevista de 90 mins).

Avisos que comentei na aula de 02 de abril



Lembretes ....

- Na maioria dos slides anteriores, onde aparece X, leia-se X^{μ}, não incluído para não complicar demais os desenhos
- ... similarmente, onde aparece y_{alvo}, leia-se $y_{alvo}{}^{\mu}$. Idem para os Eq, leia-se Eq^{μ}
- Nos itens de cadeia de derivadas (f_A') e (f_C') , atenção para os valores dos argumentos, que devem ser os mesmos de f_A e f_C na cadeia original que leva w_{1A} a Eq.
- ... lembrando ... na cadeia original tinhamos ...
  - para f_C: $f_C(w_{AC} \cdot f_A(w_{1A}.x_1 + w_{2A}.x_2 + ... + w_{0A}) + w_{BC} \cdot f_B(w_{1B}.x_1 + w_{2B}.x_2 + ... + w_{0B}) + w_{0C})$
  - para f_A: $f_A(w_{1A}.x_1 + w_{2A}.x_2 + ... + w_{0A})$
- Similarmente, para o bloco "quadrado", cuja derivada é a função "2 vezes erro", o argumento é $[y_{rede}(X,W) - y_{alvo}(X)]$

Lembretes ....

- O mesmo que foi feito para w_{1A} deve ser feito agora para os demais 10 pesos: $w_{2A}, w_{3A}, w_{0A}, w_{1B}, w_{2B}, w_{3B}, w_{0B}, w_{AC}, w_{BC}$, e w_{0C} !
- Assim compomos um gradiente de 11 dimensões, com as derivadas de Eq^{μ} com relação aos 11 diferentes pesos w: $Grad_W(Eq^{\mu})$
- Essas 11 fórmulas devem ser aplicadas repetidamente aos M exemplares numéricos de X^{μ} e y_{alvo}^{μ}, calculando M gradientes!
- Com eles, se obtém o gradiente médio dos M pares empíricos: $Grad_W(Eqm) = [\Sigma_{\mu} Grad_W(Eq^{\mu})] / M$
- Esse gradiente médio é a Bussola do Gradiente!

Método do Gradiente Aplicado aos nossos MLPs: a partir de um W#0, temos aproximações sucessivas ao Eqm mínimo, por repetidos pequenos passos DeltaW, sempre contrários ao gradiente ...

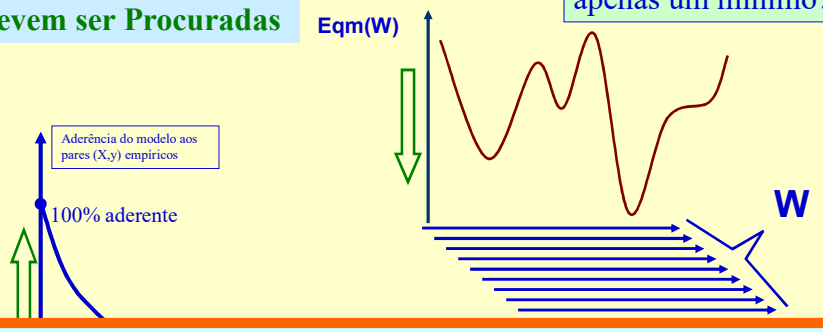
- "Chute" um W inicial para o "Wcorrente", ou "W melhor até agora"
- Em loop até obter Eqm zero, ou baixo o suficiente, ou estável:
  - Determine o vetor gradiente do Eqm, nesse espaço de Ws
  - Em loop varrendo todos os M exemplos ($X^{\mu};y^{\mu}$),
    - Calcule o gradiente de Eq^{μ} associado a um exemplo μ, e vá varrendo μ e somando os gradientes de cada Eq^{μ}, para compor o vetor gradiente de Eqm, assim que sair deste loop em μ ;
    - Cada cálculo como esse, envolve primeiro calcular os argumentos de cada tangente hiperbólica e depois usar esses argumentos na regra da cadeia das derivadas necessárias
  - Dê um passo Delta ΔW nesse espaço, com direção e magnitude dados por $-\eta$*vetor gradiente médio para os M Exemplos ($X^{\mu};y^{\mu}$) de treino

O que devemos mirar quando exploramos o espaço de pesos W buscando que a RNA seja um bom modelo?

Devemos mirar Maximização da aderência = Minimo Eqm possível

As Setas Verdes Indicam Situações que Devem ser Procuradas

Será que temos apenas um mínimo??



Atacando a possibilidade de múltiplos mínimos locais:

... Executamos o gradiente descendente repetidamente, com a mesma arquitetura neural e os mesmos dados empíricos de treino, mas com os pesos iniciais randômicos sendo distintos em cada rodada; anotamos todos os pesos otimizados e valores de Eqm final de todas as rodadas; mantemos os w´s do melhor dos resultados de treino entre eles (aquele ensaio com menor valor do Eqm)!

O Ciclo completo da modelagem:

0) Formalização do problema, mapeamento quantitativo em um modelo neural inicial e ... 0b) coleta de pares empíricos (X,y)

1) Fase de TREINO da RNA (MLP): com conhecimento dos X e dos y, que são ambos usados na calibração do modelo

2) Fase de TESTE / Caracterização da qualidade da RNA para generalizar: temos novos pares X e y, com y guardado "na gaveta", usado apenas para avaliação, não para re-calibração. É como um ensaio de uso final do modelo, com possibilidade de medir a sua qualidade com o y que foi guardado na gaveta.

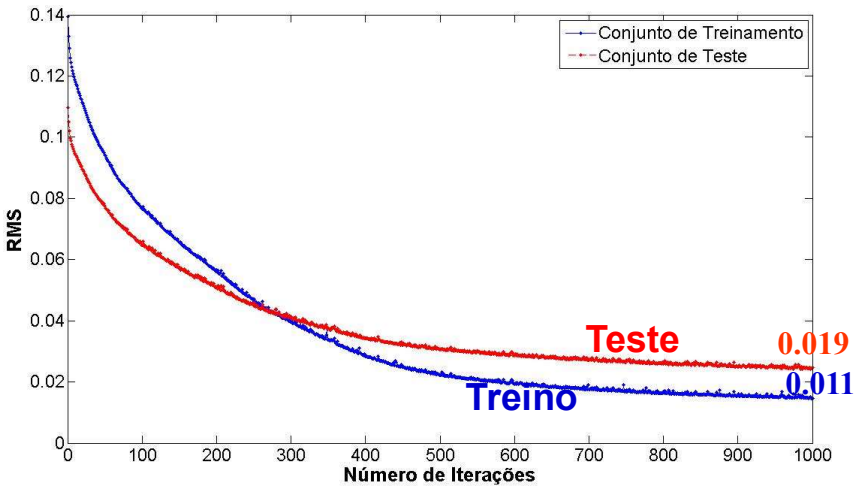
[Fase de refinamentos da RNA, dados e modelo, em ciclos, desde 0]

3) Fase de USO FINAL da RNA, com y efetivamente não conhecido, e estimado com conhecimento dos X + uso do modelo calibrado.

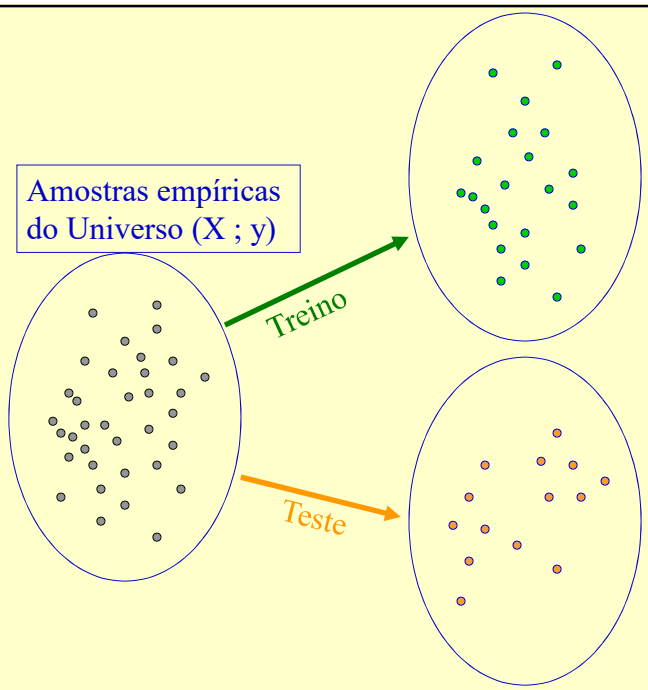
.... Diferenças e semelhanças entre 1, 2 e 3

107

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

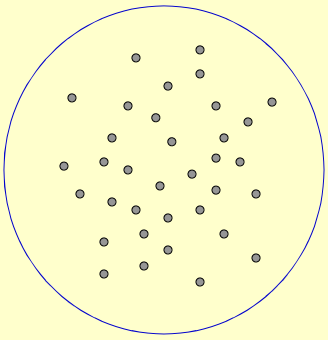


108



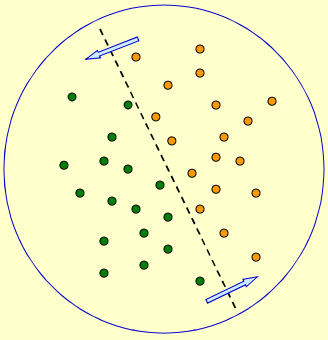
109

Conjunto total de observações empíricas



110

Cross validation / Validação cruzada ... e teremos muitos mais ensaios, se tivermos mudança na partição de observações



2 fold cross validation:
50% treino e
50% teste

111

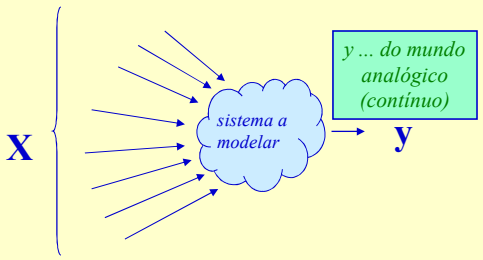
k-fold Cross Validation:
O conjunto total é "retalhado"
em k partes, e uma delas apenas
é usada para teste, com k
ensaios distintos

Conceito geral que engloba
as discussões anteriores
...
Reamostragem /
Data Resampling

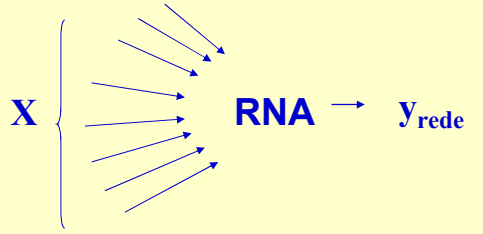
Pergunta ... Que impacto isto tem nas medidas de qualidade de regressores? Que impacto isto tem nas medidas de classificadores? Que informação ao cliente / usuário podemos fornecer com base neste conceito?

Lembrete ... Para que queremos a rede neural com pesos otimizados?

Modelagem de um sistema por função de mapeamento X -> y (a RNA como regressor contínuo não linear multivariável)

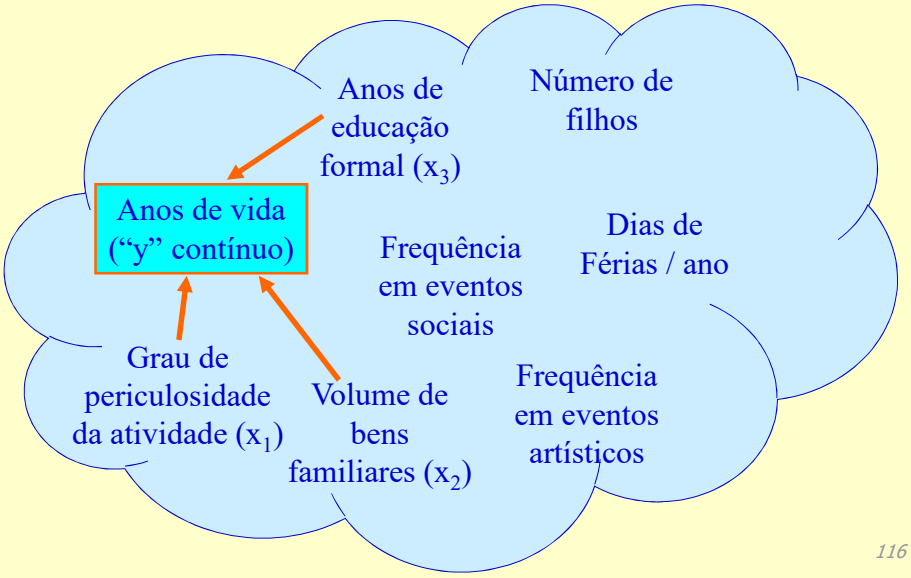


Assumimos que a variável y do sistema a modelar é uma função (normalmente desconhecida e possivelmente não linear) de diversas outras variáveis desse mesmo sistema

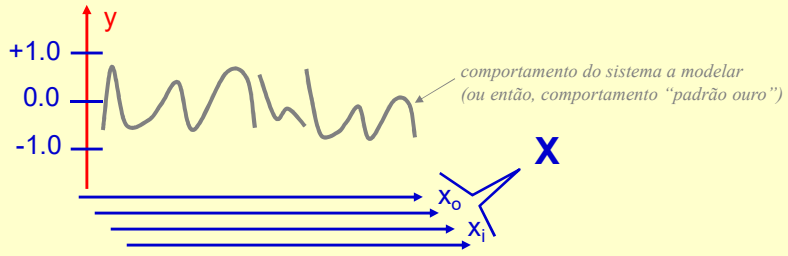


A RNA , para ser um bom modelo do sistema, deve reproduzir essa relação entre X e y, tão bem quanto possível

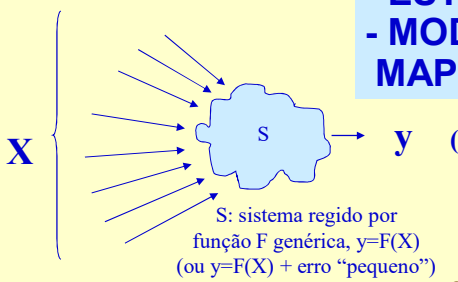
Um hipotético universo de variáveis inter-dependentes, passível de modelagem/ens



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



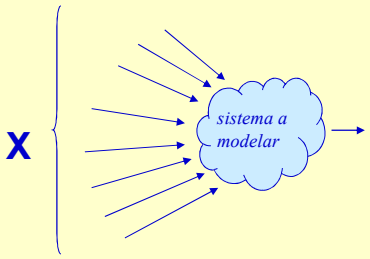
- ESTIMADOR ;
- MODELAGEM POR MAPEAMENTO X→y



(.... VALORES CONTÍNUOS!!!)

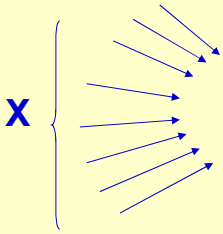
S: sistema regido por função F genérica, y=F(X) (ou y=F(X) + erro "pequeno")

RNAs como reconhecedor / detetor de padrões ...



y ... responde à seguinte pergunta:
X é padrão notável? **Sim ( ) Não ( )**

y ... é do universo discreto / binário (SIM versus NÃO), y não é analógico

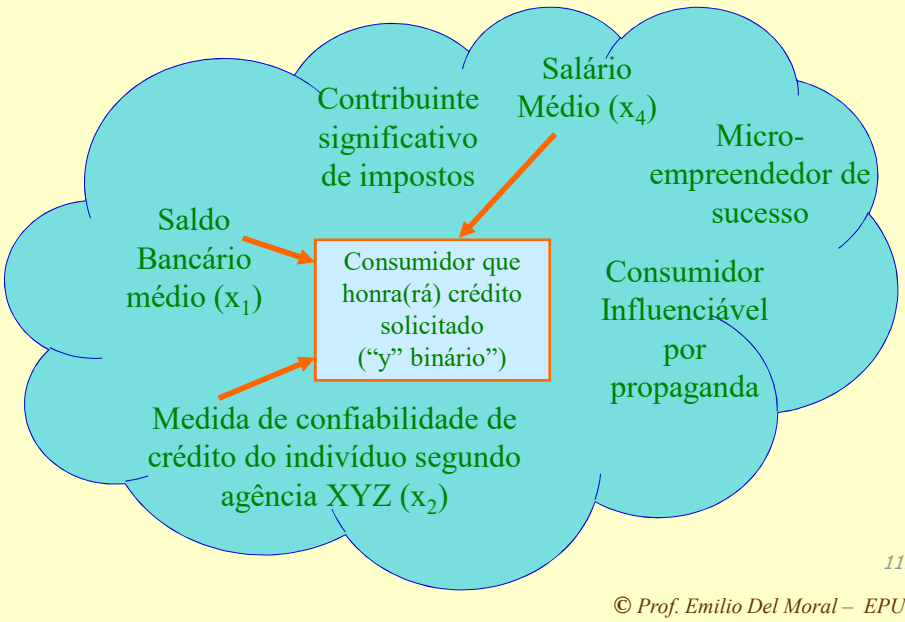


RNA → y_{rede} **... +1 ( )** *versus* **- 1 ( )**

... ou alternativamente: y_{rede} **... 0.5 ( )** *versus* **- 0.5 ( )**
... ou alternativamente: ... $y_{rede} > 0$ **( )** *versus* $y_{rede} < 0$ **( )**

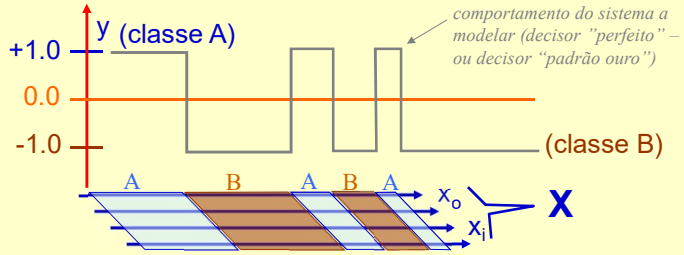
118

Um hipotético universo de variáveis inter-dependentes, passível de modelagem/ens

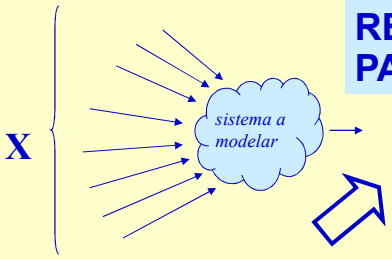


119

Caso de classificação binária / reconhecimento de padrões, será do tipo ...



CLASSIFICADOR; RECONHECEDOR DE PADRÕES NOTÁVEIS



y (... restrita a valores +1 e –1)

120

121

E será que conseguiremos que a RNA calcule essas duas funções de X complicadas representadas nos slides anteriores (uma com y contínuo e a outra com y binário)?

121

recordando

De onde vem o grande poder do MLP?

Kurt Hornik showed in 1991[2] that it is not the specific choice of the a
assumed to be linear. For notational convenience, only the single out

recordando

Formal statement [edit]

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

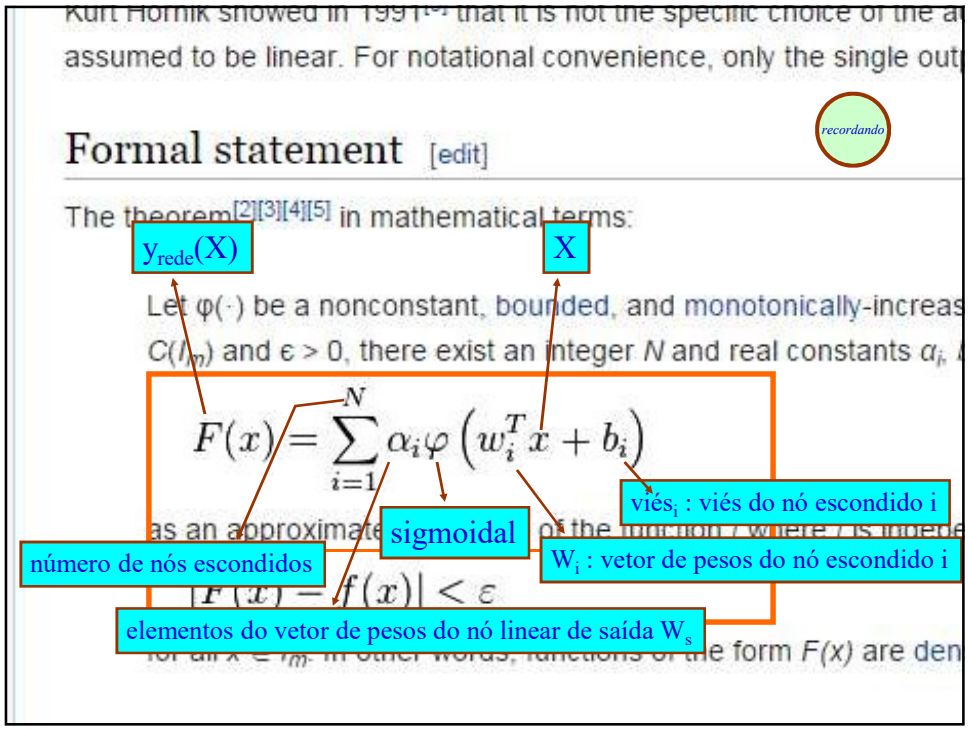
Let $\varphi(\cdot)$ be a nonconstant, bounded, and monotonically-increas
$C(I_m)$ and $\epsilon > 0$, there exist an integer N and real constants α_i,

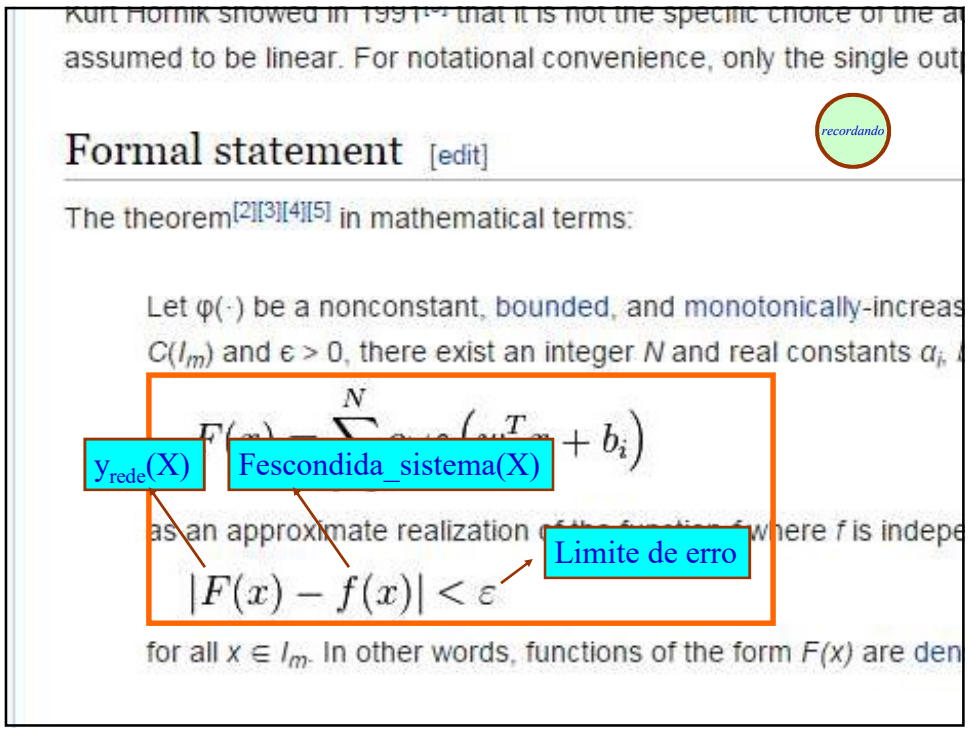
$$F(x) = \sum_{i=1}^{N} \alpha_i \varphi\left(w_i^T x + b_i\right)$$

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

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

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


assumed to be linear. For notational convenience, only the single out
Formal statement [edit]
recordando
The theorem[2][3][4][5] in mathematical terms:
$y_{rede}(X)$
X
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,
$F(x) = \sum_{i=1}^{N} \alpha_i \varphi\left(w_i^T x + b_i\right)$
$viés_i$: viés do nó escondido i
sigmoidal
número de nós escondidos
W_i : vetor de pesos do nó escondido i
$|F(x) - f(x)| < \varepsilon$
elementos do vetor de pesos do nó linear de saída W_s
for all x ∈ I_m. In other words, functions of the form F(x) are den


assumed to be linear. For notational convenience, only the single out
Formal statement [edit]
recordando
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,
$y_{rede}(X)$
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 ∈ I_m. In other words, functions of the form F(x) are den

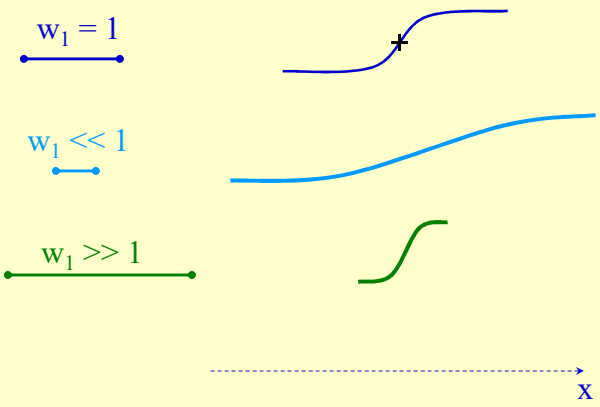
Questões intrigantes, p/ esta aula e p/ pensar em casa ...

- *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 rede neural?*
- *O que ganhamos e o que perdemos se escolhermos usar MUITOS nós na construção da rede neural?*

... Um parênteses em lousa para discutirmos um pouco a aproximação universal com sigmoides / funções em formato de "S", num caso simples e bem particular ... aproximação de uma função y de uma única variável x_1:

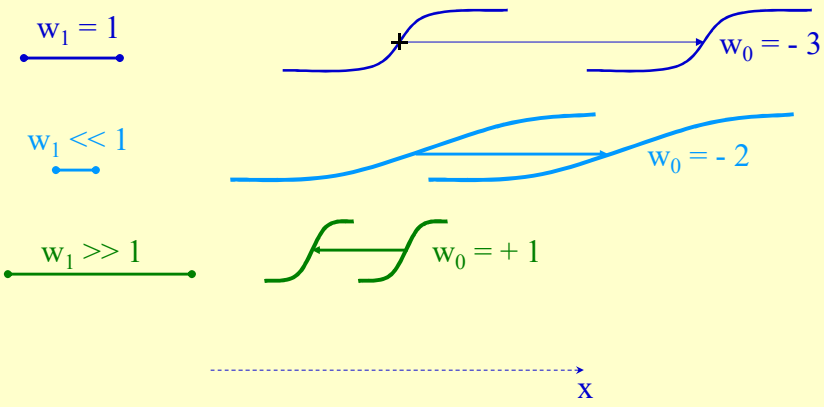
$$y(x_1)$$

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

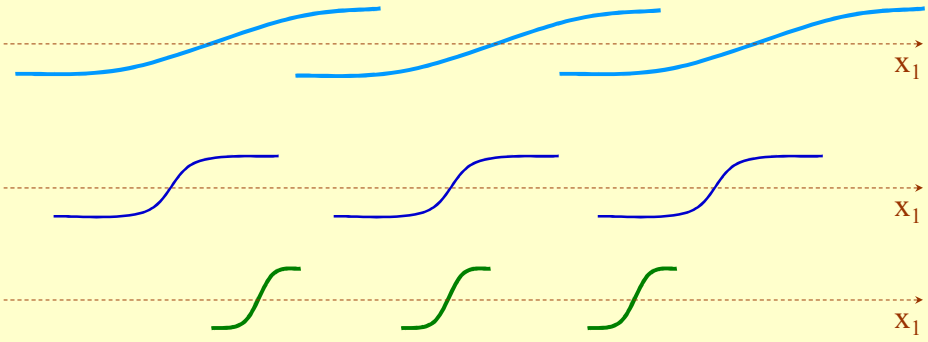


O que conseguimos fazer com um único neurônio sigmoidal $y(w_1.x_1 + w_0.1)$, c/ escalamento de x_1 via w_1

... e também com o viés, via viés w_0



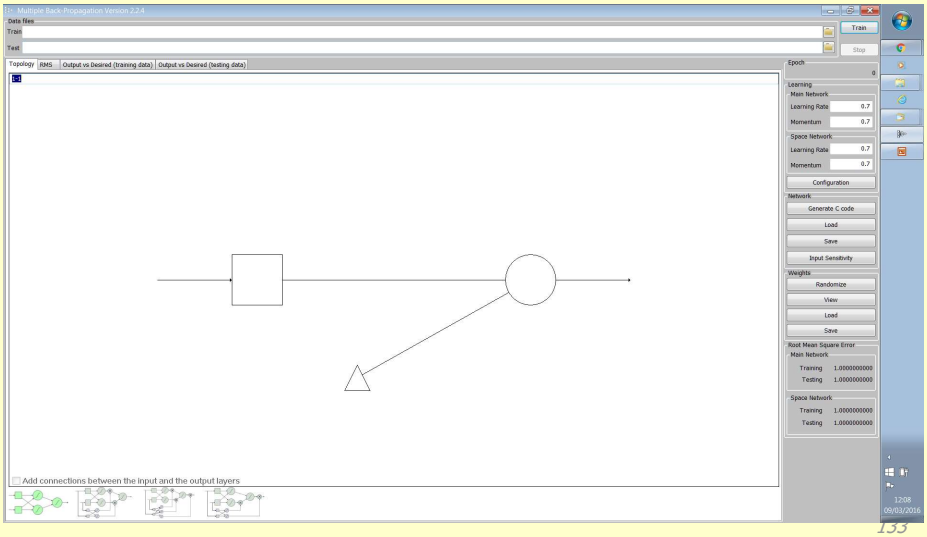
O que conseguimos fazer com um único neurônio sigmoidal, no caso de regressões ("y(x_1) contínuo")?



$$y = tgh\ (w_1.x_1 + viés)$$

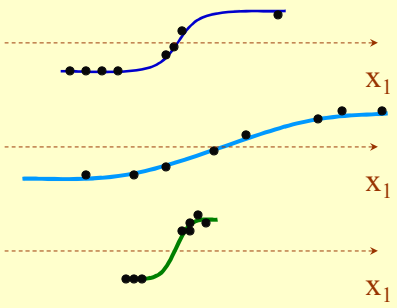
132

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

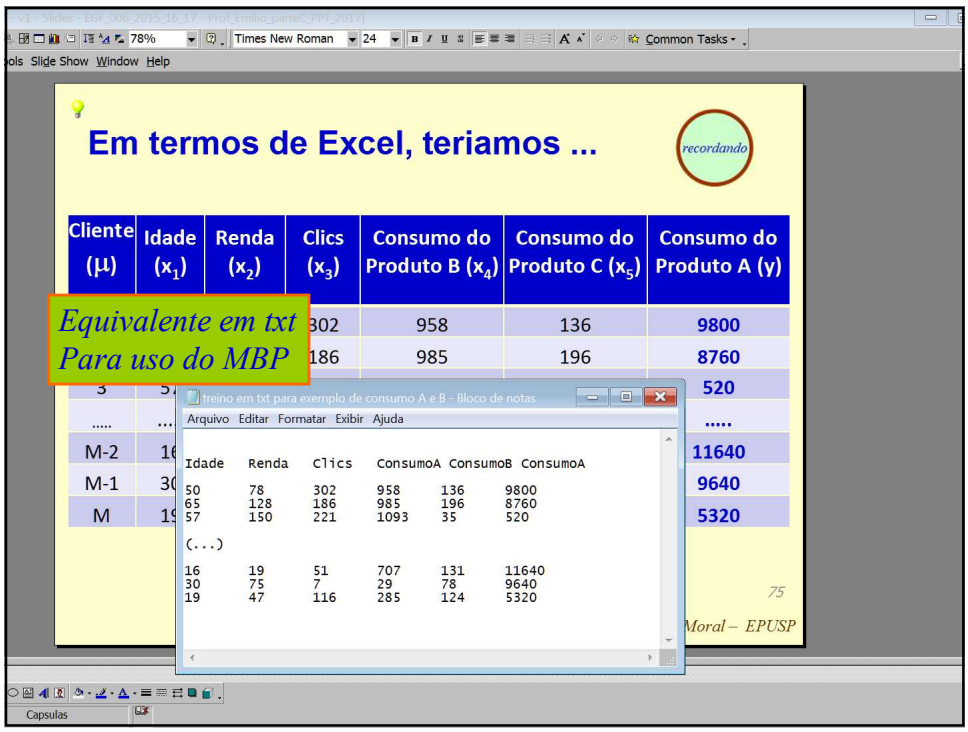


133

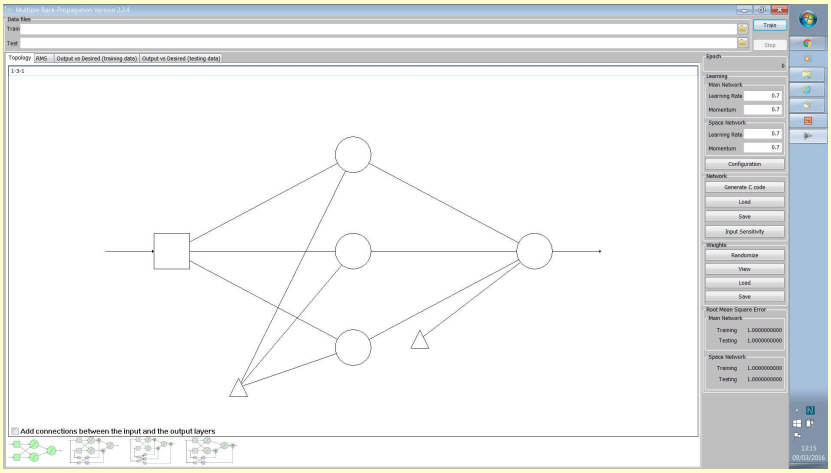
Que tipo de dados empíricos modelamos com um único neurônio sigmoidal em regressões ("y(x_1) contínuo")?



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

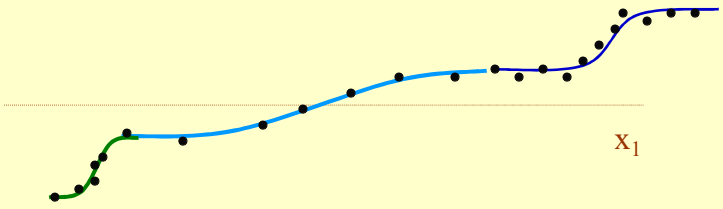


Regressão univariada com Cybenko “café com leite” de 3 nós na primeira camada ...



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

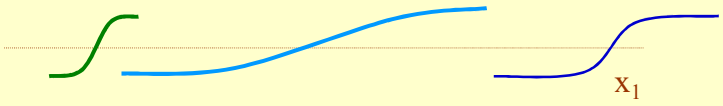
... superposição de várias sigmóides deslocadas e escaladas



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?

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

... várias sigmóides deslocadas e escaladas



Acima ... Comportamento individual dos 3 nós "tgh" na primeira camada, com com 3 viéses distintos e 3 escaladores de x_1 distintos.

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 rede neural?***
- ***O que ganhamos e o que perdemos se escolhermos usar MUITOS nós na construção da rede neural?***

recordando

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