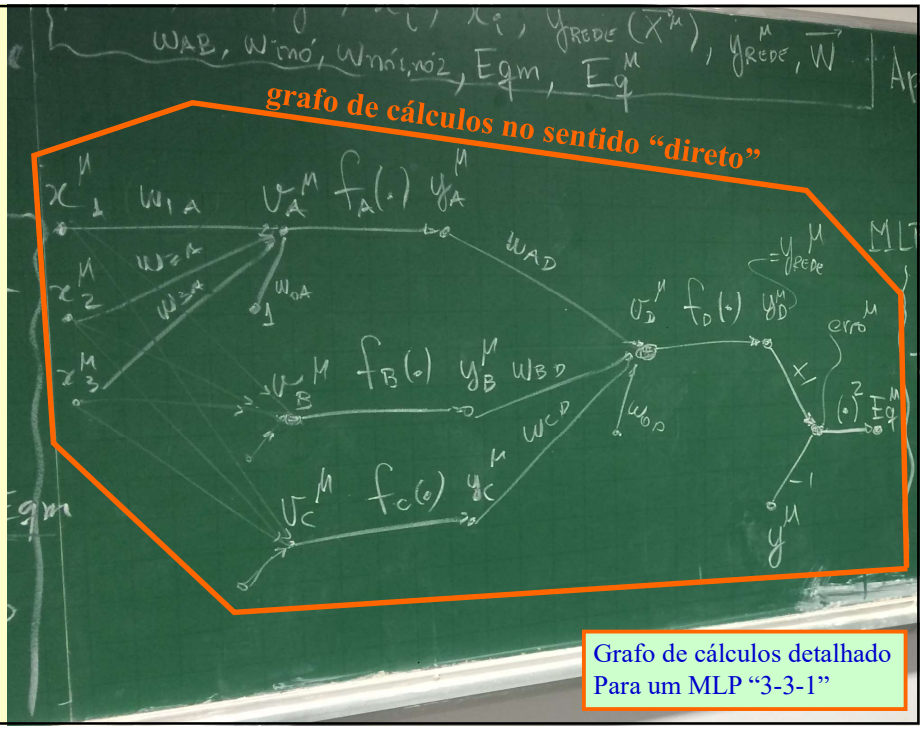
Temas da 3ª semana – PSI3471-2020 – Prof Emilio

#5 (16/março – 2ªf) Foco da semana: aprendizado da Rede Neural MLP – O Gradiente descendente e a otimização de pesos sinápticos com base no conjunto de treino e EBP; dedução das fórmulas do EBP, em sala de aula em conjunto com os alunos: trabalho focado num peso sináptico específico da rede, escolhido pelo professor para máxima complexidade da dedução.
#6 (18/março – 4ªf) ... Discussão das extensões das deduções já feitas (para um peso no EBP) para os demais pesos sinápticos; redundâncias nos cálculos dos diversos pesos da rede neural e otimização do esforço computacional. Regra "Delta" de aprendizado de Widrow, para neurônio isolado; Aprendizado por EBP recursivo, camada a camada.

59



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

60

Como escolhemos os valores dos diversos w´s ?

61

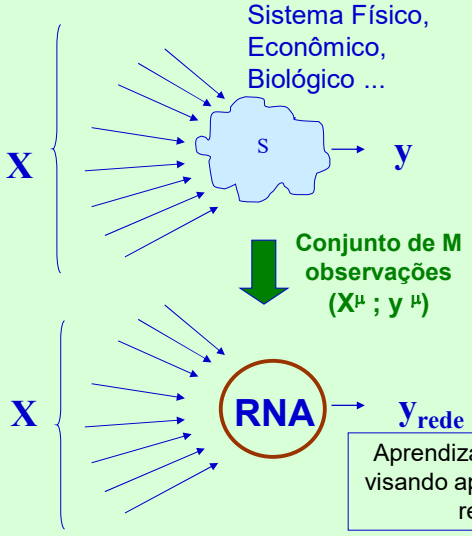
61

Aprendizado em RNAs do tipo MLP – Multi Layer Percetron – através do algoritmo Error Back Propagation

(método do gradiente usado na otimização de w´s do MLP)

62

Conjunto de treino em arquiteturas supervisionadas (ex. clássico: MLP com Error Back Propagation)



A computação desejada da rede pode ser definida simplesmente através de observações / exemplares do comportamento requerido

$$Eqm = \frac{1}{M}\sum_{\mu=1}^{M}(y_{rede}(\vec{X}^{\mu},\vec{W}) - y^{\mu})^2$$

... em loop ...

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

Aprendizado: Espaço de pesos W é explorado visando aproximar ao máximo a computação da rede da computação desejada

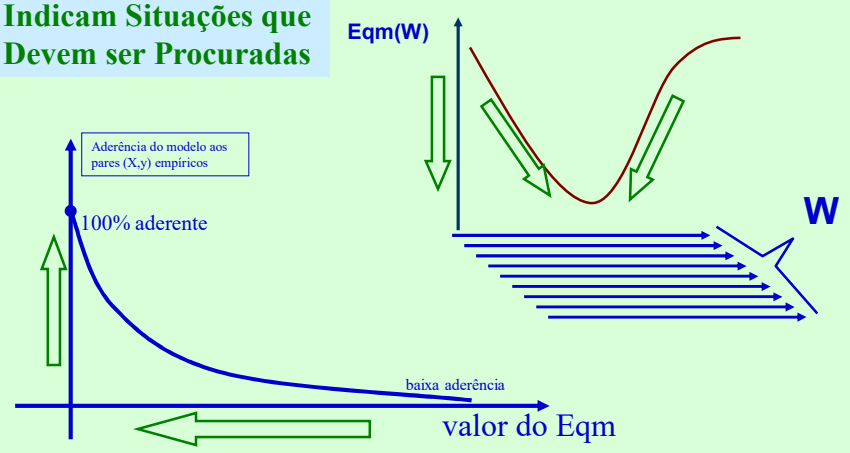
63

63

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

Devemos buscar Maximização da aderência = Minimo Eqm possível

As Setas Verdes Indicam Situações que Devem ser Procuradas



64

64

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
  - Tire a média dos M gradientes individuais e dê um passo Delta ΔW nesse espaço, com direção e magnitude dados por $-\eta$*vetor gradiente (Eqm)

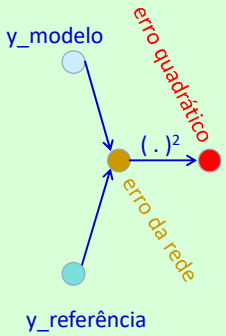
65

Processo de refinamentos graduais a cada iteração ...

W#0	Eqm#0	GradEqm(W#0)	DeltaW#0 = - n.GradEqm(W#0)
W#1 (= W#0 + DeltaW#0)	Eqm#1 (< Eqm#0)	GradEqm(W#1)	DeltaW#1 = - n.GradEqm(W#1)
W#2 (= W#1 + DeltaW#1)	Eqm#2 (< Eqm#1)	GradEqm(W#2)	DeltaW#2 = - n.GradEqm(W#2)
W#3 (= W#2 + DeltaW#2)	Eqm#3 (< Eqm#2)	GradEqm(W#3)	DeltaW#3 = - n.GradEqm(W#3)
W#4 (= W#3 + DeltaW#3)	Eqm#4 (< Eqm#3)	GradEqm(W#4)	DeltaW#4 = - n.GradEqm(W#4)
...	...	...	...
W#k (= W#k-1 + DeltaW#k-1)	Eqm#k (< Eqm#k-1)	GradEqm(W#k)	DeltaW#k = - n.GradEqm(W#4)
...	...	...	...

66

Modelo Neural no modo de Treinamento



74

... erro da rede com relação ao conjunto de treinamento como um todo; simbologia (X^{μ} ; y^{μ}); Erro quadrático de exemplar (Eq^{μ}); Erro quadrático médio (Eqm)

$$Eqm = \frac{1}{M}\sum_{\mu=1}^{M}(y_{rede}(\vec{X}^{\mu},\vec{W}) - y^{\mu})^2$$

μ *identifica um de M exemplos de treinamento*

75

Deduzindo as Equações do Aprendizado em RNAs do tipo MLP – Multi Layer Percetron – com o algoritmo Error Back Propagation (Gradiente Descendente)

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

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

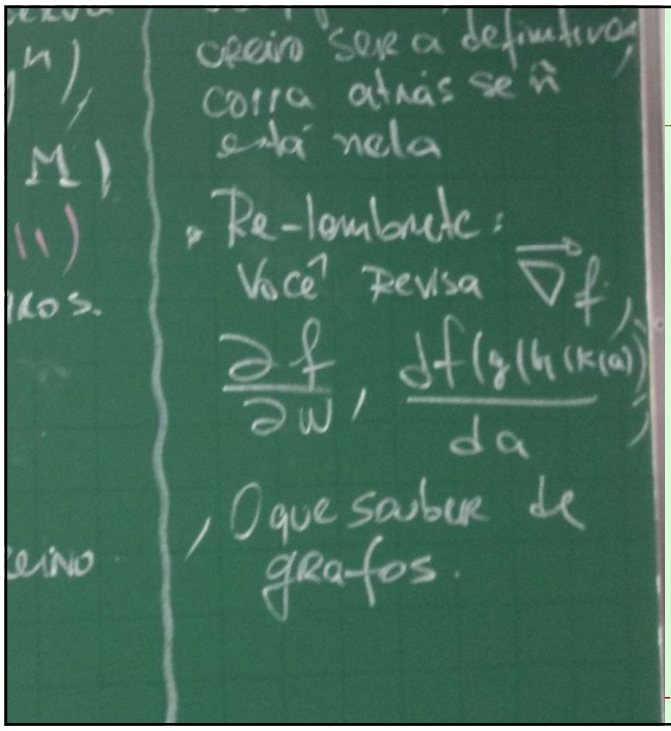
  $\partial f(a,b,c)/\partial a \qquad \partial f(a,b,c)/\partial b \qquad \partial f(a,b,c)/\partial c$

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

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

- **Regra da cadeia, necessária ao cálculo de derivadas quando há encadeamento de funções:**

  $\partial f(g(h(a)))/\partial a = \partial f/\partial g \,.\, \partial g/\partial h \,.\, \partial h/\partial a$



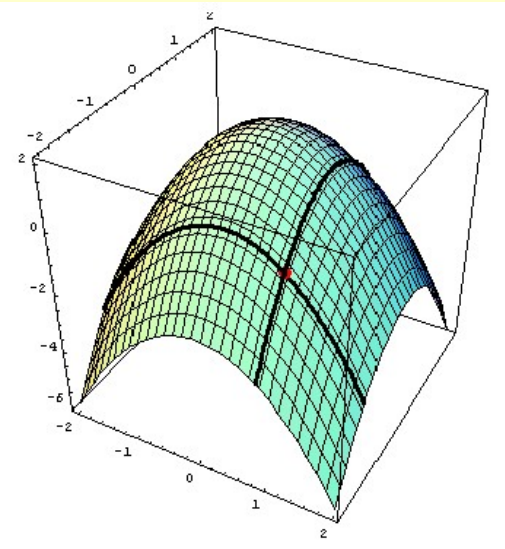
Faça sua própria revisão com base no aprendizado anterior !!

(ou aprenda por si, se saltou tópicos de cursos anteriores como cálculos, em que esses temas são vistos e/ou usados regularmente)

83

83

Derivada parcial - ilustração p/ função de 2 variáveis apenas



Faça sua própria revisão com base no aprendizado anterior !!

(ou aprenda por si, se saltou tópicos de cursos anteriores como cálculos, em que esses temas são vistos e/ou usados regularmente)

imagem exemplo extraída da internet

84

84

Derivada parcial- ilustração p/ função de 2 variáveis apenas

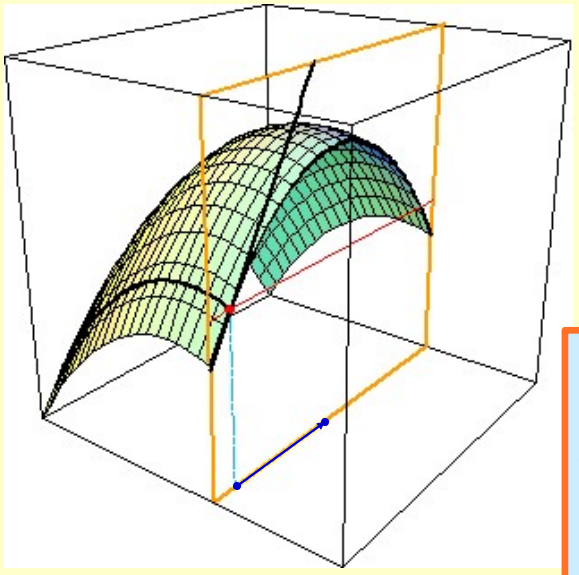


imagem base extraída da internet

Faça sua própria revisão com base no seu aprendizado anterior !!

(ou aprenda por si, se saltou tópicos de cursos anteriores como cálculos, em que esses temas são vistos e/ou usados regularmente)

85

Formação do vetor gradiente a partir de duas derivadas parciais

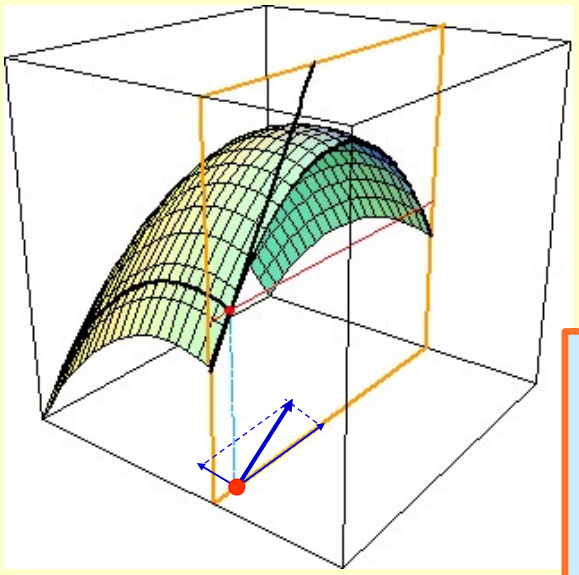


imagem base extraída da internet

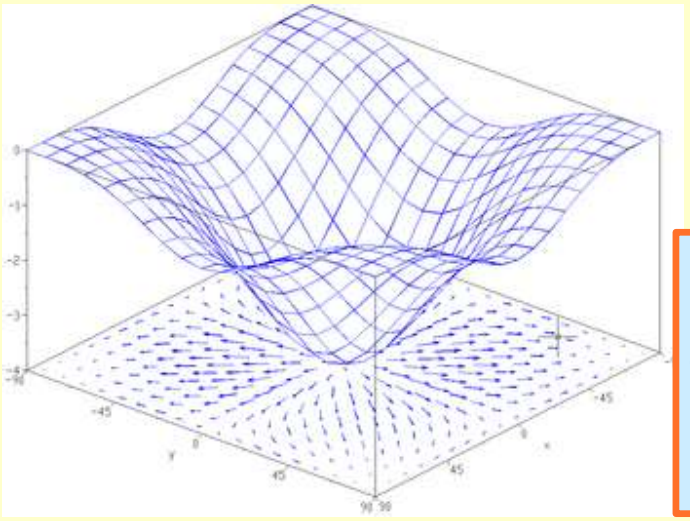
Faça sua própria revisão com base no seu aprendizado anterior !!

(ou aprenda por si, se saltou tópicos de cursos anteriores como cálculos, em que esses temas são vistos e/ou usados regularmente)

86

http://en.wikipedia.org/wiki/Gradient

... O vetor gradiente indica a direção ascendente e seu módulo a magnitude de crescimento da função escalar – ilustração p/ função de 2 variáveis apenas



Faça sua própria revisão com base no seu aprendizado anterior !!

(ou aprenda por si, se saltou tópicos de cursos anteriores como cálculos, em que esses temas são vistos e/ou usados regularmente)

87

Aprendizado do MLP por Error Back Propagation ...

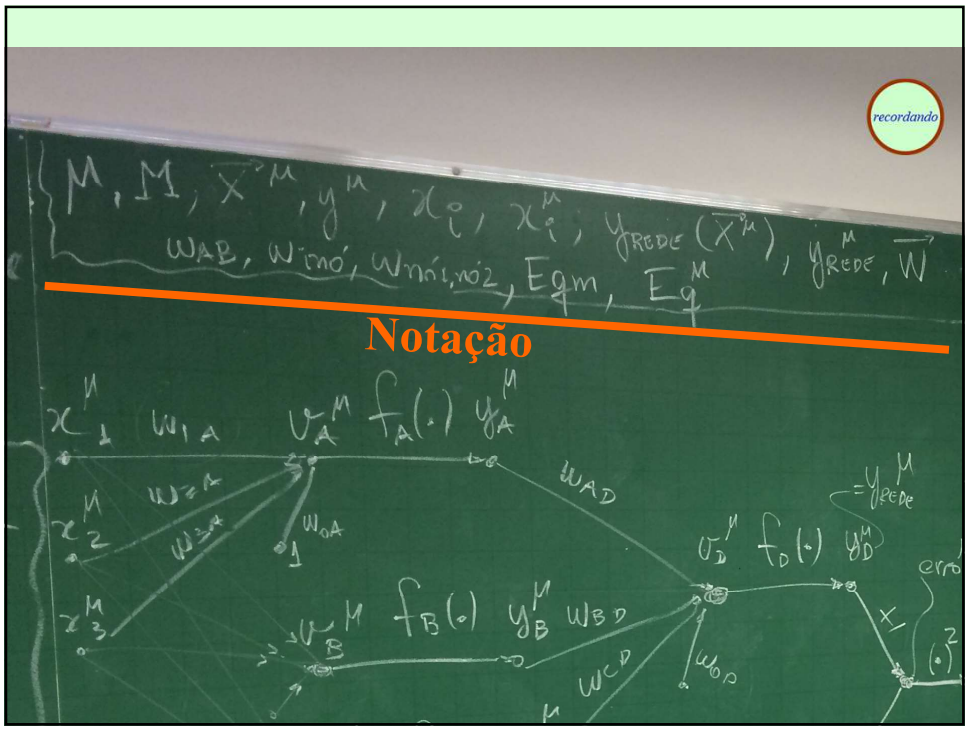
$$\Delta\vec{W} = -\eta.\vec{\nabla}Eqm\_$$

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

88

88



91

Entendendo símbolos que temos usado em nossos grafos da lousa e em alguns dos slides:

92

- X -- vetor X de entradas num MLP, com os valores genéricos nas suas componentes x_i
- x_i -- componente "i" do vetor de entradas de um MLP, com valor genérico, um número real entre " - infinito e + infinito"
- X^μ -- vetor X das entradas num MLP, mas com os valores de cada um dos x_i específicos da observação empírica μ
- μ -- identificador / indexador inteiro, entre 1 e M, usado para especificar um dos exemplares (uma das observações empíricas, aquela de "número" μ) que compõem o conjunto de pares empíricos (X,y).
- M -- número total de exemplares empíricos que compõem o conjunto de treino; cardinalidade do conjunto de treino
- x_i^μ -- componente "i" do vetor de entradas de um MLP, com o valor específico referente à observação empírica μ
- y^μ -- saída alvo para a rede neural sendo treinada com aprendizado supervisionado, com o valor específico da observação empírica μ. Poderia ser chamada também de y_{alvo}^μ, explicitando mais claramente o significado

92

92

Entendendo símbolos que temos usado em nossos grafos da lousa e em alguns dos slides:

$y_{rede}(X^{\mu})$ -- saída da rede neural quando a sua entrada corresponde aos valores empíricos X^{μ}

Ou ... y_{rede}^{μ} -- outra forma de representar $y_{rede}(X^{\mu})$

W -- vetor de todos os pesos sinápticos de uma rede MLP, incluindo todos os vieses de todos os neurônios

w_{AB} -- peso sináptico específico que conecta a saída do neurônio A com uma das entradas do neurônio B

w_{iA} -- peso sináptico específico que conecta a entrada x_i da rede neural com o neurônio A da primeira camada do MLP

Eqm -- Erro quadrático médio; média dos M valores de Eq^{μ}

Eq^{μ} -- Erro quadrático individual, referente especificamente à observação empírica μ ----- fórmula: $Eq^{\mu} = [y_{rede}(X^{\mu}) - y^{\mu}]^2$

93

Um Exemplo Ilustrativo para o Conceito de Conjunto de Treinamento e dos M pares (X,y)…

95

Exemplo de regressão multivariada para estimação contínua usando MLP

- O valor do y contínuo ... neste exemplo corresponde ao volume de consumo futuro num dado tipo de produto "A" a ser ofertado pela empresa a um cliente corrente já consumidor de outros produtos da empresa ("B" e "C"), volume esse previsto com base em várias medidas quantitativas que caracterizam tal indivíduo. ... Assim, y = Consumo do Produto A = $F(x_1, x_2, x_3, x_4, x_5)$.
- Consideremos 4 variáveis de entrada no modelo preditivo neural, ou seja, temos 5 medidas em X:
  - x_1: Idade do indivíduo
  - x_2: Renda mensal do indivíduo
  - x_3: Volume de clicks do indivído no website de exibição de produtos oferecidos pela empresa
  - x_4: Volume de consumo desse cliente observado para outro Produto B da mesma empresa
  - x_5: Volume de consumo desse cliente Produto C da mesma empresa
- Problema: desenvolver uma MLP para regressão contínua multivariada que permita estimar esse volume de consumo futuro y com base no conhecimento dos X e numa base de dados de aprendizado com esses dados X e y para 350 já clientes de universo populacional similar ao do novo consumidor potencial.

Exemplo de dados empíricos tabulados em Excel ...

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)
1	50	78	302	958	136	**9800**
2	65	128	186	985	196	**8760**
3	57	150	221	1093	35	**520**
.....	.....	.....	.....	.....	.....	**.....**
M-2	16	19	51	707	131	**11640**
M-1	30	75	7	29	78	**9640**
M	19	47	116	285	124	**5320**

97

Exemplo de dados empíricos tabulados em Excel ...

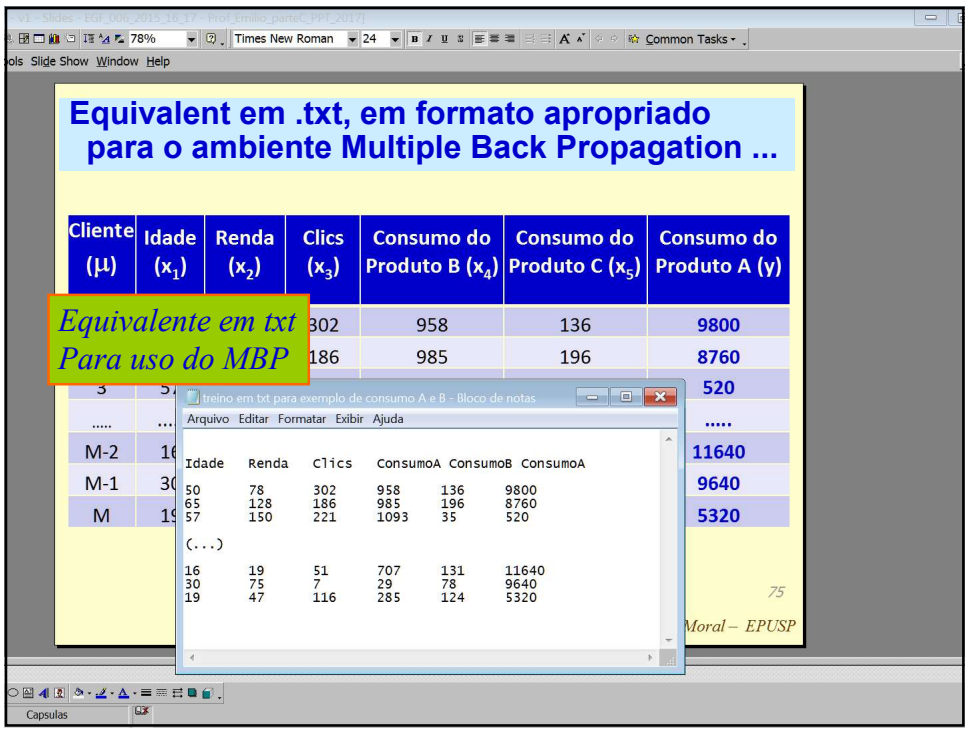
Cliente (μ)	Idade (x_1)	Renda (x_2)	Clics (x_3)	Consumo do Produto B (x_4)	Consumo do Produto C (x_5)	Consumo do Produto A (y)
1	50	78	302	958	136	9800
2	65	128	186	985	196	8760
3	57	150	221	1093	35	520
.....	.....	.....	.....	.....	.....	.....
M-2	16	19	51	707	131	11640
M-1	30	75	7	29	78	9640
M	19	47	116	285	124	5320

M vetores X^{μ} (cada um deles é 5 dimensional) de entrada do MLP, referentes cada um deles a uma das M observações empíricas

M alvos y^{μ}

98

98



99

... erro da rede com relação ao conjunto de treinamento como um todo; simbologia (X^μ ; y^μ); Erro quadrático de exemplar (Eq^μ); Erro quadrático médio (Eqm)

$$Eqm = \frac{1}{M}\sum_{\mu=1}^{M}(y_{rede}(\vec{X}^{\mu},\vec{W}) - y^{\mu})^2$$

μ *identifica um de M exemplos de treinamento*

102

Inicialmente, 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:

Grad(Eqm) =

Grad(Σ_μ Eq^μ) / M

Σ_μ Grad (Eq^μ) / M

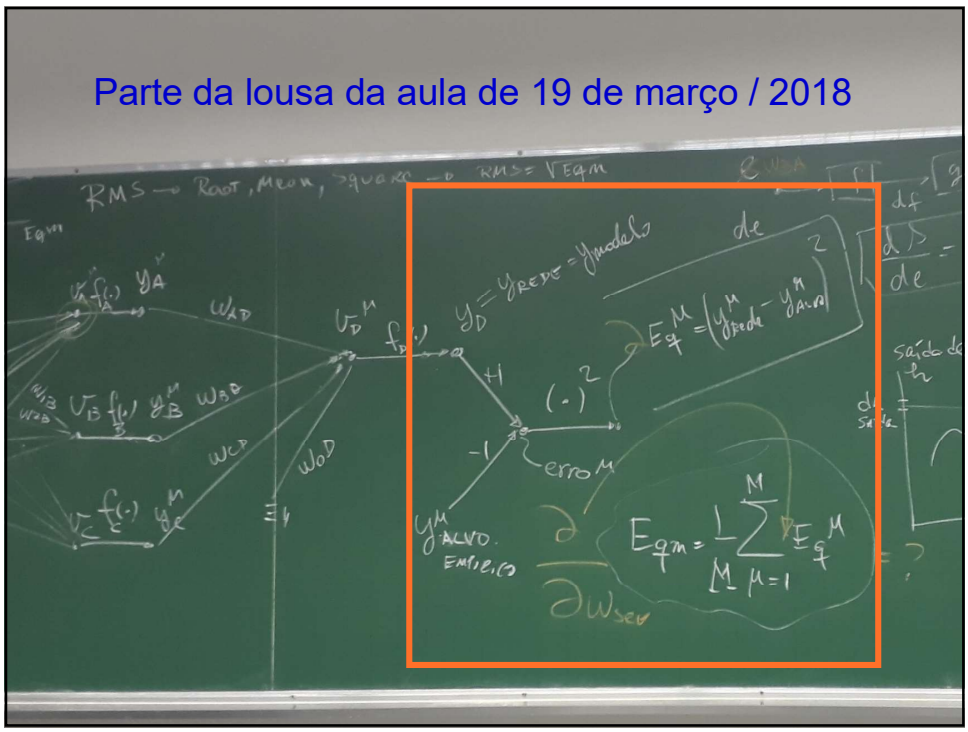
103

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^μ)*** – a seguinte propriedade simples e sua velha conhecida ...

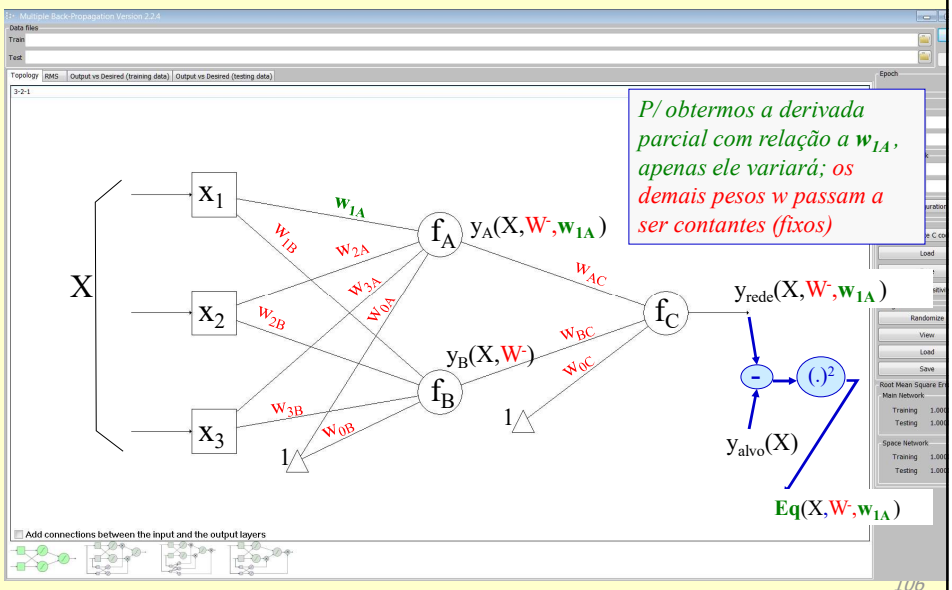
$$d(f_1(x)+f_2(x)) / dx = df_1(x)/dx + df_2(x)/dx$$

104

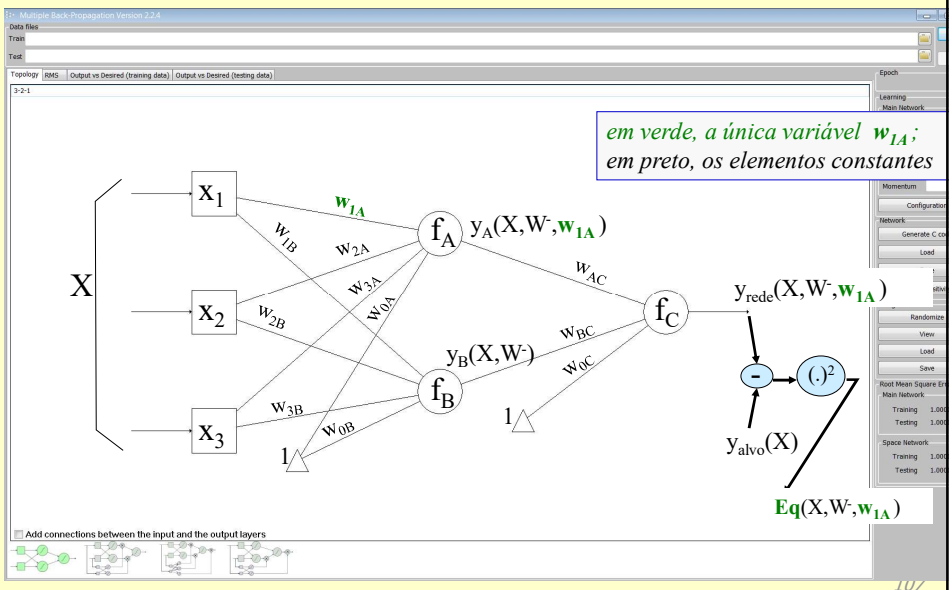
Parte da lousa da aula de 19 de março / 2018



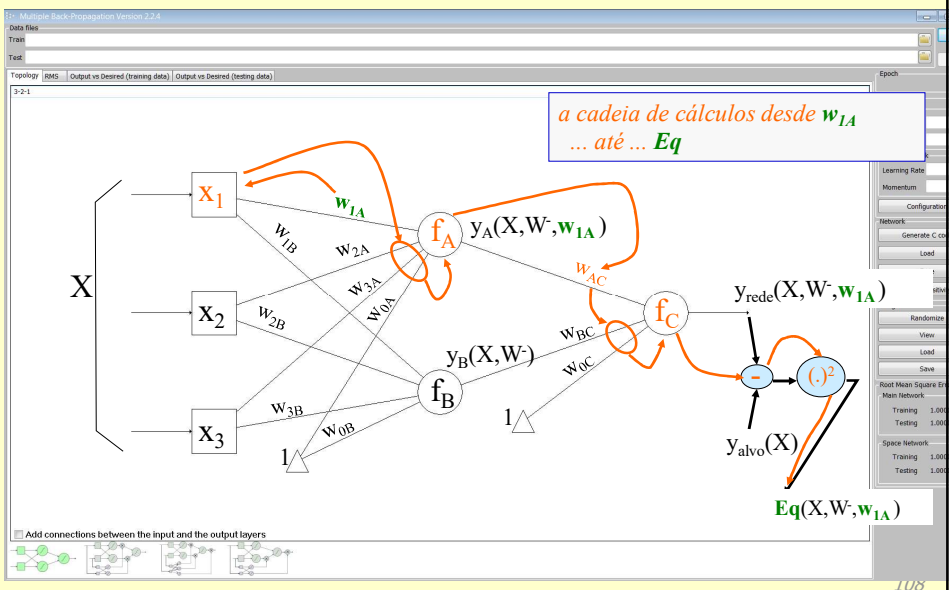
105



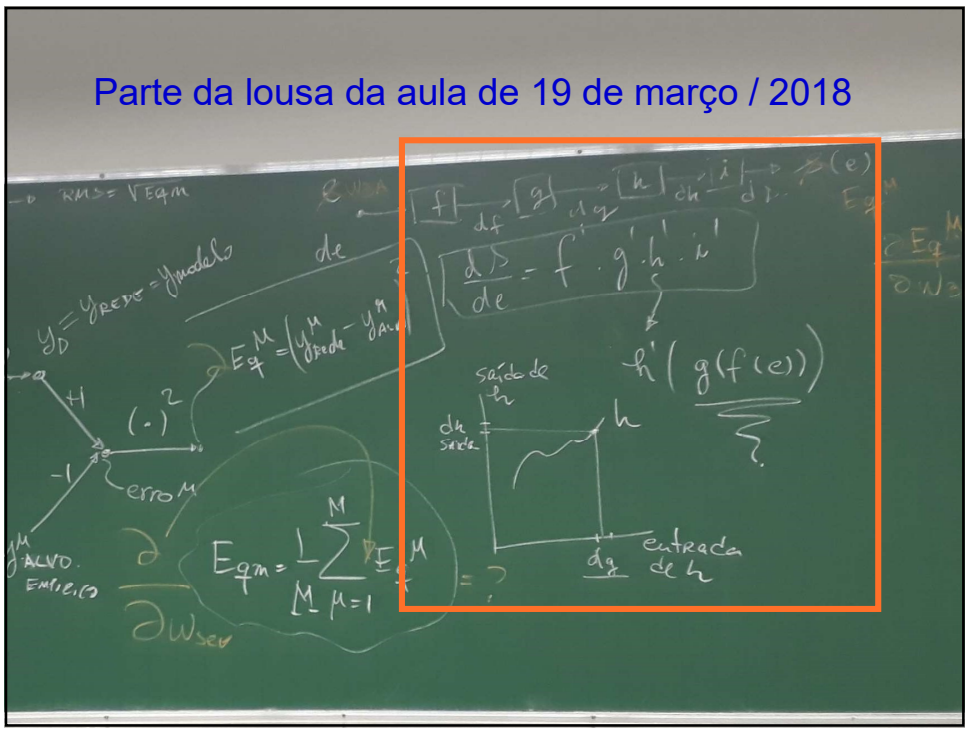
106



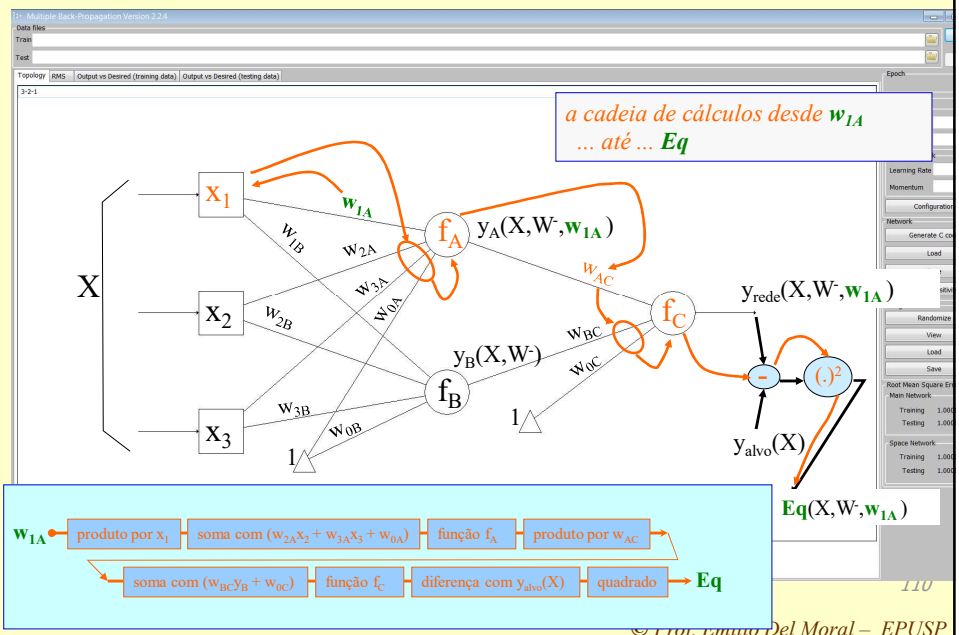
107


a cadeia de cálculos desde w_{1A}
... até ... Eq
X
x_1
x_2
x_3
1
w_{1A}
w_{1B}
w_{2A}
w_{2B}
w_{3A}
w_{3B}
w_{0A}
w_{0B}
f_A
f_B
f_C
$y_A(X,W^-,w_{1A})$
$y_B(X,W^-)$
w_{AC}
w_{BC}
w_{0C}
$y_{rede}(X,W^-,w_{1A})$
$y_{alvo}(X)$
$Eq(X,W^-,w_{1A})$
Add connections between the input and the output layers
108
© Prof. Emilio Del Moral – EPUSP

108


Parte da lousa da aula de 19 de março / 2018

109



110

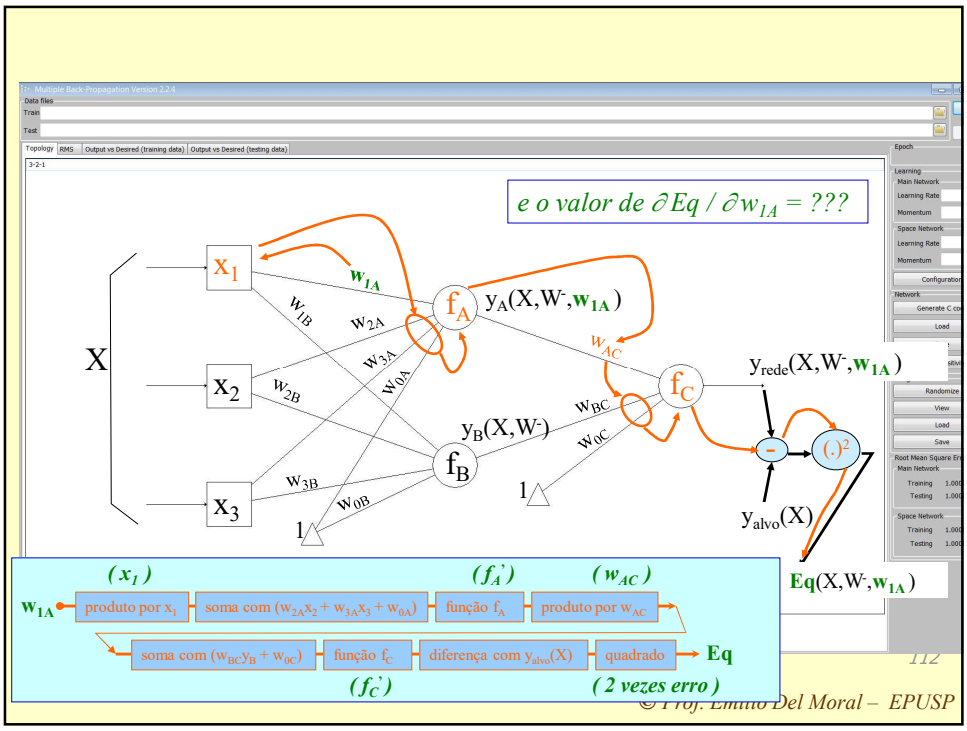
Note que aqui temos uma cadeia com muitos estágios que levam da váriável w_{1A}, à variável Eq^{μ}, e para a qual podemos calcular a derivada da saída (Eq^{μ}) com relação à entrada (w_{1A}) aplicando de forma repetida a seguinte propriedade simples e sua velha conhecida ...

$$d(f_1(f_2(x)) / dx = df_1(x)/df_2 \,.\, df_2(x)/dx$$

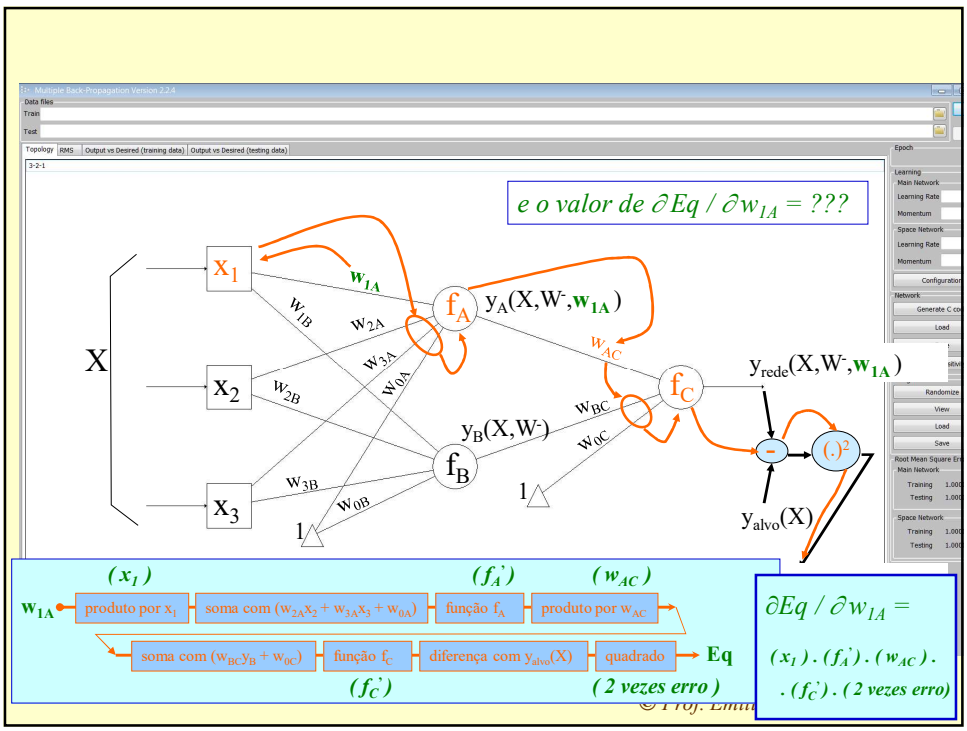
..., ou seja, calculando isoladamente o valor da derivada para cada estágio da cadeia, e finalizando o cálculo de derivada de ponta a ponta nessa cadeia toda através do produto dos diversos valores de cada estágio.

111

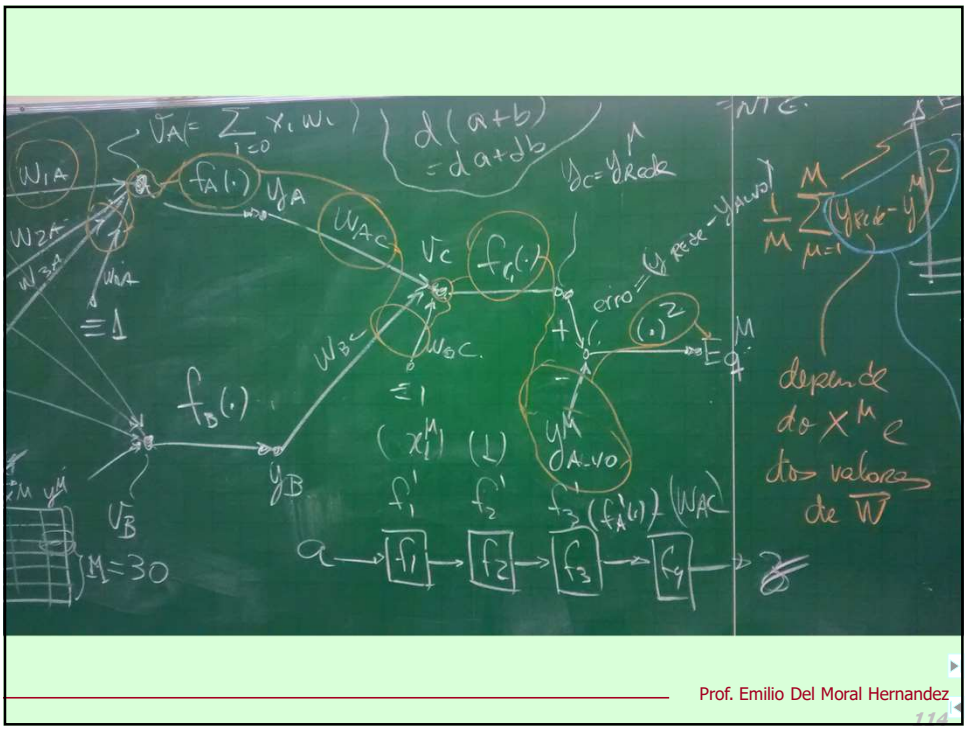
111


e o valor de ∂Eq / ∂w1A = ???
X
x1
x2
x3
w1A
w1B
w2A
w3A
w0A
w2B
w3B
w0B
1
fA
yA(X,W-,w1A)
wAC
fB
yB(X,W-)
fC
wBC
w0C
1
yrede(X,W-,w1A)
-
(.)2
yalvo(X)
Eq(X,W-,w1A)
(x1)
(fA')
(wAC)
w1A
produto por x1
soma com (w2Ax2 + w3Ax3 + w0A)
função fA
produto por wAC
soma com (wBCyB + w0C)
função fC
diferença com yalvo(X)
quadrado
Eq
(fC')
(2 vezes erro)
112
© Prof. Emilio Del Moral – EPUSP

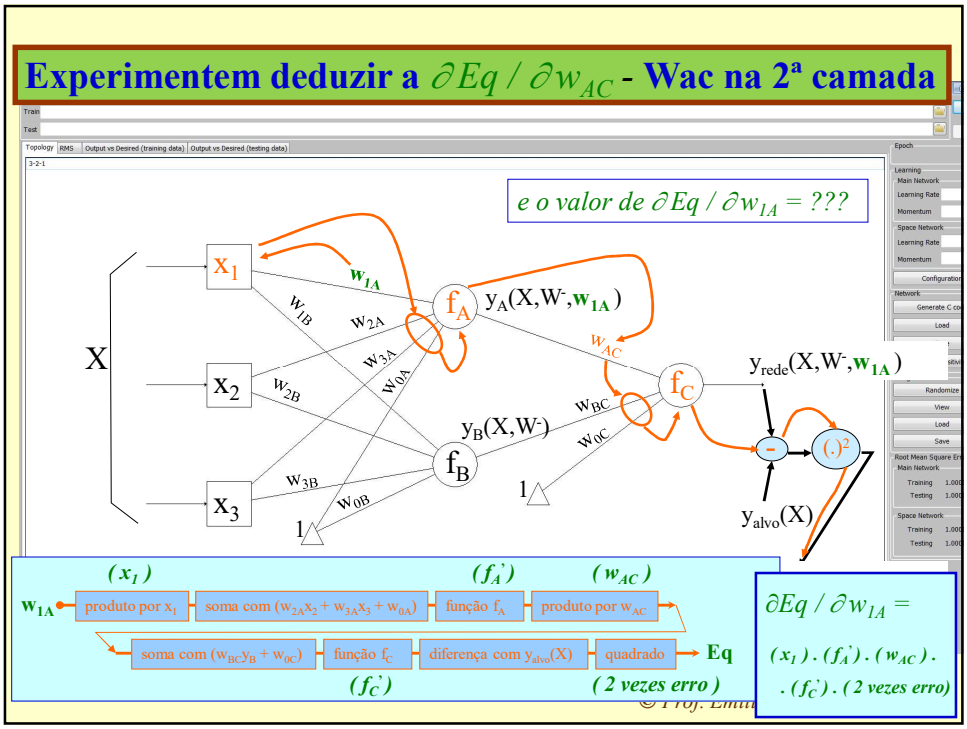
112


e o valor de ∂Eq / ∂w1A = ???
X
x1
x2
x3
w1A
w1B
w2A
w3A
w0A
w2B
w3B
w0B
1
fA
yA(X,W-,w1A)
wAC
fB
yB(X,W-)
fC
wBC
w0C
1
yrede(X,W-,w1A)
-
(.)2
yalvo(X)
(x1)
(fA')
(wAC)
w1A
produto por x1
soma com (w2Ax2 + w3Ax3 + w0A)
função fA
produto por wAC
soma com (wBCyB + w0C)
função fC
diferença com yalvo(X)
quadrado
Eq
(fC')
(2 vezes erro)
∂Eq / ∂w1A =
(x1) . (fA') . (wAC) .
. (fC') . (2 vezes erro)

113


Prof. Emilio Del Moral Hernandez
114

114


Experimentem deduzir a ∂Eq / ∂w_AC - Wac na 2ª camada
e o valor de ∂Eq / ∂w_1A = ???
X
x1
x2
x3
w1A
w1B
w2A
w2B
w3A
w3B
w0A
w0B
fA
yA(X,W-,w1A)
fB
yB(X,W-)
wAC
wBC
w0C
fC
yrede(X,W-,w1A)
(.)2
yalvo(X)
1
(x1)
(fA')
(wAC)
w1A
produto por x1
soma com (w2Ax2 + w3Ax3 + w0A)
função fA
produto por wAC
soma com (wBCyB + w0C)
função fC
diferença com yalvo(X)
quadrado
Eq
(fC')
(2 vezes erro)
∂Eq / ∂w1A =
(x1) . (fA') . (wAC) .
. (fC') . (2 vezes erro)

118

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 $[y_{rede}(X,W) - y_{alvo}(X)]$

123

123

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 $y_{alvo}{}^{\mu}$, 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!

124

124

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
  - Tire a média dos M gradientes individuais e dê um passo Delta ΔW nesse espaço, com direção e magnitude dados por $-\eta$*vetor gradiente (Eqm)

Prof. Emilio Del Moral Hernandez

126