12 -Outros critérios de erro

Erro médio quadrático vs Erro eficaz - o que usar ?

EMQ **E eficaz**

$$\varepsilon_{mq} = \underset{\forall\ par}{E}\,(y-\tilde{y})^2$$

$$\varepsilon_{ef} = \sqrt{\varepsilon_{mq}} = \sqrt{\underset{\forall\ par}{E}\,(y-\tilde{y})^2}$$

ε_{mq} - mais simples de operar algebricamente

ε_{ef} - mesmas unidades da variável e do erro

– mais fácil de interpretar

Conclusão:

faça a álgegra com ε_{mq} e

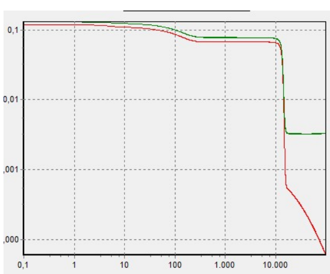
apresente os resultados com ε_{ef}

Como apresentar os resultados (ε_{ef} *vs* passo treinamento) ?

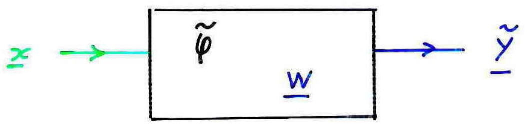
ε_{ef} inicial típico: .3 ε_{ef} final típico: 10^{-2} ... 10^{-4}

longos degraus são típicos

Conclusão: usar escala log-log



Composição de F como erro



$$F(\underline{y}, \underline{\tilde{y}}) \qquad \underline{\tilde{y}}(\underline{w}) \qquad \frac{\partial F}{\partial w_i} = \frac{\partial F}{\partial \tilde{y}} \frac{\partial \tilde{y}}{\partial w_i}$$

onde $\frac{\partial F}{\partial \tilde{y}}$ depende da função erro utilizada e

$\frac{\partial \tilde{y}}{\partial w_i}$ depende das equações da rede

no caso de saídas múltiplas $\frac{\partial F}{\partial w_i} = \sum_{l=1}^{m} \frac{\partial F}{\partial \tilde{y}_l} \frac{\partial \tilde{y}_l}{\partial w_i}$

Caso geral: Erro médio quadrático ponderado

$$F = \underset{k}{E}\left\{ \sum_{l} a(k,l)\ \varepsilon_l^2 \right\}$$

onde a(l) peso variável por saída l

a(k) peso variável por par k

como $\varepsilon_l = y_l - \tilde{y}_l$

$$\frac{\partial F}{\partial \tilde{y}} = -2\ \underset{k}{E}\left\{ \sum_{l} a(k,l)\ \varepsilon_l \right\}$$

No treinamento tipo regra delta o acréscimo na sinapse w_{ij} sera dado por

$$\Delta w_{ij} = 2\alpha\, v_j\, \delta_i \quad onde \quad \delta_i = \sum_{l=1}^{m} a(k,l)\,\varepsilon_l\, g_{li}$$

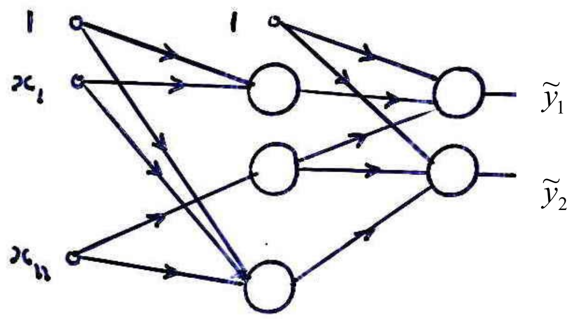
onde

v_j **é o sinal na entrada da sinapse na fase de propagação do sinal para a frente e**

$g_{li} = \dfrac{\partial \tilde{y}_l}{\partial u_i}$ **é o ganho linearizado da saída da sinapse w_{ij} à saída *l* da rede**

Ex 1 : Ponderando diferentemente saídas diferentes

A rede tem duas saídas, mas
o erro em y_1 é 3 vezes mais importante que o erro em y_2



a(k,1) = 3

a(k,2) = 1

$$F = \underset{k}{E}\left(3\varepsilon_1^2 + \varepsilon_2^2\right)$$

$$\Delta w_{ij} = 2\alpha\, v_j\, \delta_i \quad onde \quad \delta_i = 3\varepsilon_1 g_{1i} + \varepsilon_2 g_{2i}$$

Ex 2 : Ponderando diferentemente em função da saída desejadas

Falsos negativos x falsos positivos com pesos diferentes

$$F = \frac{1}{4P}\sum_{\forall p} a(y)\varepsilon^2 = \frac{1}{4P}\sum_{\forall p}\left[(M-1)y + (M+1)\right](y-\tilde{y})^2 =$$

$$= \frac{1}{4P}\left[\sum_{\substack{\forall p|y=1 \\ \forall\, pares\ positivos}} 2M(y-\tilde{y})^2 + \sum_{\substack{\forall p|y=-1 \\ \forall\, pares\ negativos}} (y-\tilde{y})^2\right]$$

Tipo	y	$\tilde{y}$	contribuição para F
verdadeiros positivos	1	1	0
verdadeiros negativos	-1	-1	0
falsos positivos	-1	1	1
falsos negativos	1	-1	2M

casos:
doenças humanas,
tuberculose bovina,
etc.

Ex 3: Erro relativo médio quadrático

Erro <u>relativo</u> médio quadrático nas variáveis originais, <u>não escaladas</u>

$$F(\underline{w}) = E\left\{\sum_l \left[\frac{Y_l - \tilde{Y}_l}{Y_l}\right]^2\right\}$$

$$y_l = \frac{1}{\sigma_{Y_l}}\left(Y_l - \mu_{Y_l}\right) \qquad Y_l = \sigma_{Y_l} y_l + \mu_{Y_l} \qquad \tilde{Y}_l = \sigma_{Y_l}\tilde{y}_l + \mu_{Y_l}$$

$$F(\underline{w}) = E\left\{\sum_l \left[\frac{Y_l - \tilde{Y}_l}{Y_l}\right]^2\right\} = E\left\{\sum_l \frac{1}{\left(y_l + \frac{\mu_{Y_l}}{\sigma_{Y_l}}\right)^2}(y_l - \tilde{y}_l)^2\right\} \qquad a(k,l) = \frac{1}{\left(y_l + \frac{\mu_{Y_l}}{\sigma_{Y_l}}\right)^2}$$

No treinamento tipo BP o acréscimo na sinapse w_{ij} é dado por

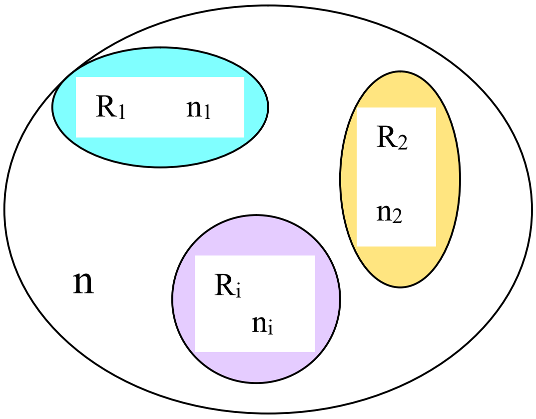
$$\Delta w_{ij} = 2\alpha\, v_j\, \delta_i \quad onde \quad \delta_i = \sum_{l=1}^{m} \frac{1}{\left(y_l + \frac{\mu_{Y_l}}{\sigma_{Y_l}} \right)^2} \varepsilon_l g_{li}$$

Ex 4: Correção do efeito da população local

Cada par pertençe a uma região R_i

Cada região R_i tem uma população n_i

A população total é $n = \sum_{\forall i} n_i$



$$F = \sum_{\forall\, região\, R_i} \boxed{\sum_{\forall\, par \in R_i} f(y, \tilde{y})}$$

peso: n_i pares

peso: 1 par, independe de n_i

$$F = \sum_{\forall\, região\, R_i} \boxed{\frac{1}{n_i} \sum_{\forall\, par \in R_i} f(y, \tilde{y})}$$

a(i,l) = a(i) = 1/n_i

n_i – população da região onde o par pertence.

12.2 Erros não quadráticos

12.2.1 Erros em potências maiores

$$F(\underline{w}) = E\left\{ \sum_l (y_l - \tilde{y}_l)^{2n+2} \right\} \qquad n = 1,2,\ldots$$

prioriza a redução dos maiores erros, mas F(w) é mais abrupta.

E o minimante ? Caso uma saída

$$F = E(y - \tilde{y})^{2n+2}$$

$$\frac{\partial F}{\partial \tilde{y}} = (2n-2)\, E(y - \tilde{y})^{2n+2} = 0$$

Usando propriedades da expansão binomial é possível provar que a equação acima é satisfeita (para 2n+1 impar) se $E(y) = E(\tilde{y})$

12.2.2 Erro absoluto

$$F(\underline{w}) = \underset{k}{E}\left\{ |\, \vec{y}_k - \tilde{\vec{y}}_k \,| \right\} = \underset{k}{E}\left\{ \sum_l |\, y_{kl} - \tilde{y}_{kl} \,| \right\} = \underset{k}{E}\left\{ \sum_l |\, \varepsilon_{kl} \,| \right\}$$

não prioriza a redução dos maiores erros como o emq

- não derivável na origem, necessita procedimentos especiais

$$F(\underline{w}) = \underset{k}{E}\left\{ |\vec{y}_k - \tilde{\vec{y}}_k| \right\} = \underset{k}{E}\left\{ \sum_l |y_{kl} - \tilde{y}_{kl}| \right\} = \underset{k}{E}\left\{ \sum_l |\varepsilon_{kl}| \right\}$$

$$|\varepsilon_{kl}| = \begin{cases} \varepsilon_{kl} = y_{kl} - \tilde{y}_{kl} & se \quad \varepsilon_{kl} > 0 \\ -\varepsilon_{kl} = -(y_{kl} - \tilde{y}_{kl}) & se \quad \varepsilon_{kl} > 0 \end{cases}$$

$$\frac{\partial}{\partial w}|\varepsilon_{kl}| = \begin{cases} -\dfrac{\partial \tilde{y}_{kl}}{\partial w} & se \quad \varepsilon_{kl} > 0 \\ \dfrac{\partial \tilde{y}_l}{\partial w} & se \quad \varepsilon_{kl} > 0 \end{cases}$$

$$\frac{\partial}{\partial w}|\varepsilon_{kl}| = -a(k,l)\frac{\partial \tilde{y}_{kl}}{\partial w} = -sign(\varepsilon_{lk})\frac{\partial \tilde{y}_{kl}}{\partial w}$$

12.2.3 Erro MAPE – Mean Absolute Percentual Error

$$F(\underline{w}) = \underset{k}{E}\left\{ \frac{\left|\vec{Y} - \tilde{\vec{Y}}\right|}{\left|\vec{Y}\right|} \right\}$$

variáveis não escaladas

não prioriza a redução dos maiores erros percentuais, como o epmq.

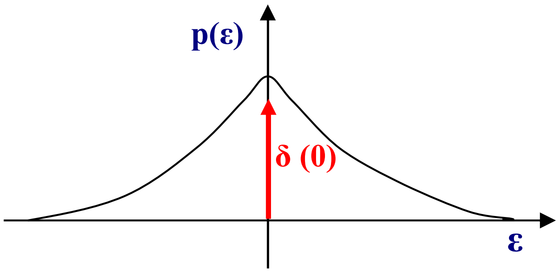
- não derivável na origem, necessita procedimentos especiais

12.3 Entropia

O objetivo na minimização do erro ε é

obter o histograma de p(ε) com média nula (fácil) e o mais estreito possível

ideal: p(ε) = δ (0) (delta de Dirac, erro nulo)

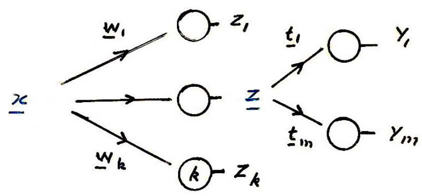


Se a distribuição do erro p(ε) for Gaussiana, o objetivo é alcançado minimizando o erro médio quadrático (a variância de p(ε))

Se a distribuição do erro p(ε) não for Gaussiana, o objetivo é alcançado minimizando a entropia (J.C. Príncipe)

Entropia Relativa (à minimizar)

$$F_{ER} = \underset{\forall\, pares}{E} \sum_{l} \left[\left(1+y_l\right) \ln \frac{1+y_l}{1+\tilde{y}_l} + \left(1-y_l\right) \ln \frac{1-y_l}{1-\tilde{y}_l} \right]$$



Por BP regra delta

$$\Delta w = -\alpha \frac{\partial f}{\partial w} = -\alpha \frac{\partial f}{\partial \tilde{y}} \frac{\partial \tilde{y}}{\partial u} \frac{\partial u}{\partial w}$$

Camada de saída

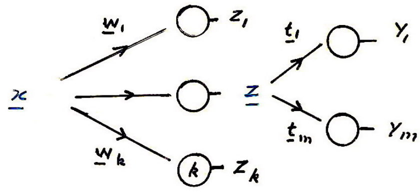
$u \rightarrow \tilde{y} = tgh\, u$

Para a entropia relativa, após alguma álgebra

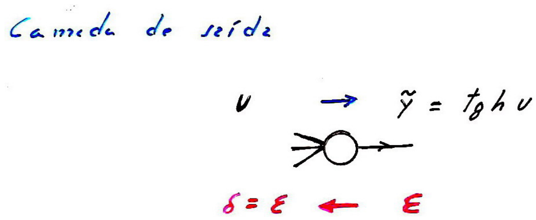
$$\Delta w_{ER} = -\alpha \frac{\partial}{\partial w}\left[(1+y_l)\ln\frac{1+y_l}{1+\tilde{y}_l} + (1-y_l)\ln\frac{1-y_l}{1-\tilde{y}_l}\right] =$$

$$= -\alpha \frac{\partial}{\partial \tilde{y}}\left[(1+y_l)\ln\frac{1+y_l}{1+\tilde{y}_l} + (1-y_l)\ln\frac{1-y_l}{1-\tilde{y}_l}\right]\frac{\partial \tilde{y}}{\partial u}\frac{\partial u}{\partial w} =$$

$$= -\alpha\left[-2\frac{y-\tilde{y}}{1-\tilde{y}^2}\right](1-\tilde{y}^2)\frac{\partial u}{\partial w} = 2\alpha\varepsilon\frac{\partial u}{\partial w}$$



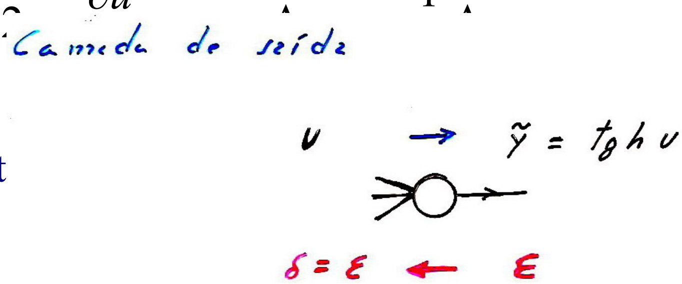
$$\Delta w_{ER} = 2\alpha\varepsilon\frac{\partial u}{\partial w}$$



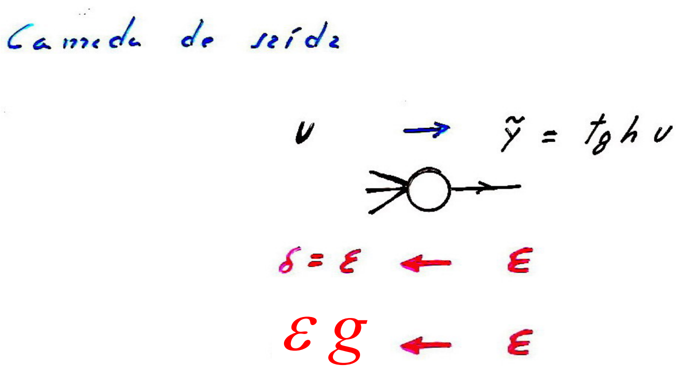
Comparando com o EMQ

Retropropagação do erro

$$\Delta w_{EMQ} = 2\alpha\varepsilon g\frac{\partial u}{\partial w} \qquad \Delta w_{ER} = \quad \partial u \quad 1$$

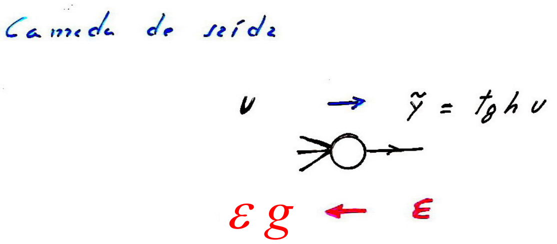


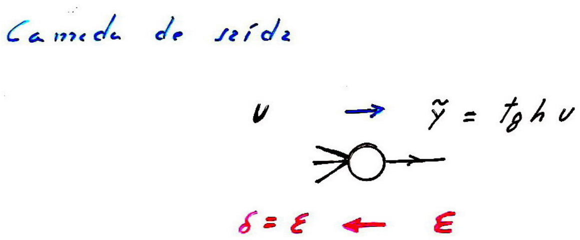
Como $\Delta w_{EMQ} = g\,\Delta w_{ER}$ o minimant
convergência é diferente.



Retropropagação do erro: a única diferença é que para minimizar a ER usa-se
g =1 (mas apenas para a camada de saída)

EMQ





Retropropagação do erro – ganho dos neurônios

Redes com uma camada

Camadas de saída com neurônios tgh

Ganho dos neurônios da camada de saída $g_i = 1$

Camada de saída com neurônios lineares

Ganho dos neurônios da camada de saída $g_i = \dfrac{1}{\left(1 - \tilde{y}_i^2\right)}$

Redes com duas camadas

Camadas intermediária e de saída com neurônios tgh

Ganho dos neurônios de saída $g_i = 1$

Ganho dos neurônios da camada intermediária $g_i = \left(1 - z_i^2\right)$

Camada intermediária neurônios tgh e camada de saída com neurônios lineares

Ganho dos neurônios de saída $g_i = \dfrac{1}{\left(1 - \tilde{y}_i^2\right)}$

Ganho dos neurônios da camada intermediária $g_i = \left(1 - z_i^2\right)$

12.4 - Erro lógico (já visto em classificadores)

$$y \in \{-1,+1\} \qquad \tilde{y} = tgh(u) \in (-1,+1)$$

$$\tilde{y}_{\log ico} = sign(\tilde{y})$$

Erro médio quadrático $\dfrac{1}{N}\sum_i \left(y_i - \tilde{y}_i\right)^2$

Erro de classificação $\dfrac{1}{4N}\sum_{i=1}^{N} \left(y_i - \tilde{y}_{i\ lógico}\right)^2$

% de classificações erradas em y_i

Obs: cuidado ao usar escala log porque $\varepsilon_{classificação} = 0$ não é incomum.