Aprendizagem de Máquina

Alessandro L. Koerich

Programa de Pós-Graduação em Informática
Pontifícia Universidade Católica do Paraná (PUCPR)

Redes Neurais Artificiais

Plano de Aula

- Introdução
- Motivação Biológica
- *Perceptron*
- Superfície de Decisão
- Descida do Gradiente
- Redes Multicamadas (*Multilayer*)
- Retropropagação (*Backpropagation*)
- Generalização e Sobreajuste
- Resumo

Referências

- Duda R., Hart P., Stork D. Pattern Classification 2ed. Willey Interscience, 2002. Capítulo 6
- Mitchell T. Machine Learning. WCB McGraw–Hill, 1997. Capítulo 4.
- Haykin S. Neural Networks: A Comprehensive Foundation (2nd Edition) 842 pages Prentice Hall; 2nd edition (July 6, 1998) ISBN: 0132733501
- Bishop C. Neural Networks for Pattern Recognition. 504 pages. Oxford University Press (January 1996) ISBN: 0198538642

Introdução

- Redes Neurais Artificiais (RNAs) fornecem um método geral e prático para a aprendizagem de funções de valor real e de valor discreto a partir de exemplos.

- Algoritmos como o *backpropagation* (retro-propagação) utilizam a “descida do gradiente” para ajustar os parâmetros das redes para melhor adaptar um conjunto de treinamento de pares entrada – saída (ou vetor de atributos – valor do conceito alvo).

- A aprendizagem de redes neurais é robusta a erros e ruídos nos dados de treinamento.

Introdução

- Modelo inspirado na aprendizagem de sistemas biológicos → redes complexas de neurônios interconectados.
- Mas é um modelo impreciso: muitas complexidades não são modeladas pelas RNAs.
- Também chamadas de Artificial Neural Networks (ANN) ou simplesmente Neural Networks (NN).

Motivação Biológica

- Considere os seres humanos:
  - Tempo de chaveamento do neurônios: 0.001 seg
  - Número de neurônios: 10.000.000.000 (10^{10})
  - Conexões por neurônio: 10.000 a 100.000 (10^4 – 10^5)
  - Tempo para o reconhecimento de uma cena: 0.1 seg
  - muita computação paralela !!

- Propriedade de redes neurais artificiais (RNAs):
  - Muitas unidades de limiar similares aos neurônios
  - Muitas interconexões ponderadas entre unidades
  - Processo altamente paralelo e distribuído
  - Ênfase em ajustar pesos automaticamente

Quando Considerar RNAs

- Entrada discreta ou de valor real de alta dimensão
- Saída discreta ou de valor real
- Saída for um vetor de valores
- Dados possivelmente ruidosos
- Forma da função alvo é desconhecida
- Leitura humana dos resultados não é importante

Problemas Apropriados para RNAs

- Instâncias são representadas por pares atributo-valor.
  - A função alvo a ser aprendida é descrita por um vetor de características (números reais).
- O valor do conceito alvo (i.e. a saída da função alvo) pode ser valores discretos, valores reais ou vetores de valores discretos e reais.
  - Geralmente o valor de cada saída é um número real entre 0 e 1 que corresponde a confiança da predição.

Problemas Apropriados para RNAs

- Os exemplos de treinamento podem conter erros.
- Grandes tempos de treinamento são aceitáveis.
  - Algoritmos de treinamento de redes geralmente consomem muito tempo (de poucos segundos a muitas horas.
  - Depende do número de pesos, número de exemplos de treinamento e outros parâmetros.

Problemas Apropriados para RNAs

- Avaliação rápida da função alvo aprendida.
  - Uma vez treinada, a avaliação da rede, dada uma nova instância, é geralmente muito rápida.

- A habilidade dos humanos entenderem a função alvo aprendida não é importante.
  - Geralmente é difícil interpretarmos os pesos aprendidos pelas redes neurais.
  - Fazem menos sentido do que um conjunto de regras (C4.5).

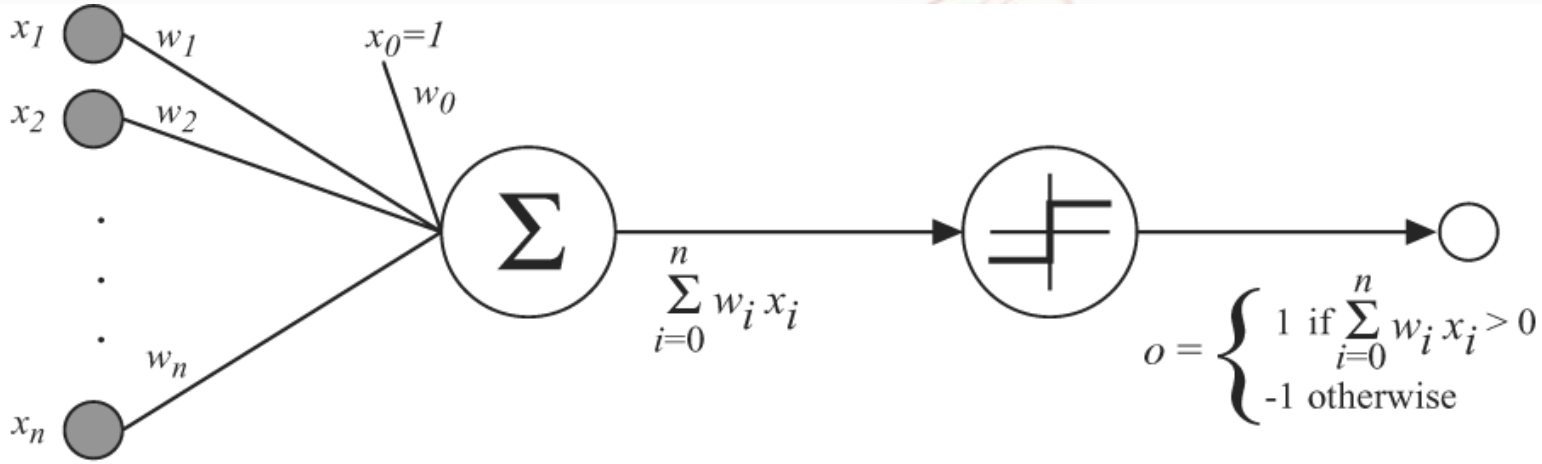
Perceptron

- Rede neural elementar baseada em uma unidade chamada *Perceptron* criada por Rosenblatt em 1958.

- Um *perceptron*:
  1. Recebe um vetor de entradas de valor real
  2. Calcula uma combinação linear destas entradas
  3. Fornece na saída:
     - "+1" se o resultado é maior que algum limiar
     - "–1" caso contrário.

- Mais precisamente, fornecidas as entradas $x_1 \ldots x_n$, a saída $o(x_1 \ldots x_n)$ calculada pelo *perceptron* é . . .

Perceptron



$$o(x_1, \ldots, x_n) = \begin{cases} 1 & \text{if } w_0 + w_1 x_1 + \cdots + w_n x_n > 0 \\ -1 & \text{otherwise.} \end{cases}$$

Algumas vezes utilizaremos notação vetorial simplificada:

$$o(\vec{x}) = \begin{cases} 1 & \text{if } \vec{w} \cdot \vec{x} > 0 \\ -1 & \text{otherwise.} \end{cases}$$

Perceptron

- onde:
  - Cada elemento w_i é uma constante de valor real, ou peso, que determina a contribuição da entrada x_i na saída do *perceptron*.
- A aprendizagem do *perceptron* envolve:
  - A escolha dos valores dos pesos w_o a w_n.
- A camada de entrada deve possuir uma unidade especial conhecida como *bias*, que é um termo constante que não depende de nenhum valor de entrada.
- O bias altera a posição da superfície de separação.

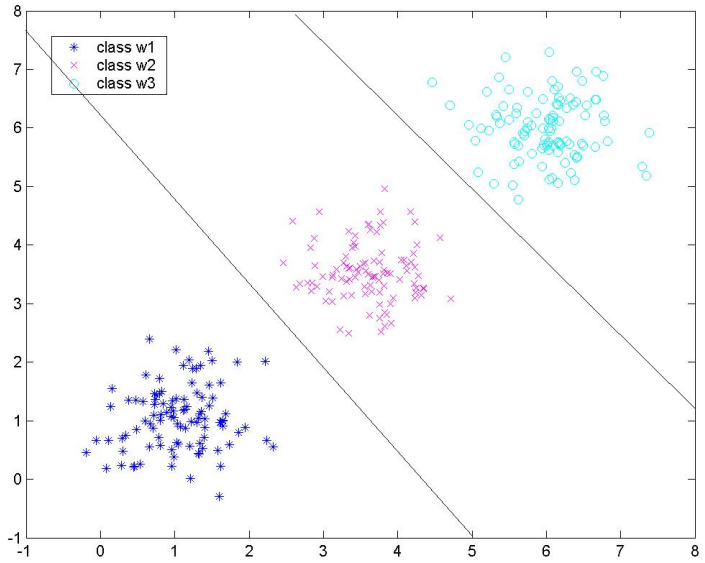
Superfícies de Decisão

- Podemos “ver” o *perceptron* como uma superfície de separação em um espaço n-dimensional de instâncias.

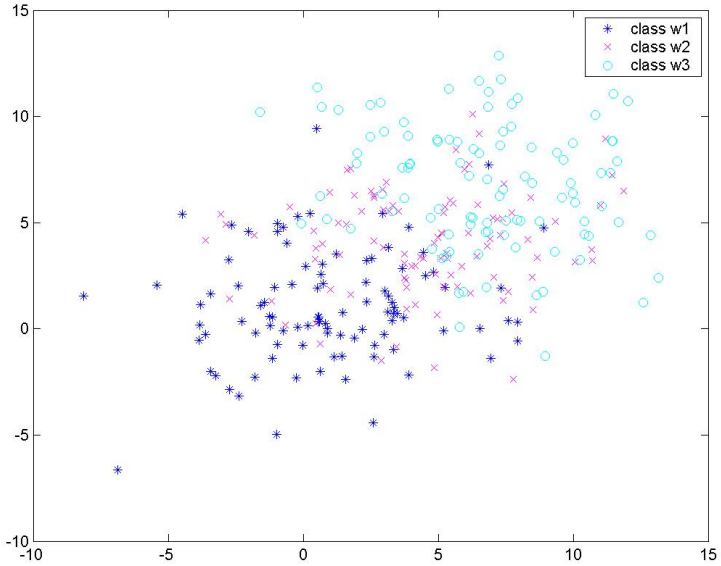
- O *perceptron* fornece “1” para instâncias dispostas em um lado do hiperplano e “-1” para instâncias dispostas no outro lado.

- Um único *perceptron* consegue separar somente conjuntos de exemplo linearmente separáveis.

Superfícies de Decisão

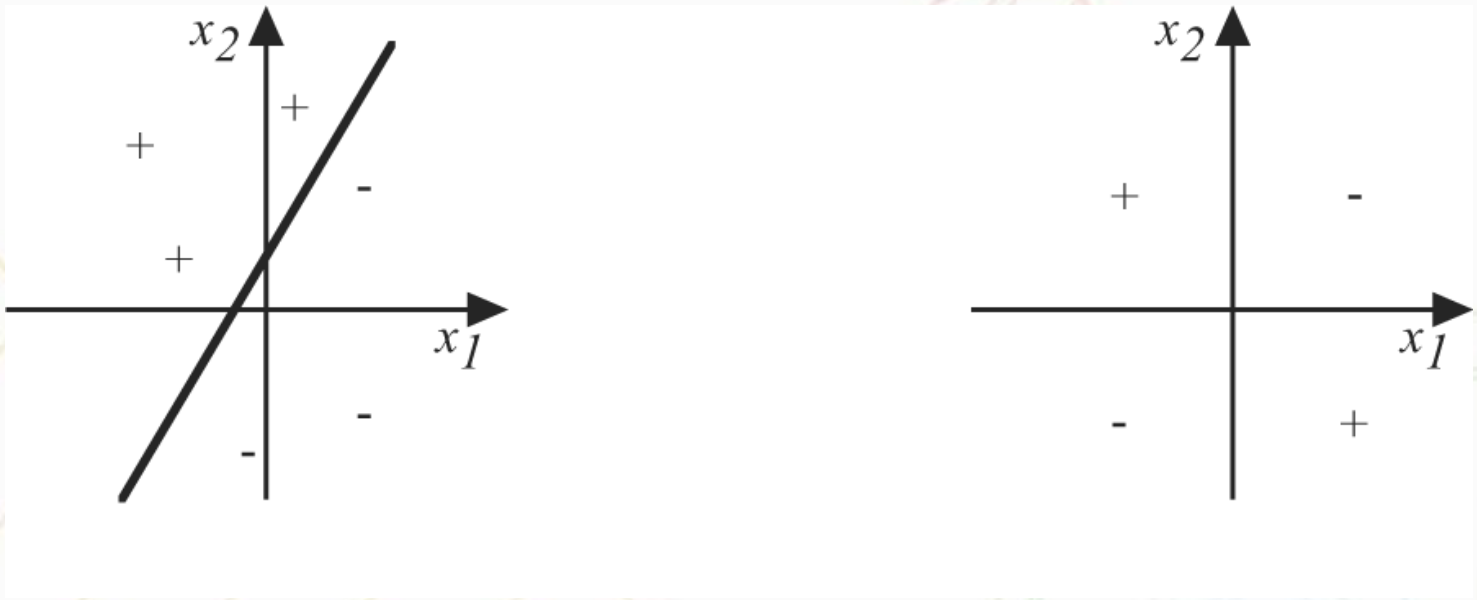


Linearmente Separável



Linearmente Não–Separável

Superfícies de Decisão



- Superfície de separação para um *perceptron* de duas entradas (x_1, x_2).

- Mas algumas funções não são representáveis
  - Por exemplo, classes não linearmente separáveis

Superfícies de Decisão

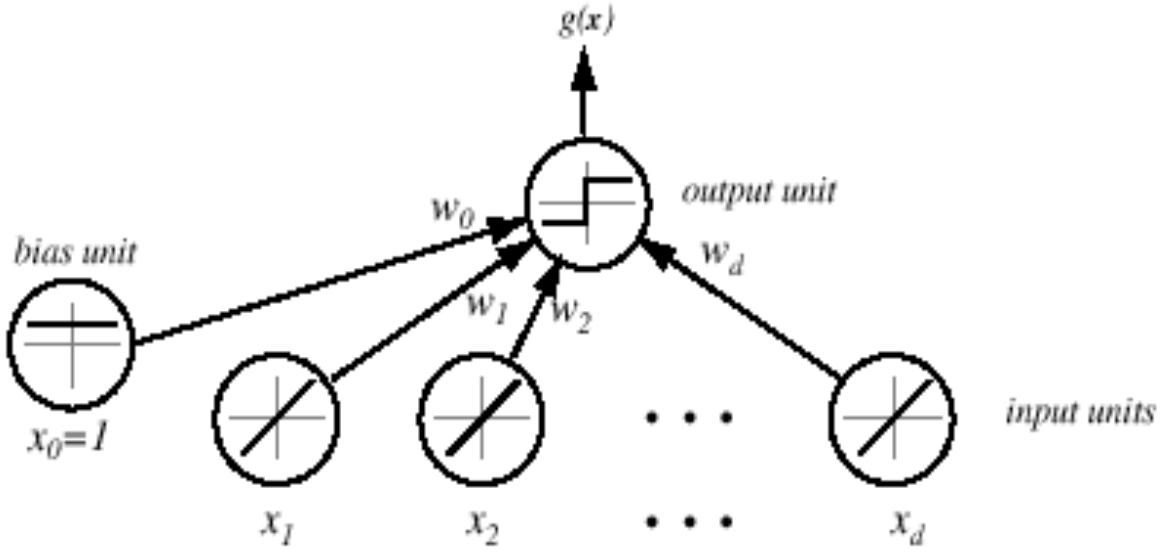


FIGURE 5.1. A simple linear classifier having d input units, each corresponding to the values of the components of an input vector. Each input feature value x_i is multiplied by its corresponding weight w_i; the effective input at the output unit is the sum all these products, $\sum w_i x_i$. We show in each unit its effective input-output function. Thus each of the d input units is linear, emitting exactly the value of its corresponding feature value. The single bias unit unit always emits the constant value 1.0. The single output unit emits a $+1$ if $\mathbf{w}^t\mathbf{x} + w_0 > 0$ or a -1 otherwise. From: Richard O. Duda, Peter E. Hart, and David G. Stork, *Pattern Classification*. Copyright © 2001 by John Wiley & Sons, Inc.

Superfícies de Decisão

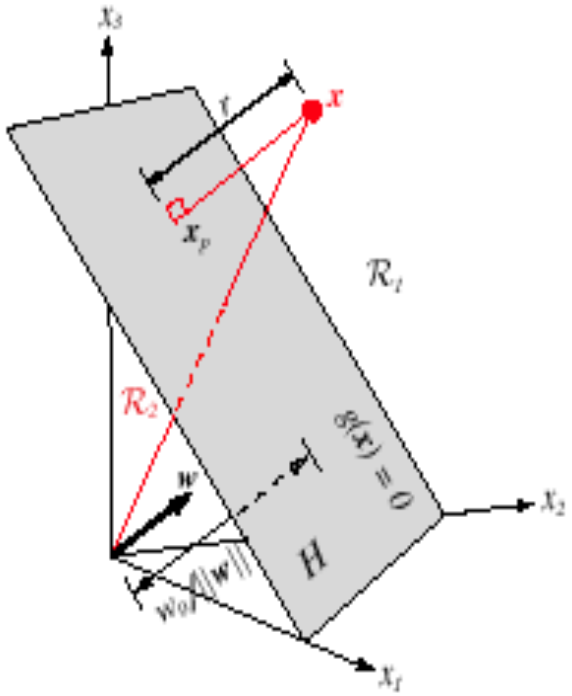


FIGURE 5.2. The linear decision boundary H, where $g(\mathbf{x}) = \mathbf{w}^t\mathbf{x}+w_0 = 0$, separates the feature space into two half-spaces $\mathcal{R}_1$ (where $g(\mathbf{x}) > 0$) and $\mathcal{R}_2$ (where $g(\mathbf{x}) < 0$). From: Richard O. Duda, Peter E. Hart, and David G. Stork, *Pattern Classification*. Copyright © 2001 by John Wiley & Sons, Inc.

Regra de Treinamento *Perceptron*

- Como aprender os pesos para um *perceptron*?
  - Problema: determinar um vetor de pesos que faça o *perceptron* produzir a saída correta (–1 ou +1) para cada um dos exemplos de treinamento.
  - Solução: Começar com um vetor de pesos aleatórios e aplicar iterativamente a regra *perceptron* para cada exemplo de treinamento, modificando os pesos cada vez que ele classificar um exemplo erroneamente.
    - Este processo é repetido várias vezes até que o *perceptron* classifique todos os exemplos de treinamento corretamente.

Regra de Treinamento *Perceptron*

- Os pesos do *perceptron* são modificados a cada passo de acordo com a regra de treinamento do *perceptron*, que modifica o peso w_i associado a entrada x_i de acordo com a regra:

$$w_i \leftarrow w_i + \Delta w_i$$

onde

$$\Delta w_i = \eta(t - o)x_i$$

  - t é o valor alvo para o exemplo de treinamento.
  - o é a saída gerada pelo *perceptron*.
  - η é uma constante pequena (0.1) chamada de taxa de aprendizagem.

Regra de Treinamento *Perceptron*

- Se o exemplo de treinamento é classificado corretamente:

$$(t - o) = \text{zero} \rightarrow \Delta w_i = 0$$

- Se o exemplo de treinamento é classificado incorretamente, o valor de Δw_i é alterado:
  - Se $x_i = 0.8$, $\eta = 0.1$, $t = 1$, $o = -1$
  - A atualização do peso será:

$$\Delta w_i = \eta(t - o)x_i = 0.1(1 - (-1))0.8 = 0.16$$

Regra de Treinamento *Perceptron*

- Pode–se provar que este procedimento de aprendizagem converge dentro de um número finito de passos quando:
  - As classes dos dados de treinamento são linearmente separáveis;
  - η é suficientemente pequeno.

- Porém: falha em convergir se as classes forem linearmente não-separáveis.

- Solução: algoritmo descida do gradiente

Descida do Gradiente

- Para dados não linearmente separáveis, a Regra Delta converge em direção a aproximação que melhor se ajusta ao conceito alvo.

- Ideia chave: usar a descida do gradiente para procurar o melhor vetor de pesos.

- Considerando uma unidade linear, isto é, um *perceptron* sem limiar:

Descida do Gradiente

- Considere um *perceptron* onde:

$$o = w_0 + w_1 x_1 + \ldots + w_n x_n$$

- Especificando uma medida para o erro de treinamento de uma hipótese (vetor de pesos) relativamente aos exemplos de treinamento:

$$E[\vec{w}] = \frac{1}{2} \sum_{d \in D} (t_d - o_d)^2$$

  - D é o conjunto de treinamento.
  - t_d é o valor alvo para o exemplo de treinamento d.
  - o_d é a saída da unidade linear para o exemplo d.
  - $E(w)$ é a metade do quadrado da diferença entre saída alvo e unidade linear de saída somada sobre todos os exemplos de treinamento.

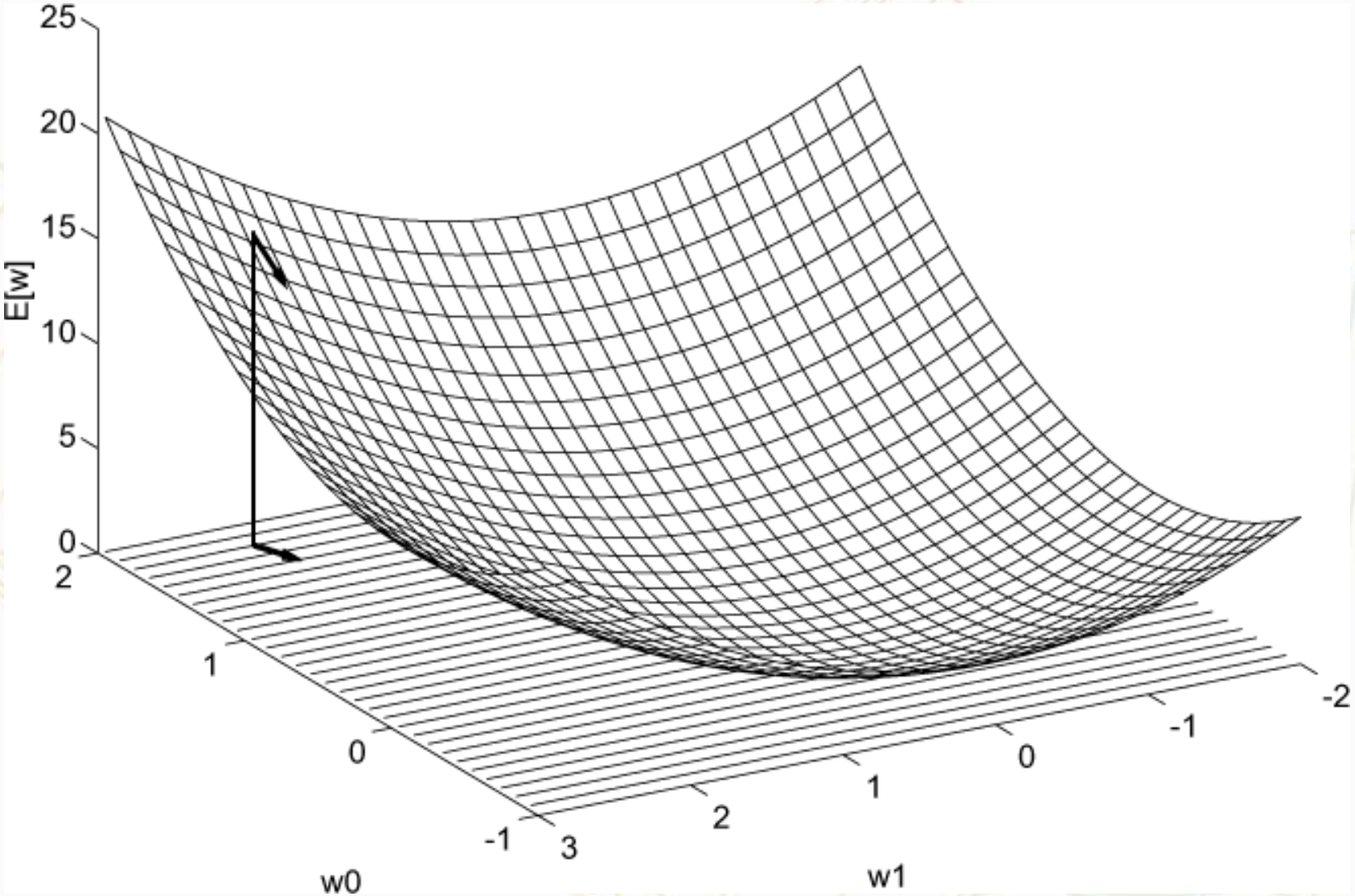
Descida do Gradiente

- A descida do gradiente determina um vetor de pesos que minimiza *E*, começando com um vetor inicial de pesos arbitrário e modificando-o repetidamente em pequenos passos.

- A cada passo, o vetor de pesos é alterado na direção que produz a maior queda ao longo da superfície de erro.

- Este processo continua até atingir um erro mínimo global.

Descida do Gradiente



Descida do Gradiente

- Gradiente: $$\nabla E[\vec{w}] = \left[\frac{\partial E}{\partial w_0}, \frac{\partial E}{\partial w_1}, \ldots, \frac{\partial E}{\partial w_n} \right]$$

- Regra de treinamento para a descida do gradiente:

onde:

$$\vec{w} \leftarrow \vec{w} + \Delta \vec{w}$$

$$\Delta \vec{w} = -\eta \nabla E[\vec{w}]$$

e

$$\Delta w_i = -\eta \frac{\partial E}{\partial w_i}$$

Descida do Gradiente

$$\frac{\partial E}{\partial w_i} = \frac{\partial}{\partial w_i}\frac{1}{2}\sum_d (t_d - o_d)^2$$

$$= \frac{1}{2}\sum_d \frac{\partial}{\partial w_i}(t_d - o_d)^2$$

$$= \frac{1}{2}\sum_d 2(t_d - o_d)\frac{\partial}{\partial w_i}(t_d - o_d)$$

$$= \sum_d (t_d - o_d)\frac{\partial}{\partial w_i}(t_d - \vec{w}_i\vec{x}_d)$$

$$\frac{\partial E}{\partial w_i} = \sum_d (t_d - o_d)(-x_{id})$$

- onde x_{id} indica um componente único de entrada x_i para o exemplo de treinamento d.

Descida do Gradiente

- Assim, a regra para atualização dos pesos para o gradiente descendente é

$$\Delta w_i = \eta \sum_{d \in D} (t_d - o_d) x_{id}$$

Descida do Gradiente

- Resumindo, o algoritmo descida do gradiente para a aprendizagem de unidade lineares:

    1. Pegar um vetor inicial aleatório de pesos;
    2. Aplicar a unidade linear para todos os exemplos de treinamento e calcular Δw_i para cada peso de acordo com a equação anterior;
    3. Atualizar cada peso w_i adicionando Δw_i e então repetir este processo.

- O algoritmo convergirá para um vetor de pesos com erro mínimo.

Descida do Gradiente

GRADIENT-DESCENT(*training_examples*, η)

Each training example is a pair of the form $\langle \vec{x}, t \rangle$, *where* $\vec{x}$ *is the vector of input values, and* t *is the target output value.* η *is the learning rate (e.g., .05).*

- Initialize each w_i to some small random value
- Until the termination condition is met, Do
  - Initialize each Δw_i to zero.
  - For each $\langle \vec{x}, t \rangle$ in *training_examples*, Do
    - Input the instance $\vec{x}$ to the unit and compute the output o
    - For each linear unit weight w_i, Do

$$\Delta w_i \leftarrow \Delta w_i + \eta(t - o)x_i \tag{T4.1}$$

  - For each linear unit weight w_i, Do

$$w_i \leftarrow w_i + \Delta w_i \tag{T4.2}$$

Regra *Perceptron* X Descida do Gradiente

- A regra de treinamento *perceptron* tem sucesso se:
  - Exemplos de treinamento são linearmente separáveis
  - Taxa de aprendizagem η for suficientemente pequena

- Regra de treinamento descida do gradiente:
  - Convergência garantida um vetor de pesos com erro quadrado mínimo
  - Dada uma taxa de aprendizagem η suficientemente pequena
  - Mesmo quando dados de treinamento contém ruído
  - Mesmo quando dados de treinamento não forem separáveis

Limitações

- O *perceptron* expressa somente superfícies de separação lineares
- Fornece somente um classificador binário (-1, +1)
- Como tratar problemas que exijem superfícies de separação não-lineares e problemas multi-classes?

Redes Multicamadas

- *Perceptrons* expressam somente superfícies de decisão linear.
- Redes multicamadas treinadas pelo algoritmo *backpropagation* são capazes de expressar uma rica variedade de superfícies de decisão não lineares, mas...
- Redes multicamadas podem representar superfícies de decisão altamente não lineares, mas...

Redes Multicamadas

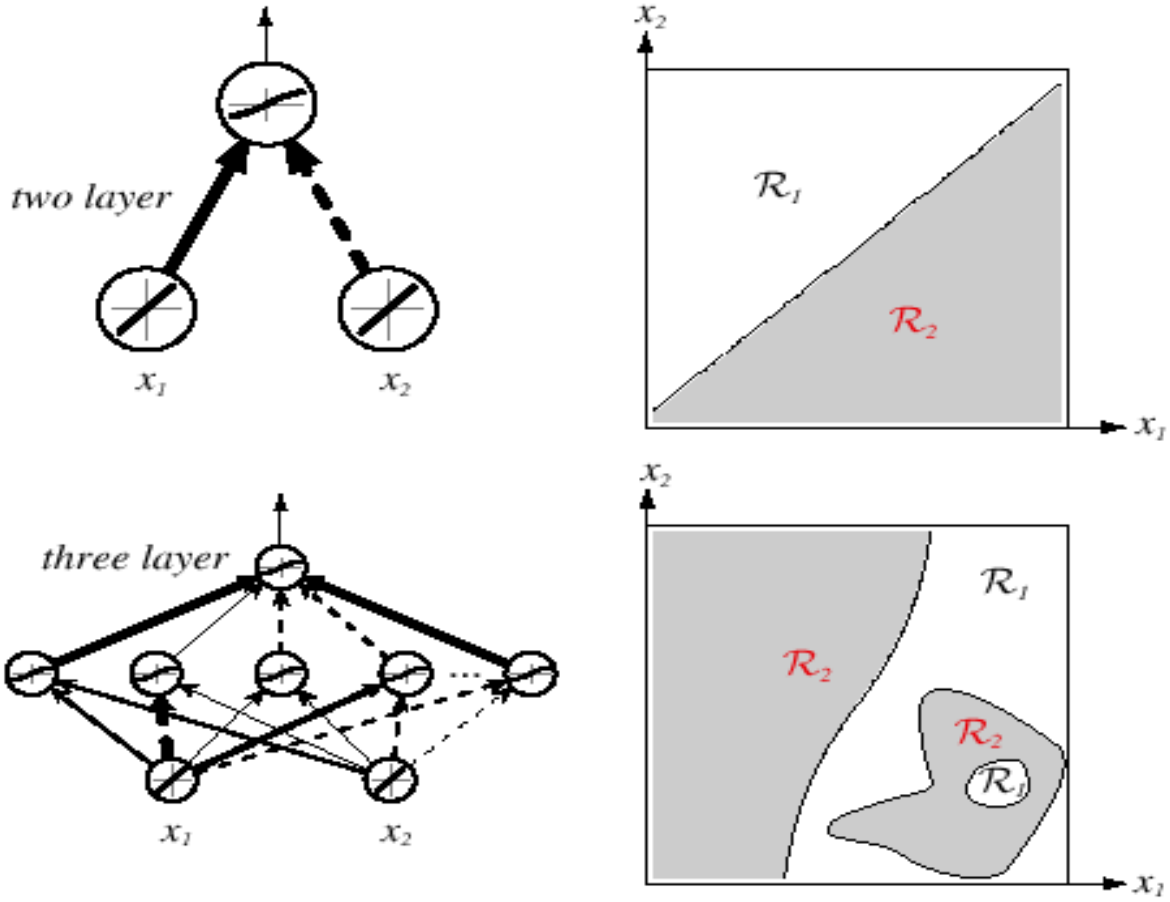
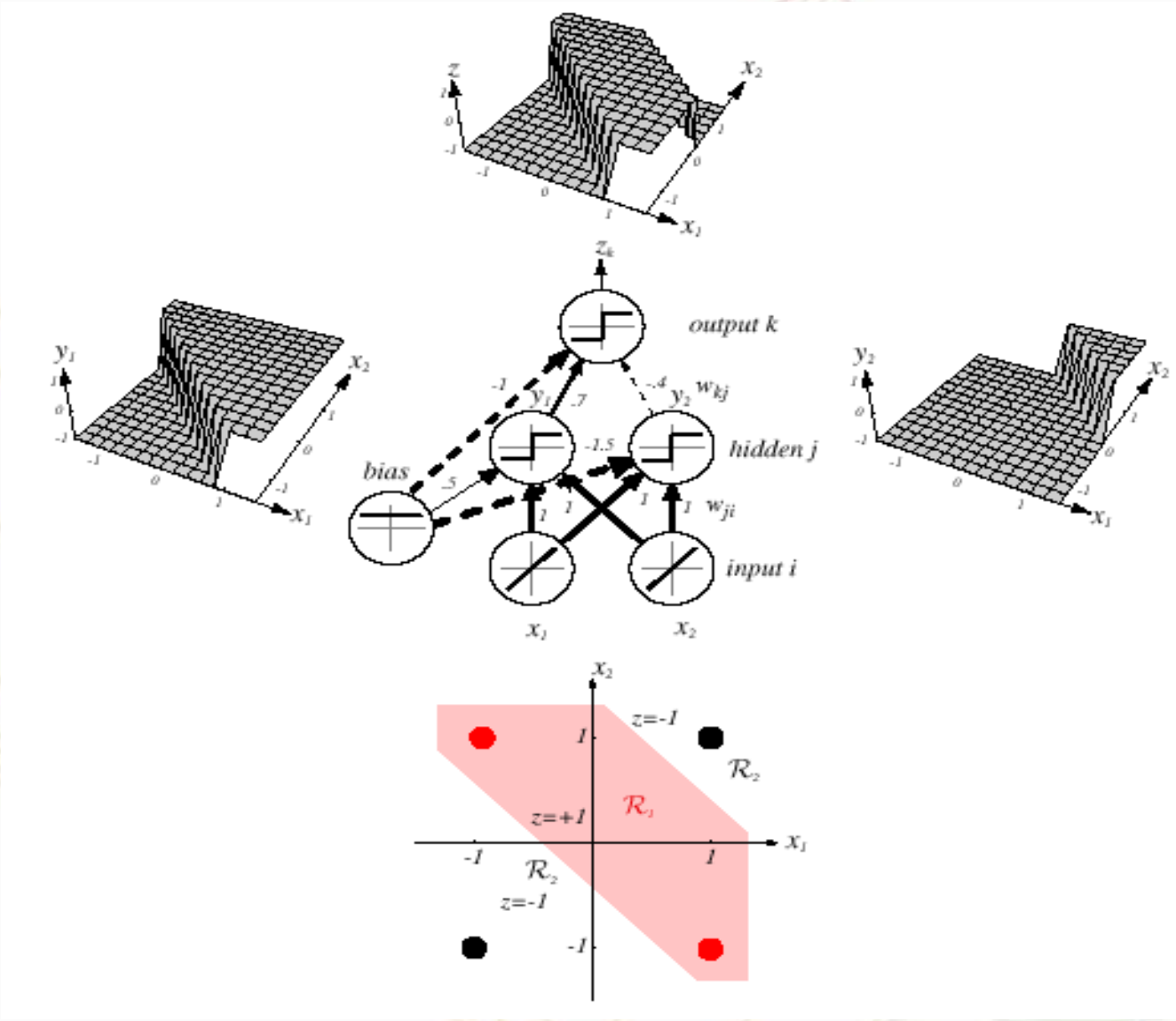


FIGURE 6.3. Whereas a two-layer network classifier can only implement a linear decision boundary, given an adequate number of hidden units, three-, four- and higher-layer networks can implement arbitrary decision boundaries. The decision regions need not be convex or simply connected. From: Richard O. Duda, Peter E. Hart, and David G. Stork, *Pattern Classification*. Copyright © 2001 by John Wiley & Sons, Inc.

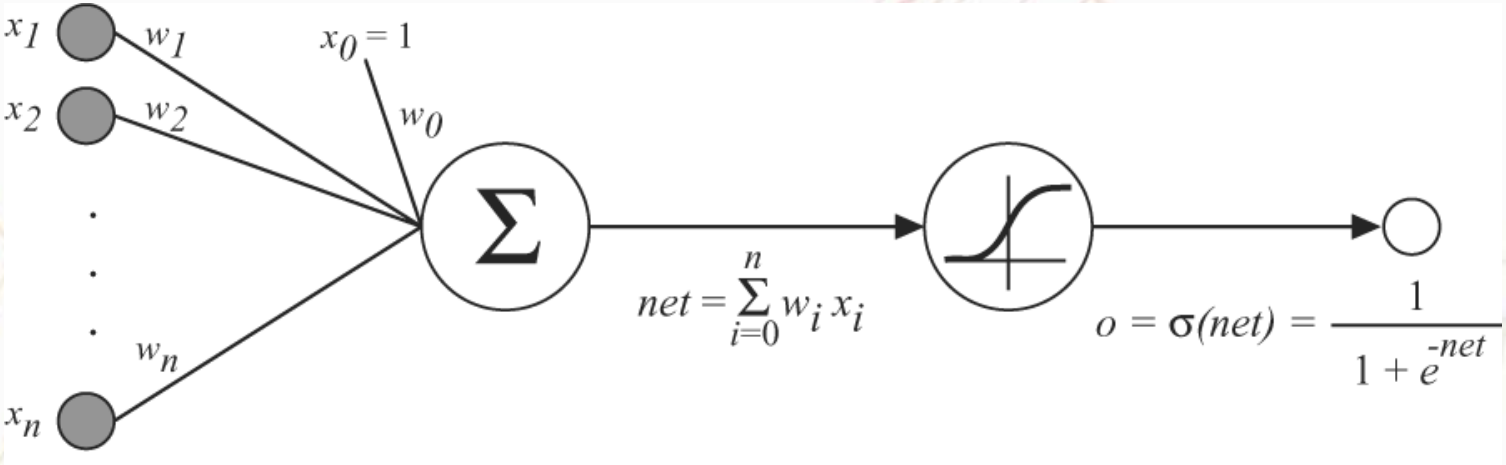
Redes Multicamadas

- Que tipo de unidades devemos utilizar como base de uma rede multicamadas?
  - Lineares ?
  - Múltiplas camadas de unidades lineares cascateadas produzem somente funções lineares (Fig.).
- Redes são capazes de representar funções altamente não lineares.
  - Unidade cuja saída seja uma função não-linear de suas entradas → unidade sigmoidal.

Redes Multicamadas



Unidade Sigmoidal

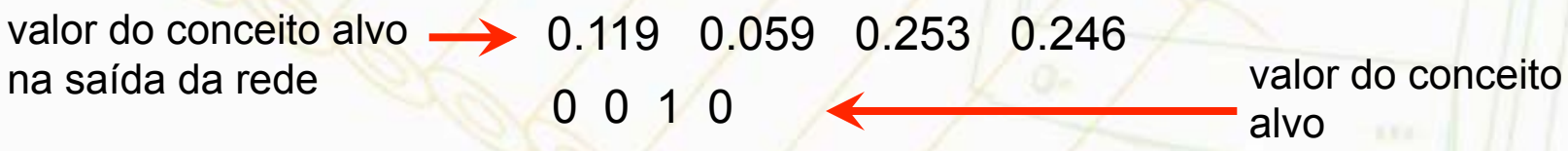


- σ (x) é a função sigmoidal: $\frac{1}{1+e^{-x}}$
- Propriedade interessante: $\frac{d\sigma(x)}{dx} = \sigma(x)(1-\sigma(x))$
- Podemos derivar regras do gradiente descendente para treinar:
  - Uma unidade sigmoidal
  - Redes multicamadas de unidades sigmoidais → *backpropagation*

Algoritmo *Backpropagation*

- Aprende os pesos para uma rede multicamadas, dada uma rede com um número fixo de unidades e interconexões.

- O algoritmo backpropagation emprega a "descida do gradiente" para minimizar o erro quadrático entre a saída da rede e os valores alvos para estas saídas.



Erro = (valor do conceito alvo real) – (valor do conceito alvo estimado)

Algoritmo *Backpropagation*

- Como temos múltiplas unidades de saída, redefinimos *E* como sendo a soma dos erros sobre todas as unidades de saída da rede:

$$E(\vec{w}) \equiv \frac{1}{2} \sum_{d \in D} \sum_{k \in outputs} (t_{kd} - o_{kd})^2$$

  - outputs é o conjunto de unidades de saída na rede
  - t_{kd} valor alvo associado com a *k*–ésima unidade de saída e exemplo de treinamento *d*.
  - o_{kd} valor de saída associado com a *k*–ésima unidade de saída e exemplo de treinamento *d*.

Algoritmo *Backpropagation*

- Problema de aprendizagem do algoritmo *backpropagation*:
  - Busca por todas as vetores de pesos possíveis para todas as unidades da rede.
  - Encontrar o vetor de pesos que minimizem o erro médio quadrático (*E*).

Algoritmo *Backpropagation*

Initialize all weights to small random numbers.
Until satisfied, Do

- For each training example, Do
  1. Input the training example to the network and compute the network outputs
  2. For each output unit k

$$\delta_k \leftarrow o_k(1 - o_k)(t_k - o_k)$$

  3. For each hidden unit h

$$\delta_h \leftarrow o_h(1 - o_h) \sum_{k \in outputs} w_{h,k}\delta_k$$

  4. Update each network weight $w_{i,j}$

$$w_{i,j} \leftarrow w_{i,j} + \Delta w_{i,j}$$

where

$$\Delta w_{i,j} = \eta \delta_j x_{i,j}$$

Algoritmo *Backpropagation*

BACKPROPAGATION($training\_examples, \eta, n_{in}, n_{out}, n_{hidden}$)

Each training example is a pair of the form $\langle \vec{x}, \vec{t} \rangle$, *where* $\vec{x}$ *is the vector of network input values, and* $\vec{t}$ *is the vector of target network output values.*

η *is the learning rate (e.g., .05).* n_{in} *is the number of network inputs,* n_{hidden} *the number of units in the hidden layer, and* n_{out} *the number of output units.*

The input from unit i *into unit* j *is denoted* x_{ji}, *and the weight from unit* i *to unit* j *is denoted* w_{ji}.

- Create a feed-forward network with n_{in} inputs, n_{hidden} hidden units, and n_{out} output units.
- Initialize all network weights to small random numbers (e.g., between $-.05$ and .05).
- Until the termination condition is met, Do
  - For each $\langle \vec{x}, \vec{t} \rangle$ in *training_examples*, Do

Propagate the input forward through the network:

1. Input the instance $\vec{x}$ to the network and compute the output o_u of every unit u in the network.

Propagate the errors backward through the network:

2. For each network output unit k, calculate its error term δ_k

$$\delta_k \leftarrow o_k(1 - o_k)(t_k - o_k) \tag{T4.3}$$

3. For each hidden unit h, calculate its error term δ_h

$$\delta_h \leftarrow o_h(1 - o_h) \sum_{k \in outputs} w_{kh}\delta_k \tag{T4.4}$$

4. Update each network weight w_{ji}

$$w_{ji} \leftarrow w_{ji} + \Delta w_{ji}$$

where

$$\Delta w_{ji} = \eta \, \delta_j \, x_{ji} \tag{T4.5}$$

Algoritmo *Backpropagation*

- Notação:
  - Um índice é atribuído a cada nó da rede, onde nó pode ser uma entrada da rede ou a saída de alguma unidade da rede.
  - x_{ji} indica a entrada a partir do nó *i* para unidade *j* e w_{ji} indica o peso correspondente.
  - δ_n indica o termo do erro associado com a unidade *n*. Similar a (*t*–*o*).

Mais sobre *Backpropagation*

- Descida do gradiente sobre o vetor de pesos inteiro da rede
- Encontrará um erro mínimo local (não necessariamente global)
  - Na prática, geralmente funciona bem (pode executar múltiplas vezes).
- Geralmente inclui peso do momento α

$$\Delta w_{i,j}(n) = \eta \delta_j x_{i,j} + \alpha \Delta w_{i,j}(n-1)$$

Mais sobre *Backpropagation*

- Minimiza o erro sobre os exemplos de treinamento
  - Generalizará bem sobre exemplos subseqüentes?
- O treinamento pode levar milhares de iterações → vagaroso
- A utilização da rede após o treinamento → muito rápida

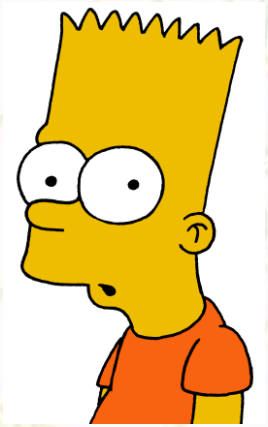
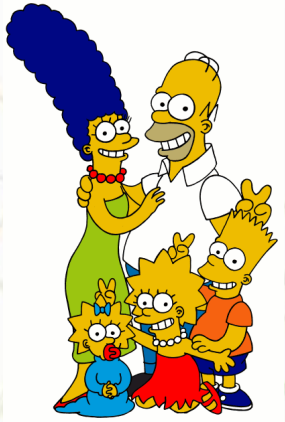
Exemplo

- Dada a imagem de um personagem, ele deve ser classificado corretamente, ou seja, se a imagem for do personagem Bart, ela deve ser classificada pelo algoritmo de aprendizagem como sendo o personagem Bart.

Classes / Valor do Conceito Alvo

Marge 0 0 0 1 **Homer 0 0 1 0** **Bart 0 1 0 0** **Família 1 0 0 0**



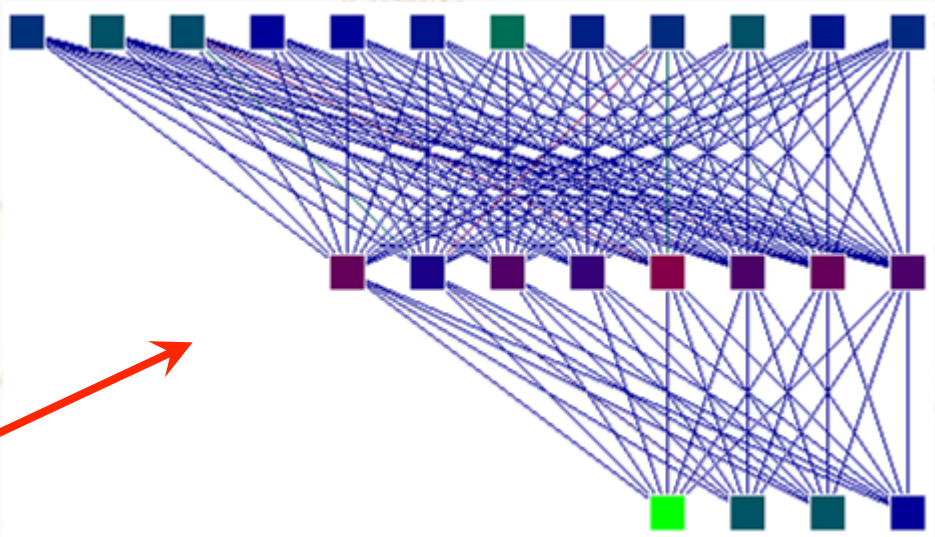



Exemplo

vetor de características

0.43 0.03 0.40 0.19 0.12 0.16 0.04 0.01 0.00 0.01 0.40 0.02
0 0 1 0

valor do conceito alvo associado ao vetor



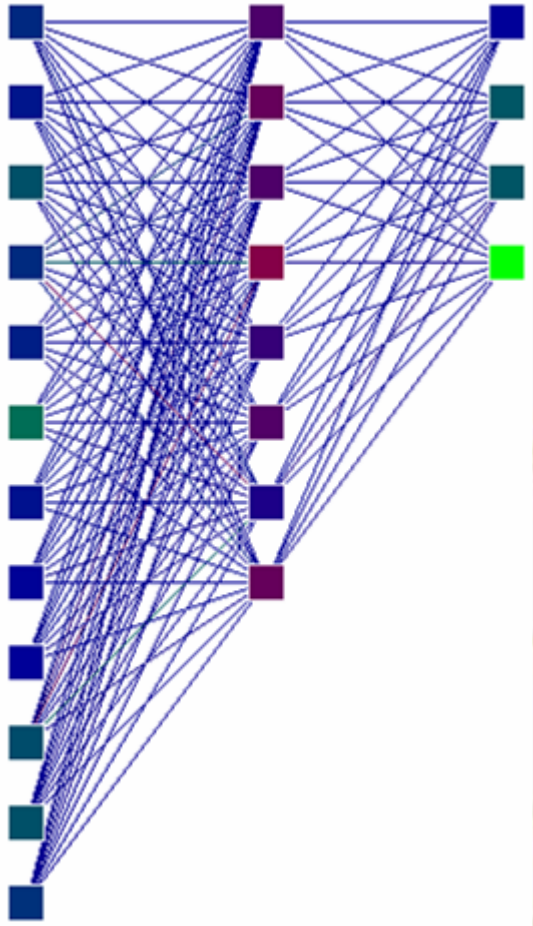
rede neural treinada

0.119 0.059 0.253 0.569

valor do conceito alvo estimado

Erro = (valor do conceito alvo real) – (valor do conceito alvo estimado)

Representação da RNA



unit definition section :

no.	typeName	unitName	act	bias	st	position	act func	out func	sites
1			0.15710	0.00200	i	2, 2, 0	Act_Identity		
2			0.08250	0.00492	i	2, 3, 0	Act_Identity		
3			0.31630	0.00955	i	2, 4, 0	Act_Identity		
4			0.16530	0.00616	i	2, 5, 0	Act_Identity		
5			0.11860	0.00476	i	2, 6, 0	Act_Identity		
6			0.43310	0.00818	i	2, 7, 0	Act_Identity		
7			0.06930	0.00605	i	2, 8, 0	Act_Identity		
8			0.00890	0.00587	i	2, 9, 0	Act_Identity		
9			0.00380	0.00916	i	2,10, 0	Act_Identity		
10			0.29860	0.00922	i	2,11, 0	Act_Identity		
11			0.31760	0.00948	i	2,12, 0	Act_Identity		
12			0.19330	0.00649	i	2,13, 0	Act_Identity		
13			-0.30391	-46.08251	h	5, 2, 0	Act_Identity		
14			-0.40381	-101.31063	h	5, 3, 0	Act_Identity		
15			-0.30793	97.62634	h	5, 4, 0	Act_Identity		
16			-0.52309	160.65987	h	5, 5, 0	Act_Identity		
17			-0.21414	-79.82547	h	5, 6, 0	Act_Identity		
18			-0.32417	135.45871	h	5, 7, 0	Act_Identity		
19			-0.10986	-53.94949	h	5, 8, 0	Act_Identity		
20			-0.39891	-55.78927	h	5, 9, 0	Act_Identity		
21			-0.00000	-0.02777	o	8, 2, 0	Act_Identity		
22			0.33768	165.30469	o	8, 3, 0	Act_Identity		
23			0.33482	380.65833	o	8, 4, 0	Act_Identity		
24			1.03949	260.54959	o	8, 5, 0	Act_Identity		

Representação da RNA: Pesos

connection definition section :
target | site | source:weight
13 | | 1:-0.13528, 2: 0.04893, 3:-0.32746, 4:-0.08375, 5:-0.34121, 6:-0.13578, 7: 0.20121, 8:-0.15724, 9: 0.00921, 10: 0.11034, 11:-0.19678, 12:-0.21812
14 | | 1:-0.55645, 2:-0.11265, 3:-0.49058, 4: 1.67717, 5:-0.26903, 6: 0.20352, 7: 0.21803, 8: 0.06280, 9: 0.28881, 10:-1.36910, 11:-0.10725, 12:-0.26972
15 | | 1: 0.04909, 2: 0.09190, 3: 0.15870, 4: 0.75809, 5:-0.17918, 6:-0.17613, 7: 0.19631, 8:-0.03773, 9:-0.03989, 10:-0.84408, 11:-0.32844, 12:-0.27992
16 | | 1:-0.24912, 2:-0.01193, 3: 0.33076, 4: 3.28676, 5:-0.07993, 6: 0.07680, 7: 0.27817, 8: 0.23728, 9: 0.19800, 10:-3.18631, 11:-0.43482, 12:-0.46441
17 | | 1:-0.20847, 2:-0.00470, 3:-0.37552, 4:-0.00523, 5:-0.25660, 6:-0.02310, 7: 0.13830, 8:-0.09997, 9: 0.06508 10: 0.09921, 11:-0.07738, 12:-0.13684
18 | | 1: 0.20934, 2: 0.16882, 3: 0.24766, 4: 0.26525, 5:-0.23404, 6:-0.32802, 7: 0.22154, 8:-0.12143, 9:-0.14677, 10:-0.46730, 11:-0.41720, 12:-0.30646
19 | | 1: 0.46806, 2: 0.29771, 3:-0.38300, 4:-3.76655, 5:-0.56653, 6:-0.68893, 7: 0.16836, 8:-0.61828, 9:-0.44112, 10: 3.38461, 11:-0.24341, 12:-0.07532
20 | | 1:-0.30391, 2: 0.00103, 3:-0.38042, 4: 0.73989, 5:-0.34855, 6:-0.02838, 7: 0.24377, 8:-0.07767, 9: 0.11745, 10:-0.61078, 11:-0.21432, 12:-0.28737
21 | | 13: 0.00300, 14:-0.00225, 15:-0.00017, 16: 0.00214, 17:-0.00030, 18:-0.00311, 19: 0.00052, 20:-0.00014
22 | | 13:-0.18994, 14:-0.34258, 15: 0.02922, 16: 0.01102, 17:-0.20919, 18: 0.09017, 19:-0.07391, 20:-0.25315
23 | | 13:-0.28049, 14: 0.06916, 15:-0.15606, 16: 0.20976, 17:-0.16213, 18:-0.30594, 19:-0.96542, 20:-0.17005
24 | | 13:-0.34923, 14:-0.43133, 15:-0.33860, 16:-0.53908, 17:-0.24484, 18:-0.36651, 19:-0.19405, 20:-0.44290

MLP 3 Camadas

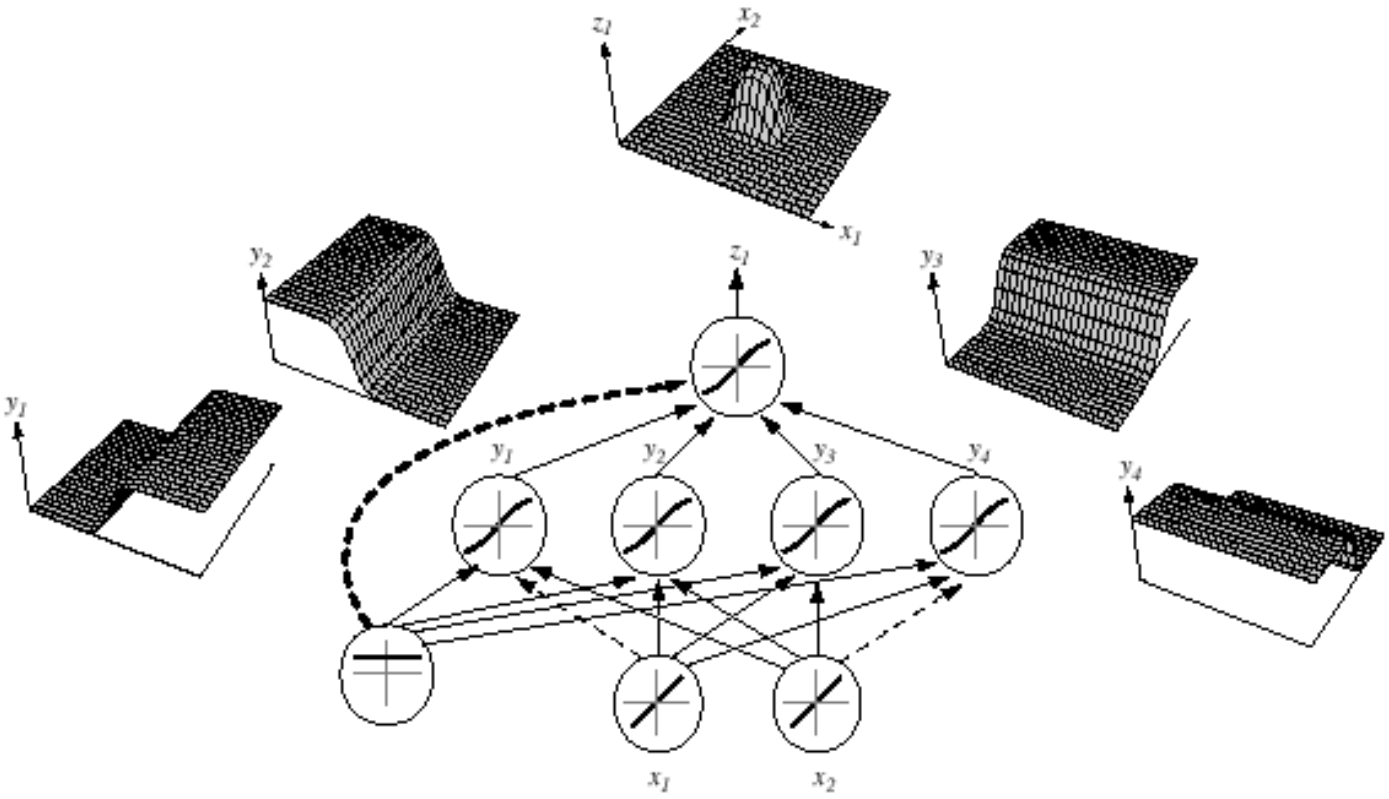


FIGURE 6.2. A 2-4-1 network (with bias) along with the response functions at different units; each hidden output unit has sigmoidal activation function $f(\cdot)$. In the case shown, the hidden unit outputs are paired in opposition thereby producing a "bump" at the output unit. Given a sufficiently large number of hidden units, any continuous function from input to output can be approximated arbitrarily well by such a network. From: Richard O. Duda, Peter E. Hart, and David G. Stork, *Pattern Classification*. Copyright © 2001 by John Wiley & Sons, Inc.

Convergência do Backpropagation

- Descida do gradiente para algum mínimo local
  - Talvez não seja um mínimo global
  - Adicionar momento
  - Descida do gradiente estocástico
  - Treinar múltiplas redes com pesos iniciais diferentes
- Natureza da convergência
  - Inicializar pesos próximo de zero
  - Portanto, redes iniciais quase lineares
  - Progressivamente para funções não lineares com o progresso do treinamento

Capacidades Expressivas de RNAs

- Funções Booleanas
  - Cada função booleana pode ser representada por redes com uma única camada escondida
  - Mas podem necessitar unidades escondidas exponenciais (em número de entradas)

- Funções Contínuas
  - Cada função contínua limitada pode ser aproximada pela rede com uma camada escondida, com um erro arbitrário pequeno.

  - Qualquer função pode ser aproximada por uma rede com duas camadas escondidas para uma precisão arbitrária.

Generalização e Sobreajuste

- A condição de parada do algoritmo *backpropagation* foi deixada em aberto.

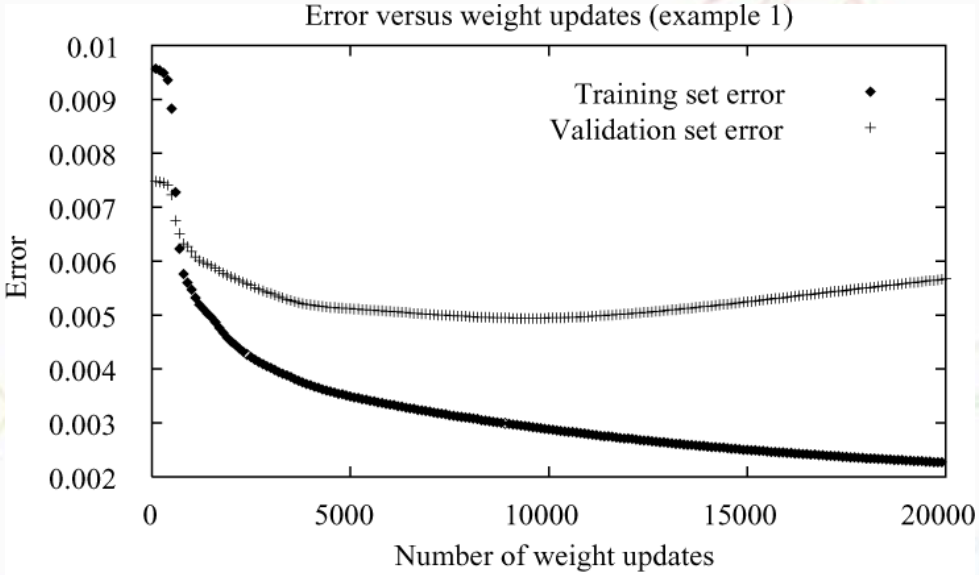
- Quando devemos parar o treinamento, i.e. parar de atualizar os pesos?
  - Escolha óbvia: continuar o treinamento até que o erro (E) seja menor do que um valor pré-estabelecido.
  - Porém, isto implica em sobre ajuste (*overfitting*) !!!

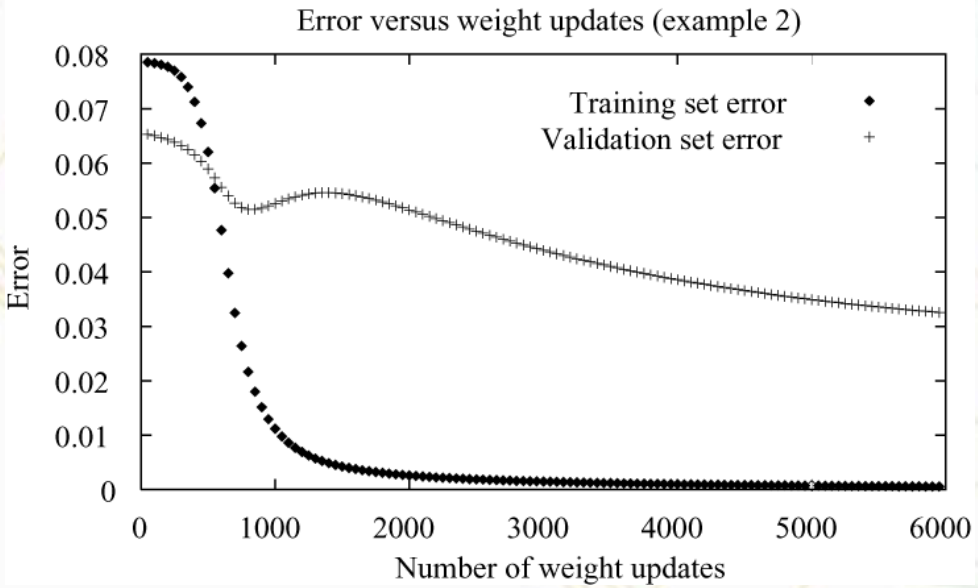
Generalização e Sobreajuste

- O algoritmo *backpropagation* é susceptível a sobre ajustar a rede aos exemplos de treinamento ao preço de reduzir a generalização sobre exemplos novos.

- A Figura a seguir ilustra o perigo de minimizar o erro sobre os dados de treinamento em função do número de iterações (atualização dos pesos).

Generalização e Sobreajuste





Generalização e Sobreajuste

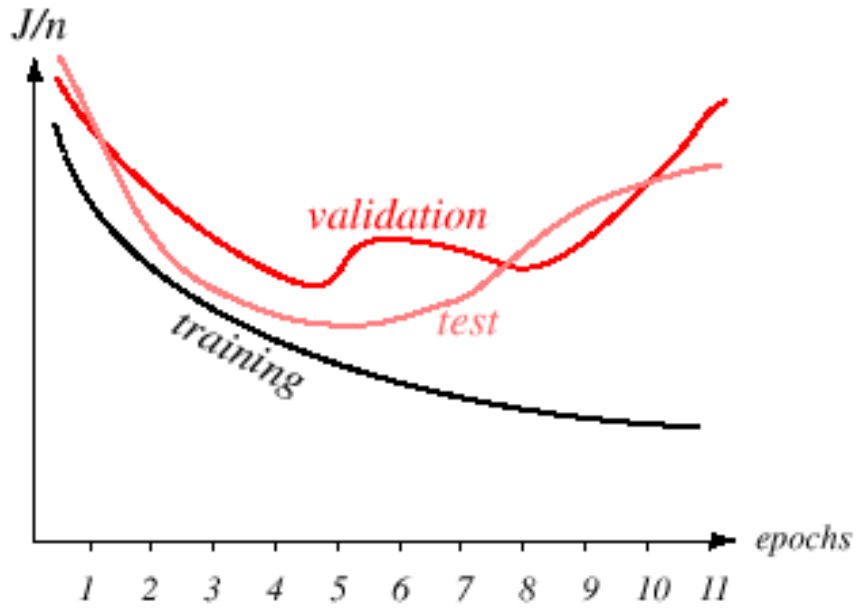


FIGURE 6.6. A learning curve shows the criterion function as a function of the amount of training, typically indicated by the number of epochs or presentations of the full training set. We plot the average error per pattern, that is, $1/n \sum_{p=1}^{n} J_p$. The validation error and the test or generalization error per pattern are virtually always higher than the training error. In some protocols, training is stopped at the first minimum of the validation set. From: Richard O. Duda, Peter E. Hart, and David G. Stork, *Pattern Classification*. Copyright © 2001 by John Wiley & Sons, Inc.

Generalização e Sobreajuste

- A linha inferior mostra o decréscimo do erro sobre os exemplos de treinamento em função do número de iterações de treinamento.
  - Esta linha mede o “Erro de Aprendizagem”
- A linha superior mostra o erro medido sobre exemplos de validação (não utilizados para atualizar os pesos !!!)
  - Esta linha mede a “Precisão da Generalização”
  - A precisão que a rede classifica corretamente exemplos diferentes dos utilizados no treinamento.

Resumo

- Redes Neurais: um método prático para aprendizagem de funções de valor real e vetorial sobre atributos de valor contínuo e discreto.
- Robustez a ruídos nos dados de treinamento.
- O espaço considerado pelo algoritmo *backpropagation* é o espaço de todas as funções que podem ser representadas pelos pesos.
- O *backpropagation* busca o espaço de todos os pesos possíveis usando a descida do gradiente para reduzir iterativamente o erro em uma rede (ajustar aos dados de treinamento).

Resumo

- Sobre ajuste resulta em redes que não generalizam bem. Utilizar um conjunto de validação para avaliar a generalização

- *Backpropagation* é o algoritmo de aprendizagem mais comum, porém existem muitos outros . . .