ICMC-USP

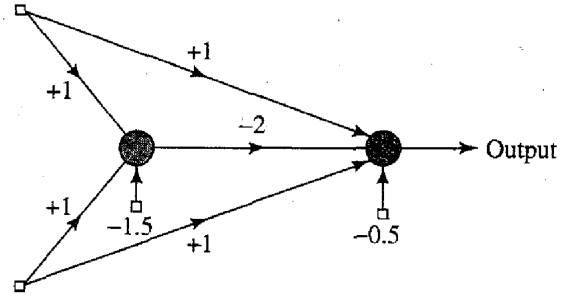
Resolução da Prova 1

SCC-5809 - Redes Neurais

2o. Semestre de 2011 - Prof. João Luís

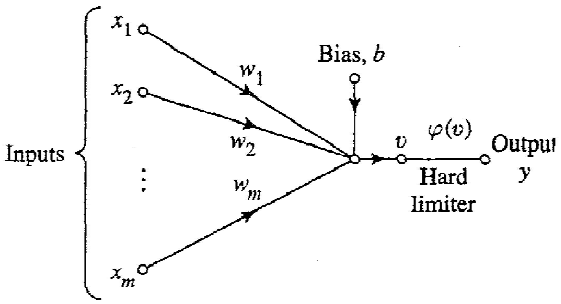
RESOLUÇÃO

(2) 1. A figura abaixo mostra uma rede neural com um único neurônio escondido. Mostre que essa rede resolve o problema do XOR construindo (a) regiões de decisão e (b) uma tabela-verdade para a rede.



Solução:

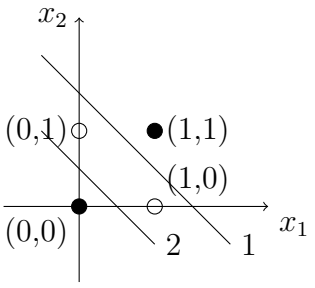
Supondo que o neurônio escondido (1) e o de saída (2) são perceptrons:



a) regiões de decisão:

Equação 1: $x_1 w_1 + x_2 w_2 = -\theta \Rightarrow x_1 + x_2 = 1.5$

Equação 2: $x_1 + x_2 - 2y_1 = 0.5 \Rightarrow 1.5 - 2y_1 = 0.5 \Rightarrow y_1 = 0.5$



 Continua na próxima página...

b) Tabela-verdade:

x_1	x_2	v_1	y_1	v_2	y_2	saída
0	0	-1.5	0	-0.5	0	0
0	1	-0.5	0	0.5	1	1
1	0	-0.5	0	0.5	1	1
1	1	0.5	1	-0.5	0	0

(1½) 2. Dada uma função qualquer $f : \mathbb{R}^n \to \mathbb{R}$ é possível encontrar sempre uma rede neural do tipo multi-camadas, constituída de apenas uma camada escondida, que aproxima esta função? Justifique sua resposta.

Solução: Não. Existe um teorema provado por Cybenko, 1989 [2] que mostrou pela primeira vez que uma rede multi-camadas com uma única camada escondida é suficiente para aproximar uniformemente qualquer função **contínua** definida num hipercubo unitário. No entanto, ele não afirmou quantos neurônios a camada escondida deverá ter para que este mapeamento seja possível de ser realizado. Esse teorema é conhecido como Teorema de Aproximação Universal. O teorema é um teorema de existência, pois ele fornece uma justificativa para a aproximação de funções contínuas (*suficiente*). Entretanto, o teorema não afirma que uma única camada é um número *ótimo*. Na prática, nem sempre dispõe-se de uma função contínua e nem de uma camada escondida de tamanho qualquer. Chester, 1990 [1] e Funahashi, 1989 [3], defendem o uso de duas camadas escondidas, pois, torna a aproximação mais maleável. Eles argumentam que as Características Locais são extraídas na primeira camada. Alguns neurônios na primeira camada são usados para particionar o espaço em várias regiões, e outros aprendem as características locais daquelas regiões. Por outro lado, as Características Globais são extraídas na segunda camada. Um neurônio na 2a. Camada combina as saídas dos neurônios da primeira que estão operando numa região particular do espaço de entrada e assim aprende características globais daquela região.

 Continua na próxima página...

(1½) 3. Qual é a diferença entre os paradigmas de aprendizado supervisionado e não supervisionado? Dê exemplos de redes neurais que utilizam aprendizado supervisionado e não supervisionado.

Solução: No aprendizado supervisionado existe a figura do professor no Processo de Aprendizado, ensinando ao sistema (controlador) qual é a resposta correta que ele deverá fornecer para o sinal de entrada que foi apresentado ao sistema. Assim, para cada sinal de entrada que o Controlador recebe, o professor fornece ao sistema a resposta correta, e o sistema vai corrigindo os seus parâmetros livres até que forneça como resposta, um valor muito próximo à saída desejada. A este processo de aprendizado dá-se o nome de aprendizado *supervisionado*. Neste caso tem-se um conjunto de exemplos de treinamento para os quais o rótulo da classe associada é conhecido. Cada exemplo pode ser visto como um vetor de características ou atributos e o rótulo da classe associada. O objetivo é construir um classificador que forneça a classe a que pertence um novo exemplo ainda não rotulado. Por outro lado, no aprendizado não supervisionado não existe a Figura do Professor, o sistema de aprendizado (indutor) analisa os exemplos de treinamento e tenta determinar se alguns deles podem ser agrupados de algum modo formando agrupamentos ou *clusters*. Após isto é necessária uma análise para determinar o que cada agrupamento significa.

Exemplos de redes que utilizam aprendizado *supervisionado*: perceptron, MLP com BP, Rede de Hamming.
Exemplos de redes que utilizam aprendizado *não-supervisionado*: Rede de Carpenter/Grossberg, Rede de Kohonen (SOM), Rede de Hopfield.

 Continua na próxima página...

(1) 4. Comente, de forma objetiva, a plasticidade do sistema nervoso central sob a óptica da Hipótese de Hebb.

Solução:

"Quando um axônio da célula A é próximo o suficiente de excitar uma célula B e repetidamente ou persistentemente toma parte em dispará-la, algum processo de crescimento ou mudança metabólica acontece em uma ou ambas as células tal que a eficiência de A, como uma das células que disparam B, é aumentada" [5].

A Hipótese de Hebb permite modelar a plasticidade do SNC, adaptando-se a mudanças do meio ambiente, através da força excitatória e inibitória das sinapses existentes e da sua topologia. Plasticidade do SNC: (a) sinapses são criadas e destruídas, (b) pesos sinápticos mudam de valor, (c) permite a aprendizagem: auto-organização interna que permite codificação de conhecimento novo e reforço de conhecimento existente.

A Hipótese de Hebb: (a) reforça-se a conexão entre dois nós que são altamente ativados ao mesmo tempo, (b) este tipo de regra é uma formalização da psicologia associacionista, que assegura que associações são acumuladas entre coisas que ocorrem juntas. A Lei de Hebb permite que uma rede conexionista aprenda correlações entre fatos.

 Continua na próxima página...

(1) 5. Explique como a escolha do parâmetro η (taxa de aprendizado) é importante para assegurar a estabilidade da convergência do processo de aprendizado iterativo.

Solução:

A escolha adequada do parâmetro η é muito importante para assegurar a estabilidade da convergência do processo de aprendizado iterativo, pois ηs pequenos permitem um aprendizado mais lento porém mais consistente, mas com o perigo de "cair" em mínimos locais, enquanto que ηs maiores permitem uma prendizado mais rápido a um custo muitas vezes de desestabilização.

Um método para aumentar η e evitar a instabilidade é usando a constante *momentum* α, que controla o *loop* de retro-alimentação agindo sobre $\Delta w_{ji}(n)$. Outra possibilidade é definir um η para cada sinapse: η_{ji}.

(1) 6. Explique a relação que existe entre o *bias* e o *threshold* de um neurônio. Qual é o significado de cada um desses parâmetros?

Solução:

O campo local induzido de um neurônio k com m entradas é dado por

$$v_k = \sum_{j=1}^{m} w_{kj} x_j + b_k \tag{1}$$

onde b_k é o *bias* do neurônio k. Mas como $b_k = w_{k0}$ e $x_0 = 1$, pode-se ter

$$v_k = \sum_{j=0}^{m} w_{kj} x_j \tag{2}$$

O *bias* representa a capacidade de um neurônio em disparar mesmo sem entradas.

Para que um neurônio dispare, é necessário que sua saída seja maior que um certo valor, o limiar (*threshold*) θ

$$\sum_{j=1}^{m} w_{kj} x_j > \theta \tag{3}$$

Ou seja, em termos matemáticos, $b_k = -\theta$.

 Continua na próxima página...

(1) 7. Comente, de forma objetiva, as seguintes critérios de parada do algoritmo *back-propagation*:

(a) O algoritmo *back-propagation* converge quando a norma Euclidiana do vetor gradiente alcança um limiar de gradiente suficientemente pequeno.

Solução:

A ideia é considerar as propriedades dos *mínimos locais ou globais* da superfície do erro. Seja o vetor de peso $\mathbf{w}^*$ denotando um mínimo, seja local ou global. Uma condição necessária para $\mathbf{w}^*$ ser mínimo é que o vetor gradiente $\mathbf{g(w)}$ (derivada parcial de primeira ordem) da superfície do erro com relação ao vetor de pesos $\mathbf{w}$ seja zero em $\mathbf{w} = \mathbf{w}^*$. A desvantagem desse critério é que, para tentativas bem sucedidas, os tempos de aprendizado podem ser longos. Além disso, requer a computação do vetor gradiente $\mathbf{g(w)}$.

(b) O algoritmo *back-propagation* converge quando a taxa absoluta de mudança no erro quadrático médio por época é suficiente pequeno.

Solução:

A função custo ou medida de erro $\xi_{av}(\mathbf{w})$ é estacionária no ponto $\mathbf{w} = \mathbf{w}^*$ (critério acima). Essa taxa é considerada pequena na faixa de 0.1 a 1% por época (as vezes, 0.01%). Este critério pode resultar numa terminação prematura do processo de aprendizado.

(1) 8. No que consiste o compartilhamento de pesos e para que serve.

Solução:

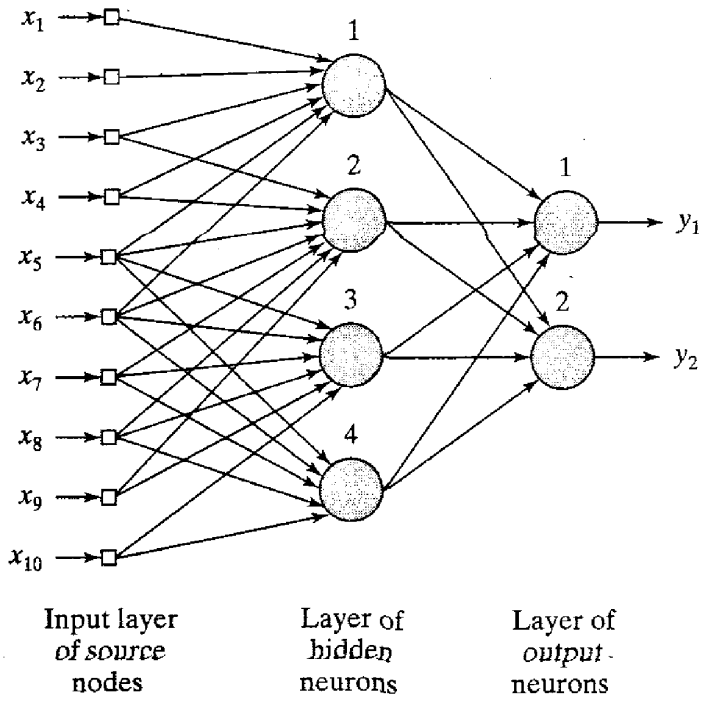
O *compartilhamento de pesos* corresponde ao uso do mesmo conjunto de pesos sinápticos para cada um dos neurônios da camada escondida. O objetivo é limitar a escolha dos pesos sinápticos. Essa técnica tem um efeito colateral benéfico: o número de parâmetros livres na rede é reduzido significativamente. Exemplo: para seis conexões locais por neurônio escondido e um total de quatro neurônios escondidos (figura 1), pode-se expressar o campo local induzido do neurônio escondido j como (soma de convolução):

$$v_j = \sum_{i=1}^{6} w_i x_{i+j-1}, \qquad j = 1, 2, 3, 4 \tag{4}$$

onde $\{w_i\}_{i=1}^{6}$ constitui o mesmo conjunto de pesos compartilhados por todos os quatro neurônios escondidos e x_k é o sinal do nó fonte $k = i + j - 1$. É por essa razão que uma rede *feedforward* usando conexões locais e compartilhamento de pesos na forma descrita é conhecida como rede convolucional.

 Continua na próxima página...

Figure 1: Ilustração do uso combinado do campo receptivo e compartilhamento de pesos [4].



References

[1] D. L. Chester, "Why two hidden layers are better than one," *International Joint Conference on Neural Networks*, vol. I, pp. 265–268, Washington, D.C., 1990.

[2] G. Cybenko, "Approximation by superpositions of a sigmoidal function," *Mathematics of Control, Signals, and Systems*, vol. 2, pp. 303–314, 1989.

[3] K. Funahashi, "On the approximate realization of continuous mappings by neural networks," *Neural Networks*, vol. 2, pp. 183–192, 1989.

[4] S. Haykin, *Neural networks - a comprehensive foundation*, 2nd. ed. Prentice Hall, 1999.

[5] D. O. Hebb, *The Organization of Behavior: A Neuropsychological Theory*, Wiley, 1949.