



Aprendizagem em Redes Neurais

Aluizio Fausto Ribeiro Araújo
Universidade Federal de Pernambuco
Centro de Informática



Conteúdo

1. Regras Básicas de Aprendizagem: Por Correção de Erros; Baseada em Memória; Hebbiana; Competitiva e Boltzmann.
2. Paradigmas de Aprendizagem: Supervisionado, Reforço e Não-supervisionado.
3. Usos de Aprendizagem: Tarefas Típicas.
4. Aspectos Probabilísticos e Estatísticos da Aprendizagem: Dilema Bias-variância, Teoria de Aprendizagem Estatística, Aprendizagem Provavelmente Aproximadamente Correta (PAC).

Introdução

- Aprendizagem é definida operacionalmente por Carbonell como a habilidade de executar novas tarefas que não podiam ser realizadas antes, ou executar melhor antigas tarefas. As novas habilidades são decorrentes das mudanças produzidas pelo processo de aprendizagem.

Introdução

- Aprendizagem envolvendo o ambiente no qual uma rede neural está imersa é fundamental.
 - Uma rede neural aprende sobre seu ambiente através de processo de ajuste de seus parâmetros (em geral, pesos sinápticos e bias).
- O entendimento do processo de aprendizagem em redes neurais compreende três premissas:
 - A rede neural é estimulada pelo meio ambiente;
 - Tal estimulação provoca modificações nos parâmetros livres das redes neurais;
 - A rede neural responde ao meio ambiente de um modo diferente após aprendizagem.

Introdução

- Definição de S. Haykin (no contexto de redes neurais):
“Aprendizagem é um processo no qual os parâmetros livres de uma rede neural são adaptados através de um processo de estimulação do meio-ambiente no qual a rede está inserida. O tipo de aprendizagem é determinado pela maneira que ocorrem as mudanças nos parâmetros.”
- Definição de M. Hassoun (no contexto de redes neurais):
“Aprendizagem é vista como processo de otimização. Isto é, aprendizagem consiste em um processo de busca de solução em um espaço multi-dimensional de parâmetros, em geral os pesos das conexões, que otimize uma dada função objetivo.”

Introdução

- Alguns marcos históricos na aprendizagem de redes neurais:
 - Experimento de condicionamento de Pavlov: Resposta condicional de cães que salivam em resposta a estímulos audíveis.
 - Conhecimento de dinâmica sináptica: Livro de Hebb (The Organization of Behavior, 1949), potenciação de longo prazo - LPT (1973 Bliss,Lomo), receptor AMPA, depressão de longo prazo - LTD, receptor NMDA.
 - Regra do vizinho mais próximo (Fix & Hodges, 1951).
 - Ideia da aprendizagem competitiva (von der Malsburg, 1973) baseada na auto-organização das células nervosas sensíveis à orientação do córtex visual.

Introdução

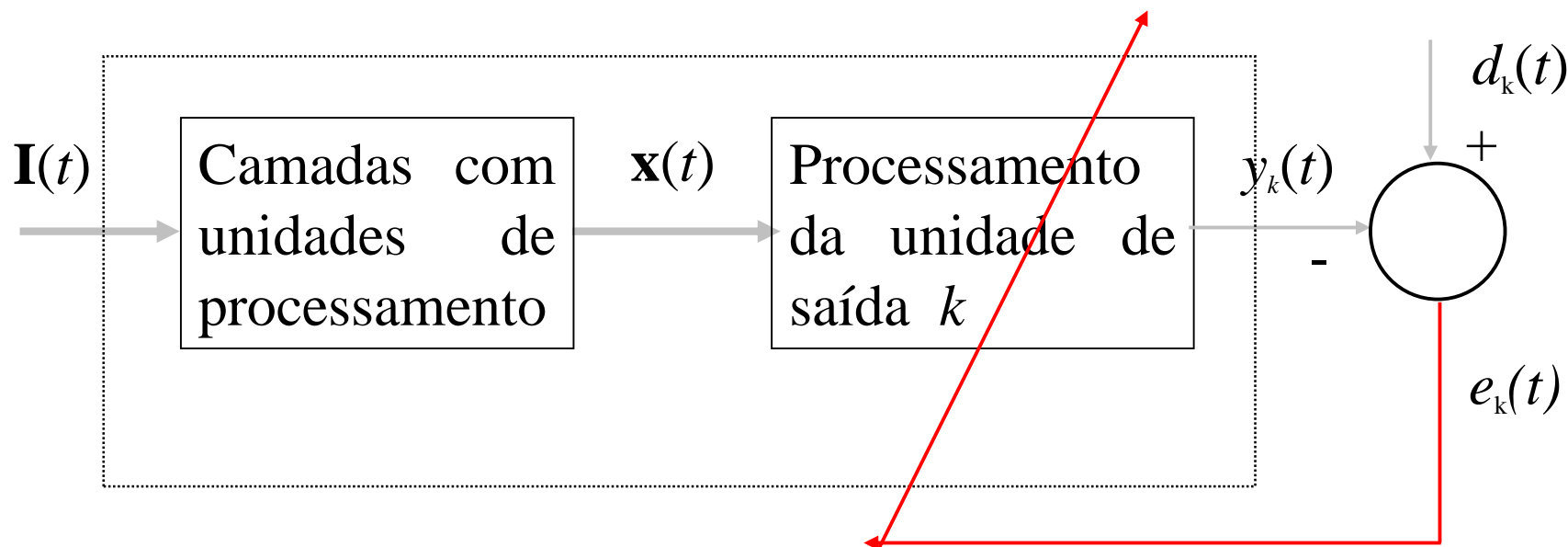
- Alguns marcos históricos na aprendizagem de redes neurais:
 - Inibição lateral tratada no trabalho de Bandas de Mach por Ernest Bach em 1865.
 - Termodinâmica estatísticas empregada no estudo de máquinas de computação por John von Neumann em seu artigo Theory and Organization of Complicated Automata, 1949.
 - Aprendizagem por reforço: Minsky 1961, Thorndike 1911.
 - Projeto de filtros lineares ótimos: Kolmogorov 1942, Wiener 1949, Zadeh 1953, Gabor 1954.

Conteúdo

1. Regras Básicas de Aprendizagem: Por Correção de Erros; Baseada em Memória; Hebbiana; Competitiva e Boltzmann.
2. Paradigmas de Aprendizagem: Supervisionado, Reforço e Não-supervisionado.
3. Usos de Aprendizagem: Tarefas Típicas.
4. Aspectos Probabilísticos e Estatísticos da Aprendizagem: Dilema Bias-variância, Teoria de Aprendizagem Estatística, Aprendizagem Provavelmente Aproximadamente Correta (PAC).

Aprendizagem Baseada em Erros

- Uma unidade de processamento k alimentada por um vetor de entrada $\mathbf{I}(\cdot)$ que gera o vetor de sinal $\mathbf{x}(\cdot)$ no instante de tempo t .



- Erro ativa um mecanismo de controle o qual realiza sequência de correções nos parâmetros de k .

Aprendizagem Baseada em Erros

- Regra Delta ou Regra de Widrow e Hoff (1960).
 - Erro: $e_k(t) = d_k(t) - y_k(t)$
- Minimizar função de custo baseada em $e_k(t)$ para aproximar o sinal obtido do sinal desejado.
- Função de custo:
$$c(t) = (1/2) \sum (e_k(t))^2$$
 - Esta função é interpretada como o valor instantâneo da energia do erro.
 - Minimização de $c(t)$ utiliza método de gradiente descendente.
 - Aprendizado chega a solução estável quando o erro atinge valores pré-estabelecidos.

Aprendizagem Baseada em Erros

- Após seleção da função de custo, aprendizado se torna um problema de otimização:
 - Uma RNA é ajustada pela minimização de $c(t)$ com respeito aos pesos da rede.
 - Estes pesos são modificados proporcionalmente ao produto do sinal de erro e do sinal de entrada da conexão considerada.
- Modelo matemático:

$$\Delta w_{ki}(t) = \eta x_i(t) e_k(t)$$

- O parâmetro η é chamado de taxa de aprendizagem e assume valores positivos. Este parâmetro pode ter valor fixo ou ajustável, através de uma função.

Aprendizagem Baseada em Erros

- Hipótese: O erro é diretamente mensurável, isto é, há disponibilidade de conjunto de saídas desejadas da rede.
- Visualização da variação do erro: Superfície multi-dimensional que traça o gráfico do valor do custo em função dos pesos.
- Busca pelo menor erro: Inicia-se em um ponto qualquer da superfície, escolhido pelo programador, e move-se em direção a um mínimo, idealmente o ótimo global.
- Ajuste do valor do peso sináptico:

$$w_{ki}(t+1) = w_{ki}(t) + \Delta w_{ki}(t)$$

Aprendizagem Baseada em Memória

- Este tipo de aprendizagem armazena experiências passadas em uma estrutura de memória contendo N exemplos corretos de pares entrada-saída (\mathbf{x}_k, d_k) , $k=1, \dots, N$.
 - Sem perda de generalidade, a saída é considerada uma grandeza escalar.
 - Dado um vetor de testes, \mathbf{x}_{teste} , a rede responde considerando os dados dentro de uma vizinhança do vetor de testes.
- Modelos possuem dois componentes principais:
 - Critério para definir vizinhança de \mathbf{x}_{teste} .
 - Regra de aprendizagem aplicada aos exemplos de treinamento na vizinhança do vetor de testes.

Aprendizagem Baseada em Memória

- Regra do vizinho mais próximo: A vizinhança local é definida como o padrão armazenado mais próximo do vetor de teste. O vetor \mathbf{x}_v que pertence ao conjunto de N vetores armazenados na rede neural ($\mathbf{x}_v \in (\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_N)$), é dito ser o vizinho mais próximo de \mathbf{x}_{teste} se:

$$\underset{k}{Min} (\mathbf{x}_k, \mathbf{x}_{teste}) = d(\mathbf{x}_v, \mathbf{x}_{teste}), \quad k = 1, \dots, N$$

onde $d(.)$ é a distância euclidiana.

- Cover e Hart (1967) usaram esta abordagem para classificação de padrões, baseada em duas hipóteses:
 - Os exemplos classificados (\mathbf{x}_i, d_i) são estatisticamente independentes e identicamente distribuídos.
 - O tamanho da vizinhança é infinitamente grande.

Aprendizagem Baseada em Memória

- Classificador dos k -vizinhos mais próximos: Variação da regra do vizinho mais próximo que possui as seguintes características:
 - Identifica k padrões mais próximos de \mathbf{x}_{teste} .
 - Assinala \mathbf{x}_{teste} para a classe que é mais instanciada entre os k vizinhos (e.g., uso de voto de maioria).
- Este classificador atua com base em médias de características ou prevalência delas.

Aprendizagem Hebbiana

- Afirmção de Hebb, no livro ‘The Organization of Behavior’ (1949, p.62) formulado como base para aprendizagem associativa no nível celular:



Donald Olding Hebb
1904-1985

- “Quando um axônio da célula A está próximo o suficiente para excitar uma célula B e repetida e persistentemente toma parte em sua emissão de sinais, algum processo de crescimento ou mudanças metabólicas ocorre em uma ou ambas células tal que aumenta a eficiência de A como uma das células que excitam B.”

Aprendizagem Hebbiana

- Stent (1973) e Changeux e Danchin (1976) estenderam o postulado de Hebb para o contexto neurobiológico:
 - Se dois neurônios em lados diferentes de uma sinapse são simultaneamente (sincronamente) ativados, então a força da sinapse que os une é seletivamente acrescida.
 - Se dois neurônios em lados diferentes de uma sinapse são assincronamente ativados, então a força da sinapse que os une é seletivamente diminuída.
- Esta sinapse é Hebbiana, mesmo que a Regra de Hebb não inclua o segundo item, pois tem mecanismo para ajustar a “força” sináptica como função da correlação entre atividades pré e pós-sináptica.
 - O mecanismo é dependente do tempo, local e muito interativo.

Aprendizagem Hebbiana

- Uma sinapse Hebbiana possui mecanismo para variar a eficiência sináptica com as seguintes propriedades:
 - Dependência do tempo: Mudanças dependem do tempo de ocorrência dos sinais pré- e pós-sinápticos.
 - Localidade: Informações localmente disponíveis (espaço-temporalmente contíguas) produzem modificações sinápticas.
 - Interatividade: Aprendizagem Hebbiana depende de interação real entre sinais pré e pós-sináptico.
 - Correlação entre sinais pré- e pós-sináptico: A ocorrência simultânea (ou dentro de um intervalo de tempo curto) entre tais sinais é suficiente para produzir modificações sinápticas.

Aprendizagem Hebbiana

- Depressão sináptica: Processo que reduz força sináptica na ausência de atividade simultânea nos dois lados da sinapse:
 - Atividade nas membranas pré- e pós-sinápticas não correlacionadas ou negativamente correlacionadas produzem enfraquecimento sináptico.
- Sinapse Hebbiana é fortalecida com atividades pré- e pós-sinápticas positivamente correlacionadas e enfraquecida para correlações negativas e descorrelacionados.
- Sinapse anti-Hebbiana é enfraquecida com atividades pré- e pós-sinápticas positivamente correlacionadas e fortalecida para correlações negativas e descorrelacionados.
- Sinapse não Hebbiana não usa mecanismos Hebbianos.

Aprendizagem Hebbiana

Pseudo-Código

Considere dois neurônios A e B.

Para todo passo de tempo t , faça:

- Compute a atividade de A
- Compute a atividade A recebida por B
- Compute a atividade de B
- Se a atividade de B > 0 E a atividade de A > 0
 - então aumente a força da conexão de A para B

(Caudill & Butler, “Understanding Neural Networks: computer explorations”, The MIT Press, 1992, vol.1, pp.61)

Aprendizagem Hebbiana

- Modelos matemáticos: Peso sináptico, $w_{ki}(t)$, entre unidades de processamento x_i e y_k é ajustado, em t , através da expressão:

$$\Delta w_{ki} = f(x_i(t), y_k(t))$$

onde $f(.,.)$ é função dos sinais pré- e pós-sinápticos e pode assumir diferentes formas.

- Hipótese de Hebb: Regra do produto de atividades:

$$\Delta w_{ki}(t) = \eta x_i(t) y_k(t)$$

onde η é a taxa de aprendizagem. A repetição da aplicação do sinal de entrada sempre causa o crescimento do sinal de saída e finalmente da conexão.

Aprendizagem Hebbiana

- Hipótese da Covariância (Sejnowski, 1997) considera a diferença dos sinais pré- e pós-sinápticos de suas médias (x_{iM} e y_{kM}) em um dado intervalo de tempo.

$$\Delta w_{ki}(t) = \eta (x_i(t) - x_{iM})(y_k(t) - y_{iM})$$

- Esta hipótese considera:
 - Convergência a estados médios das células pré- e pós-sinápticas:
 - A diferença determina o sinal da modificação sináptica.
 - Ocorrência de potenciação ($x_i(t) > x_{iM}$ e $y_k(t) > y_{iM}$) e depressão sináptica ($x_i(t) > x_{iM}$ e $y_k(t) < y_{iM}$ ou $x_i(t) < x_{iM}$ e $y_k(t) > y_{iM}$).
- Há evidências de aprendizagem Hebbiana na área do Hipocampo.

Aprendizagem Competitiva

- Cada unidade de processamento de uma rede neural compete com as demais para responder a um dado estímulo. Neste tipo de aprendizagem, uma unidade ou um grupo delas conquista o direito de responder a uma dada entrada. Rumelhart e Zipser (1985) conceberam modelo vencedor-leva-tudo com três elementos básicos:
 - Conjunto de unidades de processamento diferenciadas apenas pelo conjunto de pesos associado a cada uma delas.
 - Limite imposto na força de cada unidade.
 - Mecanismo de competição entre unidades.
- Unidades se especializam para responder a conjuntos de padrões similares, são detectores de características.

Aprendizagem Competitiva

- Mecanismo de competição baseado em métrica de distância.
 - Produto Interno: $\langle x, y \rangle = x_1 y_1 + x_2 y_2 + \dots + x_n y_n$
 - Distância Euclideana: $d_E(x, y) = \sqrt{(x_1 - y_1)^2 + \dots + (x_n - y_n)^2}$
 - Distância de Hamming:
 - $d_H(x, y) = d_H(0011, 1111) = 2$
 - $d_H(x, y) = d_H(0001, 1000) = 2$
 - $d_H(x, y) = d_H(0001, 0000) = 1$
 - $d_H(x, y) = d_H(0000, 1000) = 1$
 - $d_H(x, y) = d_H(0000, 1111) = 4$

Aprendizagem Competitiva

- Aprendizagem competitiva simples:
 - Uma camada de unidades completamente ligados à entrada através de conexões excitatórias.
 - Conexões laterais (de realimentação) inibitórias entre unidades na camada de processamento.
 - Normalização dos pesos excitatórios de i : $\sum (w_{ki})^2 = 1$
 - Ativação da vencedora:
$$y_k = \begin{cases} 1, & \text{se } y_k > v_i \\ 0, & \text{de outro modo} \end{cases}$$

onde v_i é o campo local induzido para todo i diferente de k . O vencedor inibe as outras unidades.

- Ajuste de pesos da unidade vencedora: $\Delta w_{ki}(t) = \eta (x_i(t) - w_{ki}(t))$

Aprendizagem Boltzmann

- Aprendizagem estocástica, inspirada na mecânica estatística.
- Deu origem ao modelo de rede neural máquina de Boltzmann, com as seguintes características:
 - Estrutura recorrente com 2 estados de ativação: 1, -1.
 - Função energia: $E = -\frac{1}{2} \sum_j \sum_k w_{kj} x_k x_j$, para $j \neq k$
 - Mudança de estado: $P(x_k \rightarrow -x_k) = \frac{1}{1 + \exp(-\Delta E_k / T)}$
- Operação: Escolhe-se aleatoriamente unidade k , muda-se seu estado de ativação ($x_i(t) \rightarrow -x_i(t)$), a uma pseudo-temperatura (T), até a máquina achar o equilíbrio térmico.

Aprendizagem Boltzmann

- A máquina de Boltzmann possui unidades visíveis (entrada e saída) e escondidas e opera de dois modos:
 - Condição de ativação mantida: Os estados de ativação das unidades visíveis ficam constantes nos valores determinados pelo ambiente.
 - Condição de ativação livre: Todas as unidades têm estados de ativação livres.
- Seja ρ_{kj}^m e ρ_{kj}^l as correlações entre as unidades j e k na condição de ativação mantida e livre, respectivamente. As correlações consideram todos os estados possíveis para o equilíbrio térmico. A regra de aprendizagem é:

$$\Delta w_{kj} = \eta (\rho_{kj}^m - \rho_{kj}^l), \quad j \neq k, \quad -1 \leq \rho_{kj}^m \leq 1, \quad -1 \leq \rho_{kj}^l \leq 1$$

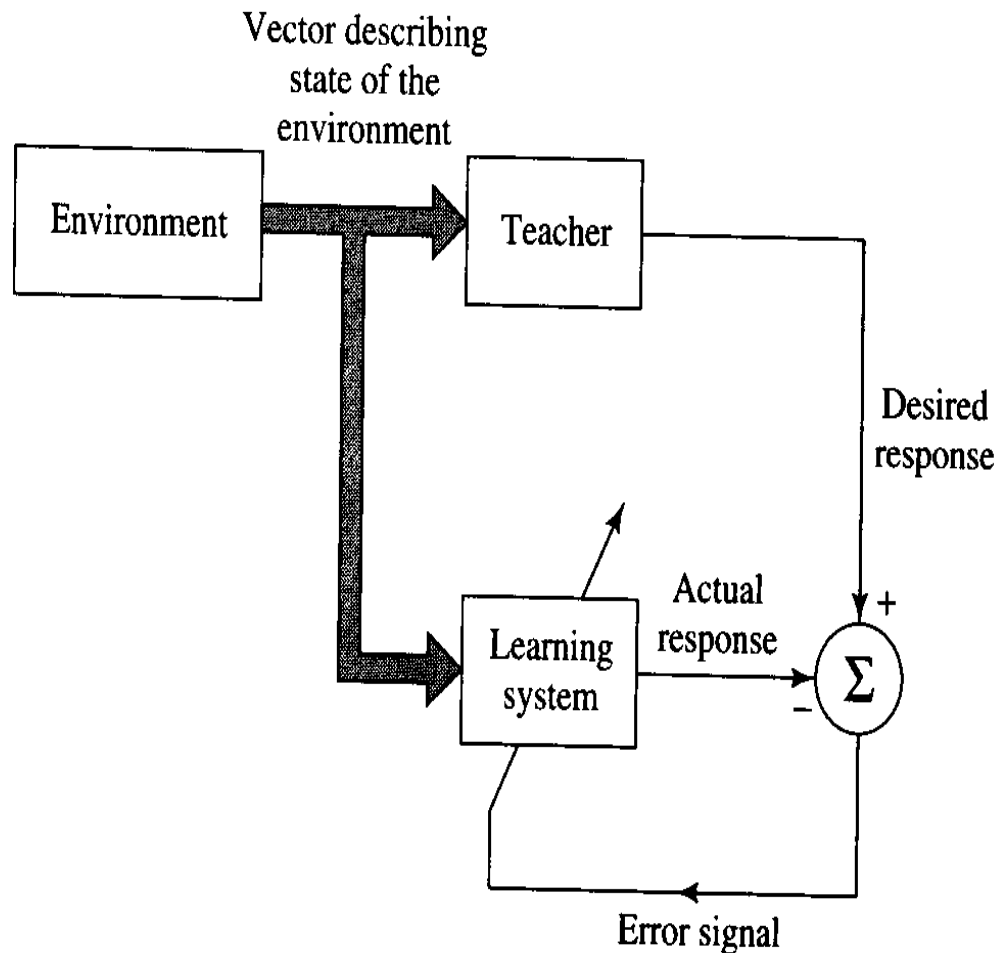
Conteúdo

1. Regras Básicas de Aprendizagem: Por Correção de Erros; Baseada em Memória; Hebbiana; Competitiva e Boltzmann.
2. Paradigmas de Aprendizagem: Supervisionado, Reforço e Não-supervisionado.
3. Usos de Aprendizagem: Tarefas Típicas.
4. Aspectos Probabilísticos e Estatísticos da Aprendizagem: Dilema Bias-variância, Teoria de Aprendizagem Estatística, Aprendizagem Provavelmente Aproximadamente Correta (PAC).

Aprendizagem com Professor

- Aprendizagem com professor ou aprendizagem supervisionada é caracterizada por:
 - Possuir conhecimento do ambiente ou do fenômeno (pares entrada-saída) em questão.
 - Ajustar os parâmetros da rede, passo a passo, com base em um sinal de erro.
 - Possuir padrões desejados de saída que são usados como professor pela rede neural.
 - Encerrar o ajuste paramétrico da rede quando o erro atingir um ponto ótimo (mínimo).

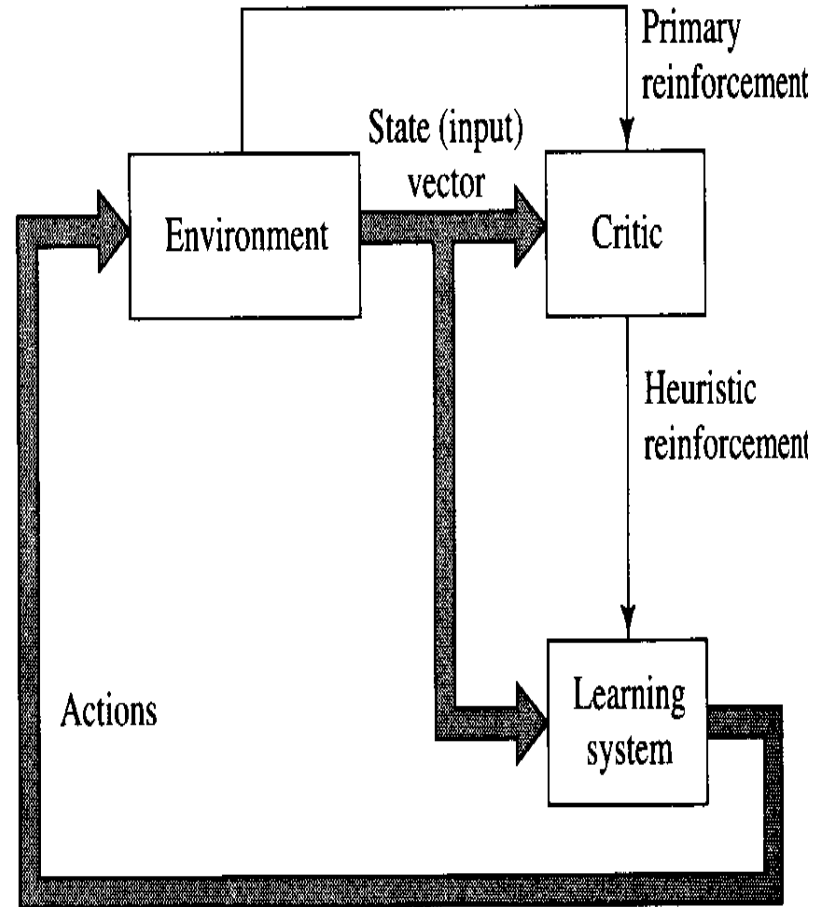
Aprendizagem com Professor



Aprendizagem por Reforço

- Aprendizagem por reforço é caracterizada por:
 - Possuir um índice de desempenho escalar para a rede neural que avalia a qualidade de suas saídas com respeito a uma dada tarefa.
 - Ajustar os parâmetros da rede, passo a passo, com base no tal índice, que não é um erro, mas sim a repercussão da saída na rede sobre o meio-ambiente, um sinal de reforço.
 - Estimar o sinal de reforço de cada estado do ambiente, com base em várias visitas a cada estado. A estimativa avalia as chances de sucesso para o objetivo desejado.
 - Encerrar o ajuste paramétrico quando as mudanças na estimativa do reforço ficarem menor que valor dado.

Aprendizagem por Reforço



Aprendizagem sem Professor

- Aprendizagem sem professor ou aprendizagem não-supervisionada é caracterizada por:
 - Não possuir professor ou sinal de reforço do ambiente.
 - Ajustar os parâmetros da rede, passo a passo, com base no “sinal de erro não-supervisionado”.
 - Encerrar o ajuste paramétrico da rede quando o erro não-supervisionado não variar significativamente.

Aprendizagem sem Professor

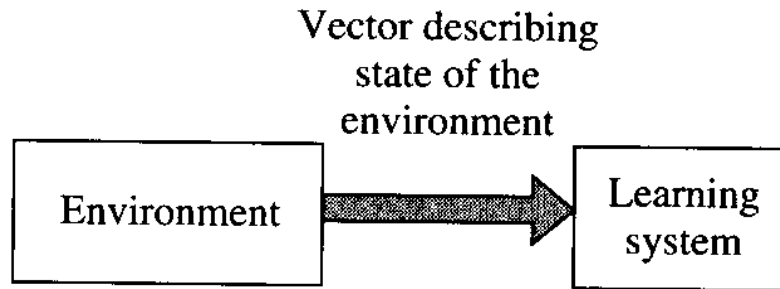


FIGURE 2.8 Block diagram of unsupervised learning.

Conteúdo

1. Regras Básicas de Aprendizagem: Por Correção de Erros; Baseada em Memória; Hebbiana; Competitiva e Boltzmann.
2. Paradigmas de Aprendizagem: Supervisionado, Reforço e Não-supervisionado.
3. Usos de Aprendizagem: Tarefas Típicas.
4. Aspectos Probabilísticos e Estatísticos da Aprendizagem: Dilema Bias-variância, Teoria de Aprendizagem Estatística, Aprendizagem Provavelmente Aproximadamente Correta (PAC).

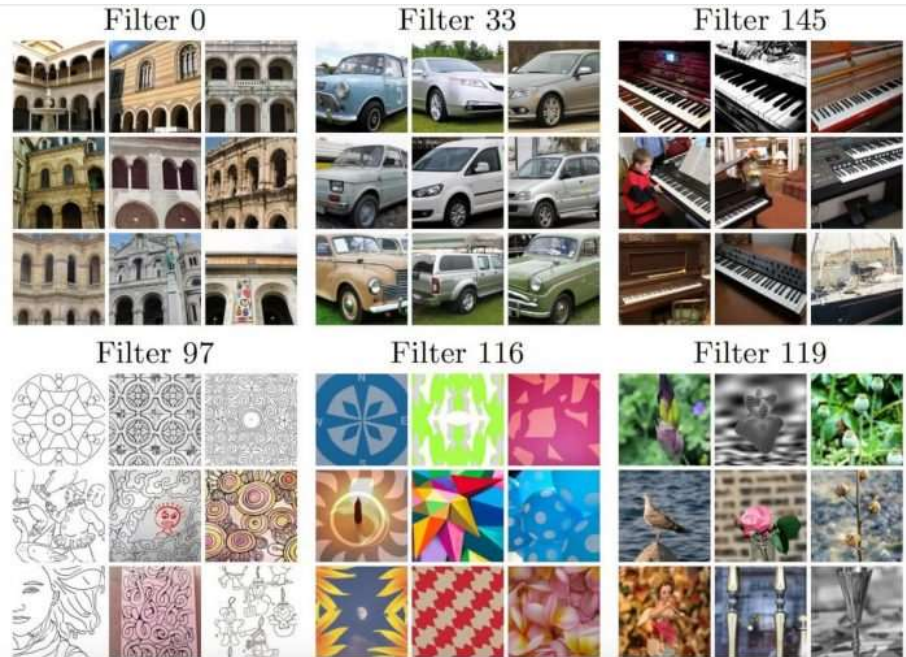
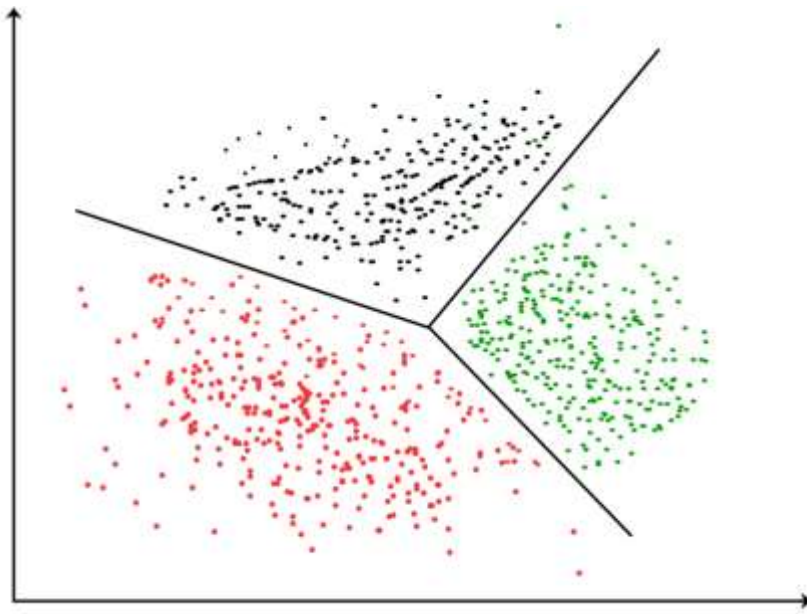
Tarefas de Aprendizagem

- Associação de padrões - Memória distribuída, inspirada no cérebro incorpora conhecimento por associação:
 - Auto-associação: Padrão armazenado após apresentações à rede. Recuperação de padrão parcialmente informado à rede.
 - Hetero-associação: Neste caso um conjunto de entrada é associado a um conjunto de saída distinto. A aridade pode ser qualquer.



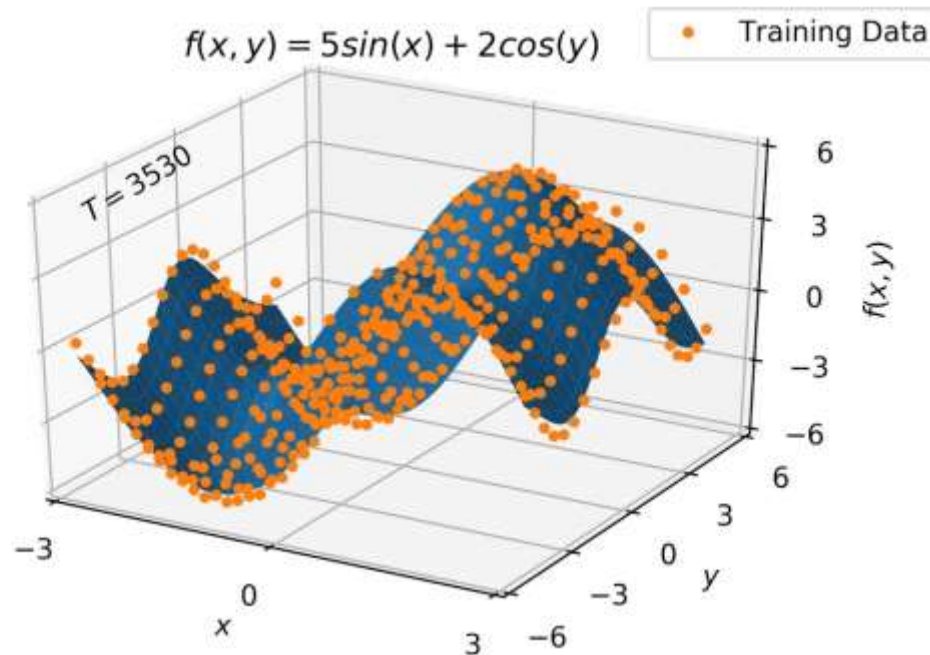
Tarefas de Aprendizagem

- Reconhecimento de padrões - Um padrão ou sinal de entrada é incluído em uma entre possíveis categorias:
 - Classificação: Se a aprendizagem for supervisionada.
 - Agrupamento: Se a aprendizagem for não-supervisionada.



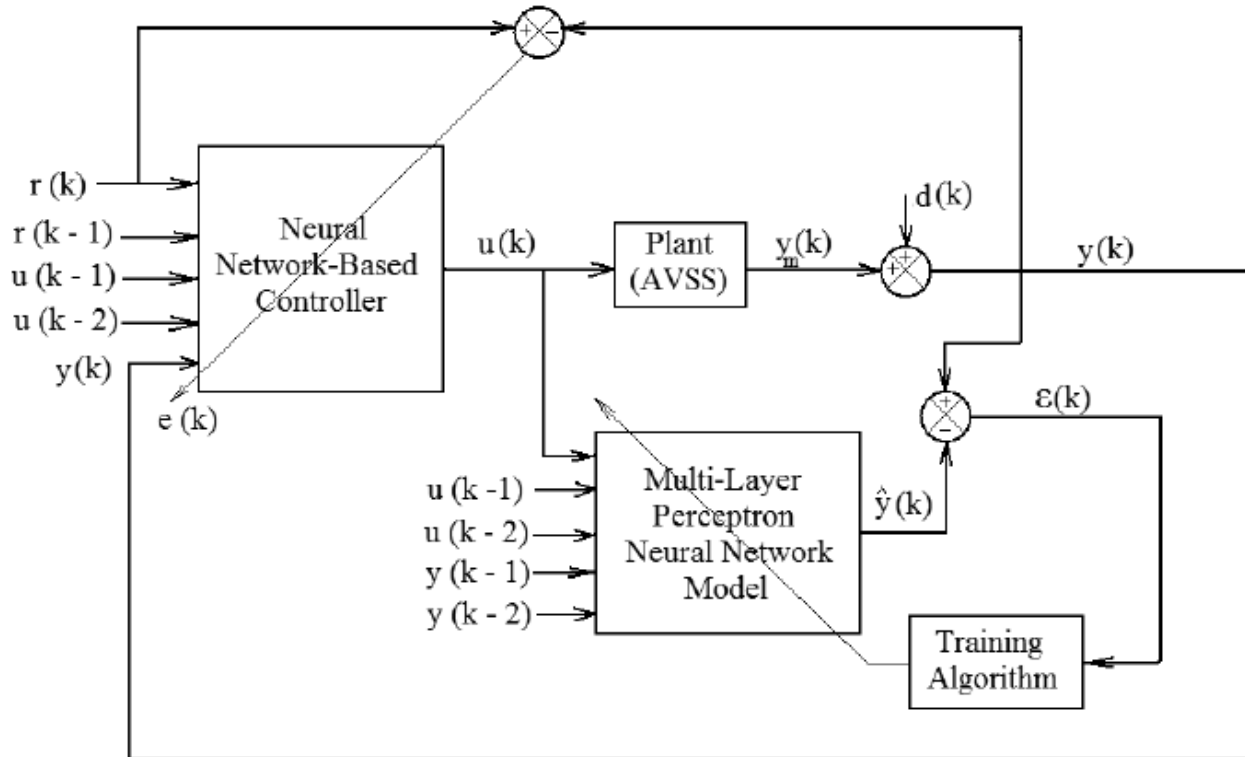
Tarefas de Aprendizagem

- Determinação de funções – Produz rede neural que define uma função que replique um dado mapeamento entrada-saída:
 - Interpolação: Mapeamento que interpola pontos aos existentes.
 - Aproximação: Mapeamento que aproxima a relação (figura).



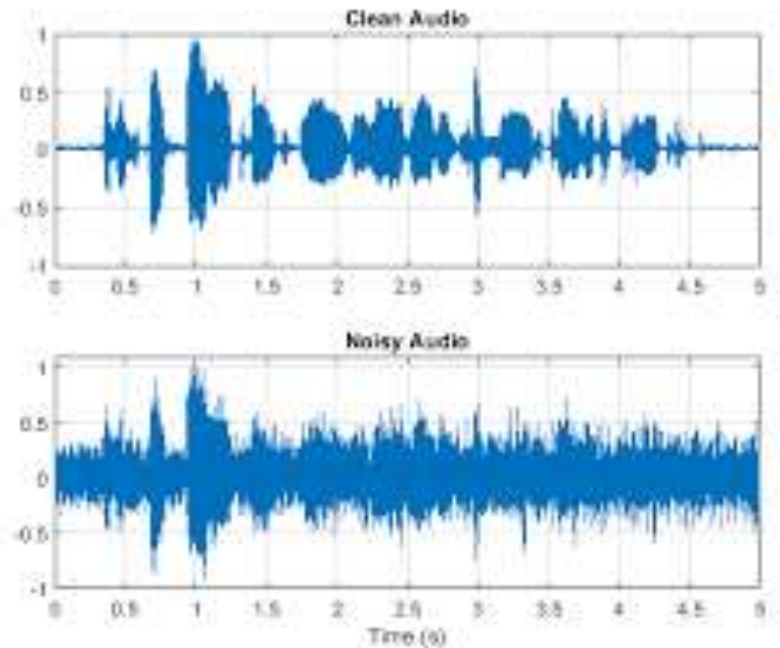
Tarefas de Aprendizagem

- Controle de processos - Rede neural para realizar identificação ou controle de um dado processo ou sistema:
 - Identificação de sistemas: Modela operação de um sistema.
 - Sistema inverso: Modela o controlador do sistema.



Tarefas de Aprendizagem

- Filtragem de sinais – Separação de componentes de um sinal.
 - Filtragem: Elimina partes indesejáveis de um sinal.
 - Suavização: Retira variações bruscas de um sinal.
 - Predição: Antevê o comportamento futuro de um dado sinal.
 - Exemplo: Retirando ruído com redes neurais profundas.



Conteúdo

1. Regras Básicas de Aprendizagem: Por Correção de Erros; Baseada em Memória; Hebbiana; Competitiva e Boltzmann.
2. Paradigmas de Aprendizagem: Supervisionado, Reforço e Não-supervisionado.
3. Usos de Aprendizagem: Tarefas Típicas.
4. Aspectos Probabilísticos e Estatísticos da Aprendizagem: Dilema Bias-variância, Teoria de Aprendizagem Estatística, Aprendizagem Provavelmente Aproximadamente Correta (PAC).

Aspectos Probabilísticos e Estatísticos da Aprendizagem

- Não há conhecimento da exata relação entre entrada e saída (\mathbf{x}_i , d_i), assim estima-se $d_i = f(\mathbf{x}_i) + \varepsilon$, por um modelo regressivo.
- O valor médio do erro ε , para qualquer realização de \mathbf{x}_i , é zero.
- O erro ε é não correlacionado com a função $f(\cdot)$.

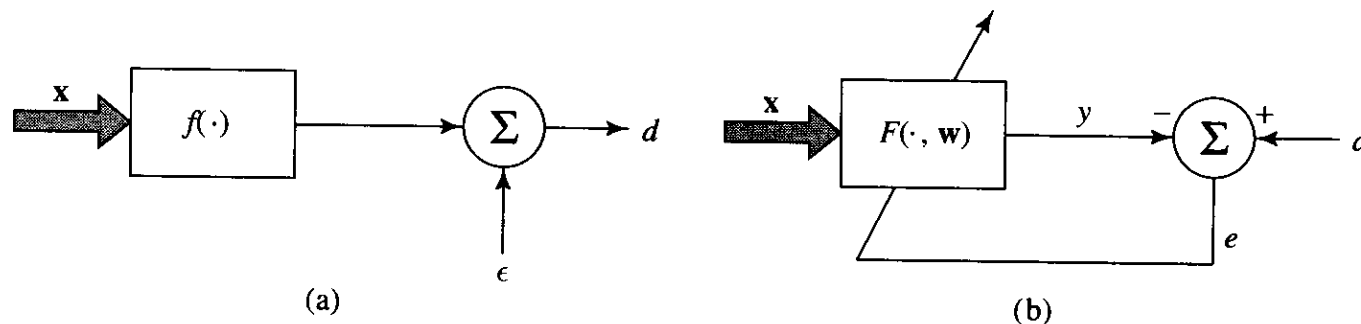


FIGURE 2.20 (a) Regressive model (mathematical).
(b) Neural network model (physical).

Aspectos Probabilísticos e Estatísticos da Aprendizagem

- Dilema Bias/variação: $L_{av}(f(\mathbf{x}), F(\mathbf{x}, T)) = B^2(\mathbf{w}) + V(\mathbf{w})$

- Onde $L_{av}(...)$ é uma função de custo; $f(\mathbf{x}) = E[D|\mathbf{X}=\mathbf{x}]$ é uma função de regressão, expressando a média condicional do modelo de saída; $F(\mathbf{x}, T)$ é a resposta da RN para uma entrada \mathbf{x} , $E_T[.]$ é o operador para calcular média, T é o conjunto de treinamento todo, \mathbf{X} é o conjunto de vetores de entrada e D de saídas.
- $B(\mathbf{w}) = E_T[F(\mathbf{x}, T)] - E[D|\mathbf{X}=\mathbf{x}]$ (erro de aproximação). Este é o bias do valor médio da função de aproximação, medido com respeito à $f(\mathbf{x})$.
- $V(\mathbf{w}) = E_T[(F(\mathbf{x}, T) - E_T[F(\mathbf{x}, T)])^2]$ (erro de estimação). Esta é a variância da função de aproximação $F(\mathbf{x}, \mathbf{w})$ medida sobre a amostra toda de treinamento (T).
- NN normalmente possui pequeno bias e alta variância (por possuir amostra de treinamento de tamanho fixo).
 - A introdução do bias pode reduzir a variância.

Aspectos Probabilísticos e Estatísticos da Aprendizagem

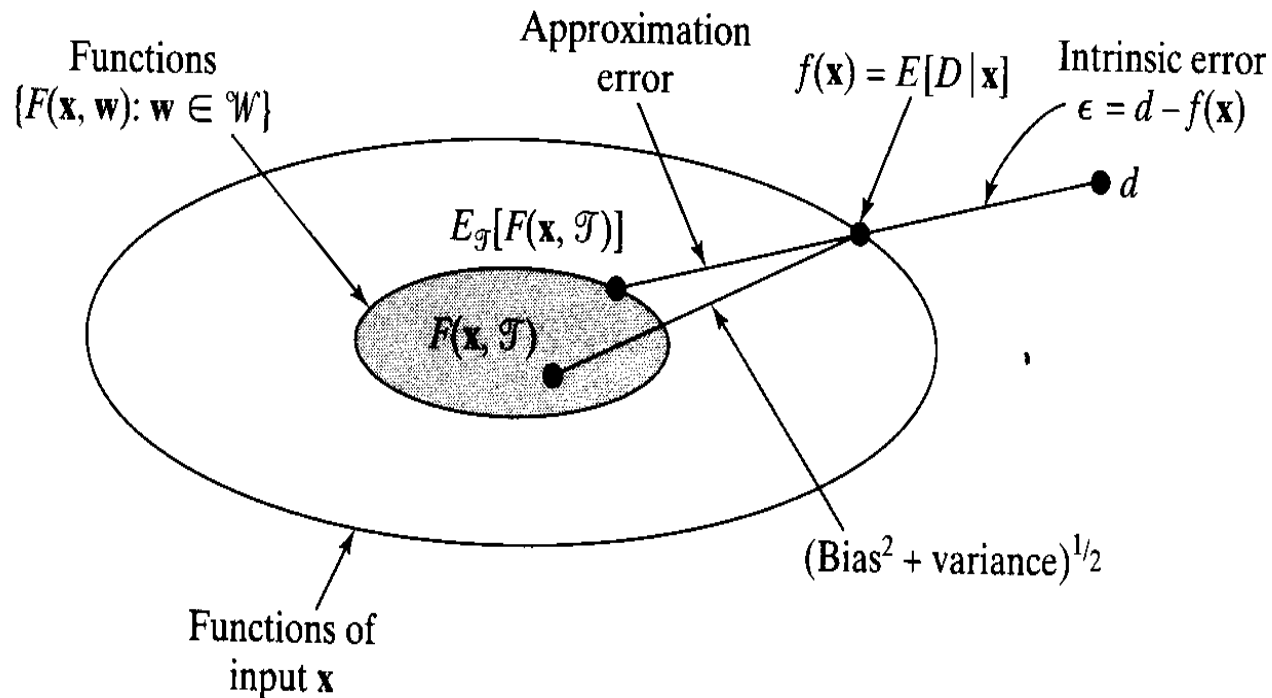


FIGURE 2.21 Illustration of the various sources of error in solving the regression problem.

Aspectos Probabilísticos e Estatísticos da Aprendizagem

- A dimensão de Vapnic-Chervonenkis (VC) é uma medida da capacidade ou poder expressivo de uma família de funções de classificação encontradas por modelos de aprendizagem de máquinas.
- A dimensão VC de um conjunto de funções de classificação é o número máximo de exemplos de treinamento que pode ser aprendido sem erro pela máquina para todos os possíveis rótulos binários das mencionadas funções.

Aspectos Probabilísticos e Estatísticos da Aprendizagem

- Seja RN uma rede neural arbitrária de alimentação direta (fluxo de processamento entrada-saída) cujas unidades de processamento (neurônios) têm função de ativação do tipo limiar. A dimensão VC de RN é $O(W \log W)$, onde W é o número total de parâmetros livres na rede.
- Seja RN outra rede neural arbitrária de alimentação direta cujas unidades de processamento têm função de ativação sigmoideal ($f(v) = 1/(1 + \exp(-v))$). A dimensão VC de RN é $O(W^2)$, onde W é seu número total de parâmetros livres.

Aspectos Probabilísticos e Estatísticos da Aprendizagem

- Minimização do risco estrutural:

$$V_{\text{guarant}}(\mathbf{w}) = V_{\text{train}}(\mathbf{w}) + \varepsilon_1(N, h, \alpha, v_{\text{train}})$$

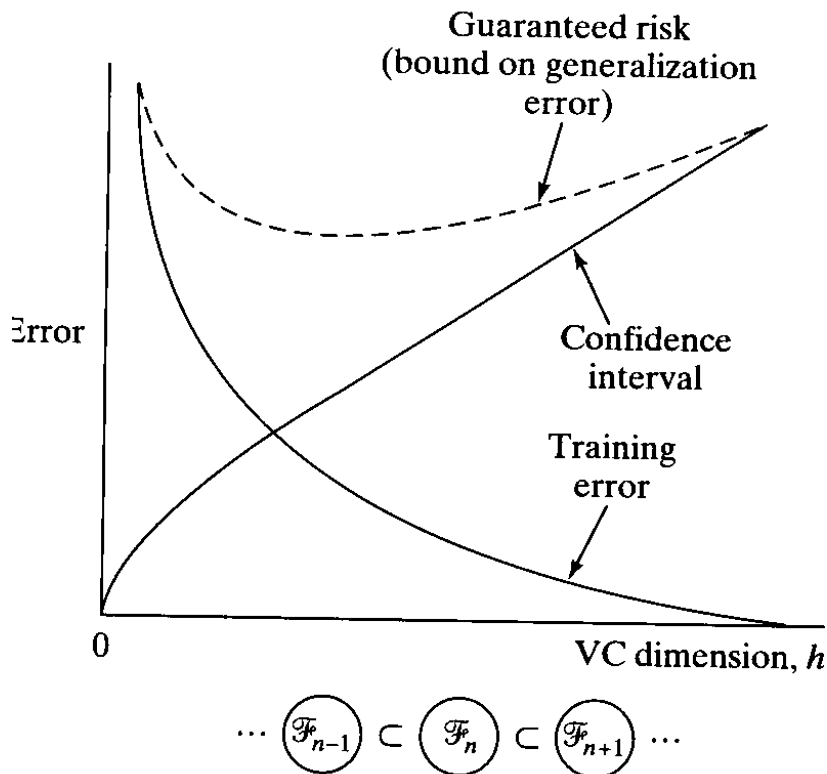


FIGURE 2.25 Illustration of the relationship between training error, confidence interval, and guaranteed risk

Referências

- Haykin, S. (2008). *Neural Networks and Learning Machines*. Third Edition. Pearson.