Paradigmas de Aprendizado

Sarajane Marques Peres

março de 2018

Material baseado em:
Haykin, S. Neural Networks: A comprehensive foundation. 2nd Edition. Prentice Hall. 1999
Russel, S.; Norvig, P. Inteligência Artificial. 2a Edição (tradução), Campus/Elsevier, 2004
Lima, C. A. M. Comitê de Máquinas: uma abordagem unificada empregando máquinas de vetores-suporte. Tese de doutorado. Universidade Estadual de Campinas, 2004

Paradigmas - definições de dicionários

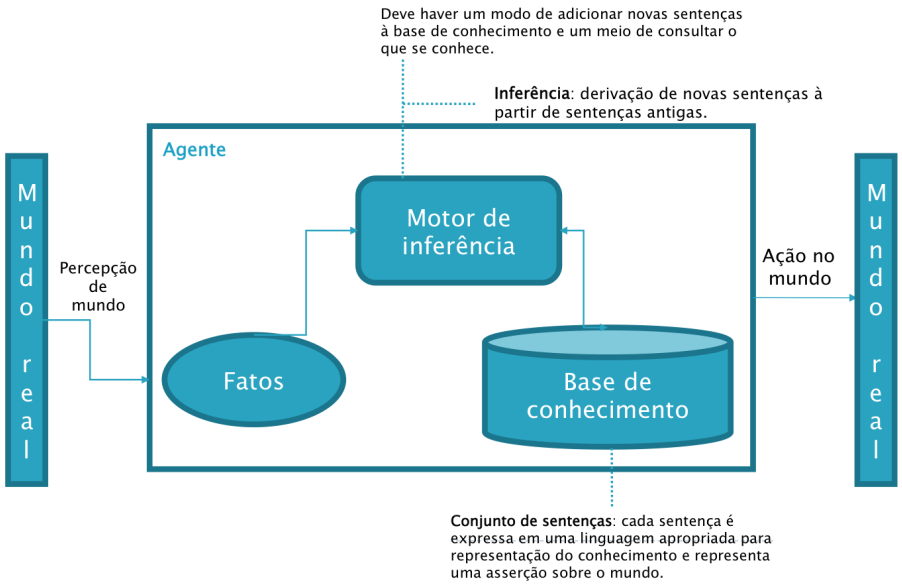
- um exemplo, um padrão ou um modelo a ser seguido;
- etimologia: tem origem na palavra grega *paradeigma*, que significa exatamente um modelo ou padrão que corresponde a algo que será seguido em uma determinada situação;
- também pode ser visto como um conjunto de normas orientadoras que estabelecem limites e que determinam que ações devem ser executadas dentro destes limites;
- pode ser visto como um princípio, um teoria, ou como conhecimento originado da pesquisa em um campo científico;
- pode ser visto como uma referência inicial que servirá de modelo para novas pesquisas.

* um pouco de leveza: `https://www.youtube.com/watch?v=g5G0qE7Lf0A`

Paradigmas de aprendizagem

Russel e Norvig ....

Sob a perspectiva de agentes inteligentes, Russel e Norvig colocam a aprendizagem em termos de percepções que devem ser usadas não apenas como base para execução de uma ação, mas também como informação para melhorar a habilidade do agente de agir no futuro.



A aprendizagem pode ser expressa por um processo simplificado de "memorização" de uma experiência ou pela elaboração de conhecimento complexo.

No aprendizado de máquina ....

"Um programa de computador aprende se ele é capaz de melhorar seu desempenho em determinada tarefa, sob alguma medida de avaliação, a partir de experiências passadas." **(Tom Mitchell)**

Na área de aprendizado de máquina o interesse está no aprendizado indutivo (baseado no raciocínio indutivo), que ocorre a partir de observações. Dentro deste contexto, destaca-se três tipos de paradigmas de aprendizado: aprendizado supervisionado, aprendizado não-supervisionado e aprendizado por reforço.

Há ainda o aprendizado semi-supervisionado, o aprendizado por transferência, o aprendizado conjunto, o aprendizado probabilístico ....

E outros paradigmas: aprendizado baseado no raciocínio dedutivo, aprendizado baseado no raciocínio abdutivo e aprendizado baseado na evolução de uma população (esse último pode ser implementado sob o raciocínio indutivo, ou não).

Racionínio indutivo e dedutivo

Argumento dedutivo

Procede de proposições (mais) universais para proposições particulares. Conclusões são extraídas a partir de premissas.

- Todo homem é mortal.
- João é homem.
- Logo, João é mortal

A partir de um modelo (Todo homem é mortal) e um fato (João é homem), conclusões são obtidas.

Argumento indutivo

Procede de proposições particulares para proposições (mais) universais.

- O ferro conduz eletricidade.
- O ouro conduz eletricidade.
- A prata conduz eletricidade.
- O chumbo conduz eletricidade.
- Logo, todo metal conduz eletricidade.

A partir de uma série de fatos, um modelo é criado. O modelo, então, pode a partir de um novo fato, gerar conclusões que concordam com os fatos primeiros.

Aprendizado supervisionado

Um algoritmo para aprendizado supervisionado recebe como entrada o valor correto de uma função desconhecida para entradas específicas (e conhecidas), e tenta recuperar a função desconhecida ou algo perto disso.

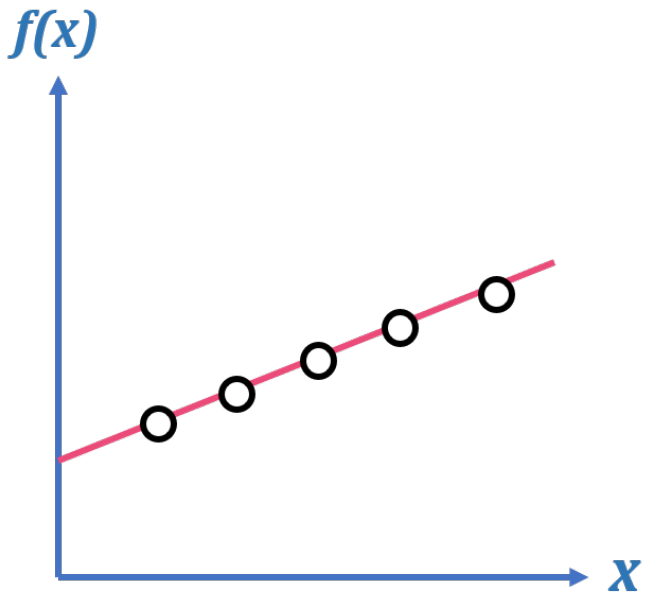
Dada uma coleção de exemplos de f, retornar uma função h que se aproxime de f, considerando que um exemplo é um par $(x, f(x))$, x é a entrada conhecida e $f(x)$ é a saída conhecida da função f aplicada a x.

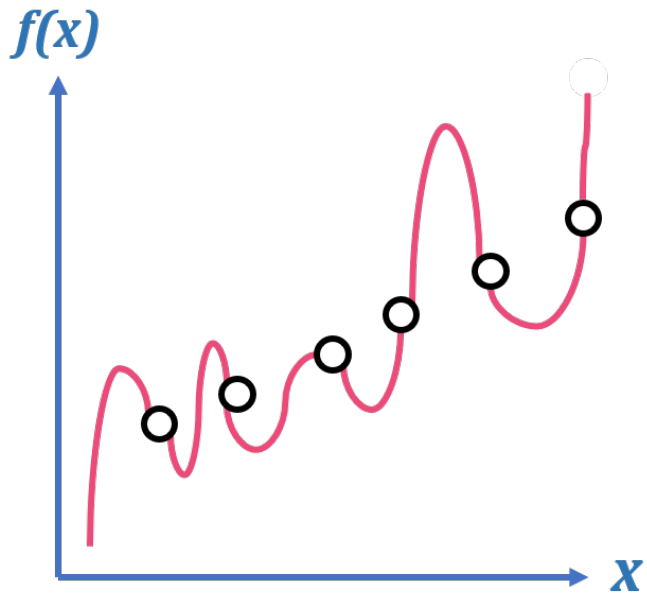
A funçao h é chamada **hipótese**.

A razão pela qual o aprendizado é difícil, de um ponto de vista conceitual, é que não é fácil saber se uma h específica é uma boa aproximação de f. Uma boa hipótese irá generalizar bem, i.e., será capaz de responder corretamente para exemplos ainda não previstos. Esse é o problema fundamental da indução.

Exemplos de Russel e Norvig

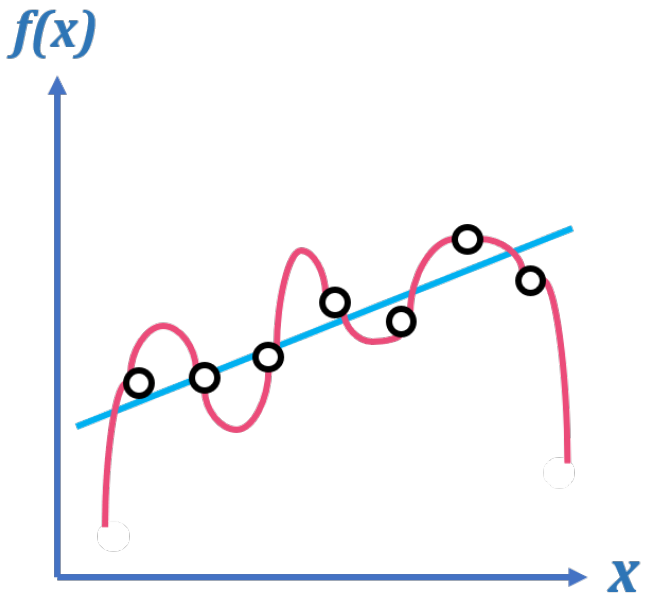
Dados com ajuste exato por um polinômio de grau um. A linha é uma hipótese consistente porque concorda com todos os dados. O polinômio de grau mais alto também representa uma hipótese consistente.





Como escolher entre várias hipóteses consistentes é uma questão importante no aprendizado indutivo. É preferível a hipótese mais simples que seja consistente. **Hipóteses mais complexas que os próprios dados estão deixando de extrair algum padrão dos dados.**

Exemplos de Russel e Norvig



Nesse caso não existe uma linha reta consistente para o conjunto de dados, mas o polinômio de grau alto também não está encontrando um padrão (um comportamento), pois ele é tão ou mais complexo que os dados. Nesse caso, talvez seja melhor aceitar a linha reta que não é exatamente consistente, mas que pode fazer previsões razoáveis.

Espaço de hipóteses

Espaço de hipóteses

A possibilidade de encontrar uma hipótese simples e consistente depende do espaço de hipóteses escolhido. Diz-se que um problema de aprendizado é **realizável** se o espaço de hipóteses contém a função verdadeira; caso contrário ele é **irrealizável**.

Nem sempre podemos dizer que um dado problema de aprendizado é realizável, pois não conhecemos a função f verdadeira. É possível usar algum conhecimento *a priori* para derivar um espaço de hipóteses no qual sabe-se que a função verdadeira reside. Outra abordagem é usar o maior espaço de hipóteses possível.

Compromisso!

Há um compromisso entre a expressividade de um espaço de hipóteses e a complexidade de encontrar hipóteses simples e consistentes dentro deste espaço. Devido a esse compromisso, geralmente o trabalho em aprendizado se concentra em representações relativamente simples.

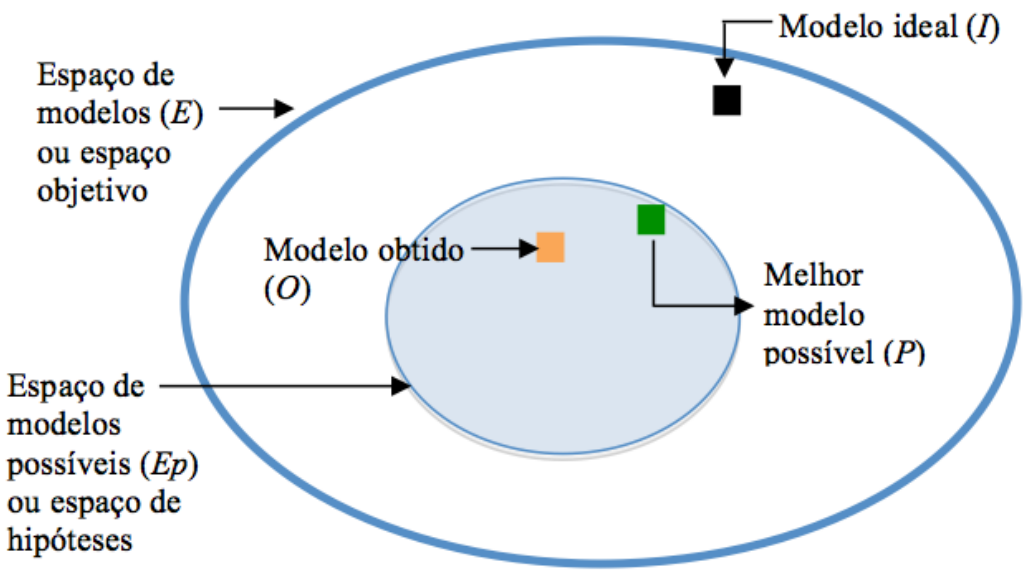
Erros *versus* Espaço de hipóteses

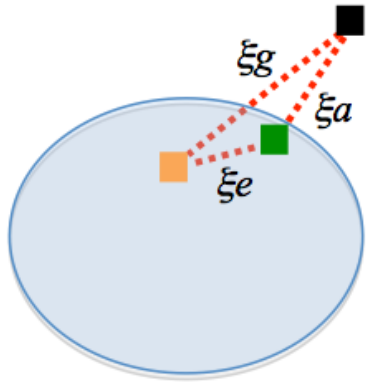
(Lima, 2004) Pensando no problema de aprendizado indutivo como uma modelagem funcional, cujo objetivo é escolher um modelo a partir de um espaço de hipóteses H que seja próximo à função fundamental $f(x_0)$ num espaço objetivo em relação a uma medida de erro, dois tipos de erro podem surgir:

- **erro de aproximação**: é uma consequência do espaço de hipóteses H_z não conter todo o espaço objetivo, de modo que a função fundamental $f_0(.)$ pode residir fora de H_z. Uma escolha ruim de H_z vai resultar em um erro de aproximação grande.
- **erro de estimação**: é um erro devido ao processo de aprendizado, que pode levar à seleção de um modelo que não seja o melhor possível dentro de H_z.

Em conjunto esses erros formam o erro de generalização.

Erros *versus* Espaço de hipóteses





ξg - Erro de generalização
ξa - Erro de aproximação
ξe - Erro de estimação

Erro empírico (Lima,2004)

No contexto discutido no slide anterior, o processo de aprendizado objetiva minimzar o **erro empírico**, dado por

$$\hat{f}_{z,n} = \underset{f \in H_z}{\operatorname{argmin}} R_{emp}(f)$$

em que N é o conjunto de treinamento (amostras/observações).

E a minimização do risco empírico deve ser consistente no sentido que:

$$\lim_{N \to \infty} R_{emp}(f) = R(f),$$

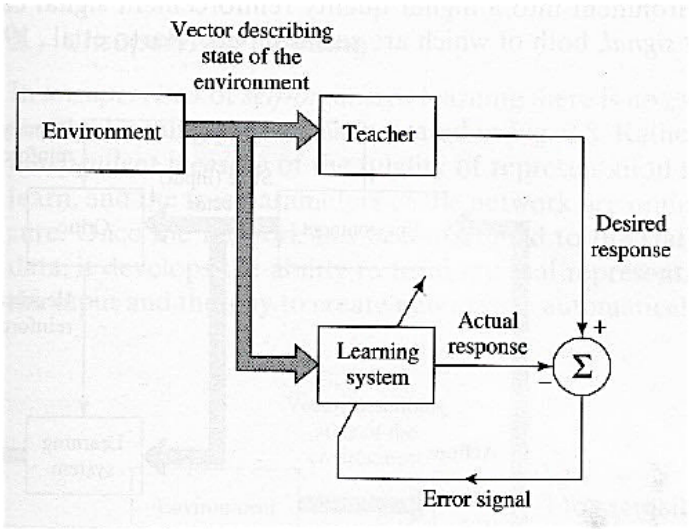
seguindo a Lei dos Grandes Números.
Contudo, devido à existência de H_z, a condição de consistência passa a ser:

$$\lim_{N \to \infty} \min_{f \in H_z} R_{emp}(f) = \min_{f \in H_z} R(f).$$

que pode não ser válida em qualquer H_z.

Paradigma de aprendizado supervisionado, por Simon Haykin

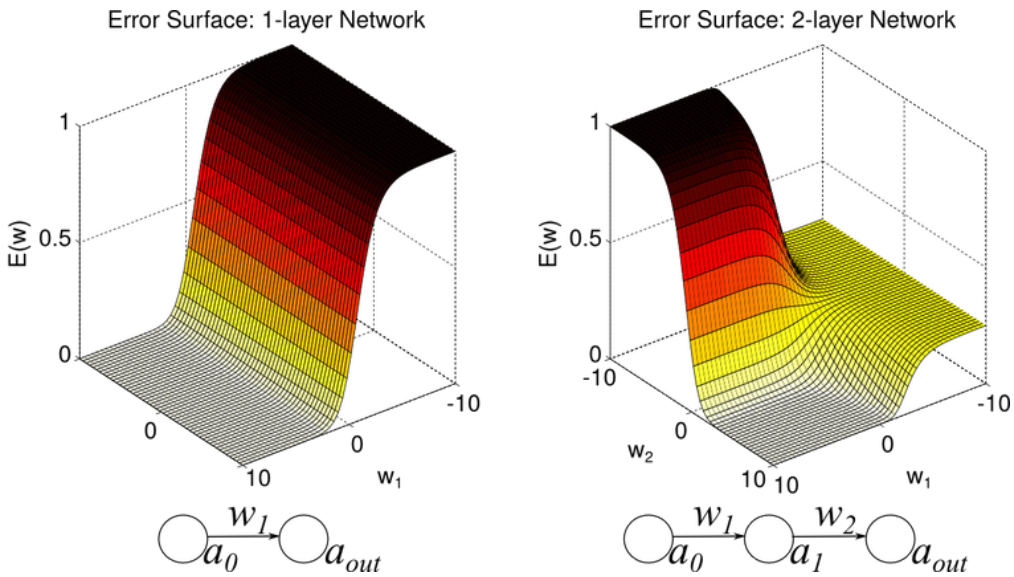
Haykin se refere ao aprendizado supervisionado como um aprendizado com professor, que está baseado na **correção do erro**. A medida de desempenho para esse aprendizado pode ser o erro quadrático médio ou o erro quadrático somado para todas os pares de treinamento.



Esses erros são definidos como uma função dos parâmetros livres do sistema.

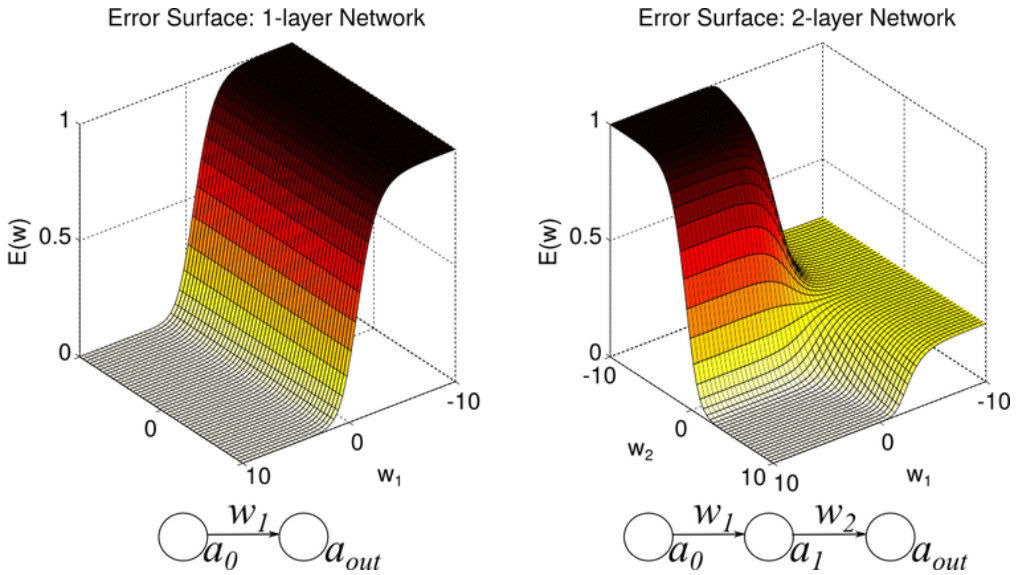
Paradigma de aprendizado supervisionado, por Simon Haykin

A função que permite definir o "erro" pode ser vista como uma superfície multidimensional de erro de desempenho, ou simplesmente como uma superfície de erro. Uma superfície de erro real seria medida sobre todos os exemplos possíveis de entrada-saída (no sistema).



https://theclevermachine.wordpress.com/2014/09/11/a-gentle-introduction-to-artificial-neural-networks/

Superfície de erro



Qualquer operação do sistema sob a supervisão *do professor* (possibilidade de cálculo de erro) é representada como um ponto nessa superfície de erro. Essas operações, durante o tempo de aprendizado, deve mover esse ponto para baixo, buscando pelo mínimo local ou global. **Isso pode ser feito, por exemplo, usando a informação sobre o gradiente da superfície de erro.**

Paradigmas de aprendizado !!

Aprendizado baseado em instâncias

Aprendizado não-paramétricos

Métodos de aprendizado não-paramétrico, segundo Russel e Norvig, permitem que a complexidade da hipótese cresça com os dados. Quanto mais dados temos, mais complicada a hipótese pode ser.

Métodos de aprendizado baseado em instâncias são métodos não-paramétricos. Exemplos:

- K-NN: K-*nearest neighbor* (K-vizinhos mais próximos)
- métodos baseados em *kernel* (núcleo)

Aprendizado em Redes Neurais Artificiais - Simon Haykin

Uma rede neural artificial (RNA) aprende sobre seu ambiente (conjunto de dados) por meio de um processo de ajuste aplicados aos seus pesos sinápticos e níveis de *bias* (viés). Idealmente, uma RNA se torna mais conhecedora de seu ambiente depois de cada iteração de seu processo de aprendizado. No contexto de RNA, o conceito de aprendizado pode ser definido como:

Aprendizado

Um processo pelo qual os parâmetros livres de uma RNA são adaptados, considerando os estímulos provenientes do ambiente no qual a RNA reside. O tipo do aprendizado é determinado pela maneira como as mudanças nos parâmetros são realizadas.

Esse processo é composto pela seguinte sequências de eventos:

- a RNA é estimulada pelo ambiente
- a RNA é submetida à mudança de seus parâmetros livres, como resultado daquele estímulo
- a RNA responde de uma maneira nova para o ambiente por conta das mudanças ocorridas na sua estrutura interna

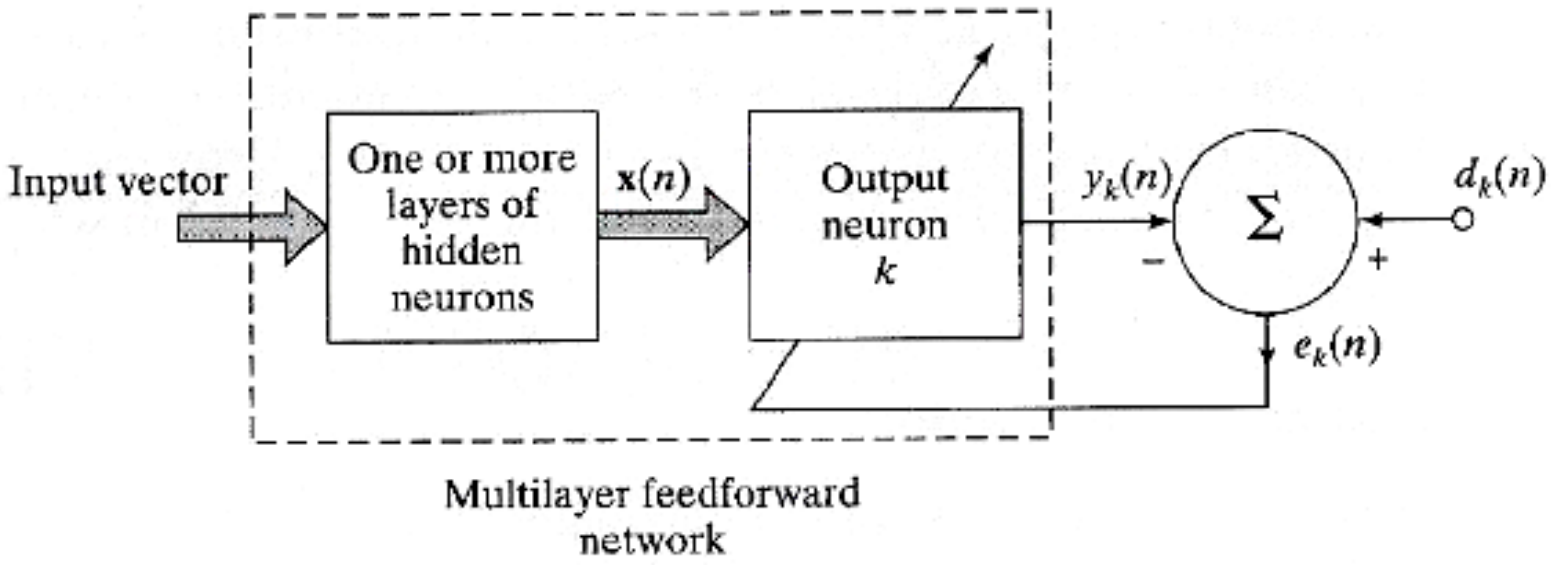
Aprendizado em RNA

Um conjunto pré determinado de regras bem definidas para solução do problema de aprendizado é chamado de algoritmo de treinamento. Basicamente, os algoritmos de treinamento se distinguem pela forma como ajustam os pesos sinápticos de um neurônio.

Importante!!!

O paradigma de aprendizado está fortemente relacionado ao modelo de ambiente no qual a rede neural opera.

Aprendizado baseado na correção do erro



O sinal de saída do neurônio k é $y_k(n)$. Este sinal de saída, representando a única saída dessa RNA, é comparado à resposta desejada ou saída alvo $d_k(n)$. O sinal de erro, $e_k(n)$ é produzido, por definição, da seguinte maneira:

$$e_k(n) = d_k(n) - y_k(n).$$

Aprendizado baseado na correção do erro

O sinal de erro $e_k(n)$ atua como um mecanismo de controle, cujo propósito é aplicar um ajuste para corrigir os pesos sinápticos do neurônio k. Essa correção é feita de maneira a ajustar os pesos para que o sinal de saída $y_k(n)$ seja mais próximo da resposta $d_k(n)$, a cada passo.

Função de custo / índice de desempenho

Este objetivo é alcançado por meio da minimização de uma função de custo, ou índice de desempenho, $\xi(n)$, definida em termos do sinal $e_k(n)$, como:

$$\xi(n) = \frac{1}{2} e_k^2(n).$$

O ajuste passo a passo dos pesos sinápticos do neurônio k continua até que o sistema alcance um estado estável, i.e. os pesos sinápticos estabilizam. Nesse ponto, o aprendizado termina.

Aprendizado baseado na correção do erro

Regra Delta ou Regra Widrow-Hoff

Minimizar a função $\xi(n)$ leva à regra conhecida como Regra Delta ou Regra Widrow-Hoff.

Nesse contexto, $w_{kj}(n)$ denota um valor de um peso sináptico w_{kj} do neurônio k, excitado pelo elemento $x_j(n)$ do sinal de entrada $\mathbf{x}(n)$, no tempo n. De acordo com a Regra Delta, o ajuste $\Delta w_{kj}(n)$, aplicado a w_{kj} no tempo n é:

$$\Delta w_{kj}(n) = \eta e_k(n) x_j(n),$$

em que η é uma constante que determina uma taxa de aprendizado. Ela tem papel importante para convergência do sistema e para o desempenho do aprendizado por correção de erro.

Aprendizado baseado na correção do erro

Ajuste

O ajuste feito no peso sináptico do neurônio é proporcional ao produto do sinal de erro e do sinal de entrada da sinapse em questão. E a Regra Delta, portanto, assume que o erro é mensurável.

Finalmente, a correção do erro é feita localmente, no neurônio k. A partir da computação de $\Delta w_{kj}(n)$, a alteração de pesos é:

$$w_{kj}(n+1) = w_{kj}(n) + \Delta w_{kj}(n).$$

Aprendizado baseado em memória

No aprendizado baseado em memória, todos (ou a maioria) das experiências passadas são explicitamente estocadas em uma grande memória de exemplos corretos de entrada-saída.

$$\{(\mathbf{x}_i, d_i)\}_{i=1}^{N},$$

em que $\mathbf{x}_i$ é o vetor de entrada e d_i é a resposta desejada.

Quando o classificador recebe um vetor de teste $\mathbf{x}_{teste}$, o algoritmo responde recuperando e analisando o dado de treinamento na vizinhança local de $\mathbf{x}_{teste}$.

Os algoritmos desta classe envolvem dois ingredientes básicos:

- o critério usado para definir a vizinhança local;
- regra de aprendizado aplicada aos exemplos de treinamento na vizinhança local de $\mathbf{x}_{teste}$.

Aprendizado baseado em memória

Regra do vizinho mais próximo

O vetor $\mathbf{x}'_N \in \{\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_N\}$ é dito ser o vizinho mais próximo de $\mathbf{x}_{teste}$ se

$$\min_i d(\mathbf{x}_i, \mathbf{x}_{teste}) = d(\mathbf{x}'_N, \mathbf{x}_{teste})$$

em que $d(\mathbf{x}_i, \mathbf{x}_{teste})$ é a distância Euclidiana (ou outra distância apropriada) entre os vetores $\mathbf{x}_i$ e $\mathbf{x}_{teste}$. A classe associada com o vetor $\mathbf{x}'_N$ é reportada como classificação para $\mathbf{x}_{teste}$.

Esta regra é independente da distribuição responsável por gerar os exemplos de treinamento. E, para o contexto de classificação de padrões:

- $(\mathbf{x}_i, d_i)$ são independentemente e identicamente distribuídos (i.i.d.), de acordo com a distribuição de $(boldx, d)$;
- o tamanho da amostra N é infinitamente grande.

O calssificador K-NN é uma variante do NN. Redes neurais que seguem esse tipo de aprendizado são as **redes de função de base radial** (RBF).

Aprendizado Hebbiano

O postulado de aprendizado de Hebb é a regra de aprendizado mais antiga, e diz que:

Postulado de Hebb

Quando o axônio de uma célula A está próximo o bastante para excitar uma célula B, e repetidamente ou persistentemente participa da ativação dele, algum processo metabólico ou processo de crescimento sináptico acontece em uma ou em ambas as células tal que a eficência de A como uma célula que excita B é aumentada.

Essa é a base do aprendizado associativo, que resulta em fortalecer/executar modificação no padrão de atividade de um conjunto de células nervosas distribuídas espacialmente. De forma procedural:

- Se dois neurônios em ambos os lados de uma sinapse (conexão) são ativado simultaneamente (de forma síncrona), então a força desta sinapse é seletivamente aumentada.
- Se dois neurônios nos dois lados de uma sinapse são ativados de forma assíncrona, então aquela sinapse é seletivamente enfraquecida ou eliminada.

* Originalmente, Hebb não menciona a segunda parte mencionada acima.

Aprendizado Hebbiano

As sinapses de Hebb são caracterizadas por:

- mecanismo dependente do tempo: as sinapses Hebbianas dependem do tempo exato de ocorrência dos sinais pré e pós sinápticos;
- mecanismo local: a sinapse é o local de transmissão do sinal onde sinais (informação) estão em uma adjacência espaço-temporal. A modificação sináptica ocorre nessa adjacência;
- mecanismo interativo: a mudança sináptica depende do sinal em ambos os lados da sinapse. O aprendizado Hebbiano depende de ocorrer uma "verdadeira interação" entre sinais pré e pós sinápticos. Isso indica que nós só podemos tirar conclusões na presença de ambos os sinais;
- mecanismo correlacional e conjuncional: a co-ocorrência do sinal pré e pós sináptico é suficiente para permitir uma mudança sináptica (conjuncional). Outra interpretação vai no sentido de dizer que a correlação sobre o tempo entre sinais pré e pós sinápticos é vista como responsável pela mudança sináptica.

Aprendizado Hebbiano

Formulando em termos matemáticos:

- w_{kj} são os pesos do neurônio k
- x_i e y_k representam respectivamente os sinais pré e pós sinápticos

O ajuste de w_{kj} no tempo n é:

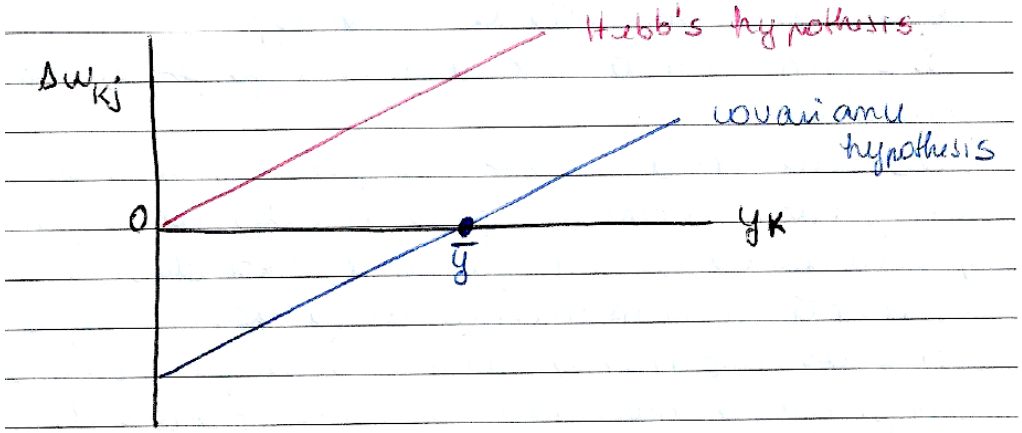
$$\Delta w_{kj}(n) = F(y_k(n), x_j(n))$$

em que $F(.)$ é uma função dos sinais pré e pós sinápticos. Essa formulação pode assumir diferentes formas:

- Hebb's hypothesis: $\Delta w_{kj}(n) = \eta y_k(n).x_j(n)$ em que η é uma taxa de aprendizado. Essa regra é também conhecida como "regra do produto de atividade".
- Covariance hypothesis: os sinais pré e pós sinápticos são substituídos pela distância deles à sua média. Seja $\overline{x}$ e $\overline{y}$ os valores "médios no tempo" (considerando um certo intervalo de tempo) dos sinais pré e pós sinápticos, o ajuste de pesos é: $\Delta w_{kj} = \eta(x_j - \overline{x})(y_k - \overline{y})$.

Aprendizado Hebbiano

Na hipótese de covariância: a convergência ocorre em $x_k = \overline{x}$ e $y_k = \overline{y}$; e há aumento e decréscimo da força sináptica.



Considerações:

- w_{kj} é fortalecido se níveis suficientes de atividades pré e pós sinápticas existem ($x_j > \overline{x}$ e $y_j > \overline{y}$);
- w_{kj} é enfraquecido se ($x_j > \overline{x}$ e $y_j < \overline{y}$) ou se ($x_j < \overline{x}$ e $y_j > \overline{y}$).

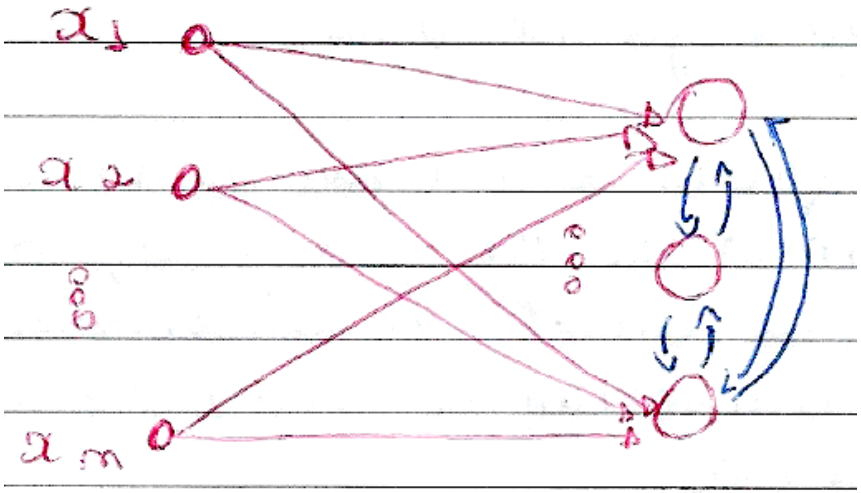
Aprendizado Competitivo

Neste aprendizado, os neurônios de uma rede neural competem entre si para se tornarem ativos, e somente um neurônio de saída estará atividado por vez. Os elementos básicos da regra de aprendizado competitivo são:

1. um conjunto de neurônios iguais, exceto por alguma distribuição aleatória de seus pesos, os quais responderão de maneira diferentes para um dado conjunto de entradas;
2. um limite imposto para a força de cada neurônio;
3. um mecanismo que permite que os neurônios compitam entre si pelo direito de responder para um dado subconjunto de entradas, tal que somente um neurônio de saída esteja ativod naquele momento. O neurônio que vence a competição é chamado winner-takes-all.

Aprendizado Competitivo

Em azul, conexões inibitórias – inibição lateral.
Em vermelho, conexões excitatórias.



Aprendizado Competitivo

Para que um neurônio k seja o neurônio vencedor, seu campo local induzido v_k para uma entrada $\mathbf{x}$ deve ser o maior para todos os neurônios na rede. O sinal de saída y_k do neurônio vencedor k é $= 1$; os demais sinais de saída são $= 0$.

$$y_k = \begin{cases} 1 \text{ se } v_k > v_j \text{ para todo } j \text{ , } j \neq k \\ 0 \text{ c.c.} \end{cases}$$

Suponha $\sum_j w_{kj} = 1$ para todo k. Um neurônio aprende por deslocar pesos sinápticos de seus nós de entrada inativos para seus nós de entrada ativos. Ou sejs, se um neurônio ganha a competição, cada nós de entrada daquele neurônio cede alguma proporção de seus pesos sinápticos, e os pesos cedidos são distribuídos igualmente aos nós de entrada ativos.

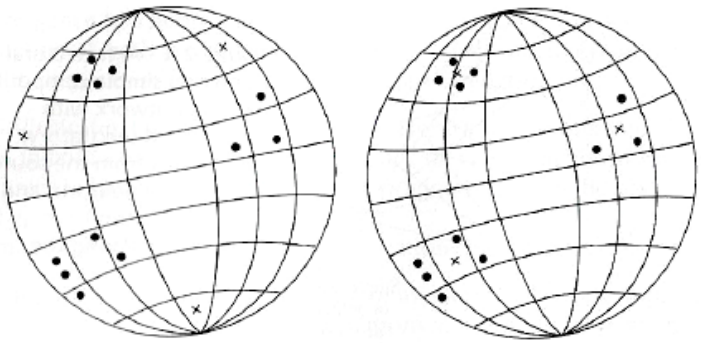
$$\Delta w_{kj} = \begin{cases} \eta(x_j - w_{kj}) \text{ se } k \text{ é o vencedor} \\ 0 \text{ c.c.} \end{cases}$$

A regra tem o efeito de mover o vetor de peso sináptico do neurônio k na direção da entrada $\mathbf{x}$.

Aprendizado Competitivo - Interpretação geométrica

Assuma que cada vetor de entrada $\mathbf{x}$ tem um comprimento constante Euclidiano, tal que nós podemos vê-lo como um ponto em uma esfera unitária N-dimensional, em que N é o número de nós de entrada. N também representa a dimensão de cada vetor de pesos sináptico $\mathbf{w_k}$.

Assumindo que todos os neurônios na rede são restritos a ter o mesmo comprimento Euclidiano (norma), $\sum_j w_{kj}^2 = 1$ para todo k, quando os pesos sinápticos são apropriadamente escalados, eles forma um conjunto de vetores que caem na mesma efera unitária dos elementos de entrada.



Aprendizado de Boltzmann

É um algoritmo de aprendizado estocástico derivado da *mecânica estatística*. A rede neural baseada neste aprendizado é a Máquina de Boltzmann.

Em uma Máquina de Boltzmann, os neurônios constituem uma estrutura recorrente e operam em um modelo binário (on $= +1$; off $=$ -1). A máquina é caracterizada por uma função energia E, cujo valor é determinado por estados particulares ocupados pelos neurônios da máquina:

$$E = \frac{1}{2} \sum_{j} \sum_{k} w_{kj} x_k x_j, j \neq k.$$

o fato de $j \neq k$ significa que não existe nenhum tipo de auto-alimentação.

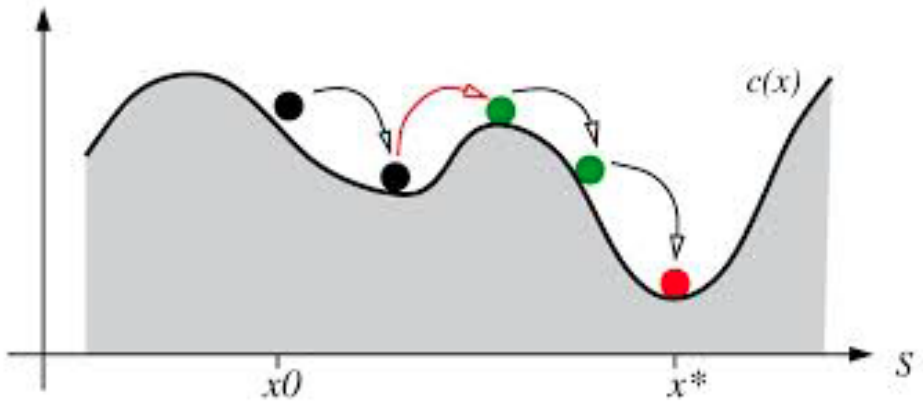
Aprendizado de Boltzmann

A máquina opera escolhendo um neurônio aleatoriamente, e trocando o estado do neurônio escolhido (k) do estado x_k para o estado $-x_k$, para uma determinada temperatura T, com probabilidade:

$$P(x_k \rightarrow -x_k) = \frac{1}{1 + exp(\frac{-\Delta E_k}{T})},$$

em que ΔE_k é a mudança de energia (mudança na função energia) resultante de tal troca. Se a regra é aplicada repetidamente, a máquina alcança o equilíbrio.

* Ideia do Simulated Annealing



Aprendizado de Boltzmann

Os neurônios dessa rede são de dois grupos funcionais:

- visível: interface entre a rede e o ambiente que ela opera;
- escondido: operam livremente.

E há dois modos de operação:

- condição fixa: os neurônios visíveis são fixados em determinados estados, de acordo com o ambiente;
- condição livre: os neurônios visíveis ou escondidos operam livremente.

A regra de alteração de pesos é: $\Delta w_{kj} = \eta(\rho_{kj}^{+} - \rho_{kj}^{-})$, com $j \neq k$, em que:

- ρ_{kj}^{+} = correlação entre os estados de neurônios j e k em operação livre;
- ρ_{kj}^{-} = correlação entre neurônios j e k em condição fixa.

A correlação considera todos os possíveis estados da rede sem sua condição de equilíbrio.

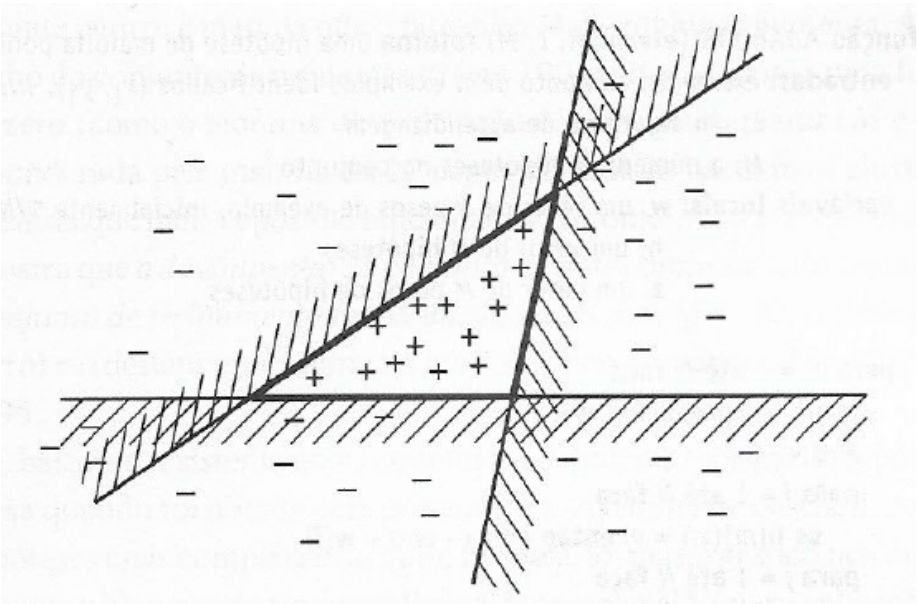
Aprendizado de Boltzmann - modelagem para o Caixeiro Viajante

Um modelo simples:

- neurônios U_ij: cidade é visitada na ordem j
- conexões: pesos representados pelas distâncias entre as cidades
- operação: ligar ou desligar neurônios
- função energia: qualidade da rota
- probabilidade de troca: vizinhança da solução (pioras aceitas via o controle baseado na temperatura)

Aprendizado por agrupamento

Russel e Norvig nomeiam de aprendizagem por agrupamento a estratégia de selecionar uma coleção inteira, ou um agrupamento, de hipóteses, a partir do espaço de hipóteses, e combinar suas previsões.



Onde todas as hipósteses concordam ....

Três hipóteses lineares. A região triangular resultante é uma hipótese que não pode ser expressa no espaço de hipóteseds original.

Aprendizado por agrupamento

Motivação

Considere um conjunto de $M = 5$ hipóteses e suponha que combinamos suas previsões usando votação por maioria simples. Para o conjunto classificar de forma incorreta um novo exemplo, pelo menos três das cinco hipóteses têm que classificar o exemplo de modo incorreto. A esperança é que isso seja muito menos provável do que obter uma classificação incorreta a partir de uma única hipótese.

Suposição

Cada hipótese h_i no conjunto tem um erro p – i.e., a probabilidade de um exemplo escolhido ao acaso ser classificado de forma incorreta por h_i é p. Também, considere que os erros cometidos por cada hipótese sejam independentes. Nesse caso, se p é pequeno, então a probabilidade de ocorrer um grande número de classificações incorretas é pequeno. Porém ....

Aprendizado por agrupamento

... a suposiçao de independência é pouco razoável, porque as hipóteses provavelmente serão '*iludidas* do mesmo modo por quaisquer aspectos enganosos dos dados de treinamento. Mas ...

... se as hipóteses forem pelo menos um pouco diferentes, reduzindo assim a correlação entre seus erros, então o aprendizado por agrupamento poderá ser muito útil.

Outra forma de compreender o aprendizado por agrupamento é considerá-lo como uma forma de amplicar o espaço de hipóteses.

Taxonomia

- Comitê de Máquinas
  - Mistura de Especialistas
  - *Ensemble*

Aprendizado por agrupamento

(Lima, 2004) Em geral, o aprendizado por agrupamento se dá em dois passos:

- geração de vários componentes (treinamento de componentes);
- combinação das saídas propostas pelos componentes.

Porém, há uma alternativa que insere uma fase intermediária de "seleção de componentes".

Quanto à geração dos componentes, as abordagens predominantes são: *bagging* e *boosting*.

Quanto à combinação dos componentes, poder feita por:

- pruralidade ou voto majoritário para problemas de classificação;
- médias simples para problemas de regressão;
- empilhamento: quando um componene final faz a combinação;
- uso de pesos na combinação (podendo ser uma medida de confiança) dos componentes com eliminação de redundância.

Aprendizado por agrupamento - *bagging*

Baseado em amostragem "*bootstrap*". São gerados vários conjuntos de treinamento a partir de uma amostragem uniforme do conjunto original de dados, com reposição, e então é obtida uma proposta de solução a partir de cada um destes conjuntos de treinamento.

Os conjuntos de treinamento têm o mesmo número de amostras do conjunto original, mas algumas amostras do conjunto original podem aparecer mais de uma vez, fazendo com que outras não sejam selecionadas.

Esta distinção aleatória dentre os vários conjuntos de treinamento confere diversidade aos modelos obtidos por cada um desses conjuntos.

Aprendizado por agrupamento - *boosting*

Os vários conjuntos de treinamento são gerados a partir de uma probabilidade de escolha de uma amostra que depende da contribuição dela para o erro de treinamento dos componentes já treinados.

Ou seja, caso uma amostra não tenha sido corretamente classificada pelos componentes já gerados, a probabilidade de escolha desta aumenta em relação às demais amostras quando do treinamento de novos componentes.

Assim, essa amostra terá uma chance maior de ser escolhida para compor o conjunto de dados do próximo componentes a ser gerado. Portanto, apenas o primeiro componente do conjunto é treinado a partir de uma amostragem uniforme do conjunto de dados original.

Nesse esquema, os vários componentes do conjunto são treinados sequencialmente, visando redefinir a probabilidade de escolha das amostras na geração dos próximos conjuntos de treinamento.

Aprendizado por agrupamento - *boosting*

Boosting é conhecido também por ACELERAÇÃO. Existem muitas variantes de *boosting* que diferem na forma como as amostras são ponderadas e como a resposta final é obtida. Uma forma de combinar as respostas é de acordo com a eficiência de cada hipótese obtida. Um algoritmo popular para a aceleração é o ADABOOST, o qual possui uma propriedade importante:

Se o algorimto de entrada L é um algoritmo de aprendizagem fraca – o que significa que L sempre retorna uma hipótese com erro ponderado sobre o conjunto de treinamento que é ligeriamente melhor que o palpite aleatório – então o ADABOOST retornará uma hipótese que classifica perfeitamente os dados de treinamento para um M grande o bastante (M = número de componentes).

ADABOOST - Russel e Norvig

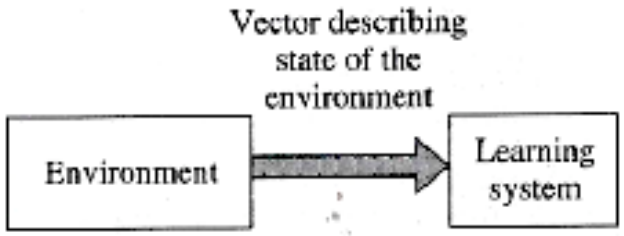
função ADABOOST(*exemplos*, *L*, *M*) **retorna** uma hipótese de maioria ponderada
  entradas: *exemplos*, conjunto de *N* exemplos identificados $(x_1, y_1), \ldots, (x_N, y_N)$
    L, um algoritmo de aprendizagem
    M, o número de hipóteses no conjunto
  variáveis locais: **w**, um vetor de *N* pesos de exemplo, inicialmente $1/N$
    h, um vetor de *M* hipóteses
    z, um vetor de *M* pesos de hipóteses

  para $m = 1$ **até** *M* **faça**
    $\mathbf{h}[m] \leftarrow L(exemplos, \mathbf{w})$
    $erro \leftarrow 0$
    para $j = 1$ **até** *N* **faça**
      se $\mathbf{h}[m](x_j) \neq y_j$ **então** $erro \leftarrow erro * \mathbf{w}[j]$
    para $j = 1$ **até** *N* **faça**
      se $\mathbf{h}[m](x_j) = y_j$ **então** $\mathbf{w}[j] \leftarrow \mathbf{w}[j]\ erro/(1 - erro)$
    $\mathbf{w} \leftarrow$ NORMALIZAR(**w**)
    $\mathbf{z}[m] \leftarrow \log\ (1 - erro)/erro$
  retornar MAIORIA-PONDERADA(**h**, **z**)

Variante ADABOOST do método de aceleração para aprendizagem por agrupamento. "Maioria-ponderada(h,z)" gera uma hipótese que retorna o valor de saída com o voto mais alto entre as hipóteses em h, sendo que os votos são ponderados por z.

Aprendizado não-supervisionado, por Simon Haykin

A provisão é feita por uma medida independente da tarefa que atesta a qualidade da representação que a rede neural (ou o algoritmo) aprendeu, e os parâmetros livres da rede são otimizados com respeito a tal medida. Uma vez que a rede neural está estável, ela desenvolve uma habilidade de formar uma representação interna que codifica características da entrada.



O aprendizado não supervisionado pode ser implementado com o aprendizado competitivo.

Adaptação, por Simon Haykin

Dimensões do aprendizado

Espaço é uma dimensão fundamental do aprendizado. O **tempo** é a outra.

- quando uma rede neural opera em um ambiente estacionário (um ambiente cujas características estatísticas não mudam com o tempo), as características estatística essenciais do ambiente podem, em teoria, ser aprendidads pela rede sob a supervisão de um professor.
- mas, frequentemente, o ambiente de interesse é não estacionário, o que significa que os parâmetros estatísticos dos sinais gerados mudam com o tempo. Nesse caso, os métodos tradicionais de aprendizado são inadequados porque a rede não é equipada com os meios necessários para perseguir as variações estatísticas do ambiente.

Adaptação

Para sobrepor este problema é desejável que a rede neural continue a adaptar os seus parâmetros livres para variações no sinal de entrada em um modelo temporal. Assim, um sistema adaptativo responde para toda entrada como se ela fosse nova. O aprendizado "nunca" pára. Isso é chamado aprendizado contínuo ou aprendizado on-the-fly.

Como uma rede neural pode se adaptar a dados de entrada que variam no tempo? Admite-se que as características estatísticas de um processo estacionário usualmente muda muito devagar, e o processo pode ser considerado pseudo-estacionário, sobre uma janela de curta duração.

Exemplo

O mecanismo responsável pela produção de um sinal de voz pode ser considerado essencialmente estacionário sobre um período de 10 a 30 milisegundos.

Adaptação

Uma abordagem dinâmica de aprendizado deve:

- selecionar uma janela pequena o bastante para os dados de entrada serem considerados pseudo-estacionários, e usar os dados para treinar a rede;
- quando um novo exemplo de dados é recebido, atualizar a janela por apagar o evento mais velho e deslocar o restante dos dados para trás uma unidade de tempo, abrindo espaço para o novo dado;
- usar a janela de dados atualizada para retreinar a rede;
- reptir o procedimento regularmente.

Profa. Dra. Sarajane Marques Peres
Programa de Pós Graduação em Sistemas de Informação - PPgSI
Escola de Artes, Ciências e Humanidades - EACH
Universidade de São Paulo - USP