Aprendizado construtivo

Índice Geral

1. Motivação para o aprendizado construtivo

- Prosseguimos, neste tópico do curso, com redes neurais não-recorrentes e treinamento supervisionado. Neste contexto, tanto os métodos construtivos como os de poda podem ser empregados na busca de arquiteturas de redes neurais dedicadas às demandas de cada aplicação, no sentido de disporem de recursos de processamento de informação na medida certa para resolver o problema de mapeamento multidimensional de entrada-saída (problema de aproximação de função).
- A motivação para o uso de métodos construtivos pode ser apresentada levando-se em conta o fato deles operarem no sentido contrário dos métodos de poda.
- Conforme descrito por KARNIN (1990), LE CUN *et al.* (1990), HASSIBI & STORK (1993) e REED (1993), os métodos de poda supõem que a arquitetura inicial da rede neural contém pelo menos tanta estrutura quanto a necessária para realizar a tarefa de aproximação.

- Por exemplo, é comum estabelecer que a arquitetura inicial apresenta uma dimensão elevada e conexões entre todos os neurônios ou, pelo menos, entre todos os neurônios de camadas adjacentes. Neste caso, os recursos considerados em excesso por não estarem sendo utilizados ativamente no processo de aproximação podem ser gradativamente desativados ou simplesmente eliminados.
- Os recursos em excesso devem ser adequadamente identificados, podendo corresponder a conexões, neurônios ou até camadas de neurônios. Após o procedimento de poda, geralmente segue um processo de reajuste da estrutura ainda ativa.
- No entanto, os métodos de poda apresentam invariavelmente os seguintes problemas (GHOSH & TUMER, 1994; KWOK & YEUNG, 1995):
  - ✓ Não existe um método prático de se determinar diretamente uma arquitetura inicial para a rede neural que contenha garantidamente tanta estrutura quanto a necessária para realizar a tarefa de aproximação. Com isso, para aumentar a

probabilidade de se escolher uma arquitetura com tal característica, geralmente adotam-se arquiteturas iniciais fortemente sobredimensionadas.

- Já que a maior parte do processo de aproximação é realizado considerando-se redes neurais sobredimensionadas, a demanda por recursos computacionais durante o treinamento é grande e parte dos recursos computacionais utilizados acaba sendo desperdiçada toda vez que a poda elimina estruturas que já passaram por alguma fase de processamento.
- Como geralmente inúmeras redes neurais de diferentes dimensões são capazes de representar soluções aceitáveis para o problema de aproximação, a aplicação de métodos de poda não favorece a escolha da solução de menor dimensão, ou seja, aquela com um menor número de componentes e operadores.
- Para que métodos de poda sejam computacionalmente factíveis, eles devem estimar o efeito da eliminação de cada recurso individualmente, mas devem

eliminar múltiplos recursos simultaneamente. Com isso, não é possível obter uma estimativa confiável do efeito que cada operação de poda possa vir a causar junto ao erro de aproximação.

- O tratamento de parte destes problemas tem conduzido a soluções específicas, como em WEIGEND *et al.* (1991), embora com base em métodos mal-condicionados e pouco eficientes computacionalmente, conforme observado por KWOK & YEUNG (1995).

- Por operarem no sentido contrário dos métodos de poda, os métodos construtivos podem evitar a ocorrência de problemas como os mencionados acima. No entanto, pelo fato de não ser possível garantir que toda inclusão de estrutura por parte do algoritmo construtivo venha contribuir para a solução do problema de aproximação, métodos de poda representam um procedimento complementar importante, no sentido de promover a eliminação de estruturas desnecessariamente incluídas (ou necessárias, quando incluídas, mas que perderam relevância durante a construção do modelo).

- Conclui-se, portanto, que um método híbrido de aproximação é o mais adequado, contendo etapas construtivas seguidas de etapas de poda. Em razão de procedimentos de poda só entrarem em operação esporadicamente junto ao modelo de aproximação, o método híbrido de aproximação é predominantemente construtivo. Sendo assim, é preservada aqui a denominação de método construtivo de aproximação, mesmo que haja etapas de poda.

2. Regressão por busca de projeção (*Projection Pursuit Regression – PPR*)

- Objetivo: Realização da automação de etapas adicionais do processo de aquisição de conhecimento, baseado em aprendizagem e generalização.
- Modelos de regressão por busca de projeção (*projection pursuit regression*):

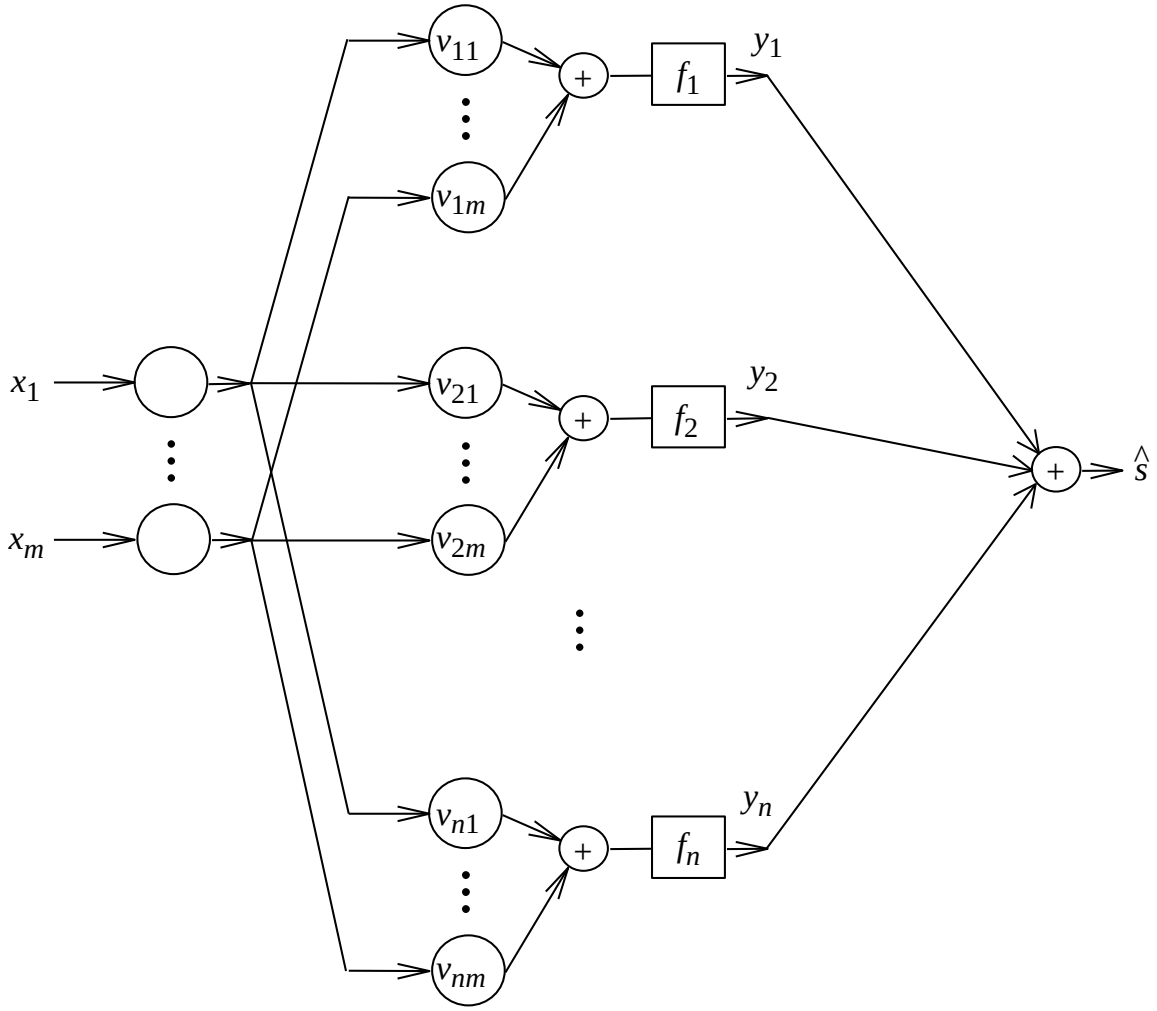
$$\hat{s} = \hat{g}_n(\mathbf{x}) = \sum_{j=1}^{n} f_j(\mathbf{v}_j^T \mathbf{x}), \qquad (1)$$

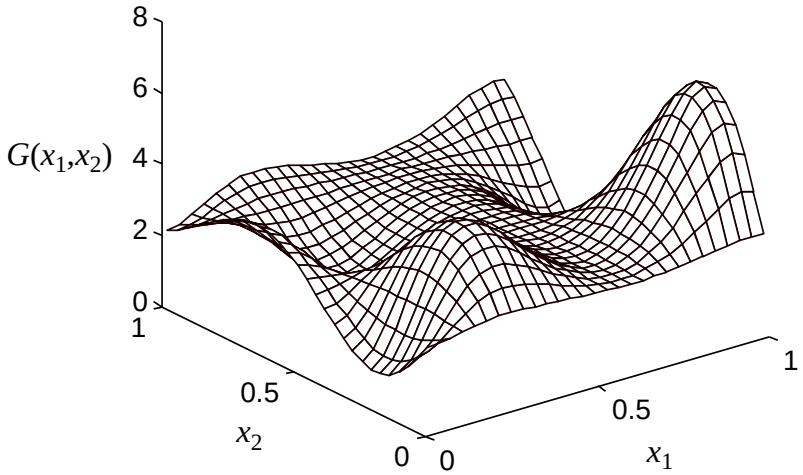
  - ✓ $\mathbf{x} \in \Re^m$ é o vetor de variáveis de entrada;

- ✓ $\mathbf{v}_j \in \Re^m$ é a direção de projeção (j=1,...,n).

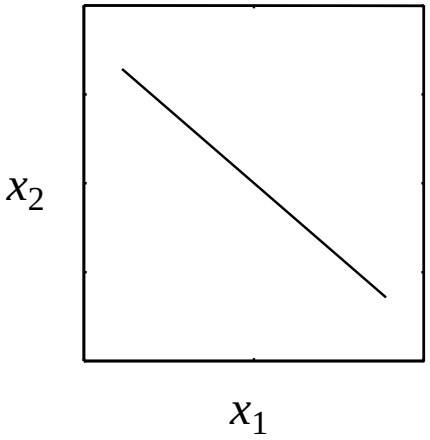
- O produto escalar $\mathbf{v}_j^T \mathbf{x}$ pode ser tomado, a menos de um fator de escala, como uma projeção de **x** na direção $\mathbf{v}_j$.
- O j-ésimo termo $f_j(\cdot)$ do somatório é constante para **x** em hiperplanos do $\Re^m$ na forma $\mathbf{v}_j^T \mathbf{x} = c$, com $c \in \Re$ constante (função de expansão ortogonal ou *ridge function*).
- A utilização de modelos na forma da equação (1) conduz a processos de aproximação por expansão ortogonal aditiva (FRIEDMAN & STUETZLE, 1981). Neste modelo, os termos da composição aditiva correspondem a funções escalares de expansão ortogonal a projeções unidirecionais.
- A projeção consiste de operações lineares em que uma estrutura de uma determinada dimensão tem suprimidas algumas de suas particularidades de modo a tornar possível sua representação em espaços de menor dimensão.

Modelo de regressão por busca de projeção (*Projection Pursuit Regression*)

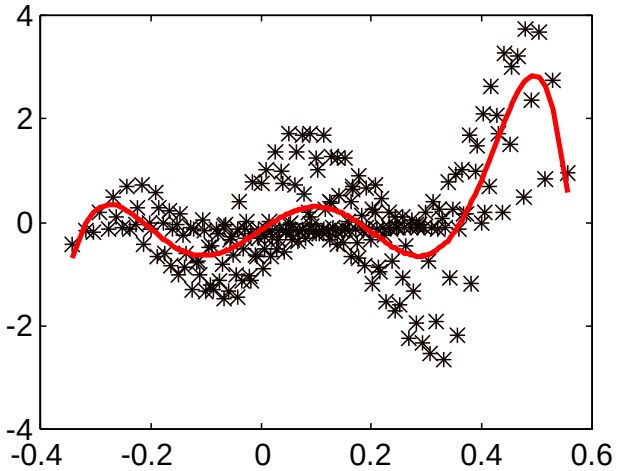




Mapeamento original de onde se amostram dados



Direção de projeção



Dados projetados e melhor representação

- A estrutura projetada pode ser considerada como uma "sombra" da estrutura original, fazendo com que as projeções *mais interessantes* sejam aquelas que preservam parcelas representativas da estrutura original.
- A busca dessas direções de projeção envolve uma série de manipulações do conjunto de dados de entrada/saída disponíveis. Baseadas nos dados de entrada-saída, as direções de projeção devem enfatizar as relações, possivelmente não-lineares, existentes entre as variáveis do problema de aproximação.
- A questão que surge é: Como obter de forma automática essas direções de projeção? Uma alternativa foi apresentada por FRIEDMAN & TUKEY (1974), em que a direção de projeção corresponde à solução que maximiza um determinado índice numérico de projeção. A partir de então, uma série de índices foram apresentados na literatura, cada qual evidenciando um conjunto particular de características a serem atendidas pelos dados projetados.

- Em virtude da inexistência de um índice de projeção que se aplique a todos os casos, e como geralmente não existe um conhecimento prévio a respeito das características presentes no conjunto de dados de aproximação, a escolha prévia de um índice de desempenho na determinação da direção de projeção, ao invés de determiná-la arbitrariamente, permite assegurar apenas um aumento da probabilidade de se encontrar direções de projeção *interessantes*. Em boa parte dos problemas de aproximação, este aumento da probabilidade é bastante significativo (HUBER, 1985).
- Não serão considerados, no entanto, índices de projeção neste tópico do curso, pois uma das contribuições da adaptação dos conceitos de PPR em redes neurais artificiais é a aplicação de técnicas de otimização não-linear no ajuste dos pesos da camada intermediária, neurônio a neurônio, os quais representam a direção de projeção para os dados de entrada. A direção de projeção inicial será, assim, aleatória e faz-se uma busca exploratória, definindo múltiplas direções iniciais candidatas.

- Uma vez definida a direção de projeção, funções monovariáveis são então determinadas de tal forma que sua expansão em direções ortogonais à direção de projeção forneça a melhor aproximação possível com base nos dados disponíveis.
- A interdependência entre as funções de expansão ortogonal e a correspondente direção de projeção acaba conduzindo a processos iterativos de aproximação. Para que estes processos sejam computacionalmente eficientes, é geralmente necessário que:

    1. Cada passo do processo iterativo demande a menor quantidade de cálculo possível, o que geralmente conduz à necessidade de se recorrer a informações variacionais;
    2. Propriedades teóricas que garantam redução de dimensão estejam presentes, fazendo com que a aproximação em espaços multidimensionais apresente taxas de convergência típicas de problemas de aproximação de menor dimensão (STONE, 1982).

- Dentre as vantagens de se utilizar projeções unidirecionais têm-se a manutenção de uma maior simplicidade do processo de aproximação e a possibilidade de visualizar graficamente o comportamento da função de aproximação na direção de projeção. Além disso, é possível explicitar o tipo de associação não-linear existente entre as variáveis (na direção de projeção) e uma série de outras informações que não estão diretamente disponíveis considerando-se a dimensão completa do espaço de aproximação.

- No entanto, apesar de não ser o objetivo deste estudo, é importante salientar que a interpretação dos dados projetados geralmente não se apresenta como uma tarefa simples. Estruturas interessantes presentes na projeção dos dados não necessariamente correspondem à projeção de estruturas interessantes, da mesmo forma que estruturas interessantes presentes nos dados podem conduzir a nenhuma projeção interessante (JONES & SIBSON, 1987).

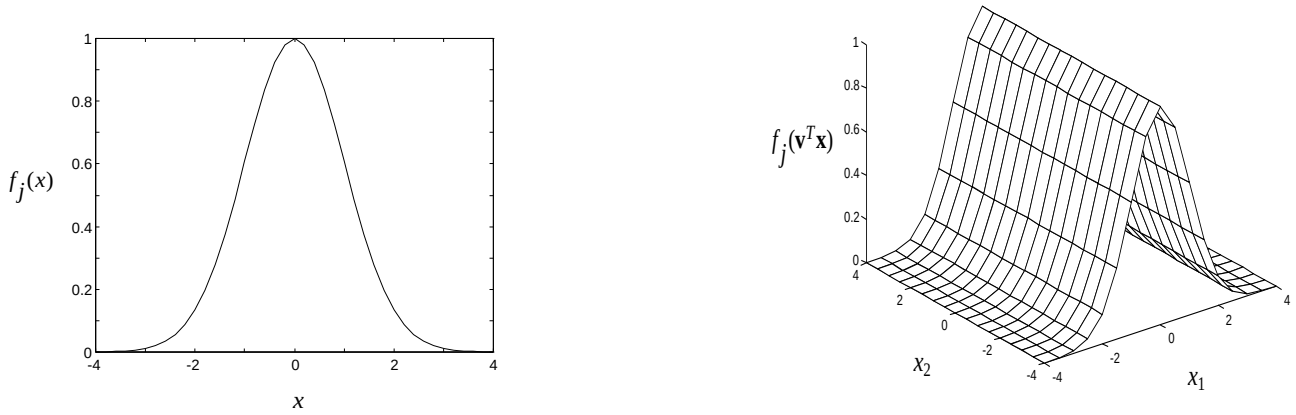
- Além disso, é improvável que exista apenas uma única direção de projeção capaz de explicitar todo tipo de informação a respeito do conjunto multidimensional de dados de entrada-saída, o que seria equivalente a supor que o problema de aproximação é monovariável.
- Mesmo que este seja o caso, nem sempre é possível garantir a determinação adequada desta direção de projeção. Portanto, justifica-se o estabelecimento de uma sequência de direções de projeção, cada qual explicitando a maior parcela possível de informação necessária para o sucesso da tarefa de aproximação.
- Com isso, após a definição de uma direção de projeção e da correspondente função de expansão ortogonal, **uma transformação deve ser aplicada ao conjunto de dados para que a informação já representada seja removida**, permitindo o reinício do processo a partir de uma nova direção de projeção e com base em um novo problema de aproximação: o problema original, menos o que já foi representado.

- Logo, a busca sequencial de direções de projeção pode ser implementada na forma (FRIEDMAN *et al.*, 1984):

    1. Encontra-se um direção de projeção *ótima* (segundo algum critério);
    2. Remove-se do conjunto de dados a estrutura resultante da projeção dos dados nesta direção;
    3. Reinicia-se o processo até que nenhuma outra projeção revele qualquer estrutura, ou seja, até que o modelo de aproximação concorde com os dados amostrados em todas as projeções.

- Este procedimento iterativo e construtivo, em que cada novo sub-problema de aproximação deve representar apenas informações não representadas pelos sub-problemas de aproximação anteriores, produz funções de aproximação multivariáveis utilizando composição aditiva de funções monovariáveis expandidas ortogonalmente, na forma da equação (1).

- Como as funções $f_j(\cdot)$ (j=1,...,n) são constantes para valores de **x** em hiperplanos do $\Re^m$, elas são denominadas funções de expansão ortogonal a uma determinada direção – *ridge functions* (DAHMEN & MICCHELLI, 1987). Tomando m = 2 e $f_j(\cdot)$ arbitrário, as Figuras 1(a) e 1(b) permitem verificar esta propriedade. Observe que a expansão é ortogonal à direção de projeção $\mathbf{v} = [1\ 0]^T$.



(a) $f_j(x) = e^{-0,5 \cdot x^2}$ (b) $f_j(\mathbf{v}^T\mathbf{x}) = e^{-0,5 \cdot \left([1\ 0] \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} \right)^2}$

Figura 1 – Função de expansão ortogonal em que $\mathbf{v} = [1\,0]^T$ e $\mathbf{x} = [x_1\ x_2]^T$

3. O problema de aproximação resultante

- O problema de aproximação por expansão ortogonal aditiva pode ser completamente descrito na forma (FRIEDMAN & STUETZLE, 1981; HUBER, 1985):
  - ✓ Seja X uma região compacta do $\Re^m$ e seja $g: X \subset \Re^m \rightarrow \Re$ a função a ser aproximada.
  - ✓ O conjunto de dados de aproximação $\{(\mathbf{x}_l, s_l) \in \Re^m \times \Re\}_{l=1}^N$ é gerado considerando-se que os vetores de entrada $\mathbf{x}_l$ estão distribuídos na região compacta $X \subset \Re^m$ de acordo com uma função densidade de probabilidade fixa $d_P: X \subset \Re^m \rightarrow [0,1]$ e que os vetores de saída s_l são produzidos pelo mapeamento definido pela função g na forma:

$$s_l = g(\mathbf{x}_l) + \varepsilon_l, \quad l = 1, \ldots, N,$$

onde $\varepsilon_l \in \Re$ é uma variável aleatória de média zero e variância fixa.

- ✓ A função g que associa a cada vetor de entrada $\mathbf{x} \in X$ uma saída escalar $s \in \Re$ pode ser aproximada com base no conjunto de dados de aproximação $\{(\mathbf{x}_l, s_l) \in \Re^m \times \Re\}_{l=1}^N$ por uma composição aditiva de funções de expansão ortogonal na forma:

$$g(\mathbf{x}) \approx \hat{g}_n(\mathbf{x}) = \sum_{j=1}^{n} f_j(\mathbf{v}_j^T \mathbf{x}),$$

sendo que as funções de expansão ortogonal $f_j(\cdot)$, por serem constantes em direções ortogonais ao plano de projeção, são consideradas como uma generalização de funções lineares. Por motivações de ordem numérica e por analogia com operadores de projeção, é interessante, sempre que possível, tomar direções de projeção de norma unitária tal que $\mathbf{v}_j^T \mathbf{v}_j = 1$ (j=1,...,n).

- ✓ Considere que os primeiros n–1 termos já foram determinados, ou seja, os vetores $\mathbf{v}_j$ e as funções $f_j(\cdot)$ (j=1,...,n–1). Sejam:

$$d_l = s_l - \sum_{j=1}^{n-1} f_j(\mathbf{v}_j^T \mathbf{x}_l),\ l=1,...,N,$$

os resíduos do processo de aproximação. Obtenha a direção de projeção $\mathbf{v}_n$ e a função $f_n(\cdot)$, soluções do seguinte problema de otimização com restrição de suavidade:

$$\min_{\mathbf{v}_n, f_n} \frac{1}{N} \sum_{l=1}^{N} \left(d_l - f_n(\mathbf{v}_n^T \mathbf{x}_l) \right)^2 + \lambda_n \phi(f_n). \qquad (2)$$

- ✓ Faça $n = n+1$ e repita o processo a partir do cálculo dos novos resíduos, enquanto o nível de aproximação desejado ainda não foi atingido.

- Este processo de aproximação tem algumas propriedades importantes:

  1. A aproximação por expansão ortogonal aditiva apresenta um bom nível de robustez a dados não-informativos (HUBER, 1985).

2. Considerando que a função g a ser aproximada é quadraticamente integrável, uma condição quase sempre satisfeita em regiões compactas de espaços multidimensionais, HUBER (1985) conjeturou a convergência absoluta da aproximação dada pela equação (1), o que mais tarde foi demonstrado por JONES (1987). Além disso, HALL (1989) demonstrou que a taxa de convergência do processo é $\sqrt{n}$-consistente e independente da dimensão m do espaço de entrada.
3. Obviamente, a convergência é tanto mais rápida quanto mais estruturalmente aditiva for a não-linearidade presente nas associações entre as variáveis do problema de aproximação.

4. Determinação da função de expansão ortogonal

- Uma vez definida a direção de projeção $\mathbf{v}_n \in \Re^m$, o problema de aproximação regularizado apresentado na equação (2) tem por objetivo aproximar uma versão suave da projeção dos resíduos da função desconhecida g: $\Re^m \rightarrow \Re$ na direção $\mathbf{v}_n$.

- O fator fundamental que continua caracterizando todo o processo de implementação é a tentativa de aproximar a função em regiões onde não se dispõe de informação suficiente para implementar um processo totalmente não-paramétrico.
- Para $\mathbf{v}_n$ ($n \geq 1$) fixo, é possível renomear as projeções unidirecionais dos dados de entrada na forma:

$$z_l = \mathbf{v}_n^T \mathbf{x}_l, \; l=1,...,N. \tag{3}$$

- Com isso, a função monovariável $f_n(\cdot)$ deve resolver o seguinte problema de aproximação regularizado:

$$\min_{f_n} \frac{1}{N} \sum_{l=1}^{N} \left[d_l - f_n(z_l) \right]^2 + \lambda_n \phi(f_n), \tag{4}$$

onde d_l são os resíduos do processo de aproximação, dados por:

$$d_l = s_l - \sum_{j=1}^{n-1} f_j(\mathbf{v}_j^T \mathbf{x}_l), \; l=1,...,N.$$

- Mesmo que o tipo de suavidade da função g seja compatível com aquele imposto pela função de regularização $\phi(\cdot)$ às funções $f_j(\cdot)$ (j=1,...,n–1), o comportamento dos pontos $\{(z_l, d_l)\}_{l=1}^{N}$ pode ser bastante errático devido à variação de $g(\mathbf{x}) - \sum_{j=1}^{n-1} f_j(\mathbf{v}_j^T \mathbf{x})$ em outras direções que não $\mathbf{v}_n$.
- Com isso, o valor ótimo do parâmetro de regularização λ_n não pode ser determinado a priori, sendo função do conjunto de dados projetados $\{(z_l, d_l)\}_{l=1}^{N}$.
- Solução paramétrica: Uso de bases de funções ortonormais (HWANG *et al.*, 1994).
- Solução não-paramétrica: Splines polinomiais suavizantes (VON ZUBEN, 1996).

4.1 Solução paramétrica empregando polinômios de Hermite

- Dados N pontos no plano, na forma $\{x_l, s_l\}_{l=1}^{N}$, é possível obter uma fórmula fechada para os coeficientes dos polinômios de Hermite de ordem P definidos abaixo, de tal forma a se obter a melhor aproximação segundo o método dos quadrados mínimos.

- Teorema 1: Dados f_1, $f_2 \in C[a,b]$ (onde $C[a,b]$ é o espaço das funções contínuas no intervalo $[a,b]$), se $w(x)$ é uma função integrável em $[a,b]$, então a integral

$$\langle f_1, f_2 \rangle_w = \int_a^b w(x) f_1(x) f_2(x) dx$$

  define um produto interno em $C[a,b]$. A função $w(x)$ é denominada função de ponderação, sendo geralmente tomada como sendo positiva em $[a,b]$.

- Teorema 2: Dadas as funções f_1, $f_2 \in C[a,b]$, elas são funções ortogonais se:

$$\langle f_1, f_2 \rangle = \begin{cases} 0 & \text{se } f_1 \neq f_2 \\ 1 & \text{se } f_1 = f_2 \end{cases}.$$

- De acordo com a escolha do intervalo $[a,b]$ e da função de ponderação $w(x)$, inúmeras funções ortonormais podem ser obtidas em $C[a,b]$ com base no processo de ortonormalização das potências 1, x, x^2, ... Neste estudo, são consideradas apenas as

funções ortonormais geradas a partir de polinômios de Hermite, que são obtidos tomando-se $a = -\infty$, $b = +\infty$ e $w(x) = e^{-x^2}$.

- Os polinômios de Hermite são definidos recursivamente como segue:
  - ✓ $p_0(x) = 1$;
  - ✓ $p_1(x) = 2x$;
  - ✓ $p_{i+1}(x) = 2xp_i(x) - 2ip_{i-1}(x), \ i > 0$.

Estes polinômios são ortogonais com base no seguinte produto interno:

$$\int_{-\infty}^{\infty} e^{-x^2} p_i(x) p_j(x) dx = \begin{cases} 0 & \text{se } i \neq j \\ \sqrt{\pi} 2^i i! & \text{se } i = j \end{cases},$$

de tal forma que as funções:

$$h_i(x) = \frac{e^{\frac{-x^2}{2}}}{\sqrt{\sqrt{\pi} 2^i i!}} p_i(x), \ i = 0,1,... \qquad (5)$$

são ortonormais em (–∞,∞) por produzirem:

$$\int_{-\infty}^{\infty} h_i(x)h_j(x)dx = \begin{cases} 0 & \text{se } i \neq j \\ 1 & \text{se } i = j \end{cases}. \qquad (6)$$

Charles Hermite

- O problema da melhor aproximação dos *N* pontos no plano, $\{x_l, s_l\}_{l=1}^{N}$, por polinômios de Hermite até ordem *P*, pode ser colocado na forma de um problema de quadrados mínimos como segue:

$$\min_{c_0,\ldots,c_P} \sum_{l=1}^{N}\left(s_l - \sum_{i=0}^{P} c_i h_i(x_l)\right)^2 . \qquad (7)$$

- Tomando a norma euclidiana $\|\cdot\|_2$ e construindo a matriz **H** e os vetores **c** e **s** na forma:

$$\mathbf{H} = \begin{bmatrix} h_0(x_1) & h_1(x_1) & \cdots & h_P(x_1) \\ h_0(x_2) & \ddots & & \vdots \\ \vdots & & & \\ h_0(x_N) & \cdots & & h_P(x_N) \end{bmatrix}, \quad \mathbf{c} = \begin{bmatrix} c_0 \\ c_1 \\ \vdots \\ c_P \end{bmatrix} \quad \text{e} \quad \mathbf{s} = \begin{bmatrix} s_1 \\ s_2 \\ \vdots \\ s_N \end{bmatrix},$$

a equação (7) pode ser reescrita como segue:

$$\min_{\mathbf{c}} \|\mathbf{s} - \mathbf{H}\mathbf{c}\|_2^2 = \min_{\mathbf{c}} (\mathbf{s} - \mathbf{H}\mathbf{c})^T (\mathbf{s} - \mathbf{H}\mathbf{c}) = \mathbf{s}^T\mathbf{s} + \min_{\mathbf{c}} \left(-\mathbf{s}^T\mathbf{H}\mathbf{c} - \mathbf{c}^T\mathbf{H}^T\mathbf{s} + \mathbf{c}^T\mathbf{H}^T\mathbf{H}\mathbf{c}\right).$$

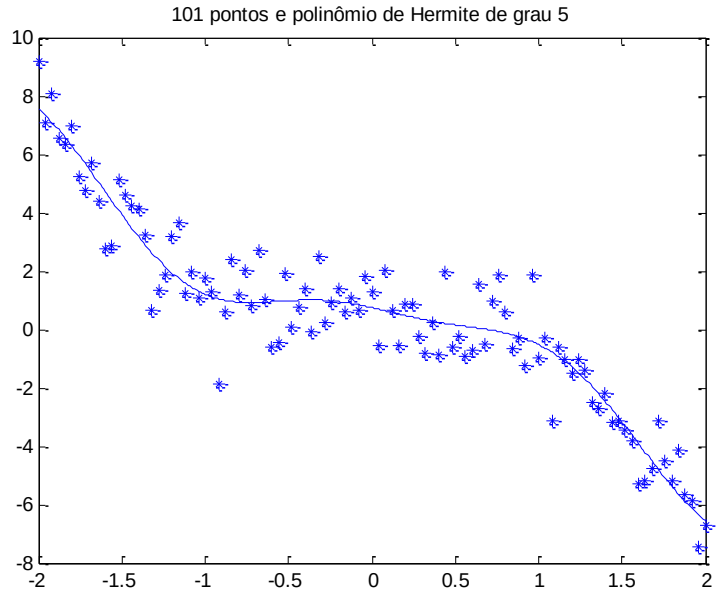
- Fazendo $J(\mathbf{c}) = -\mathbf{s}^T\mathbf{H}\mathbf{c} - \mathbf{c}^T\mathbf{H}^T\mathbf{s} + \mathbf{c}^T\mathbf{H}^T\mathbf{H}\mathbf{c}$, a seguinte condição necessária deve ser atendida no ponto de mínimo:

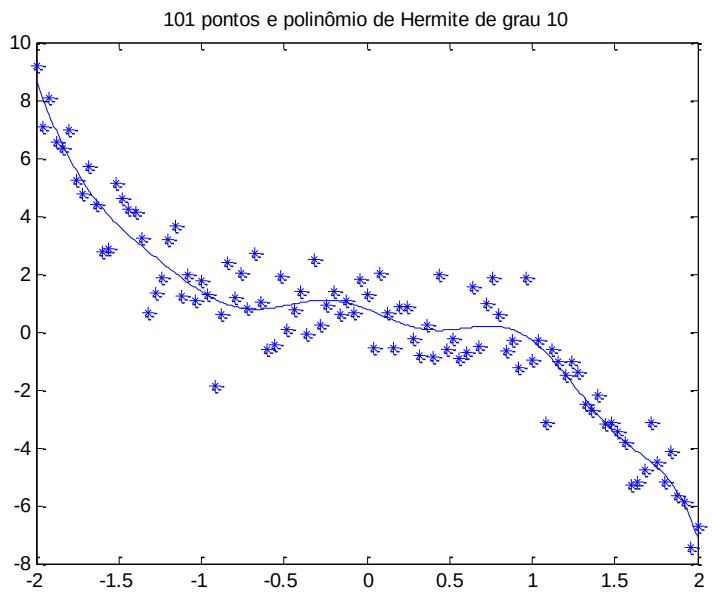
$$\frac{\partial J(\mathbf{c})}{\partial \mathbf{c}} = 0 \Rightarrow -2\mathbf{H}^T\mathbf{s} + 2\mathbf{H}^T\mathbf{H}\mathbf{c} = 0 \Rightarrow \mathbf{H}^T\mathbf{H}\mathbf{c} = \mathbf{H}^T\mathbf{s}.$$

- Considerando $P < N$ (ou seja, o grau do polinômio de maior grau menor que o número de amostras), a matriz $\mathbf{H}$ terá posto completo se todos os valores de x forem distintos, fazendo com que $\mathbf{H}^T\mathbf{H}$ seja inversível. Assim, a solução ótima, no sentido dos quadrados mínimos, é denominada $\mathbf{c}^*$ e pode ser expressa na forma:

$$\mathbf{c}^* = \left(\mathbf{H}^T\mathbf{H}\right)^{-1}\mathbf{H}^T\mathbf{s}. \qquad (8)$$

- Não é usual regularizar o problema de quadrados mínimos, pois é possível suavizar as funções-base que serão combinadas linearmente.
- Exemplos (sugere-se, na prática, o uso de um grau em torno de $P = 10$):





5. O processo de ajuste retroativo

- O processo de aproximação por expansão ortogonal aditiva segue, então, os seguintes passos básicos:

- ✓ Dado o conjunto de dados de aproximação $\{(\mathbf{x}_l, s_l) \in X \times \Re\}_{l=1}^{N}$, com $X \subset \Re^m$, e seja g: $X \subset \Re^m \rightarrow \Re$ a função a ser aproximada.
- ✓ Partindo de $n = 1$, e construindo o vetor $\mathbf{d} = [d_1 \ldots d_N]^T$ na forma:

$$d_l = s_l - \sum_{j=1}^{n-1} f_j(\mathbf{v}_j^T \mathbf{x}_l),\ l=1,\ldots,N,$$

resolva o seguinte problema de otimização:

$$\min_{\mathbf{v}_n, f_n} \frac{1}{N} \sum_{l=1}^{N} \left(d_l - f_n(\mathbf{v}_n^T \mathbf{x}_l) \right)^2 + \lambda_n \phi(f_n),$$

pela obtenção sucessiva de valores ótimos para f_n e $\mathbf{v}_n$ na forma:

1. Defina um valor inicial para $\mathbf{v}_n$;
2. Para $\mathbf{v}_n$ fixo, resolva o seguinte problema de otimização:

$$\min_{f_n} \frac{1}{N} \sum_{l=1}^{N} \left(d_l - f_n(\mathbf{v}_n^T \mathbf{x}_l) \right)^2 + \lambda_n \phi(f_n);$$

3. Para f_n fixo, partindo do valor atual de $\mathbf{v}_n$, resolva iterativamente o seguinte problema de otimização (este é um problema de otimização paramétrica):

$$\min_{\mathbf{v}_n} \frac{1}{N} \sum_{l=1}^{N} \left(d_l - f_n(\mathbf{v}_n^T \mathbf{x}_l) \right)^2;$$

4. Enquanto não houver convergência (medida por algum critério de parada), retorne ao passo 2.

✓ Faça $n = n+1$ e repita o processo a partir do cálculo dos novos valores para o vetor de resíduos **d** enquanto o nível de aproximação desejado ainda não foi atingido (medido por algum critério de parada).

- Desse processo de aproximação resulta, então, um modelo de aproximação por composição aditiva de funções de expansão ortogonal, na forma:

$$g(\mathbf{x}) \approx \hat{g}_n(\mathbf{x}) = \sum_{j=1}^{n} f_j(\mathbf{v}_j^T \mathbf{x}).$$

- No entanto, o processo de construção deste modelo de aproximação apresenta uma limitação advinda da estratégia de aproximação empregada, a qual é descrita a seguir:

  1) Com base nos dados de aproximação referentes ao problema de aproximação original;

  2) Encontre uma única direção de projeção e uma única função de expansão ortogonal a esta direção (sujeita a restrições de suavidade) que melhor aproxime os dados;

  3) Remova do conjunto de dados a informação representada no passo 2;

  4) Enquanto o nível de aproximação desejado ainda não foi atingido, retorne ao passo 2.

- Observe que cada um dos *n* termos da composição aditiva resultou de um processo de aproximação que tinha por objetivo representar *toda* a informação presente nos dados de aproximação e que *ainda não tinham sido representadas* pelos termos anteriores.

- Isso implica que cada novo termo da composição aditiva não leva em conta a possibilidade de que, posteriormente, *novos* termos possam vir a compor o processo de aproximação.
- Tomando qualquer termo da composição aditiva, com exceção do *n*-ésimo termo, e denominando-o *k* ($1 \leq k < n$), surge a seguinte questão: O que ocorreria com f_k e $\mathbf{v}_k$ se, na solução do problema:

$$\min_{\mathbf{v}_k, f_k} \frac{1}{N} \sum_{l=1}^{N} \left(d_l - f_k(\mathbf{v}_k^T \mathbf{x}_l) \right)^2 + \lambda_k \phi(f_k),$$

em lugar de $\mathbf{d} = [d_1 \ldots d_N]^T$ tal que

$$d_l = s_l - \sum_{j=1}^{k-1} f_j(\mathbf{v}_j^T \mathbf{x}_l),\ l=1,\ldots,N,$$

se tomasse

$$d_l = s_l - \sum_{\substack{j=1 \\ j \neq k}}^{n} f_j(\mathbf{v}_j^T \mathbf{x}_l),\ l=1,\ldots,N\ ?$$

- Se as duas escolhas para o vetor de resíduos **d** produzirem dados distintos, então f_k e $\mathbf{v}_k$ podem ser (e geralmente são) diferentes em cada caso. Conclui-se, portanto, que a solução produzida pelo processo de aproximação descrito acima pode deixar de ser ótima para os termos já calculados sempre que um termo adicional for incorporado.
- Sendo assim, é recomendável a aplicação de um processo de ajuste retroativo (*backfitting*) na forma:

    1. Para cada j $(1 \leq j \leq n)$, omite-se $f_j(\mathbf{v}_j^T \mathbf{x}_l)$ do somatório e determinam-se novos valores ótimos para f_j e $\mathbf{v}_j$ (os já obtidos são condição inicial). Repita o processo de ajuste retroativo até convergência, medida por algum critério de parada.

- Uma demonstração de convergência do processo de ajuste retroativo foi apresentada por BREIMAN & FRIEDMAN (1985). Vale salientar também que o processo de ajuste retroativo foi originalmente proposto para reajustar apenas a função f_j, mantendo-se fixa a direção $\mathbf{v}_j$ (FRIEDMAN & STUETZLE, 1981).

- Além disso, o processo de ajuste retroativo é indispensável na implementação de métodos complementares de poda.

6. Aprendizado por busca de projeção

- Deixando de lado alguns detalhes, como o tratamento de múltiplas saídas, apresentamos a seguir o algoritmo construtivo denominado aprendizado por busca de projeção, que vai produzir a rede neural da figura a seguir.
- Ela é muito similar a uma rede neural MLP, com a diferença fundamental de ter uma função de ativação distinta para cada neurônio da camada intermediária, cujo formato é definido durante o treinamento e de acordo com as demandas da aplicação, e também levando-se em conta que o número de neurônios desta única camada intermediária é definido de modo automático pelo algoritmo construtivo.

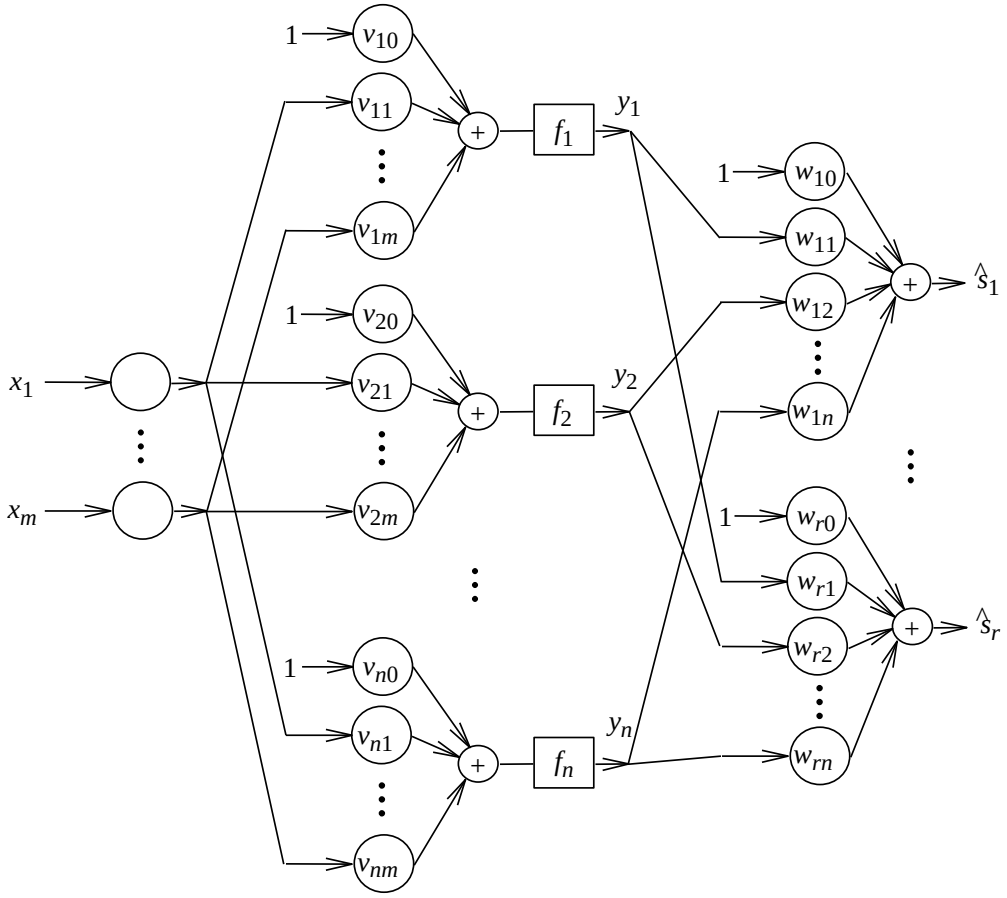


Figura 2 – Rede neural com uma camada intermediária e funções de ativação distintas, resultante do processo de aprendizado construtivo a ser apresentado.

Algoritmo de aproximação construtivo para múltiplas saídas

1. Dados $\mathbf{X} \in \Re^{m \times N}$ e $\mathbf{S} \in \Re^{r \times N}$, tome $j = 0$, $\mathbf{D} = \mathbf{S}$;

2. Faça $j = j+1$ e atribua um valor inicial para $\mathbf{v}_j \in \Re^m$, uma forma inicial para f_j e um valor inicial para $\mathbf{w}_j \in \Re^r$;

3. Utilizando **X** e **D**, resolva os seguintes problemas em sequência até convergência (medida por algum critério de parada):

3.1. Fixe f_j e obtenha um valor ótimo para $\mathbf{v}_j$;

3.2. Fixe $\mathbf{v}_j$, obtenha um f_j ótimo via técnicas de regularização e retorne ao passo **3.1**;

4. Obtenha um valor ótimo para $\mathbf{w}_j$ pelo método dos quadrados mínimos;

5. Para cada b tal que $1 \leq b < j$, calcule:

$$\mathbf{D} = \mathbf{S} - \sum_{\substack{k=1 \\ k \neq b}}^{j} \mathbf{w}_k \, f_k\left(\mathbf{v}_k^T \mathbf{X}\right),$$

e repita os passos 3 e 4, com $j = b$;

6. Por avaliação da participação de cada neurônio $\mathbf{w}_k f_k\left(\mathbf{v}_k^T \mathbf{X}\right)$, $k = 1, ..., j$, na representação da matriz **S**, aplique um procedimento de poda de neurônios que não apresentem um nível de participação mínima;

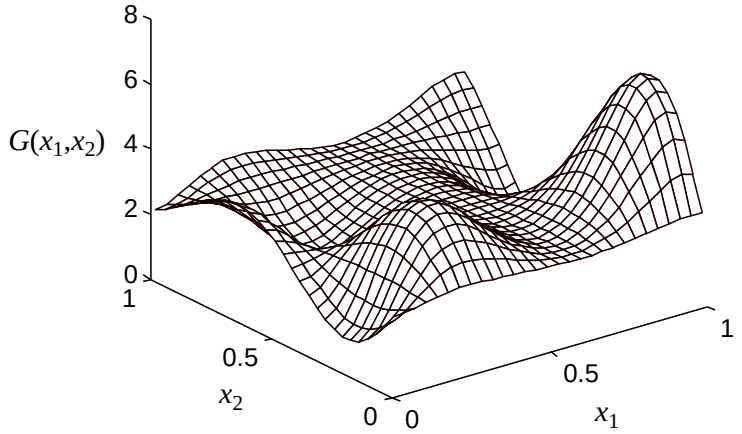
7. Enquanto um determinado nível de aproximação não for atingido (medido por algum critério de parada, por exemplo, utilizando validação cruzada), retorne ao passo 2.

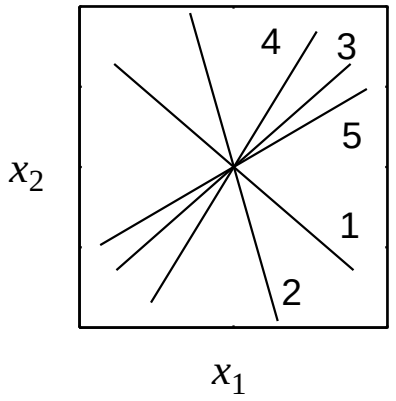
7. Outras abordagens construtivas

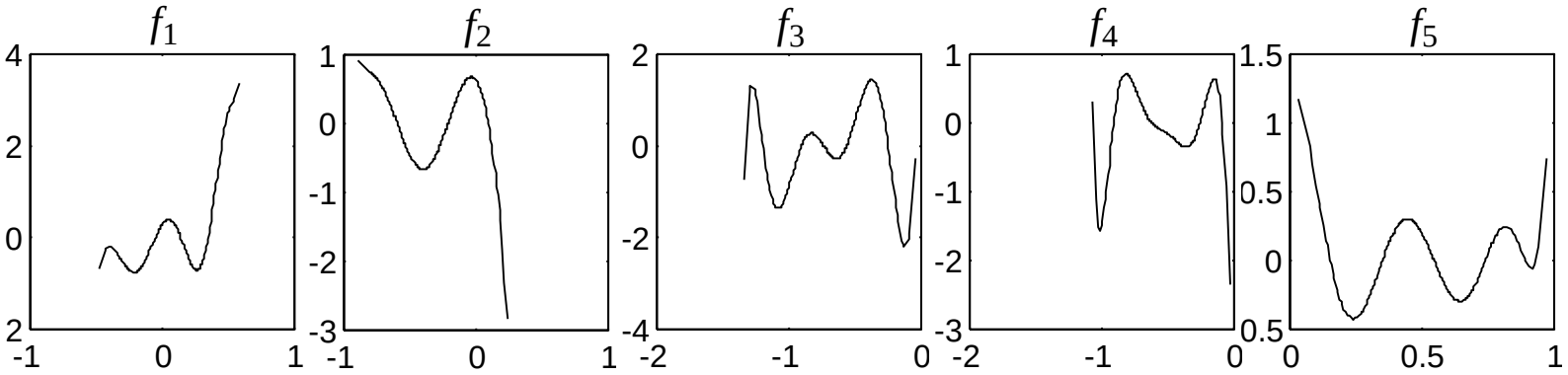
- Existem muitos outros métodos construtivos alternativos, como os propostos por FAHLMAN & LEBIERE (1990) (Cascade correlation), FREAN (1990) (Upstart), GALLANT (1993) (Tower / Pyramid), NABHAN & ZOMAYA (1994), PAREKH (1998) (Tiling / M-Tiling) e SANGER (1991).
- Uma visão mais recente encontra-se em FRANCO & JEREZ (2009).

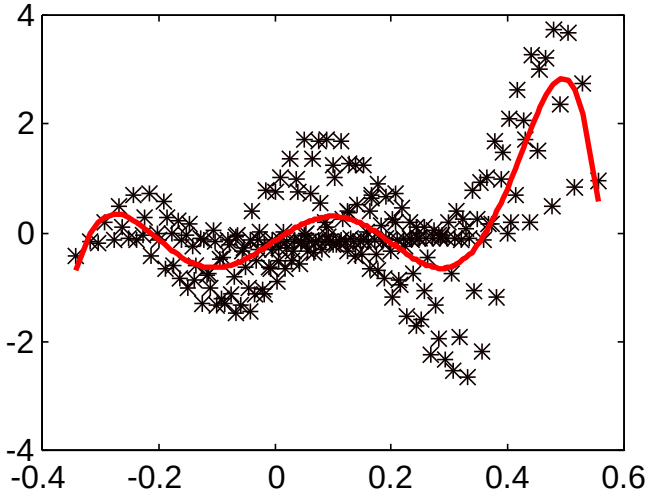
8. Exemplo de aplicação

$$G(x_1, x_2) = 1.9 \cdot \left[1.35 + e^{x_1 - x_2}\,\text{sen}\left(13(x_1 - 0.6)^2\right) \cdot \text{sen}(7x_2)\right]$$



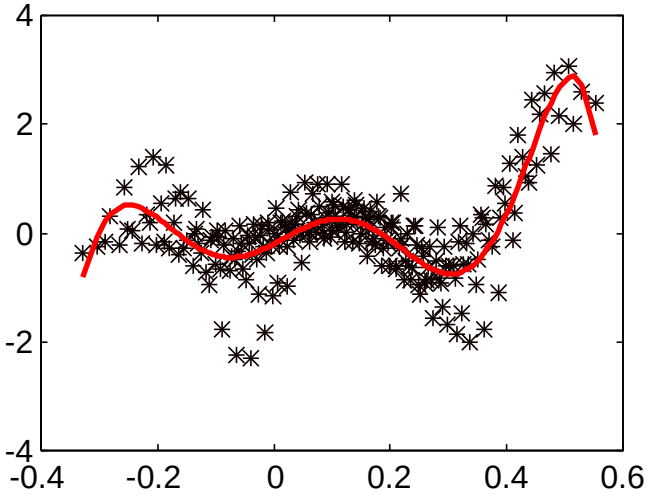


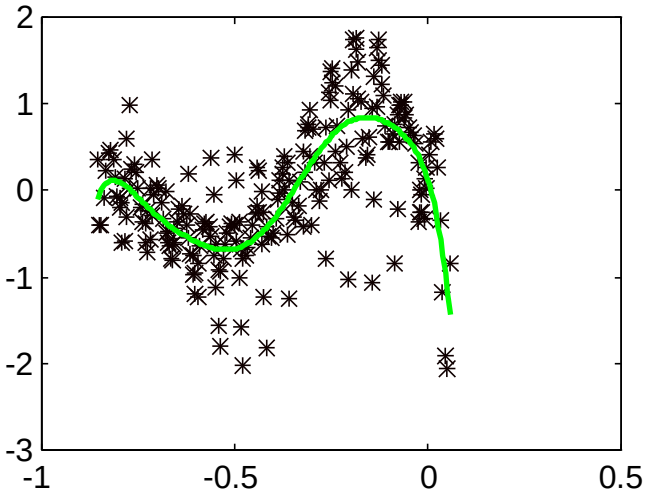




1 neurônio na

camada intermediária

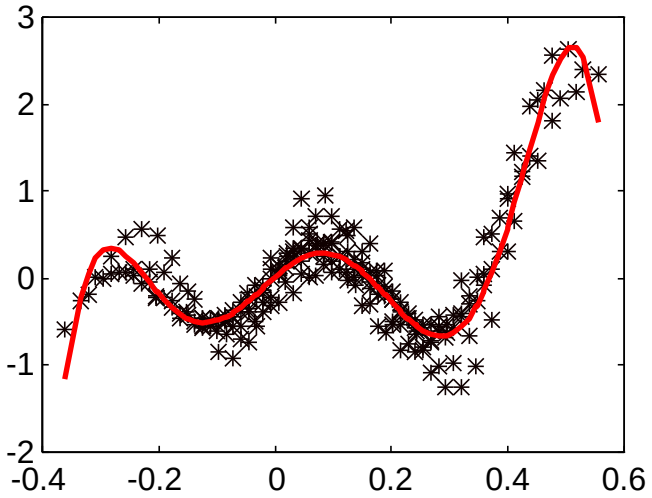


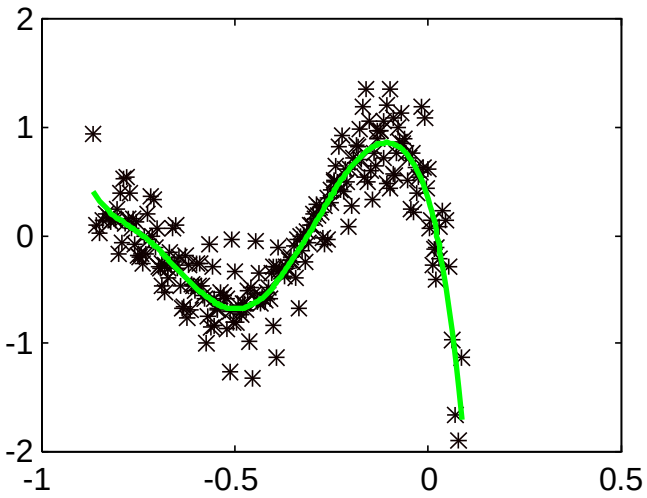


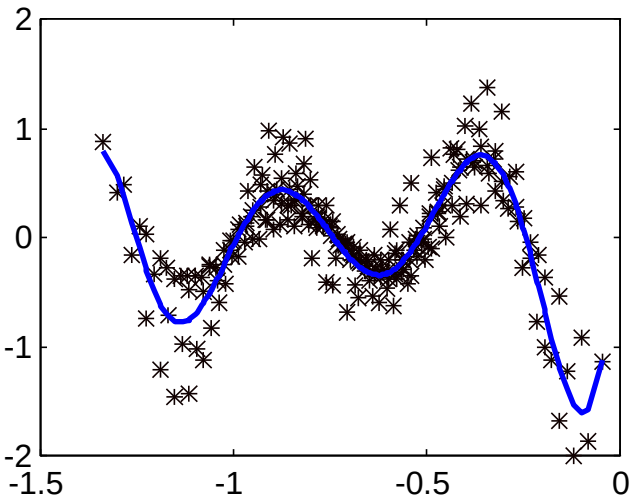
2 neurônios

na camada

intermediária







3 neurônios na camada intermediária

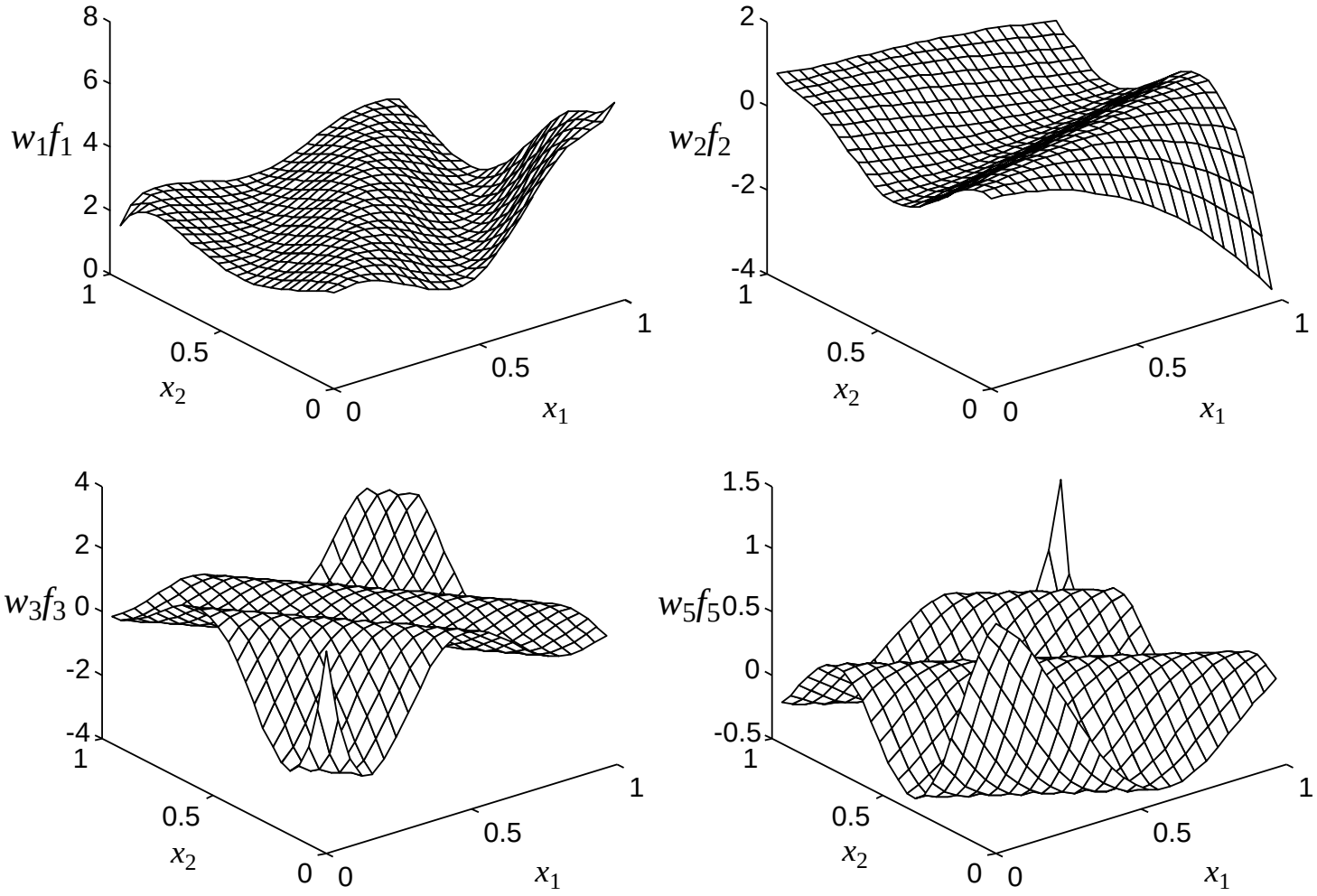


Figura 3 – Contribuição individual de 4 dos 5 neurônios da camada intermediária

9. Referências bibliográficas

ASH, T. Dynamic node creation in backpropagation neural networks. *Connection Science*, vol. 1, no. 4, pp. 365-375, 1989.

BÄRMANN, F., BIEGLER-KÖNIG, F. On a Class of Efficient Learning Algorithms for Neural Networks. *Neural Networks*, vol. 5, no. 1, pp. 139-144, 1992.

CRAVEN, P. & WAHBA, G. Smoothing Noisy Data with Spline Functions: Estimating the Correct Degree of Smoothing by the Method of Generalized Cross-Validation. *Numerische Mathematik*, vol. 31, Fasc. 4, pp. 377-403, 1979.

DAHMEN, W. & MICCHELLI, C.A. Some remarks on ridge functions. *Approximation Theory and its Applications*, 3(2-3): 139-143, 1987.

BREIMAN, L. & FRIEDMAN, J.H. Estimating Optimal Transformations for Multiple Regression and Correlation (with discussion). *Journal of the American Statistical Association (JASA)*, 80(391): 580-619, 1985.

BREIMAN. L., FRIEDMAN, J.H. Predicting multivariate responses in multiple linear regression. *Journal of the Royal Statistical Society B*, vol. 59, no. 1, pp. 3-54, 1997.

FAHLMAN, S.E., LEBIERE, C. The Cascade-Correlation Learning Architecture. *in* D.S. Touretzky (ed.) *Advances in Neural Information Processing Systems 2*, pp. 524-532, San Mateo, CA: Morgan Kaufmann Publishers, 1990.

FRANCO, L. & JEREZ, J.M. *Constructive Neural Networks*. Springer, 2009.

FREAN, M. The upstart algorithm: A method for constructing and training feedforward neural networks. *Neural Computation*, vol. 2, no. 2, pp. 198-209, 1990.

FRIEDMAN, J.H. An overview of predictive learning and function approximation. *in* J.H. Friedman, H. Wechsler (eds.) From Statistics to Neural Networks: Theory and Pattern Recognition Applications. *Proceedings of the NATO/ASI Workshop*, Springer-Verlag, 1994.

FRIEDMAN, J.H. SMART User's Guide. *Report LCM001*, Department of Statistics, Stanford University, 1984.

FRIEDMAN, J.H. & STUETZLE, W. Projection Pursuit Regression. *Journal of the American Statistical Association (JASA)*, 76(376): 817-823, 1981.

FRIEDMAN, J.H., STUETZLE, W. & SCHROEDER, A. Projection Pursuit Density Estimation. *Journal of the American Statistical Association (JASA)*, vol. 79, no. 387, pp. 599-608, 1984.

FRIEDMAN, J.H. & TUKEY, J. A Projection Pursuit Algorithm for Exploratory Data Analysis. *IEEE Transactions on Computers*, 23(9): 881-890, 1974.

GALLANT, S.I. Neural Network Learning and Expert Systems. A Bradford Book, 1993.

GHOSH, J. & TUMER, K. Structural adaptation and generalization in supervised feed-forward networks. *Journal of Artificial Neural Networks*, vol. 1, no. 4, pp. 431-458, 1994.

HALL, P. On Projection Pursuit Regression. *The Annals of Statistics*, 17(2): 573-588, 1989.

HASSIBI, B. & STORK, D.G. Second Order Derivatives for Network Pruning: Optimal Brain Surgeon. *in* S.J. Hanson, J.D. Cowan, C.L. Giles (eds.) *Advances in Neural Information Processing Systems 5*, pp. 164-171, San Mateo, CA: Morgan Kaufmann Publishers, 1993.

HIROSE, Y., YAMASHITA, K., HIJIYA, S. Back-propagation algorithm which varies the number of hidden units. *Neural Networks*, vol. 4, no. 1, pp. 61-66, 1991.

HUBER, P.J. Projection pursuit (with Discussion). *The Annals of Statistics*, 13(2): 435-475, 1985.

HWANG, J.-N., LAY, S.R., MAECHLER, M., MARTIN, R.D. & SCHIMERT, J. Regression modeling in back-propagation and projection pursuit learning. *IEEE Transactions on Neural Networks*, 5(3): 342-353, 1994.

INTRATOR, N. Combining Exploratory Projection Pursuit and Projection Pursuit Regression with Application to Neural Networks. *Neural Computation*, vol. 5, no. 3, pp. 443-455, 1993a.

INTRATOR, N. On the Use of Projection Pursuit Constraints for Training Neural Networks. *in* S.J. Hanson, J.D. Cowan, C.L. Giles (eds.) *Advances in Neural Information Processing Systems 5*, pp. 3-10, San Mateo, CA: Morgan Kaufmann Publishers, 1993b.

INTRATOR, N. On the combination of supervised and unsupervised learning. *Physica A*, vol. 200, nos. 1-4, pp. 655-661, 1993c.

JONES, L.K. On a conjecture of Huber concerning the convergence of projection pursuit regression. *The Annals of Statistics*, 15(2): 880-882, 1987.

JONES, M.C. & SIBSON, R. What is Projection Pursuit?. *Journal of the Royal Statistical Society A*, 150(1): 1-36, 1987.

KARNIN, E.D. A simple procedure for pruning back-propagation trained neural networks. IEEE Transactions on Neural Networks, vol. 1, no. 2, pp. 239-242, 1990.

KWOK, T.-Y. & YEUNG, D.-Y. Constructive Feedforward Neural Networks for Regression Problems: A Survey. *Technical Report HKUST-CS95-43*, Department of Computer Science, Hong Kong University of Science and Technology, Hong Kong, 1995.

LE CUN, Y., DENKER, J.S. & SOLLA, S.A. Optimal Brain Damage. *in* D.S. Touretzky (ed.) *Advances in Neural Information Processing Systems 2*, pp. 598-605, San Mateo, CA: Morgan Kaufmann Publishers, 1990.

MALTHOUSE, E.C. *Nonlinear Partial Least Squares*. Ph.D. Thesis, Northwestern University, Illinois, 1995.

MARDIA, K.V., KENT, J.T., BIBBY, J.M. *Multivariate Analysis*. London: Academic Press, 1979.

MOODY, J.E. Prediction risk and architecture selection for neural networks. *in* V. Cherkassky, J.H. Friedman, H. Wechsler (eds.) From Statistics to Neural Networks. *Proceedings of the NATO/ASI Workshop*, pp. 143-156, Springer-Verlag, 1994.

NABHAN, T.M., ZOMAYA, A.Y. Toward generating neural networks structures for function approximation. *Neural Networks*, vol. 7, no. 1, pp. 89-99, 1994.

PAREKH, R.G. Constructive learning: Inducing grammars and neural networks. Ph.D. Thesis, Iowa State University, 1998.

REED, R. Pruning algorithms - a survey. *IEEE Transactions on Neural Networks*, vol. 4, no. 5, pp. 740-747, 1993.

SANGER, T.D. A tree-structured adaptive network for function approximation in high-dimensional spaces. *IEEE Transactions on Neural Networks*, vol. 2, no. 2, pp. 245-256, 1991.

SKELTON, R.E. *Dynamic Systems Control*. New York: John Wiley & Sons, 1988.

STONE, C.J. Optimal Global Rates of Convergence for Nonparametric Regression. *The Annals of Statistics*, 10(4): 1040-1053, 1982.

VON ZUBEN, F.J. *Modelos paramétricos e não-paramétricos de redes neurais artificiais e aplicações*. Tese de Doutorado, Faculdade de Engenharia Elétrica e de Computação, Unicamp, 1996.

VON ZUBEN, F.J., NETTO, M.L.A. Aprendizado construtivo para redes neurais com uma camada intermediária. *Anais do 2º Simpósio Brasileiro de Automação Inteligente*, pp. 283-288, Curitiba, PR, 1995a.

VON ZUBEN, F.J., NETTO, M.L.A. Unit-growing learning optimizing the solvability condition for model-free regression. *Proceedings of the IEEE Int. Conference on Neural Networks*, vol. 2, pp. 795-800, 1995b.

WEIGEND, A.S., RUMELHART, D.E. & HUBERMAN, B.A. Generalization by Weight-Elimination with Application to Forecasting. *in* R.P. Lippmann, J.E. Moody, D.S. Touretzky (eds.) *Advances in Neural Information Processing Systems 3*, pp. 875-882, San Mateo, CA: Morgan Kaufmann Publishers, 1991.