Estimação de Dados

Prof. Guilherme Palermo Coelho

CET0621 - Aprendizado de Máquina na Análise de Dados

Roteiro

- Estimação de Dados:
  - Introdução;
  - Avaliação de Desempenho;
  - Métodos Paramétricos:
    - Regressão Linear;
    - Regressão Polinomial;
  - Métodos não-paramétricos;
  - Ensembles.

Nota: Este material está parcialmente baseado nas notas de aula do curso “Introdução à Mineração de Dados”, desenvolvidas pelo Prof. João Roberto Bertini Jr. (FT/UNICAMP).

Introdução

Introdução

- Como vimos nas aulas anteriores, em aprendizado supervisionado a tarefa de **predição** pode ser dividida em duas categorias:
  - **Classificação** de dados;
  - **Estimação (ou Regressão)** de dados;

Classificação:

- Consiste em predizer uma **classe** para uma amostra;
- **Classe**: conjunto finito de possibilidades.

ID	Idade	Nível da substância A no sangue	Nível da substância B no sangue	Histórico na Família	Está doente
1	Jovem	Alto	Baixo	Não	Não
2	Jovem	Alto	Alto	Não	Sim
3	Adulto	Alto	Baixo	Não	Não
4	Idoso	Médio	Baixo	Não	Não
5	Idoso	Baixo	Baixo	Sim	Sim
6	Idoso	Baixo	Alto	Sim	Não

Introdução

- Como vimos nas aulas anteriores, em aprendizado supervisionado a tarefa de **predição** pode ser dividida em duas categorias:
  - **Classificação** de dados;
  - **Estimação (ou Regressão)** de dados

Estimação/Regressão:

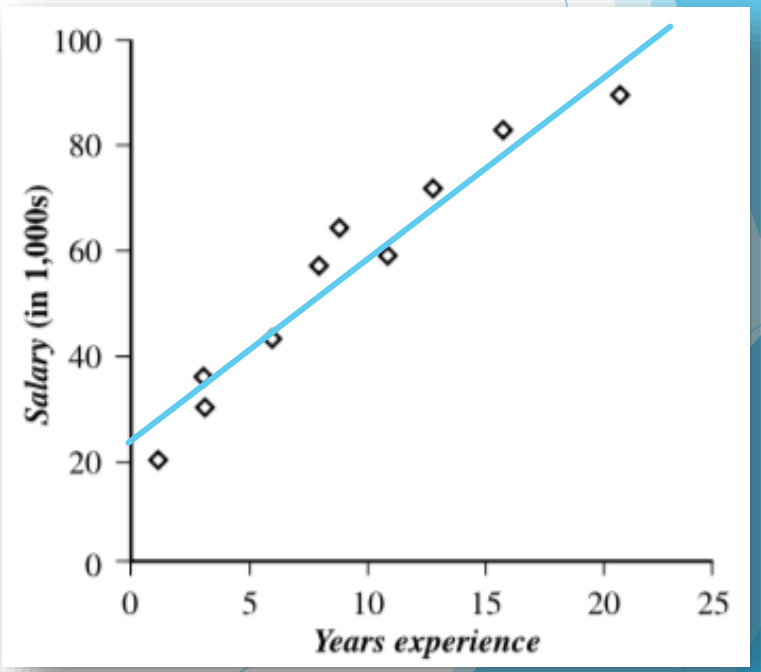
- Consiste em predizer uma **saída** para uma amostra;
- **Saída**: valor contínuo pertencente a um conjunto possivelmente infinito.

gender	*age*	*education*	*occupation*	*income*
female	23	college	teacher	$85,000
female	40	college	programmer	$50,000
female	31	college	programmer	$52,000
female	50	graduate	teacher	$90,000
female	62	graduate	CEO	$500,000
male	25	high school	programmer	$50,000
male	28	high school	CEO	$250,000
male	40	college	teacher	$80,000
male	50	college	programmer	$45,000
male	57	graduate	programmer	$80,000

Introdução

- **Ex.**: estimação de dados
  - Relação entre anos de experiência e salário

Anos de Experiência	Salário
2	21.000,00
4	30.000,00
10	60.000,00
15	70.000,00
...	...



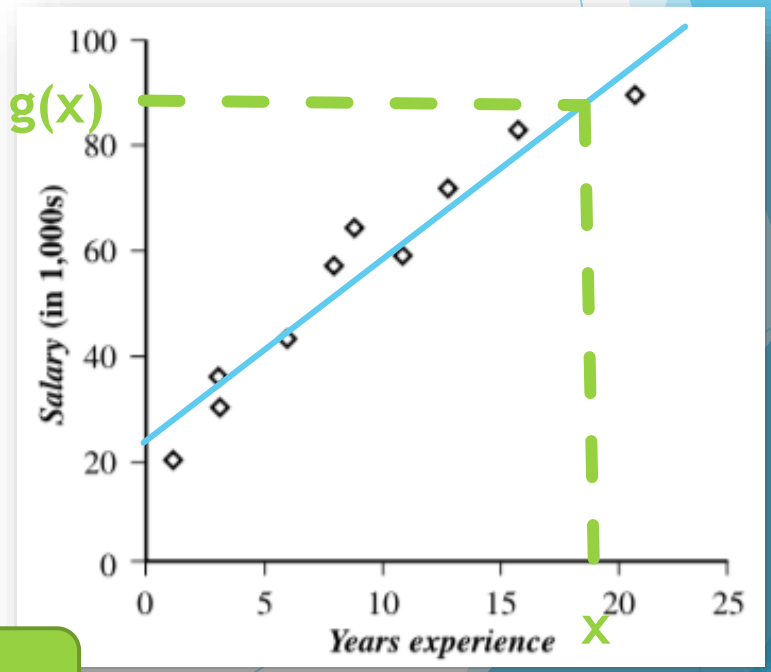
Introdução

- **Ex.**: estimação de dados
  - Relação entre anos de experiência e salário

Anos de Experiência	Salário
2	21.000,00
4	30.000,00
10	60.000,00
15	70.000,00
...	...

g(x) = ?

g(x) = 87.000,00



Atenção: nem sempre a relação entre as variáveis e a saída é linear!

Introdução

- A modelagem de um problema de **estimação** é similar à vista anteriormente para classificação de dados:
  - Apenas a saída que passa a ser um valor contínuo e não um rótulo.
- Ou seja, pode-se supor que o **mapeamento** entre os vetores de entrada $\boldsymbol{x}_i$ e as saídas esperadas y_i seja feito por uma função $g(\cdot)$ desconhecida, tal que:

$$y_i = g(\boldsymbol{x}_i) + \varepsilon_i$$

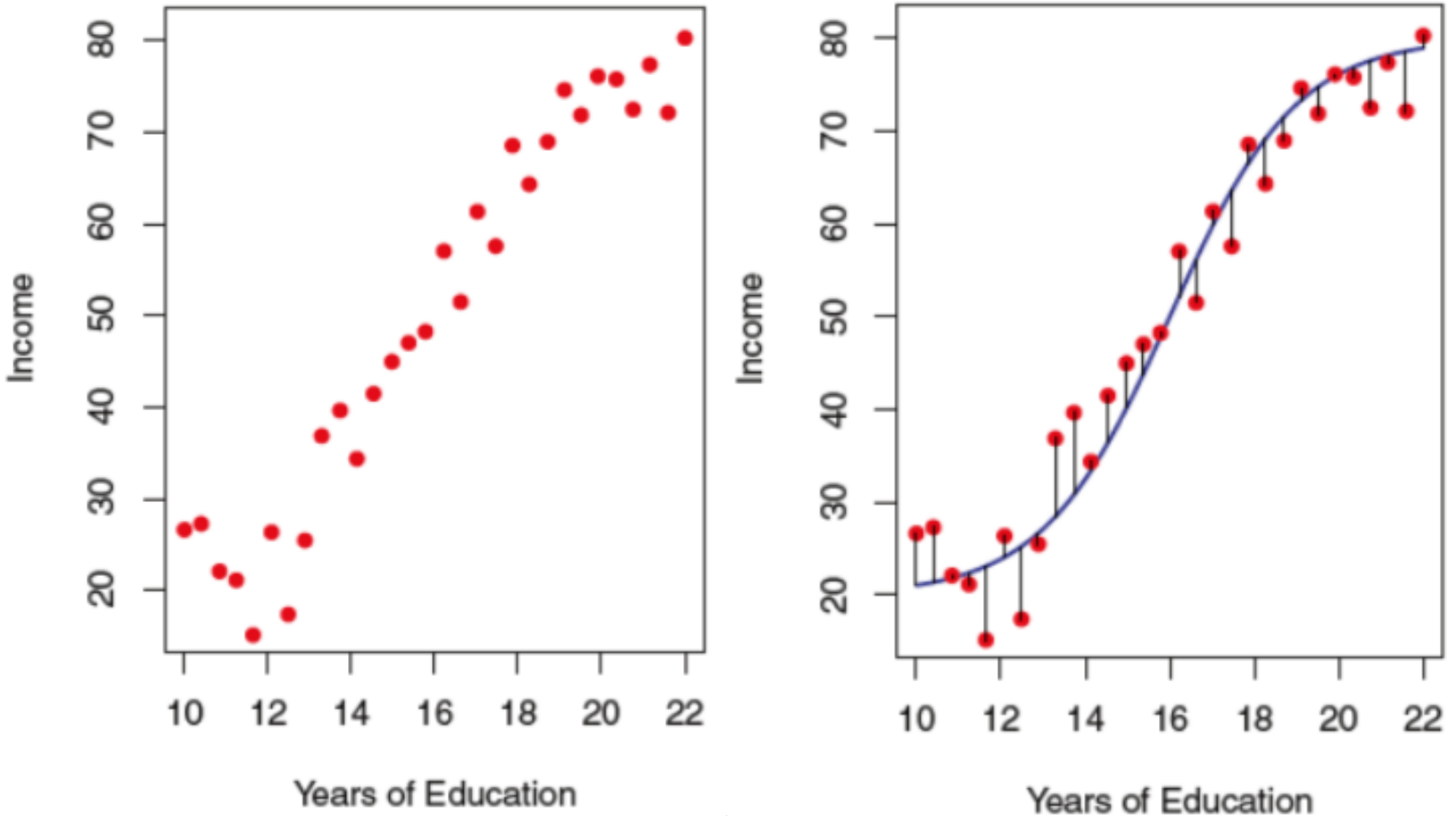
  onde ε_i é o erro intrínseco do processo de amostragem.
- Deseja-se então ajustar os parâmetros $\boldsymbol{\theta} \in \mathbb{R}^{\boldsymbol{p}}$ de um modelo, de forma a aproximar, da melhor maneira possível, uma função $\hat{g}(\boldsymbol{x}_i, \boldsymbol{\theta})$ de $g(\boldsymbol{x}_i)$;
  - Isto é feito a partir de um conjunto de dados $\{(\boldsymbol{x}_i, y_i)\}_{i=1,\cdots,n}$ com n amostras.

Os parâmetros são ajustados de forma a minimizar algum **erro de estimação.**

Introdução

- **Exemplo:** estimar o salário, y, em função dos anos de estudo x.

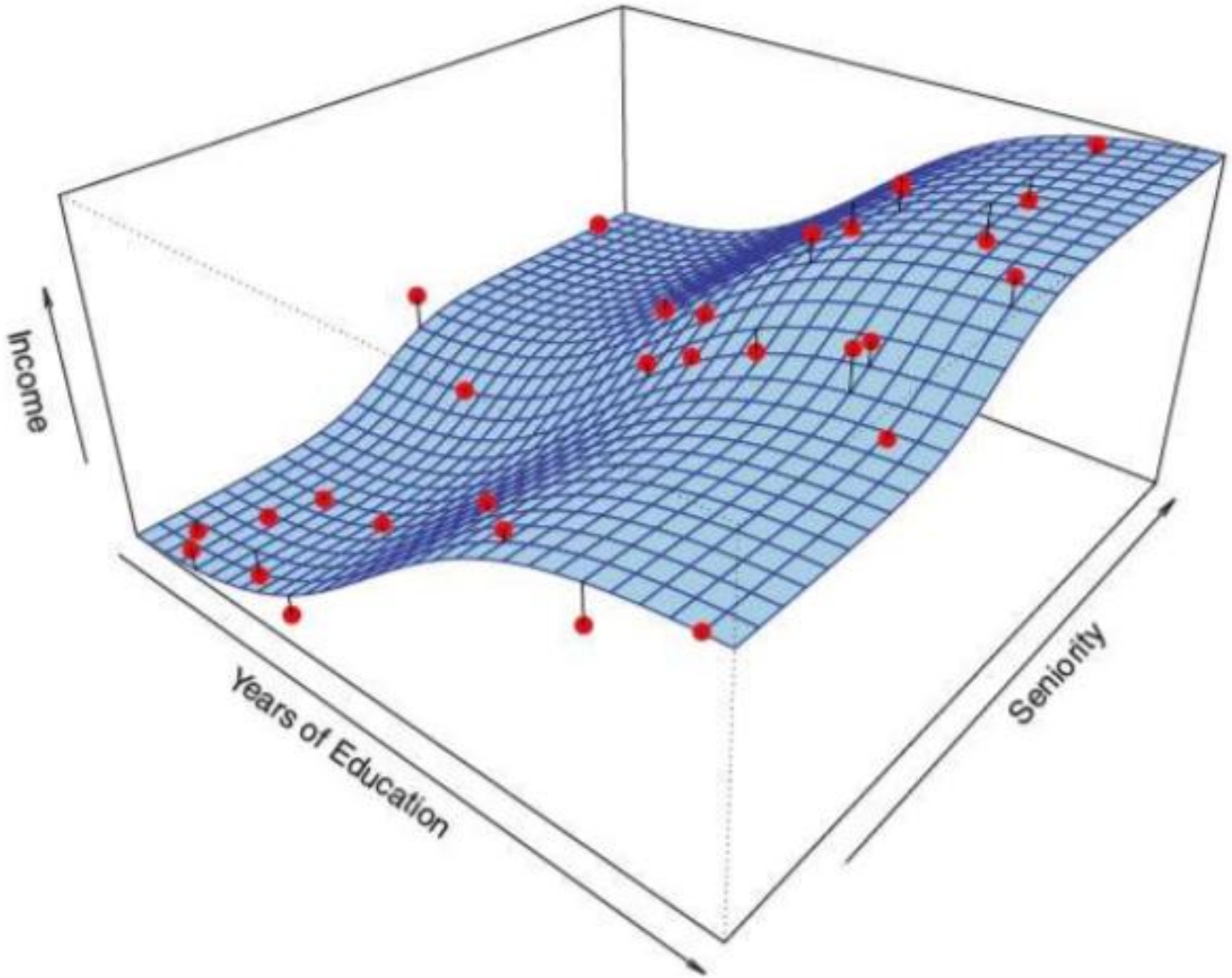
$$y = g(x) + \varepsilon$$



Introdução

- Em geral, a função envolve mais de uma variável;
- **Exemplo**: estimar o salário, y, em função dos anos de estudo x_1 e nível hierárquico na empresa x_2.

$$y = g(x_1, x_2) + \varepsilon$$



Introdução

- Como estimar $\hat{g}(\boldsymbol{x}_i, \boldsymbol{\theta})$ a partir de um conjunto de dados $\{(\boldsymbol{x}_i, y_i)\}_{i=1,\cdots,n}$?
  - **Métodos Paramétricos**: envolvem uma abordagem em duas etapas
    1. Faz-se uma suposição a respeito da **forma** de $g(\cdot)$;

       Exemplo: $g(\cdot)$ é linear

       $$g(\boldsymbol{x}) = \theta_0 + \theta_1 x_1 + \cdots + \theta_p x_p$$

    2. Adota-se um procedimento de ajuste dos parâmetros θ_i:

       $$\hat{g}(\boldsymbol{x}) \cong g(\boldsymbol{x})$$

  - Desvantagem dos métodos paramétricos:
    - Geralmente não conhecemos a **forma** exata de $g(\cdot)$;
    - Uma escolha errada pode levar a erros de estimação (efeito *bias*).

Introdução

- Como estimar $\hat{g}(\boldsymbol{x}_i, \boldsymbol{\theta})$ a partir de um conjunto de dados $\{(\boldsymbol{x}_i, y_i)\}_{i=1,\cdots,n}$?
  - **Métodos Não-Paramétricos**: não exigem suposições a respeito da **forma** de $g(\cdot)$;
    - Procuram uma estimativa que represente bem **os dados de treinamento**;
    - **Ex.**:redes neurais do tipo MLP.
  - **Vantagem**: ao evitar a suposição inicial de forma, têm o potencial de realizarem um ajuste de uma gama maior de formas possíveis para $g(\cdot)$.
  - **Desvantagem:** exigem um número maior de amostras para conseguir estimar $\hat{g}(\cdot)$.
    - A suposição de forma reduz (simplifica) o problema.

Avaliação de Desempenho

Avaliação de Desempenho

- A saída do estimador é um valor **contínuo** que deve ser o mais próximo possível do valor desejado.
  - A diferença entre os valores estimado e desejado fornece uma avaliação do **erro de estimação**;
    - Para uma amostra i dos dados:

$$e_i = \hat{g}(\boldsymbol{x}_i, \boldsymbol{\theta}) - y_i$$

  - O processo de treinamento busca corrigir este erro, minimizando uma **função objetivo** baseada em e_i.
  - Existem várias medidas que podem ser usadas para avaliar o erro de estimação:
    - Elas avaliam aspectos diferentes dos resultados de estimação;
    - Qual é a mais indicada depende de um estudo da aplicação.

Avaliação de Desempenho

- Exemplos de medidas de desempenho (onde n é o número de amostras de treinamento):
  - Erro quadrático médio (EQM):

$$EQM = \frac{1}{n}\sum_{j=1}^{n} e_j^2$$

  - Raiz do erro quadrático médio (REQM):

$$REQM = \sqrt{\frac{1}{n}\sum_{j=1}^{n} e_j^2}$$

  - Estas duas medidas tendem a amplificar grandes discrepâncias entre a saída e o valor esperado.

Avaliação de Desempenho

- Exemplos de medidas de desempenho (onde n é o número de amostras de treinamento e y_j a saída esperada para a amostra j):
  - Erro quadrático relativo (EQR):

$$EQR = \frac{1}{n}\sum_{j=1}^{n}\frac{e_j^2}{\left(y_j - \bar{y}\right)^2}, \qquad \bar{y} = \frac{1}{n}\sum_{j=1}^{n} y_j$$

  - Raiz do erro quadrático relativo (REQR):

$$REQR = \sqrt{\frac{1}{n}\sum_{j=1}^{n}\frac{e_j^2}{\left(y_j - \bar{y}\right)^2}}, \qquad \bar{y} = \frac{1}{n}\sum_{j=1}^{n} y_j$$

  - Normalizam o erro pelo erro de um estimador simples (que retorna a média).
    - Menor impacto de “erros grandes”.

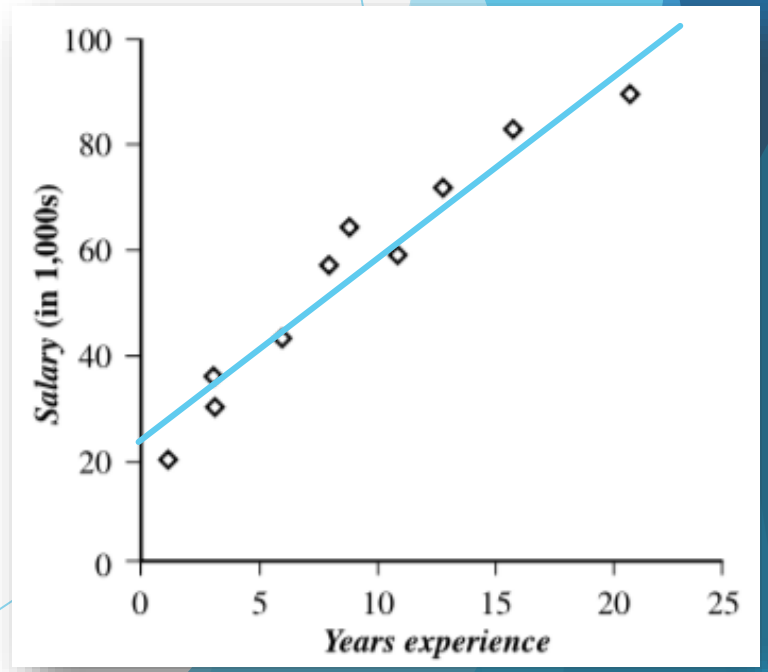
Métodos Paramétricos

Regressão Linear

- **Regressão** pode ser definida como o problema de estimar uma função a partir de pares entrada-saída:
  - **Saída**: variável dependente;
  - **Entrada**: variáveis independentes (os atributos dos dados, no caso desta disciplina);
- Na **Regressão Linear**, supõe-se que a relação entre entradas e saídas é linear (pode ser representada por uma **reta,** caso tenhamos uma única entrada);
  - Supondo que existem p atributos:

$$g(\boldsymbol{x}) = \theta_0 + \theta_1 x_1 + \cdots + \theta_p x_p + \varepsilon$$

  - A partir dos dados, devemos estimar os parâmetros $\boldsymbol{\theta}$.
    - Método paramétrico!



Regressão Linear

- Supondo que nosso conjunto de dados seja formado por um conjunto de n amostras rotuladas, tal que $\{(\boldsymbol{x}_i, y_i)\}_{i=1,\cdots,n}$:
  - $\boldsymbol{x}_i$ é o vetor de p atributos da i-ésima amostra dos dados;
  - y_i é a saída rotulada da i-ésima amostra dos dados;
  - A relação entre entradas e saídas será:

$$g(\boldsymbol{x}_i) = y_i \ = \theta_0 + \theta_1 x_{i1} + \cdots + \theta_p x_{ip} + \varepsilon_i$$

onde ε_i é o erro de amostragem da amostra i

- Nosso objetivo é obter $\hat{g}(\boldsymbol{x}_i, \boldsymbol{\beta})$ que melhor se aproxime de $g(\boldsymbol{x}_i)$:

$$\hat{g}(\boldsymbol{x}_i, \boldsymbol{\beta}) = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip} = \boldsymbol{v}_i^T \boldsymbol{\beta}$$

$$\boldsymbol{v}_i^T = [1 \quad x_{i1} \quad \cdots \quad x_{ip}]$$

$$\boldsymbol{\beta} = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_p \end{bmatrix}$$

Regressão Linear

- Uma das formas mais conhecidas de se obter o conjunto de parâmetros $\boldsymbol{\beta}$ é através do método dos *mínimos quadrados* (*least squares*);
  - Busca minimizar a **soma dos erros quadráticos** (SEQ) entre as saídas estimadas e esperadas:

$$SEQ(\boldsymbol{\beta}) = \sum_{j=1}^{n} \left(y_j - \boldsymbol{v}_j^T \boldsymbol{\beta}\right)^2$$

  - Em notação matricial:

$$SEQ(\boldsymbol{\beta}) = (\boldsymbol{y} - \mathbf{X}\boldsymbol{\beta})^T (\boldsymbol{y} - \mathbf{X}\boldsymbol{\beta})$$

onde $\mathbf{X}$ é a matriz de dados, acrescida de uma primeira coluna com valores 1, e $\boldsymbol{y}$ é o vetor com as saídas esperadas para cada uma das n amostras.

$$\mathbf{X} = \begin{bmatrix} 1 & x_{11} & \cdots & x_{1p} \\ 1 & x_{21} & \cdots & x_{2p} \\ \vdots & \cdots & \cdots & \vdots \\ 1 & x_{n1} & \cdots & x_{np} \end{bmatrix}$$

Regressão Linear

- **Exemplo**: prever o valor de fechamento de uma ação na bolsa, em um dia t, a partir dos valores de fechamento em dois dias anteriores (t-1 e t-2);

Valor em t-2	Valor em t-1	Valor em t
8,33	8,75	8,22
8,41	8,33	8,75
8,43	8,41	8,33
8,51	8,43	8,41

Vetor y

Matriz X

1	8,33	8,75
1	8,41	8,33
1	8,43	8,41
1	8,51	8,43

- Modelo linear a ser ajustado:

$$\hat{g}(\boldsymbol{x}_i, \boldsymbol{\beta}) = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2}$$

x_{i1}= valor da ação em t-2 (amostra i)

x_{i2}= valor da ação em t-1 (amostra i)

Regressão Linear

- O objetivo do método dos *mínimos quadrados* é encontrar $\boldsymbol{\beta}^*$ que minimize a SEQ;
  - Para isso, basta derivar a equação abaixo, em relação a $\boldsymbol{\beta}$ e igualar a 0:

$$SEQ(\boldsymbol{\beta}) = (\boldsymbol{y} - \mathbf{X}\boldsymbol{\beta})^T(\boldsymbol{y} - \mathbf{X}\boldsymbol{\beta})$$

  - Derivando:

$$\mathbf{X}^T(\boldsymbol{y} - \mathbf{X}\boldsymbol{\beta}^*) = \mathbf{0}$$

  - Resolvendo para $\boldsymbol{\beta}^*$ e supondo que as matrizes são não-singulares (inversíveis):

$$\boldsymbol{\beta}^* = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\boldsymbol{y}$$

  - Tendo $\boldsymbol{\beta}^*$ pode-se obter a saída do regressor para qualquer amostra $\boldsymbol{x}$ de dados:

$$\boldsymbol{s} = \boldsymbol{v}^T\boldsymbol{\beta}^*$$ (veja como obter $\boldsymbol{v}$ no Slide 19)

Regressão Polinomial

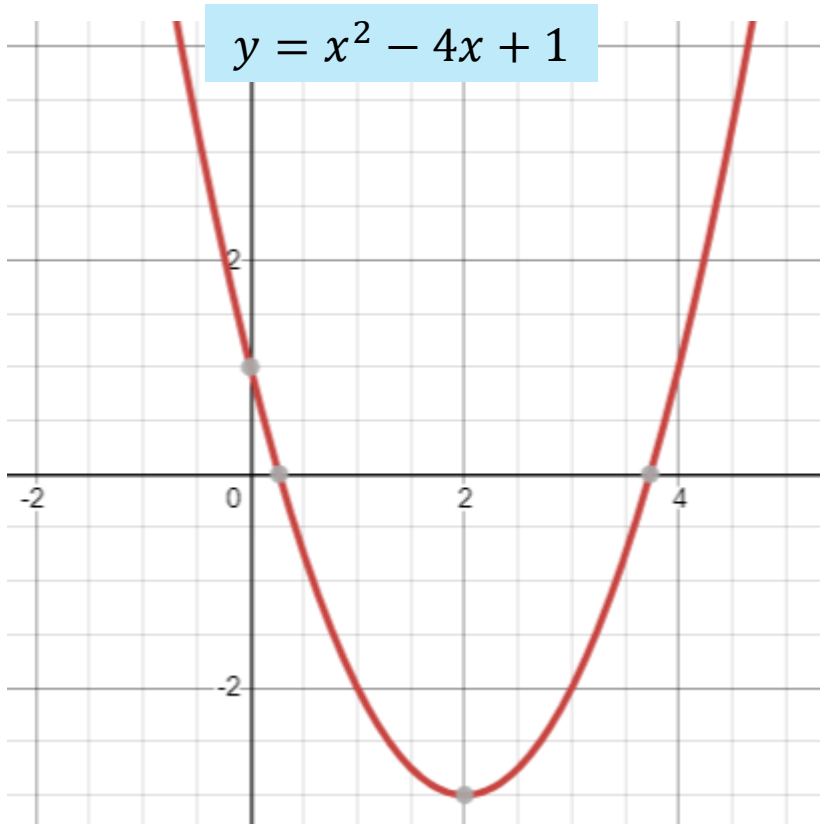
- Apesar da regressão linear ser simples, nem sempre a relação entre as variáveis independentes (atributos dos dados) e a variável dependente é linear;
- É preciso supor uma relação mais “complexa” entre as variáveis:
  - **Possibilidade simples**: supor que esta relação tem a forma de um **polinômio**
  - **Exemplo**: uma única variável independente x e um polinômio de grau k:

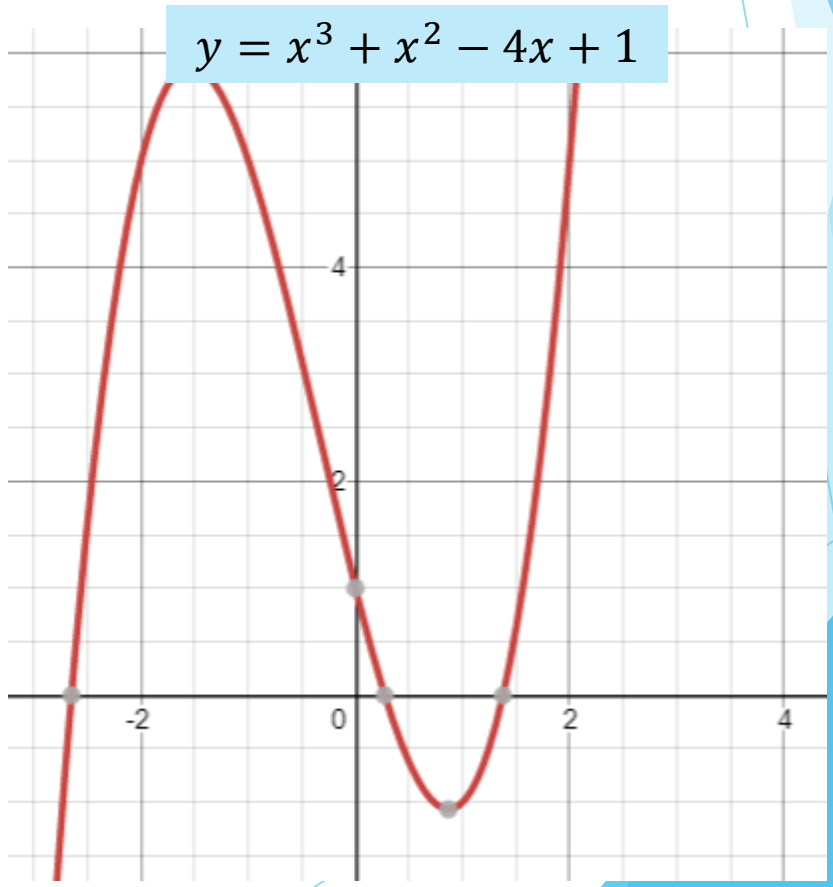
$$g(x_j) = \alpha_0 + \alpha_1 x_j + \alpha_2 x_j^2 + \alpha_3 x_j^3 + \cdots + \alpha_k x_j^k + \varepsilon$$

  - **Grau do polinômio**: maior expoente da variável independente;
  - **Coeficientes do polinômio**: termos que multiplicam a variável independente (α_i).

Regressão Polinomial

- Quanto maior o grau do polinômio, maior a "*flexibilidade*" da curva para se ajustar aos dados;

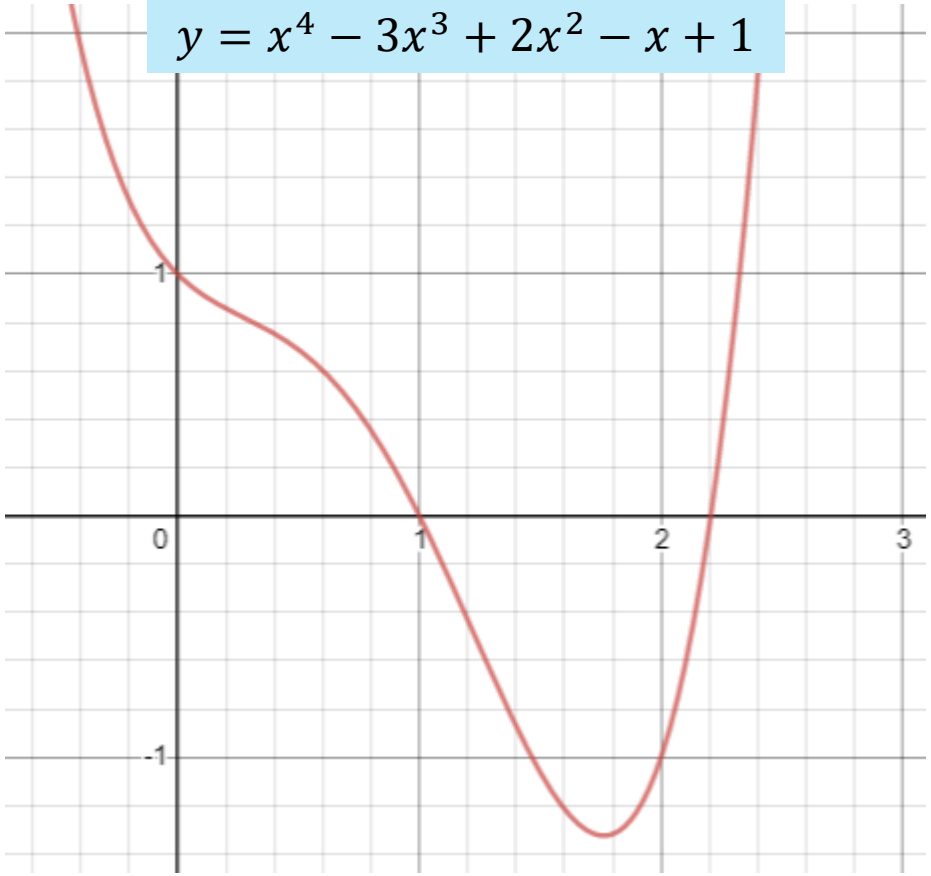


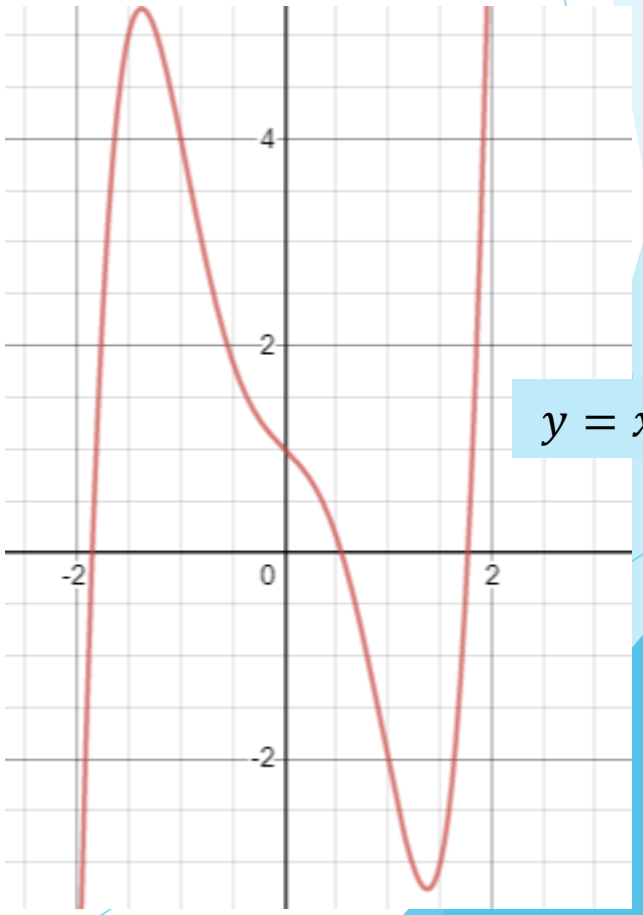


Regressão Polinomial

- Quanto maior o grau do polinômio, maior a "*flexibilidade*" da curva para se ajustar aos dados;

$y = x^4 - 3x^3 + 2x^2 - x + 1$





$y = x^5 - 3x^3 - x + 1$

Regressão Polinomial

- Para obter os parâmetros α_i do modelo polinomial, podemos fazer a seguinte substituição de variáveis:

$$x_j = z_1$$
$$x_j^2 = z_2$$
$$x_j^3 = z_3$$
$$\dots$$
$$x_j^k = z_k$$

- Dessa forma, a equação do modelo polinomial se torna:

$$g(x_j) = \alpha_0 + \alpha_1 x_j + \alpha_2 x_j^2 + \alpha_3 x_j^3 + \cdots + \alpha_k x_j^k + \varepsilon$$

$$g(x_j) = \alpha_0 + \alpha_1 z_1 + \alpha_2 z_2 + \alpha_3 z_3 + \cdots + \alpha_k z_k + \varepsilon$$

- **Mesma equação do Modelo Linear!**
- **Pode ser resolvido da mesma forma.**

Regressão Polinomial

- Embora a relação entre as variáveis independentes e dependente do problema seja não-linear, o problema de **estimação dos parâmetros** do modelo é linear;
  - A mesma estratégia de regressão linear (*mínimos quadrados*) pode ser utilizada;

- Na discussão anterior, a modelagem para regressão polinomial foi feita considerando apenas uma variável independente:

$$g(x_j) = \alpha_0 + \alpha_1 x_j + \alpha_2 x_j^2 + \alpha_3 x_j^3 + \cdots + \alpha_k x_j^k + \varepsilon$$

- No entanto, a ampliação para múltiplas variáveis independentes é análoga, inclusive com a etapa de substituição de variáveis.

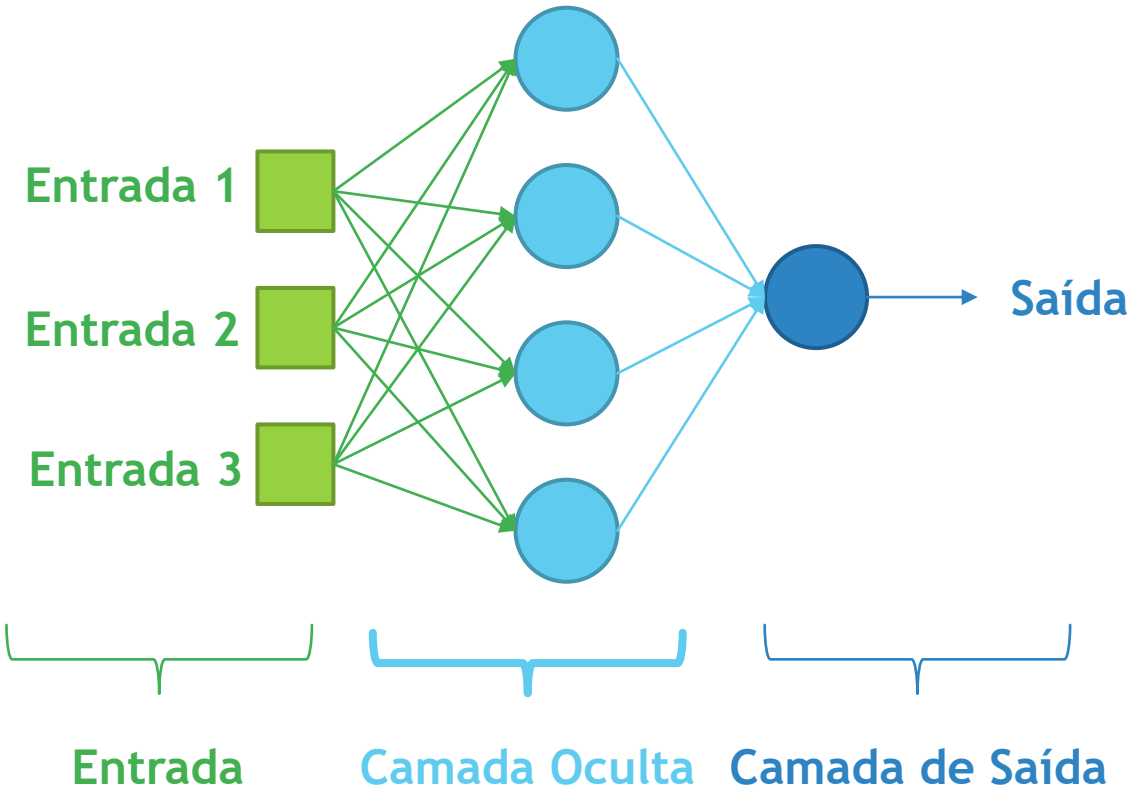
Métodos Não-Paramétricos

Métodos Não-paramétricos

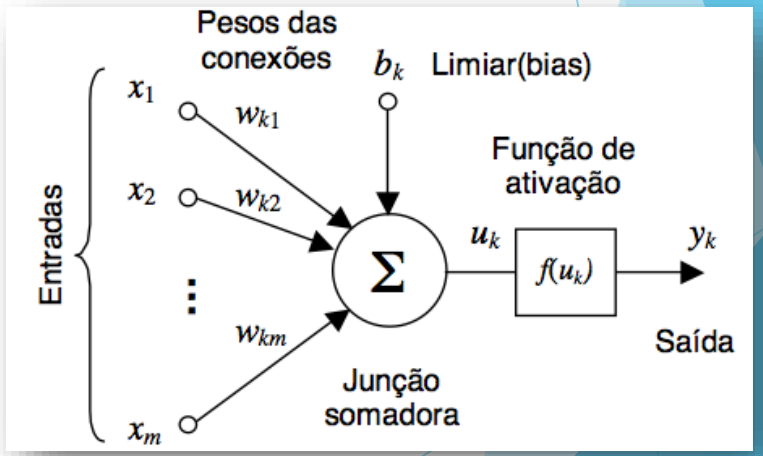
- Como visto antes, **métodos não-paramétricos** não exigem suposições a respeito da **forma** de $g(\cdot)$;
  - Ou seja, não exigem muito conhecimento prévio sobre o problema;
    - Necessário que a suposição em um método paramétrico seja feita corretamente;
- **Vantagem**: maior flexibilidade para ajustar a função aos dados;
- **Desvantagem**: exigem mais dados para um bom ajuste do modelo.
- Exemplos de métodos não paramétricos:
  - Redes Neurais MLP (já vistas);
  - Árvores de Regressão (*Classification and Regression Trees* – CART – não serão tratadas aqui);
  - Máquinas de Vetores Suporte (SVM - não serão tratadas aqui);
  - ...

Métodos Não-paramétricos

- Perceptrons multicamadas como estimadores:



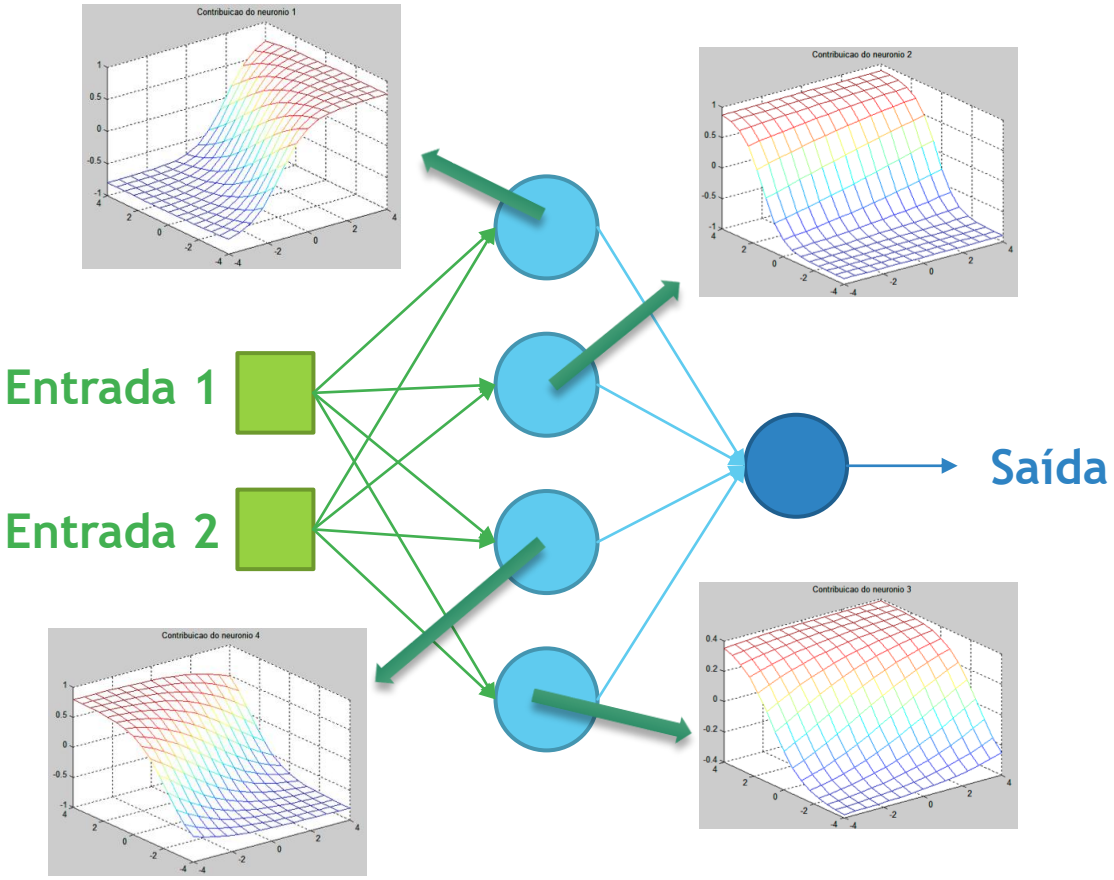
Cada "bolinha" corresponde a um *perceptron*.



A saída é um valor *contínuo* → estimação!

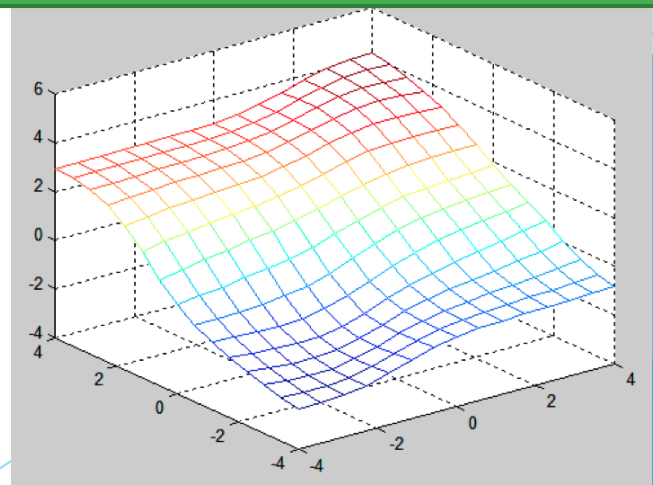
Métodos Não-paramétricos

- Perceptrons multicamadas como estimadores:



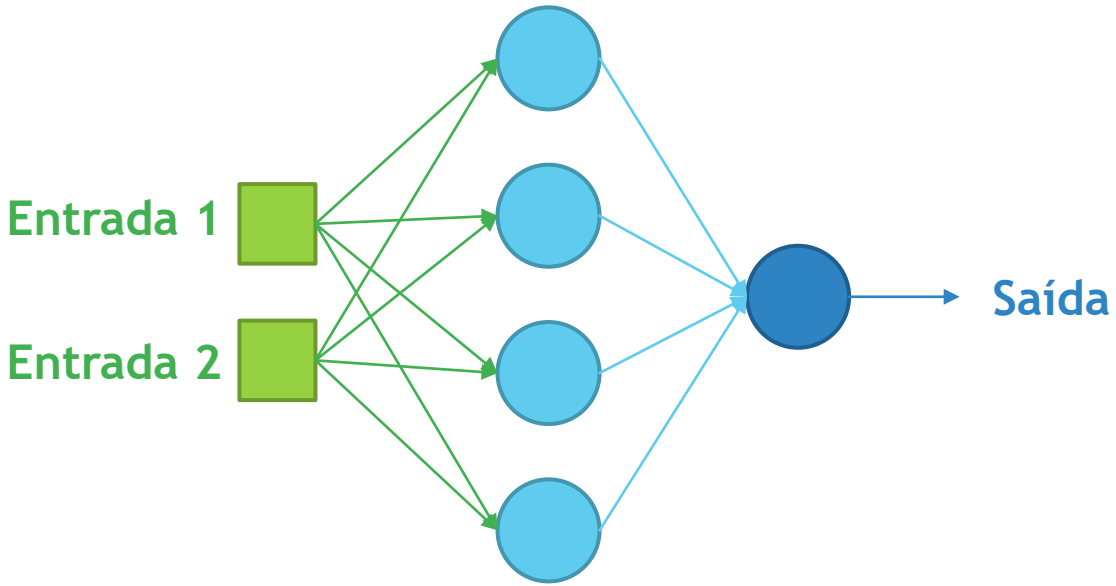
Cada neurônio, ao ter seus pesos ajustados, contribui para a saída da MLP com uma *função de expansão ortogonal*

Saída da MLP:



Métodos Não-paramétricos

- Perceptrons multicamadas como estimadores:



Em nenhum momento fizemos suposições a respeito do *formato* da função a ser aproximada.

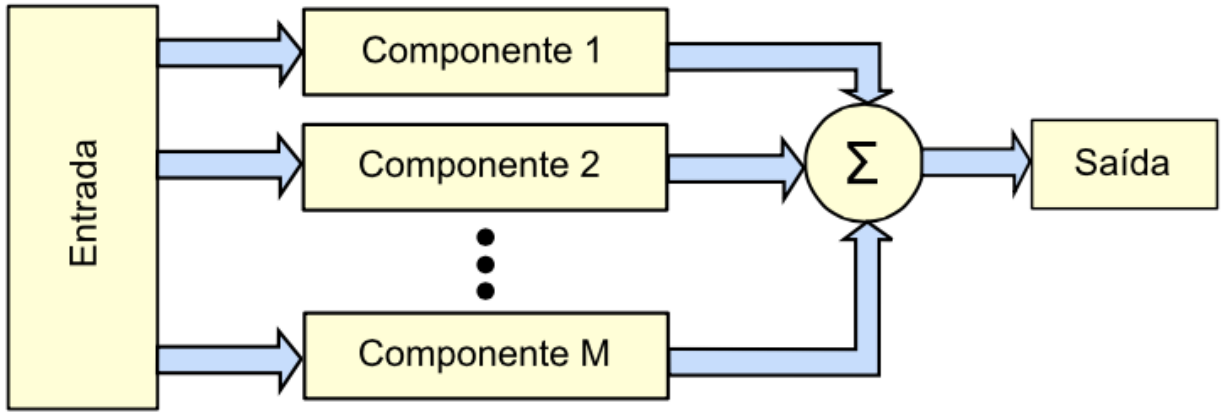
Existem decisões que podem afetar significativamente a qualidade da aproximação (num. neurônios e camadas).

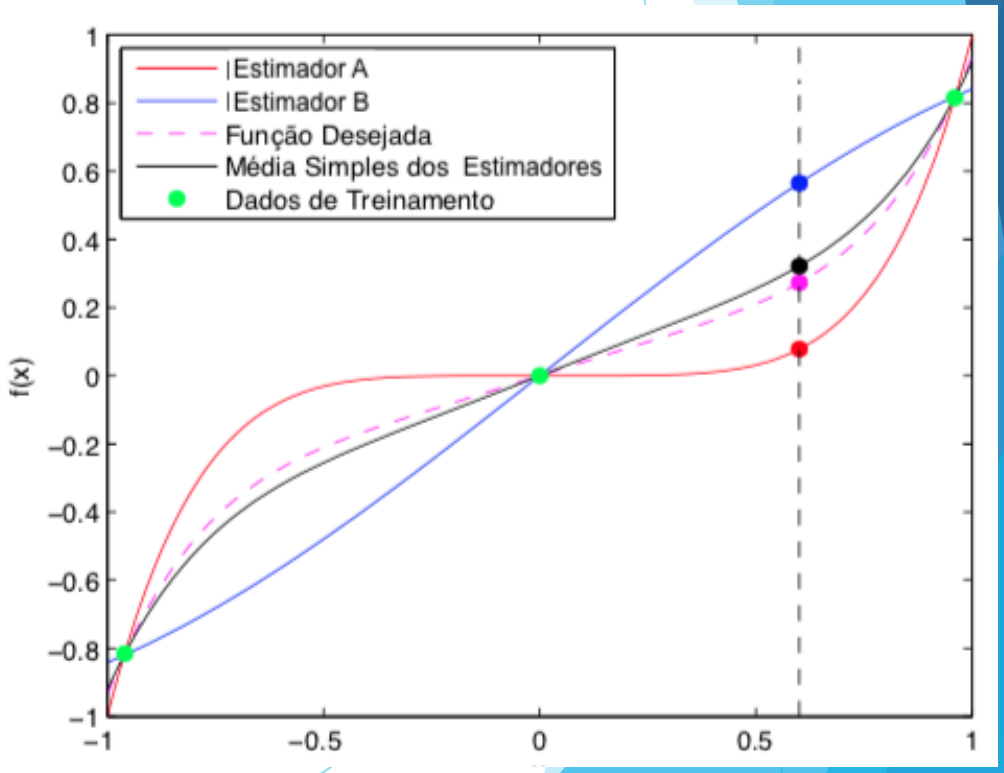
O número de parâmetros do modelo, a serem ajustados durante o treinamento, pode ser significativo.

Ensembles

Ensembles

- Não se esqueçam que *ensembles* também podem ser utilizados para estimação:
  - Basta apenas ajustar a forma de combinação das saídas.





É possível ter componentes paramétricos e não-paramétricos em um *ensemble*!

Referências Bibliográficas

Referências Bibliográficas

de Castro, L. N. & Ferrari, D. G. *Introdução à Mineração de Dados – Conceitos Básicos, Algoritmos e Aplicações*. Ed. Saraiva, 2016.

Han, J. & Kamber, M. *Data Mining: Concepts and Techniques*, Elsevier, 2006.

Haykin, S. *Neural Networks: A Comprehensive Foundation*, 2nd Ed., Prentice-Hall, 1999.

Witten, I. H., Frank E. & Hall, M. A. *Data Mining: Practical Machine Learning Tools and Techniques*, Elsevier, 2011.

Coelho, G. P. *Geração, Seleção e Combinação de Componentes para Ensembles de Redes Neurais Aplicadas a Problemas de Classificação*. Dissertação de Mestrado, Faculdade de Engenharia Elétrica e de Computação (FEEC), Unicamp, 2006.