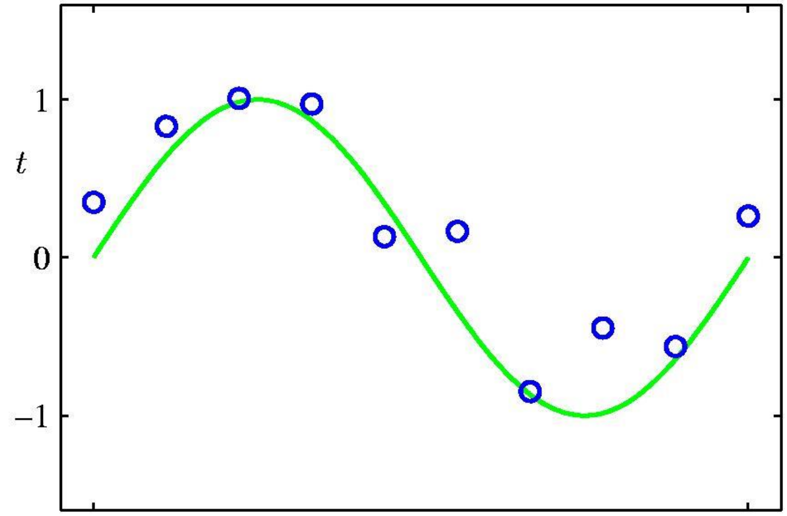
Sistemas Inteligentes

REGRESSÃO

O que é Regressão?

Técnica para contrução de models que caracterizem relações entre uma variável dependente, y, e uma ou mais variáveis independents, x_1, x_2, ….



$$y(x, \mathbf{w}) = w_0 + w_1 x + w_2 x^2 + \ldots + w_M x^M = \sum_{j=0}^{M} w_j x^j$$

Exemplo

Considere o cálculo da media final dos alunos em uma disciplina da universidade.

- A média depende de 6 notas de exercícios, duas provas e um projeto final

Poderíamos perguntar:

- Esqueci quais eram os pesos de cada nota na média final do curso. Será que conseguiria estimar isso a partir da planilha com as notas dos alunos e a média final?
- Perdi a nota da prova final dos alunos. Será que conseguiria estimar qual seria a média final dos alunos a partir das outras notas?
- Qual o nível de importância de cada componente? Será que eu conseguiria predizer se um aluno irá bem na disciplina apenas baseado nos exercícios? Ou apenas baseado nas notas das provas?

Regressão Linear

De maneira geral, a regressão assume que y é uma função das variáveis independentes $x_1, x_2, \ldots$, que podem ser agrupadas em um vetor $\mathbf{x} = [x_1, x_2, \ldots]^T$. A forma da função é definida por um conjunto de parâmetros, geralmente expressos por um vetor de parâmetros $\mathbf{w}$, ou seja

$$y = f(\mathbf{x}, \mathbf{w})$$

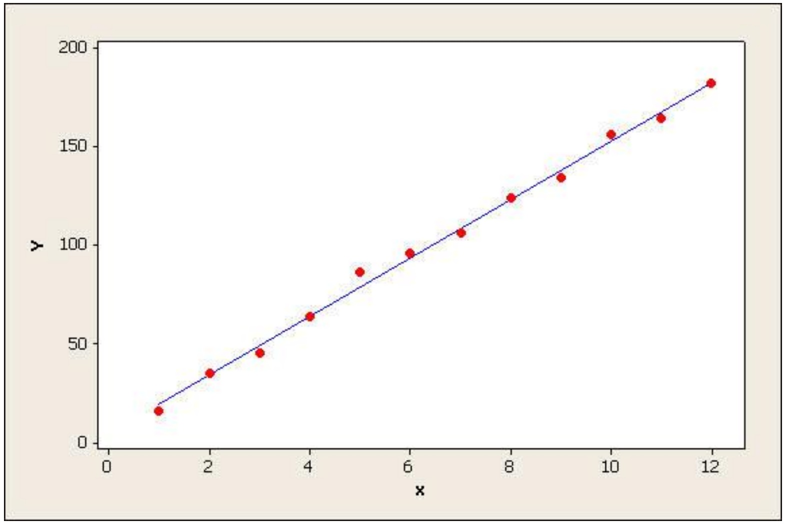
No caso particular da regressão linear

$$y = \mathbf{w}^T\mathbf{x} = [w_0 \quad w_1 \quad \cdots \quad w_N] \begin{bmatrix} 1 \\ x_1 \\ \vdots \\ x_N \end{bmatrix}$$

Quando há apenas uma variável independente, obtemos a equação de uma reta

$$y = w_1 x_1 + w_0$$

Como ajustar os parâmetros?



$$S_{xx} = \sum_i [x(i) - \bar{x}\,]^2$$

$$S_{yy} = \sum_i [y(i) - \bar{y}\,]^2$$

$$S_{xy} = \sum_i [[x(i) - \bar{x}\,]][y(i) - \bar{y}]$$

Uma vez que tenhamos os dados para extrair o modelo, pode-se empregar o método dos **mínimos quadrados** para obter os valores dos parâmetros w_1 e w_0

O objetivo, nesse caso, é obter os parâmetros de maneira que o erro de aproximação seja o menor possível. Para isso, define-se a função custo a ser minimizada, que corresponde a

$$\sum_i e_i^2 = \sum_i \left[y(i) - \big(w_0 + w_1 x(i)\big)\right]^2$$

E a solução é dada por

$$w_0 = \bar{y} - w_1 \bar{x} \text{ e } w_1 = \frac{S_{xy}}{S_{xx}}$$

E no caso de regressão múltipla?

Quando há mais de uma variável independente, a formulação matricial pode ser bastante conveniente para obter a solução de mínimos quadrados. Seja a função a ser estimada definida por $y = \mathbf{w}^T\mathbf{x}$, para cada vetor de entrada $\mathbf{x}(i)$ obteremos um valor de saída $y(i) = \mathbf{w}^T\mathbf{x}(i)$.

Podemos agrupar os *L* vetores de entrada e *L* valores de saída desejada de maneira que

$$\underbrace{\begin{bmatrix} 1 & x_1(1) & x_2(1) & \cdots & x_N(1) \\ 1 & x_1(2) & x_2(2) & \cdots & x_N(2) \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_1(L) & x_2(L) & \cdots & x_N(L) \end{bmatrix}}_{\boldsymbol{\Phi}} \underbrace{\begin{bmatrix} w_0 \\ w_1 \\ w_2 \\ \vdots \\ w_N \end{bmatrix}}_{\mathbf{w}} - \underbrace{\begin{bmatrix} y(1) \\ y(2) \\ \vdots \\ y(L) \end{bmatrix}}_{\mathbf{y}} = \underbrace{\begin{bmatrix} e(1) \\ e(2) \\ \vdots \\ e(L) \end{bmatrix}}_{\mathbf{e}}$$

Regressão – *Least Squares*

Como o objetivo é minimizar o erro quadrático, temos que

$$\sum_i e(i)^2 = \mathbf{e}^T\mathbf{e}$$
$$= (\boldsymbol{\Phi}\mathbf{w} - \mathbf{y})^T(\boldsymbol{\Phi}\mathbf{w} - \mathbf{y})$$
$$= (\mathbf{w}^T\boldsymbol{\Phi}^T - \mathbf{y}^T)(\boldsymbol{\Phi}\mathbf{w} - \mathbf{y})$$
$$= \mathbf{w}^T\boldsymbol{\Phi}^T\boldsymbol{\Phi}\mathbf{w} - \mathbf{y}^T\boldsymbol{\Phi}\mathbf{w} - \mathbf{w}^T\boldsymbol{\Phi}^T\mathbf{y} + \mathbf{y}^T\mathbf{y}$$

Como queremos encontrar o conjunto de parâmetros que minimiza a função custo, devemos obter $\mathbf{w}$ tal que

$$\nabla_{\mathbf{w}}\left(\sum_i e(i)^2\right) = 0$$

Least Squares

O cálculo do gradiente (derivada da função custo em relação a cada um dos parâmetros) pode ser feito utilizando algumas “regras de cálculo matricial”, similares às regras de derivação vistas em FVV

$$\frac{\partial}{\partial \mathbf{w}}(\mathbf{w}^T\mathbf{a}) = \frac{\partial}{\partial \mathbf{w}}(\mathbf{a}^T\mathbf{w}) = \mathbf{a}$$

$$\frac{\partial}{\partial \mathbf{w}}\mathbf{w}^T\mathbf{A}\mathbf{w} = (\mathbf{A} + \mathbf{A}^T)\mathbf{w}$$

se $\mathbf{A}$ for simétrica → $\frac{\partial}{\partial \mathbf{w}}\mathbf{w}^T\mathbf{A}\mathbf{w} = 2\mathbf{A}\mathbf{w}$

Least Squares

Assim,

$$\frac{\partial}{\partial \mathrm{w}}(\mathbf{w}^T\mathbf{\Phi}^T\mathbf{\Phi}\mathbf{w} - \mathbf{y}^T\mathbf{\Phi}\mathbf{w} - \mathbf{w}^T\mathbf{\Phi}^T\mathbf{y} + \mathbf{y}^T\mathbf{y}) = 2\mathbf{\Phi}^T\mathbf{\Phi}\mathbf{w} - 2\mathbf{\Phi}^T\mathbf{y} = \mathbf{0}$$

Ou seja,

$$\mathbf{\Phi}^T\mathbf{\Phi}\mathbf{w} - \mathbf{\Phi}^T\mathbf{y} = \mathbf{0}$$

$$\boxed{\mathbf{w} = (\mathbf{\Phi}^T\mathbf{\Phi})^{-1}\,\mathbf{\Phi}^T\mathbf{y}}$$

Regressão polinomial

O mesmo ferramental matemático se aplica para regressão com outros modelos, desde que sejam **lineares nos parâmetros**. Um exemplo é o modelo polinomial, i.e.,

$$y = a_0 + a_1 x + a_2 x^2 + \cdots + a_N x^N$$

Note que o modelo pode ser descrito na forma matricial

$$y = \underbrace{\begin{bmatrix} a_0 & a_1 & \cdots & a_N \end{bmatrix}}_{\mathbf{w}^T} \underbrace{\begin{bmatrix} 1 \\ x \\ \vdots \\ x^N \end{bmatrix}}_{\mathbf{x}}$$

Exatamente o mesmo modelo que vimos na regressão múltipla.

Regressão polinomial

Nesse caso, a matriz $\mathbf{\Phi}$ assume a seguinte forma

$$\mathbf{\Phi} = \begin{bmatrix} 1 & x_1(1) & x_1^2(1) & \cdots & x_1^N(1) \\ 1 & x_1(2) & x_1^2(2) & \cdots & x_1^N(2) \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_1(L) & x_1^2(L) & \cdots & x_1^N(L) \end{bmatrix}$$

Mas a solução continua sendo dada por

$$\mathbf{w} = \underbrace{(\mathbf{\Phi}^T \mathbf{\Phi})^{-1}\, \mathbf{\Phi}^T}_{pseudo\ inversa} \mathbf{y}$$

Regressão Linear – Funções Base

De maneira geral, a regressão assume que y é uma função das variáveis independentes $x_1, x_2, \ldots$, que podem ser agrupadas em um vetor $\mathbf{x} = [x_1, x_2, \ldots]^T$. A forma da função é definida por um conjunto de parâmetros, geralmente expressos por um vetor de parâmetros $\mathbf{w}$, ou seja

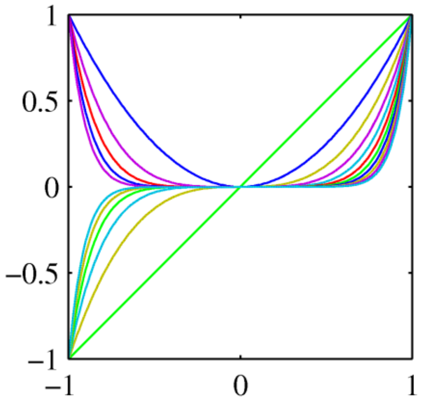
$$y = f(\mathbf{x}, \mathbf{w})$$

Assumindo que a função é linear nos parâmetros, pode-se empregar o mesmo ferramental matemático desenvolvido para uma classe mais ampla de modelos, baseados na combinação de funções base, ou seja

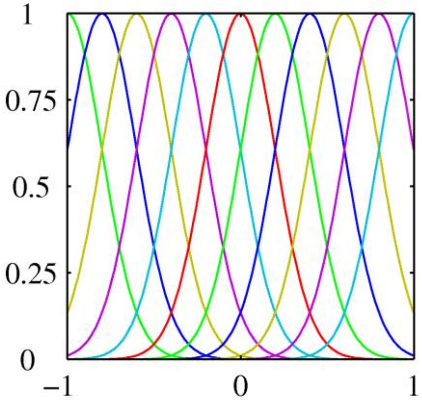
$$y = \mathbf{w}^T \boldsymbol{\phi}(\mathbf{x}) = \begin{bmatrix} w_1 & \cdots & w_N \end{bmatrix} \begin{bmatrix} \phi_1(\mathbf{x}) \\ \vdots \\ \phi_N(\mathbf{x}) \end{bmatrix}$$

onde $\boldsymbol{\phi}_i(\cdot)$ denotam funções pré-definidas (possivelmente não-lineares) do vetor $\mathbf{x}$

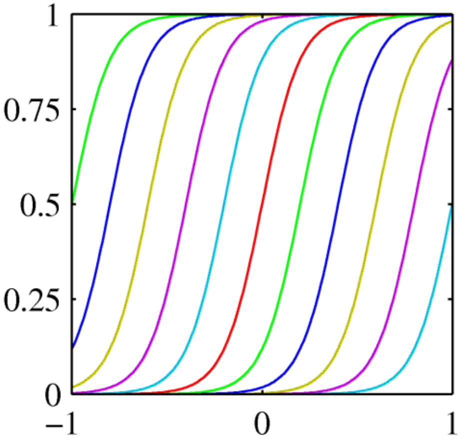
Exemplos de funções base



Polynomial basis functions



Gaussian basis functions
(relacionada a métodos de kernel)



Sigmoidal basis functions
(já vi isso antes…)

Regressão Linear com Funções Base

A matriz $\mathbf{\Phi}$ assume a seguinte forma

$$\mathbf{\Phi} = \begin{bmatrix} 1 & \phi_1(\mathbf{x}(1)) & \phi_2(\mathbf{x}(1)) & \cdots & \phi_N(\mathbf{x}(1)) \\ 1 & \phi_1(\mathbf{x}(2)) & \phi_2(\mathbf{x}(2)) & \cdots & \phi_N(\mathbf{x}(2)) \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & \phi_1(\mathbf{x}(L)) & \phi_2(\mathbf{x}(L)) & \cdots & \phi_N(\mathbf{x}(L)) \end{bmatrix}$$

$$\mathbf{w} = (\mathbf{\Phi}^T \mathbf{\Phi})^{-1}\, \mathbf{\Phi}^T \mathbf{y}$$

Solução também está relacionada a outra abordagem estatística para estimação de parâmetros denominada de **máxima verossimilhança**.

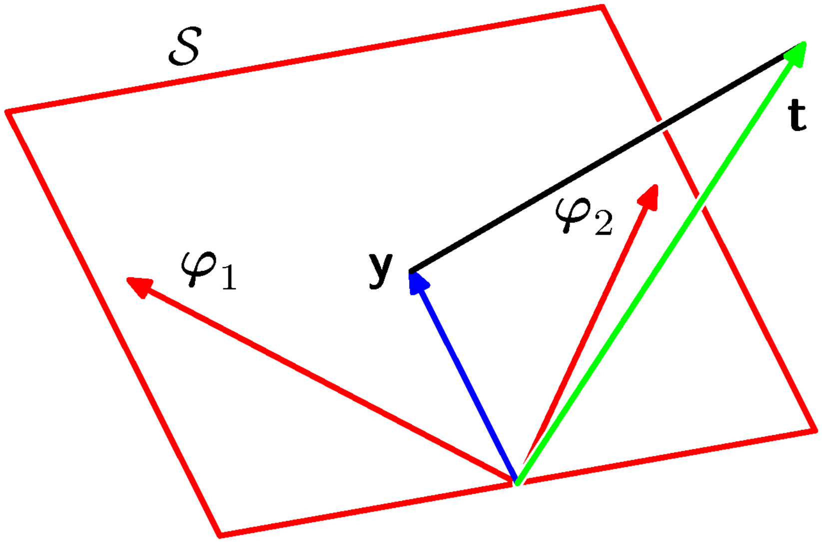
Interpretação geométrica do mínimos quadrados

Considere

$$\mathbf{y} = \boldsymbol{\Phi}\mathbf{w}_{\mathrm{ML}} = [\boldsymbol{\varphi}_1, \ldots, \boldsymbol{\varphi}_M]\, \mathbf{w}_{\mathrm{ML}}.$$

$$\mathbf{y} \in \mathcal{S} \subseteq \mathcal{T} \qquad \mathbf{t} \in \mathcal{T}$$

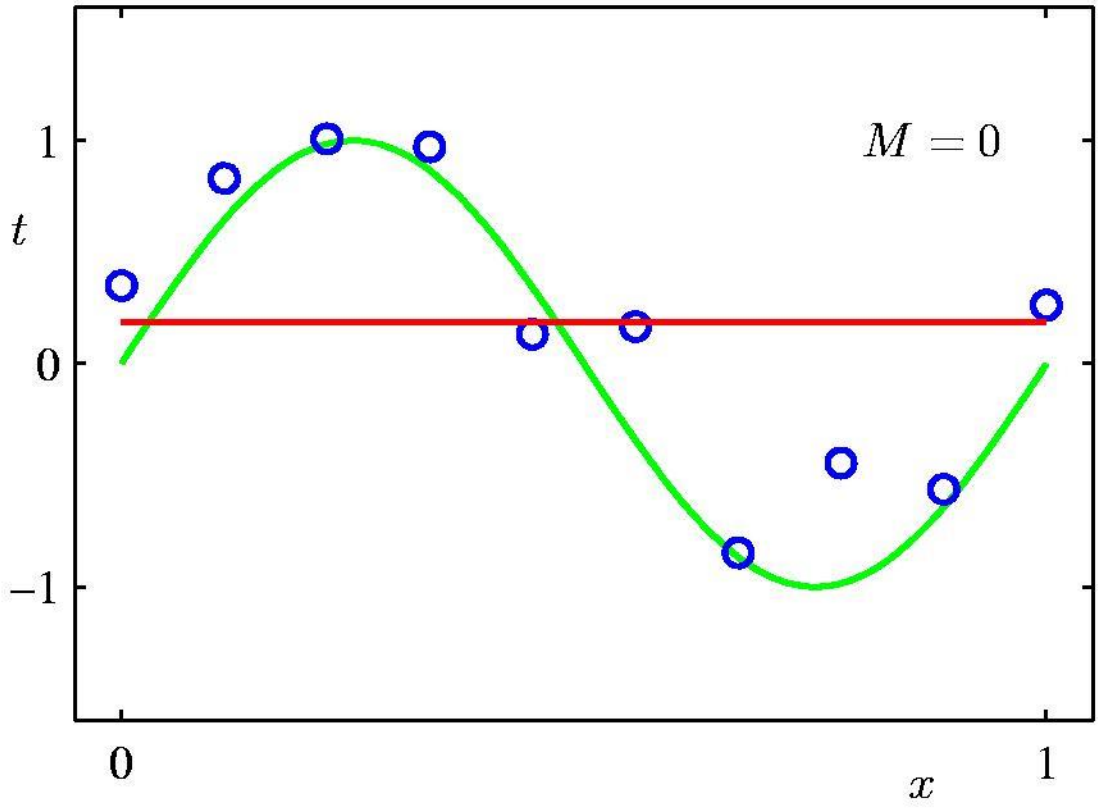
N-dimensional
M-dimensional



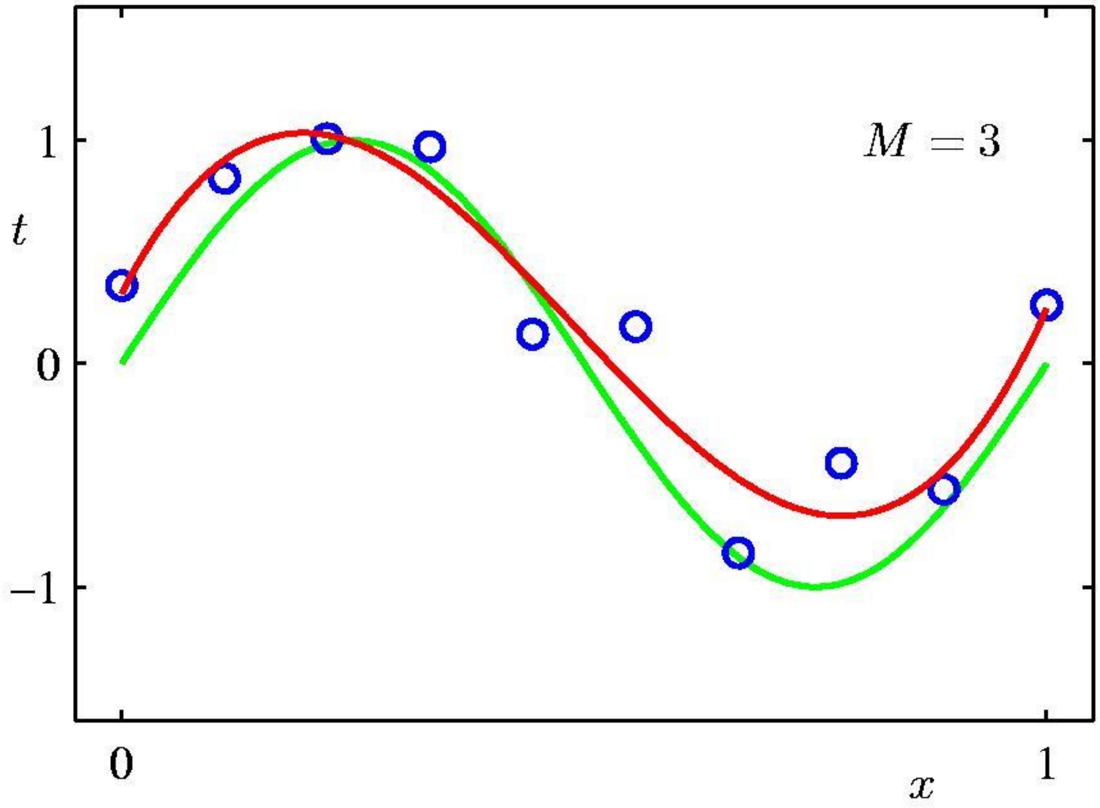
S é gerado por $\boldsymbol{\varphi}_1, \ldots, \boldsymbol{\varphi}_M$

$\mathbf{w}_{\mathrm{ML}}$ minimiza a distância entre **t** e sua projeção orthogonal.

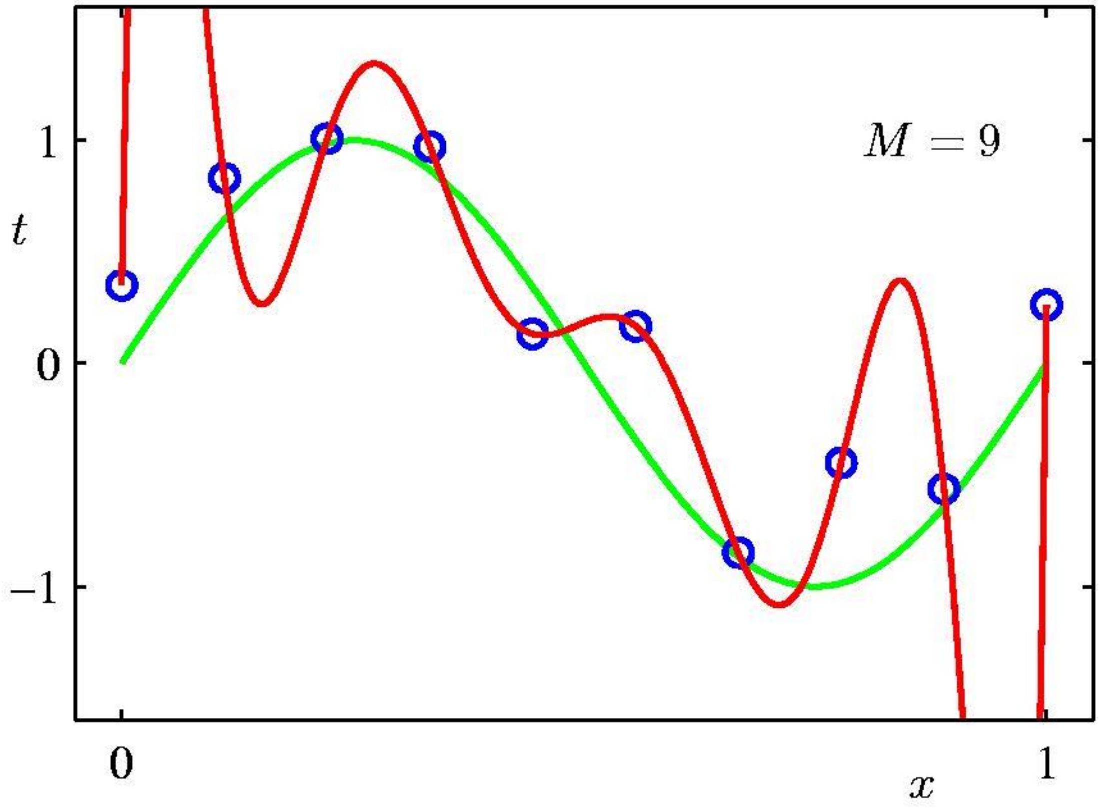
Polinômio de grau 0



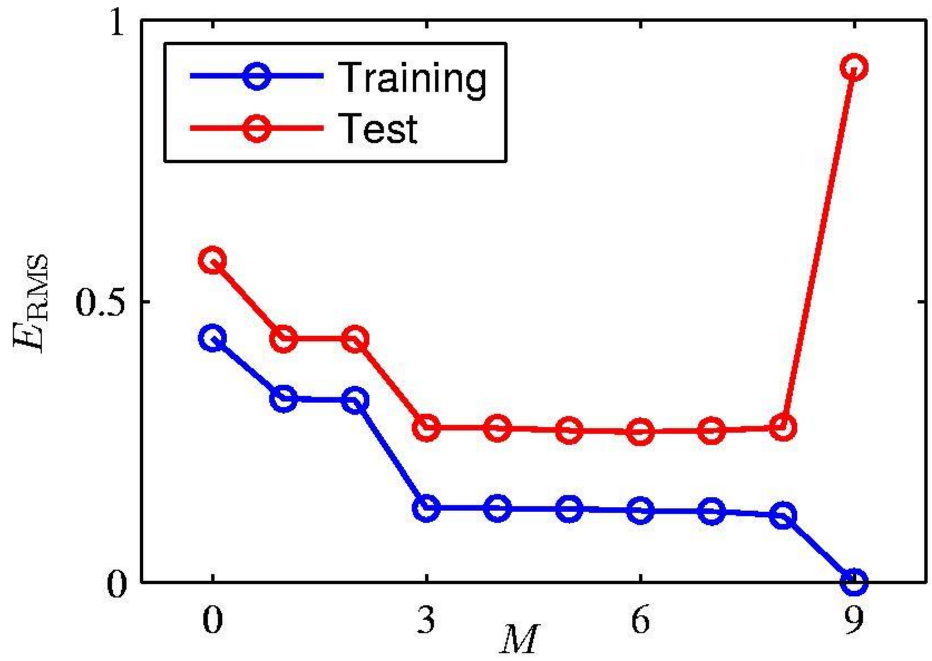
Polinômio de grau 3



Polinômio de grau 9



Overfitting



Root-Mean-Square (RMS) Error: $E_{\mathrm{RMS}} = \sqrt{2E(\mathbf{w}^\star)/N}$

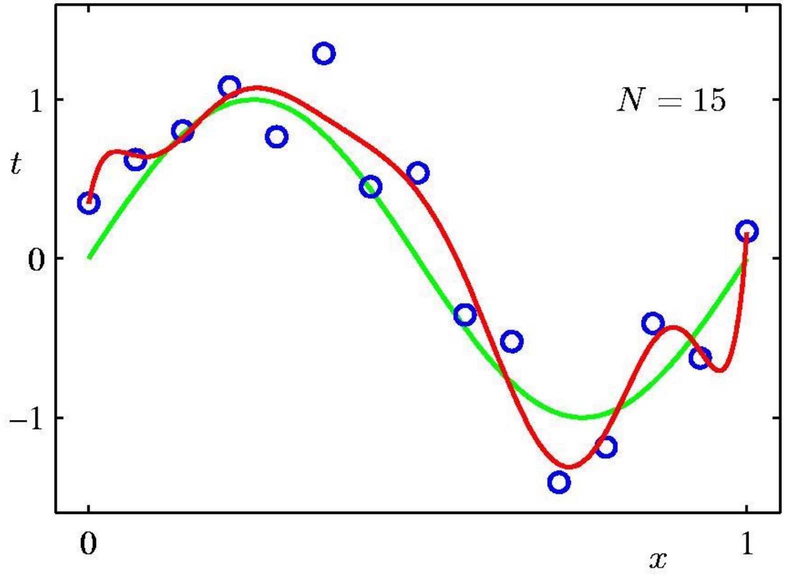
Coeficientes dos polinômios

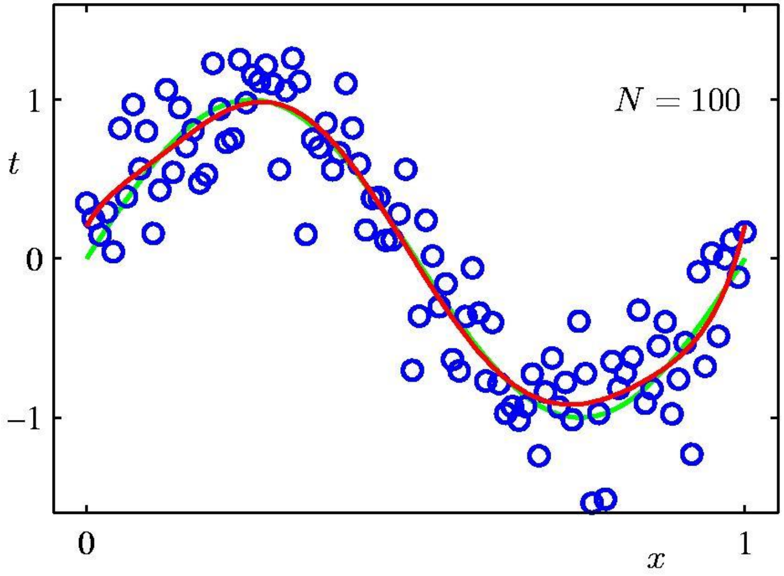
	$M=0$	$M=1$	$M=3$	$M=9$
$w_0^\star$	0.19	0.82	0.31	0.35
$w_1^\star$		-1.27	7.99	232.37
$w_2^\star$			-25.43	-5321.83
$w_3^\star$			17.37	48568.31
$w_4^\star$				-231639.30
$w_5^\star$				640042.26
$w_6^\star$				-1061800.52
$w_7^\star$				1042400.18
$w_8^\star$				-557682.99
$w_9^\star$				125201.43

Influência do tamanho do conjunto de dados

OLINÔMIO DE GRAU 9

POLINÔMIO DE GRAU 9





Como controlar o ajuste do modelo?

A fim de obter um modelo que tenha melhor desempenho na generalização, podemos tentar limitar a magnitude dos parâmetros

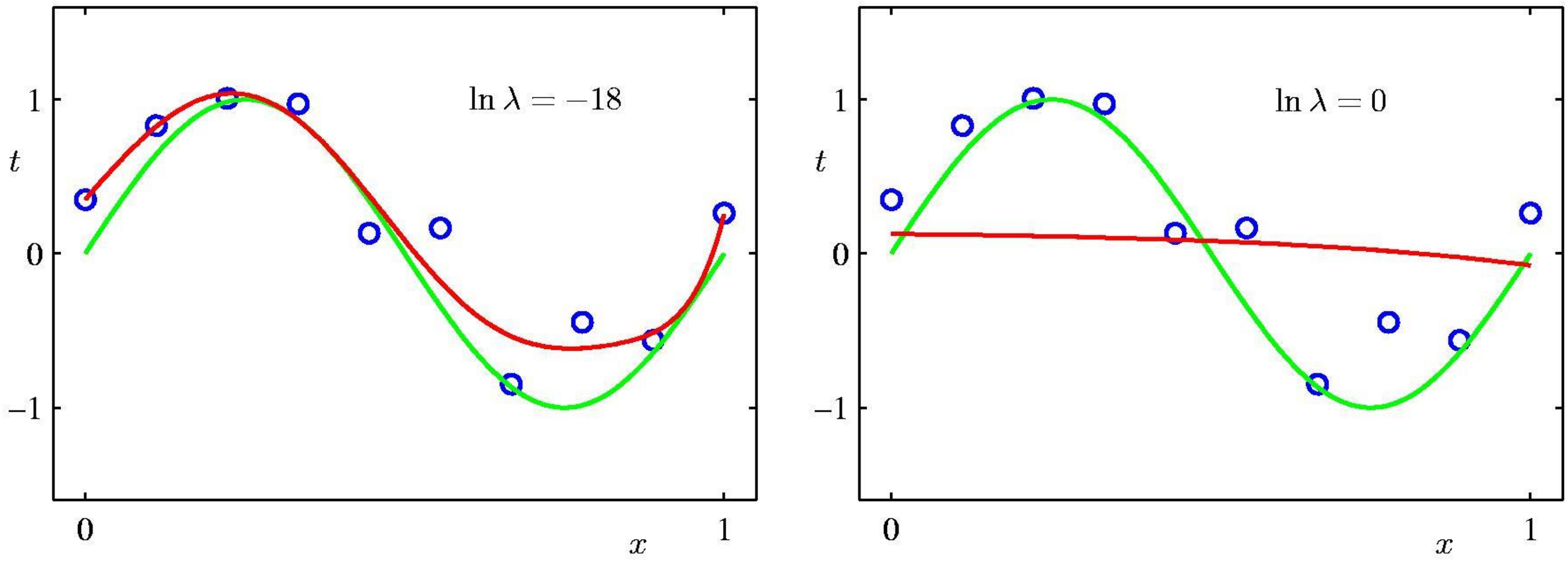
Para isso, podemos incluir um termo de penalização na função custo (termo de regularização)

$$\sum_{i} e(i)^2 + \lambda \parallel \mathbf{w} \parallel^2$$

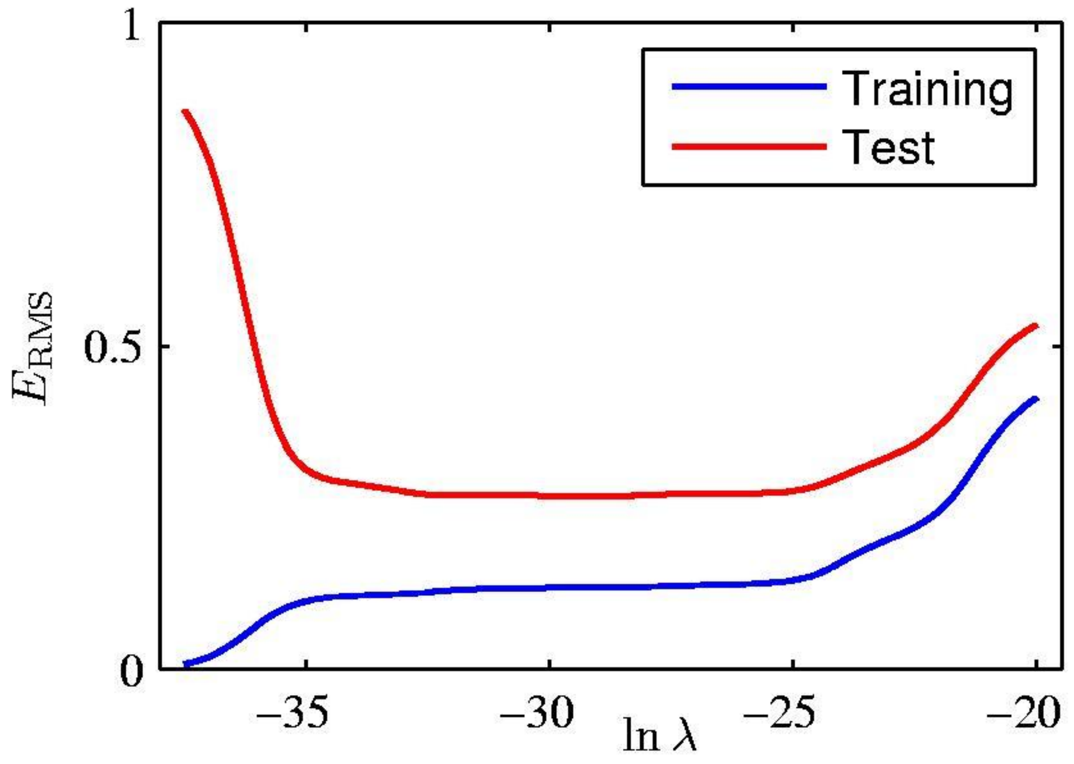
Quanto maior a magnitude dos coeficientes, maior será o valor da função custo. Portanto, a inclusão do termo de regularização tende a privilegiar soluções com a norma de $\mathbf{w}$ pequena. A solução, nesse caso, a solução é dada por

$$\mathbf{w} = \left(\lambda\mathbf{I} + \mathbf{\Phi}^{\mathrm{T}}\mathbf{\Phi}\right)^{-1}\mathbf{\Phi}^{T}\mathbf{y}$$

Regularização



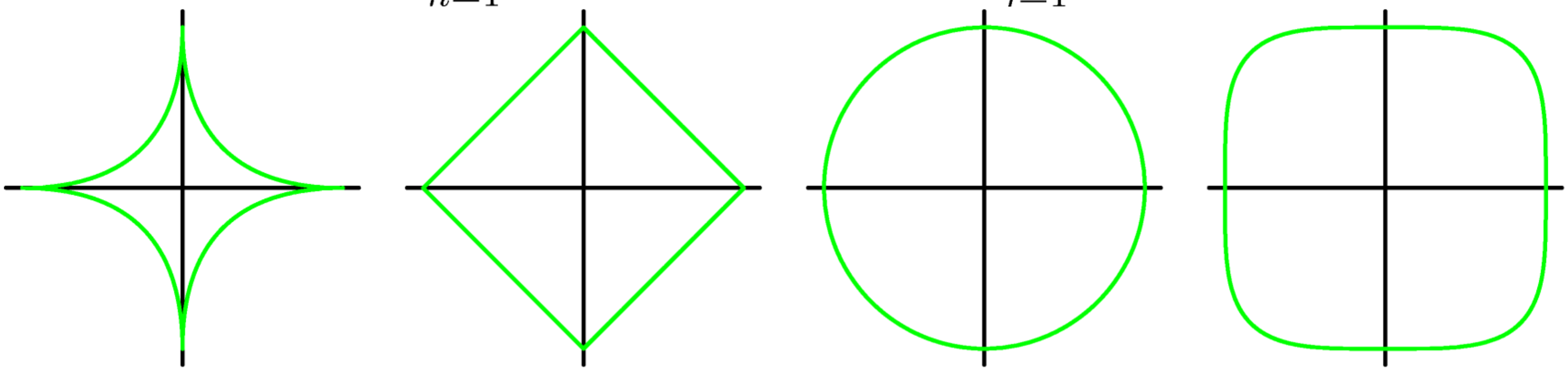
Regularização



Regularização

Podem ser considerados outros tipos e regularização

$$\frac{1}{2}\sum_{n=1}^{N}\{t_n - \mathbf{w}^{\mathrm{T}}\boldsymbol{\phi}(\mathbf{x}_n)\}^2 + \frac{\lambda}{2}\sum_{j=1}^{M}|w_j|^q$$



$q = 0.5$ $q = 1$ $q = 2$ $q = 4$

Lasso Quadratic

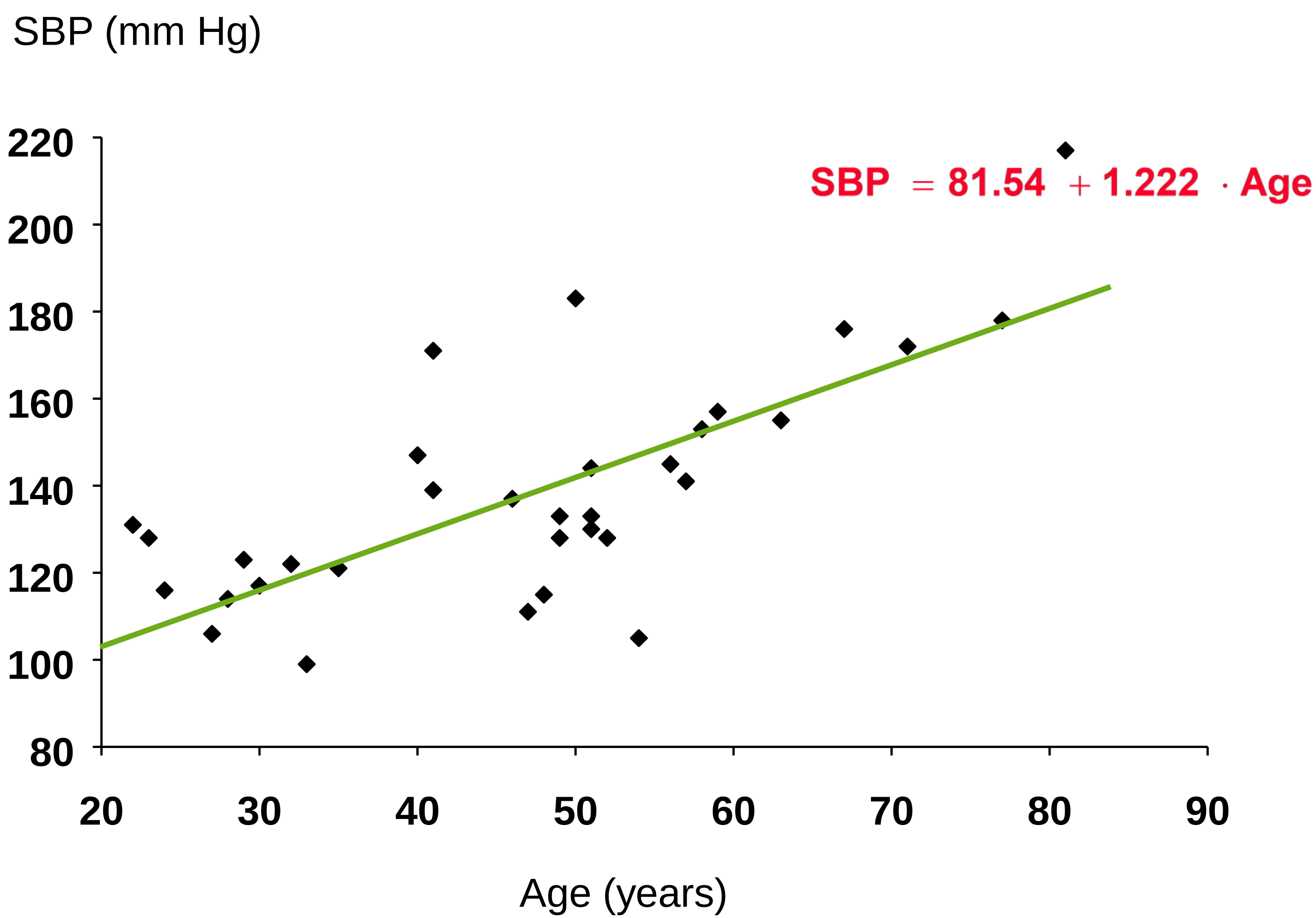
Regressão – Exemplo 1

Age	SBP
22	131
23	128
24	116
27	106
28	114
29	123
30	117
32	122
33	99
35	121
40	147

Age	SBP
41	139
41	171
46	137
47	111
48	115
49	133
49	128
50	183
51	130
51	133
51	144

Age	SBP
52	128
54	105
56	145
57	141
58	153
59	157
63	155
67	176
71	172
77	178
81	217

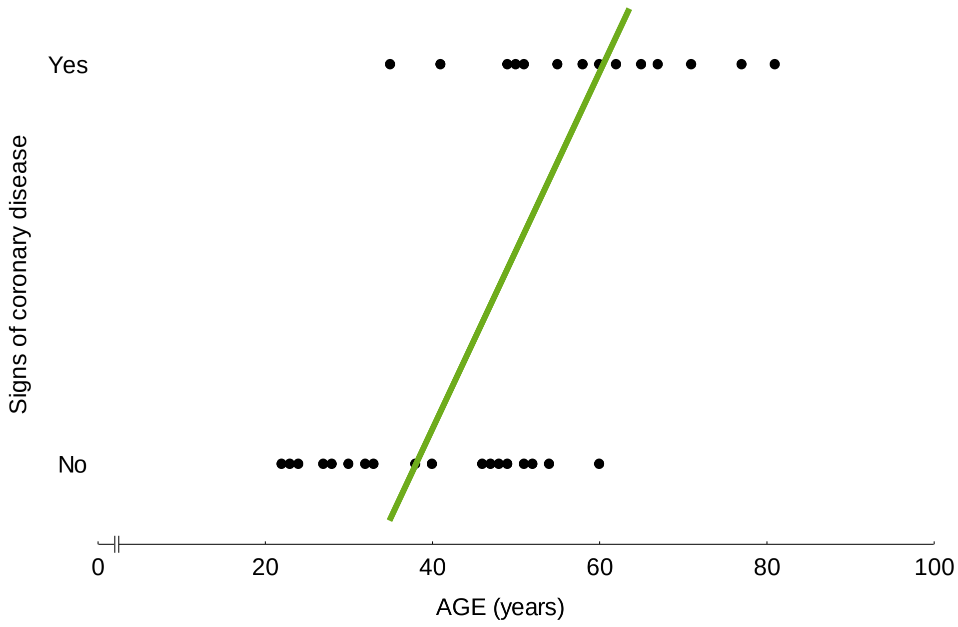
Pressão x Idade de 33 mulheres jovens


SBP (mm Hg)
220
200
180
160
140
120
100
80
20
30
40
50
60
70
80
90
Age (years)
SBP = 81.54 + 1.222 · Age

Regressão – Exemplo 2

Age	CD	Age	CD	Age	CD
22	0	40	0	54	0
23	0	41	1	55	1
24	0	46	0	58	1
27	0	47	0	60	1
28	0	48	0	60	0
30	0	49	1	62	1
30	0	49	0	65	1
32	0	50	1	67	1
33	0	51	0	71	1
35	1	51	1	77	1
38	0	52	0	81	1

Idade x Sinais de Morte por Doença Coronária



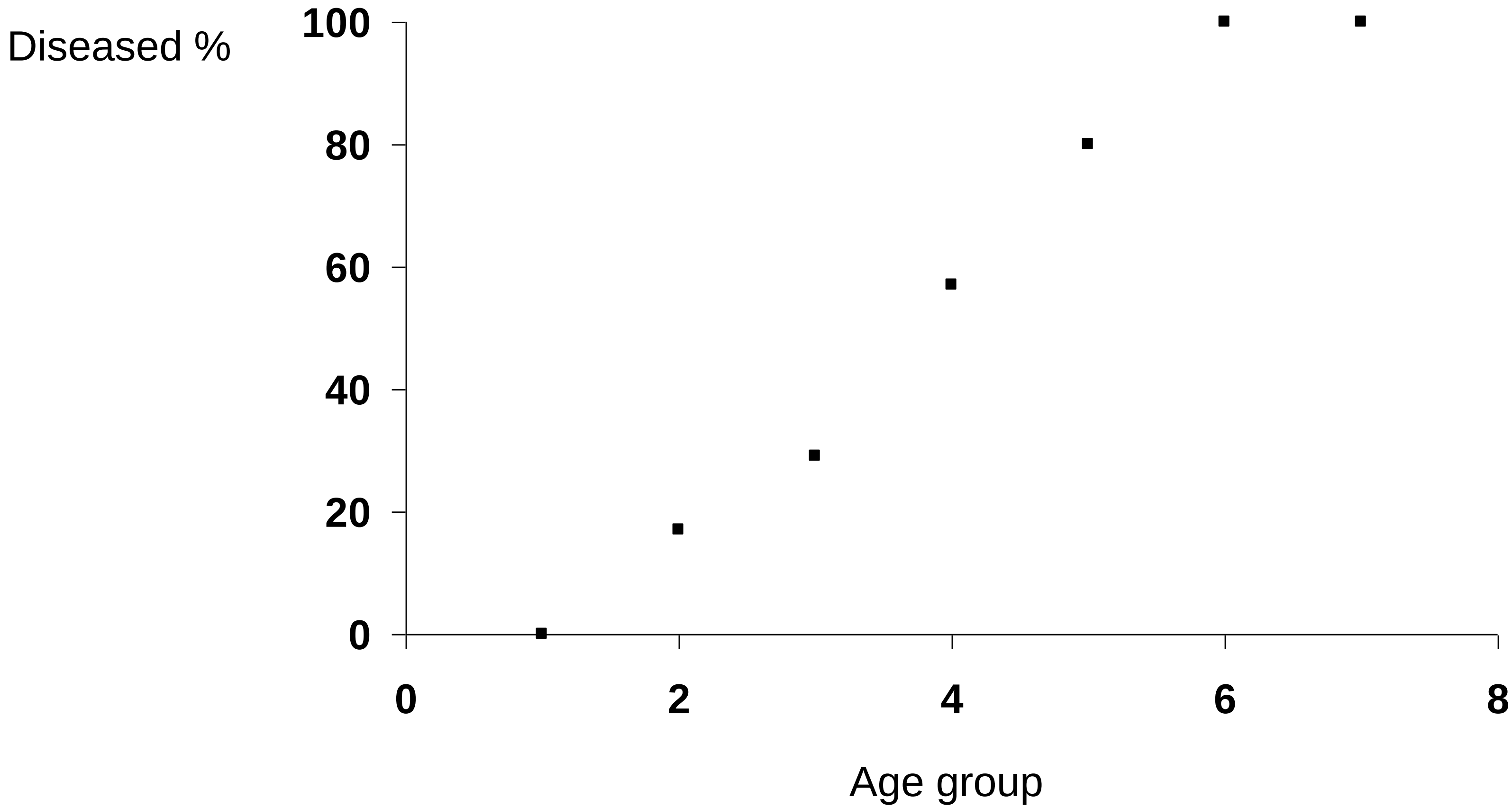
Como analisar os dados? Regressão linear?

Reorganizando os dados

		Diseased	
Age group	# in group	#	%
20 - 29	5	0	0
30 - 39	6	1	17
40 - 49	7	2	29
50 - 59	7	4	57
60 - 69	5	4	80
70 - 79	2	2	100
80 - 89	1	1	100

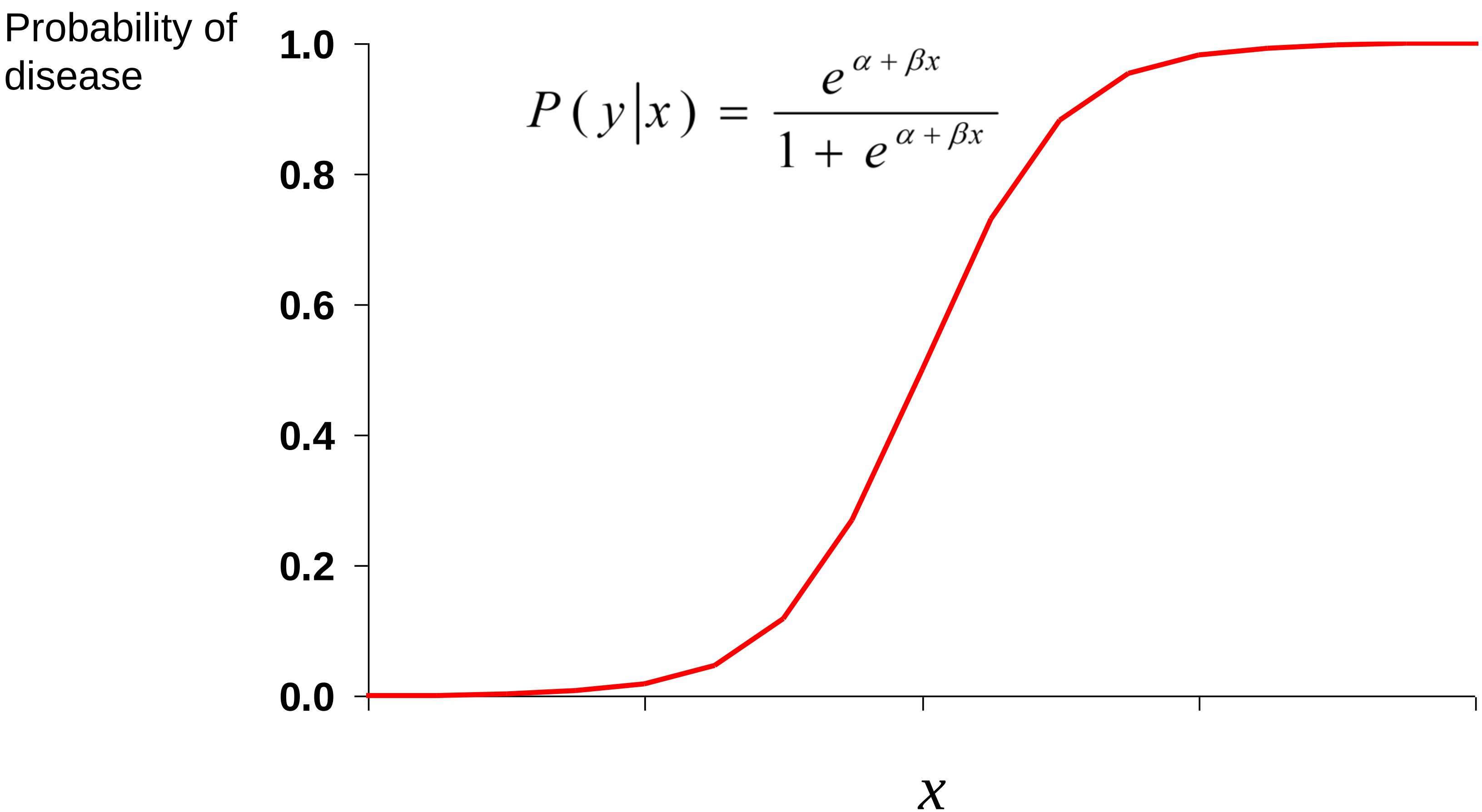
“Probabilidade de morrer por doença coronária”

Porcentagem de mortes por doença coronária (%) x Faixa Etária



Diseased %
100
80
60
40
20
0
0
2
4
6
8
Age group

Função Logística



Regressão Logística

Caso particular de regressão que é útil quando a variável dependente é binária ou multinomial

Nesse caso, modela-se a “probabilidade” de ocorrência de um determinado evento. Observe que

$$p(x) = \frac{e^{\alpha+\beta x}}{1+e^{\alpha+\beta x}} \rightarrow \frac{p(x)}{1-p(x)} = e^{\alpha+\beta x}$$

onde $\frac{p(x)}{1-p(x)}$ representa a razão entre a probabilidade de ocorrência e de não-ocorrência de y (*odds ratio*), que pode ser colocada em uma forma mais conveniente utilizando o logaritmo, i.e.

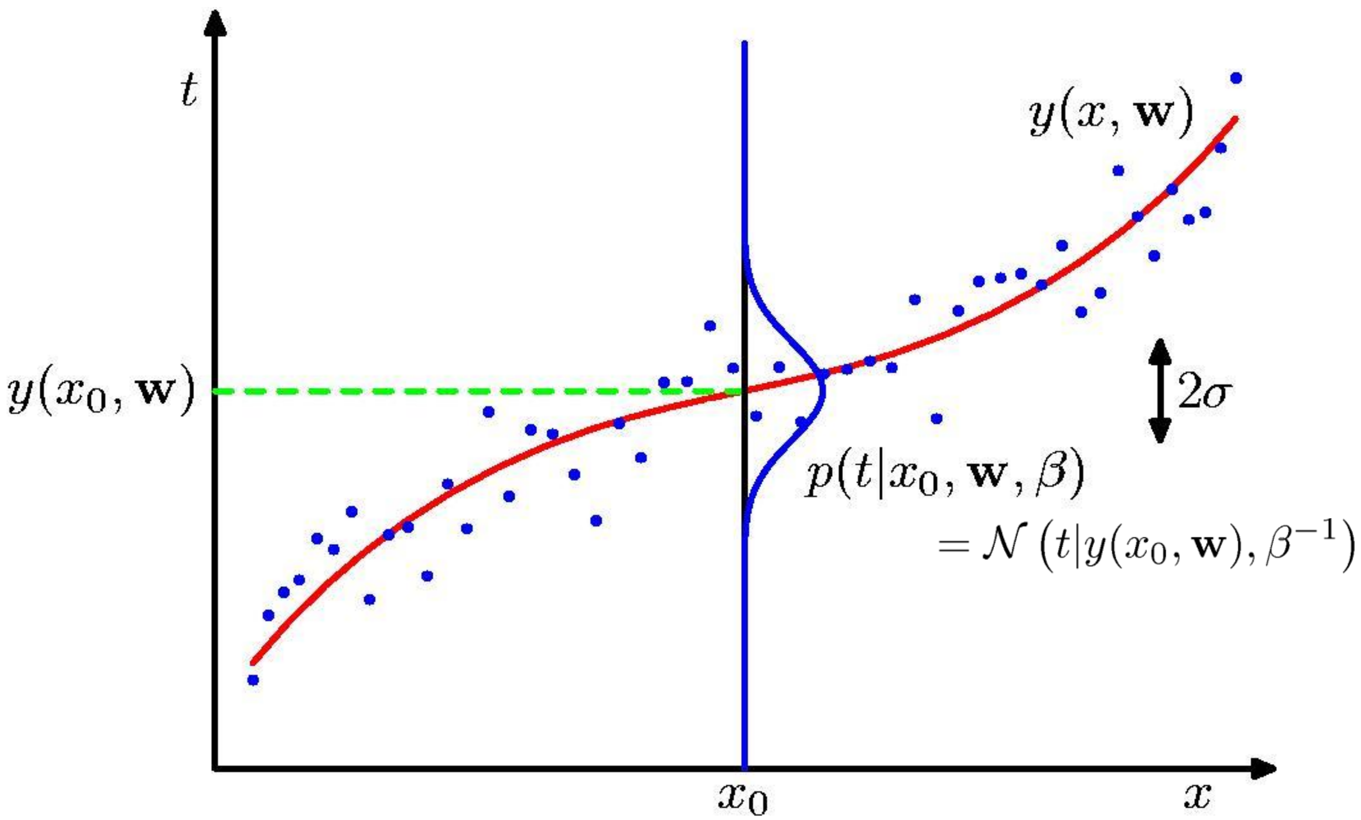
$$\ln \frac{p(x)}{1-p(x)} = \operatorname{logit}\big(p(x)\big) = \alpha + \beta x$$

Ajuste dos parâmetros por Máxima Verossimilhança

No caso da regressão logística utiliza-se o método de máxima verossimilhança → resultado coincide com o método de mínimos quadrados apenas em situações específicas (i.e., o resíduo do modelo apresenta distribuição normal)

O método consiste em determinar o conjunto de parâmetros que maximiza o valor de uma determinada função custo, denominada função de verossimilhança, que está associada ao modelo estatístico definido para os dados observados.

Ajuste de Curvas com ruído



Função de Verossimilhança

Suponha que tenhamos um modelo para os dados observados, e devido à presença de ruído/incertezas, o modelo é descrito em termos de distribuições de probabilidade

- Por exemplo, suponha que o valor observado $x[0]$ esteja relacionado a um modelo generativo do tipo

$$x[0] = \theta + w[0]$$

onde $w[0]$ corresponde a um ruído aditivo gaussiano

Nesse caso, o valor de $y[0]$ possui uma distribuição de probabilidade associada, de maneira que

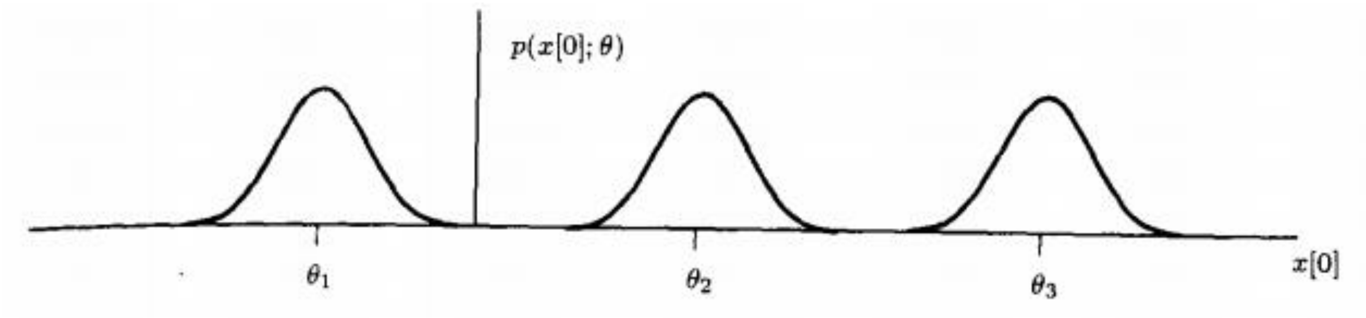
$$p(x[0];\theta) = \frac{1}{\sqrt{2\sigma^2}} \exp\left[-\frac{1}{2\sigma^2}(x[0]-\theta)^2\right]$$

i.e., depende do valor do parâmetro desconhecido θ.

Qual o valor de θ “faz mais sentido”?

Suponha que o valor observado de $x[0] = -10$.

- Neste caso, é mais razoável supor que $\theta = \theta_1$, $\theta = \theta_2$ ou $\theta = \theta_3$?



O valor de θ mais razoável é aquele que maximiza a função $p(x[0]; \theta)$, denominada de *função de verossimilhança*. Note que pode-se construir a função de verossimilhança também para o caso em que temos acesso a mais do que uma observação e também mais do que um parâmetro, i.e., $p(\mathbf{y}; \boldsymbol{\theta})$

Voltado ao caso da Regressão Logística

Considere o caso em que a variável dependente é binária ($y = \{0,1\}$)

- Seja $p(y = 1|x) = p$ e $p(y = 0|x) = 1 - p$
- Assim, a função de verossimilhança, é dada por

$$L(\alpha, \beta) = p(\mathbf{y}|\mathbf{x}, \alpha, \beta) = \prod_i p^{y_i}(1 - p)^{1 - y_i}$$

$$L(\alpha, \beta) = \prod_i \left(\frac{e^{\alpha + \beta x_i}}{1 + e^{\alpha + \beta x_i}}\right)^{y_i} \left(\frac{1}{1 + e^{\alpha + \beta x_i}}\right)^{1 - y_i} = \prod_i \frac{\left(e^{\alpha + \beta x_i}\right)^{y_i}}{1 + e^{\alpha + \beta x_i}}$$

Como a função envolve exponenciais, é conveniente trabalhar com o seu logaritmo, i.e.,

$$\log L(\alpha, \beta) = \sum_i y_i(\alpha + \beta x_i) - \log\left(1 + e^{\alpha + \beta x_i}\right)$$

Métodos Iterativos para o MLE

Como não é possível obter uma forma fechada para os parâmetros que maximizam $\log L(\alpha, \beta)$ utilizam-se métodos iterativos para a busca

- A exemplo do que foi visto no algoritmo *backpropagation*, uma possibilidade é utilizar o gradiente da função para isso, i.e.,

$$\frac{\partial \log L(\alpha, \beta)}{\partial \alpha} = \sum_i y_i - \frac{e^{\alpha + \beta x_i}}{1 + e^{\alpha + \beta x_i}}$$

$$\frac{\partial \log L(\alpha, \beta)}{\partial \beta} = \sum_i x_i y_i - \frac{x_i e^{\alpha + \beta x_i}}{1 + e^{\alpha + \beta x_i}}$$

O algoritmo iterativo de busca é descrito pelas seguintes equações de atualização

$$\alpha \leftarrow \alpha + \mu \frac{\partial \log L(\alpha, \beta)}{\partial \alpha}$$

$$\beta \leftarrow \beta + \mu \frac{\partial \log L(\alpha, \beta)}{\partial \beta}$$

Outros métodos de otimização podem ser utilizados (e.g., Gradiente Conjugado, Newton, etc)

Regressão Bayesiana

Ao invés de considerar o modelo com parâmetros fixos, podemos formular o problema considerando que os parâmetros também possuem uma distribuição de probabilidade associada.

A abordagem explora a regra de Bayes

verossimilhança

$$p(\theta|\mathbf{x}) = \frac{p(\mathbf{x},\theta)}{p(\mathbf{x})} = \frac{p(\mathbf{x}|\theta)p(\theta)}{p(\mathbf{x})}$$

Distribuição *a posteriori*

Distribuição *a priori*

Note que a distribuição $p(\mathbf{x})$ não depende dos parâmetros, e por essa razão representa apenas um fator de normalização. A estimativa dos parâmetros, nesse caso, é conhecida como solução de máxima a posteriori (MAP)

Outras abordagens

Regressão com *Decision Trees* (e *random forests*)

Support Vector Machines (*Support Vector Regression*)

Redes Neurais Artificiais

- *Deep Learning Networks*

Sistemas Fuzzy