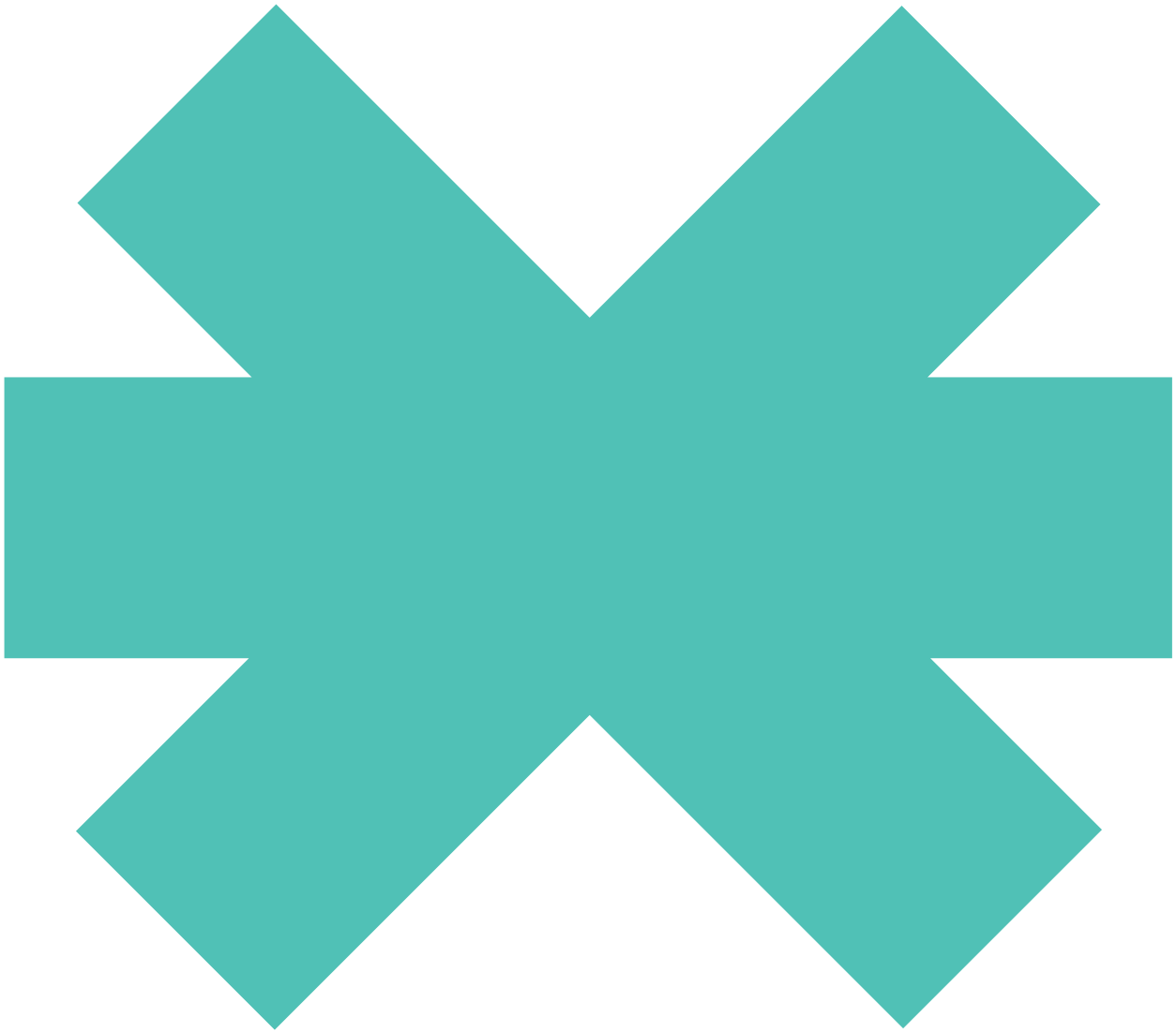
MÓDULO 2
Aprendizagem supervisionada
Especialização em Machine Learning

1

Regressão linear

New Technology School

Tokio.



1 Regressão linear

Sumário

1.1 Regressão linear simples 05

1.2 Regressão linear múltipla 08

New Technology School

Tokio.

A aprendizagem automática consiste na modelação de sistemas, através do treino de um modelo que aprende com um conjunto de dados, cujos resultados são previamente conhecidos.

Depois desse treino inicial, obtemos o modelo, os pesos, os coeficientes, etc., que poderemos usar para predizer ou descrever novos exemplos, cujos resultados ainda desconhecemos.

Assim, trata-se de treinar um modelo com perguntas, conhecendo as suas respostas para, posteriormente, poder usá-lo para resolver perguntas cujas respostas desconhecemos.

APLICAÇÕES

- Preditores ou avaliadores de preços de venda de qualquer produto, como imóveis, automóveis, bens, etc.
- Preditores de procura, oferta, níveis de produção, etc.
- Classificação de frutos segundo classes, qualidades, gamas e usos.
- Deteção de *spam* em *emails*, comentários, contas *fake*, etc.
- Manutenção preditiva, predição de falhas.
- Visão computacional.
- Reconhecimento de voz, agentes virtuais, transcrição de texto por voz.
- Processamento da linguagem natural.

1.1 Regressão linear simples

Procuramos uma forma de modelar um conjunto de dados, de propor uma função que relacione linearmente uma variável independente (os dados dos exemplos) com uma variável dependente (os seus resultados, a variável a predizer).

Essa função modela o conjunto de dados, isto é, simula o seu comportamento de uma forma matemática, que ajude a explicá-lo e a realizar predições no futuro.

Equação do modelo

A equação linear simples é a equação da reta:

$$Y = h_\Theta(x) = \Theta^T \times x = \theta_0 + \theta_1 \times x_1$$
$$x_0 = 1$$

Figura 1

Código Latex:

```
Y=h_\Theta(x)=\Theta^T \times x=\theta_0+\theta_1\times x_1 \\
x_0=1
```

A equação representa a hipótese que existe de uma relação linear, entre duas variáveis, *X* e *Y*, modelada desta forma e dependente de Theta.

- *x* poderia representar o nível socioeconómico de um bairro.
- *Y* poderia representar o nível de delinquência desse bairro.
- θ_0 seria o "parâmetro de intersecção" ou nível base do modelo.
- θ_1 seria o coeficiente da variável *x* que explica como varia o nível de delinquência, em função do nível socioeconómico do bairro.
- x_0, por conveniência, será 1; desta forma, podemos multiplicar diretamente ambos os vetores, em vez de multiplicar um vetor e realizar uma soma.

Representação gráfica

Y	x1
-12,42	-2
-7,35	-1
10,64	0
5,43	1
15,35	2
23,53	3
50,13	4
64,05	5
81,21	6
81,03	7
101,51	8
116,16	9
87,34	10
99,00	11
99,38	12
136,59	13
127,14	14
161,28	15
173,39	16
192,36	17
174,91	18

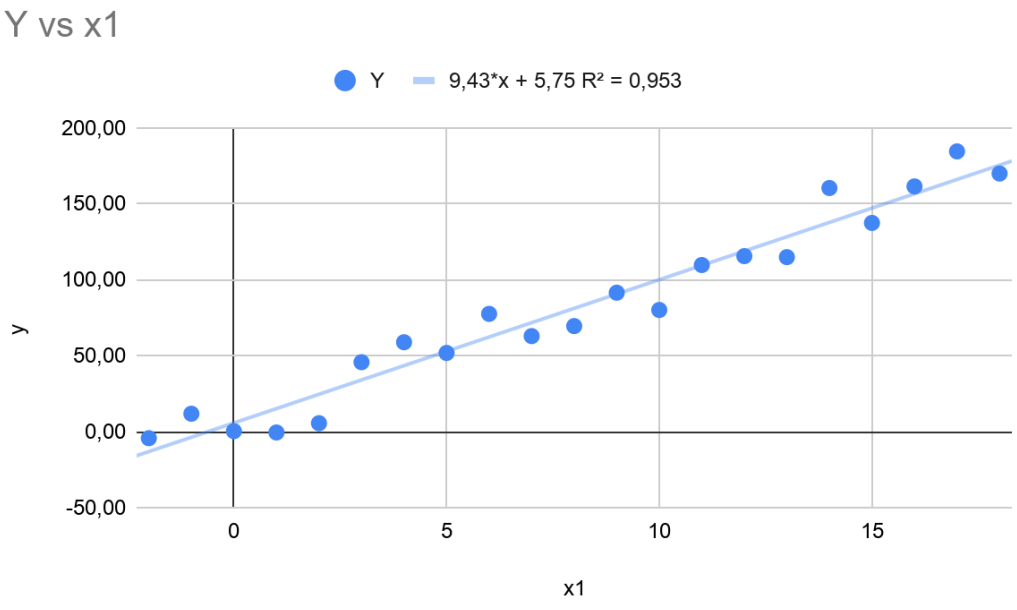


Figura 2

Tipos de variáveis

As variáveis podem ser de vários tipos:

- **Quantitativas.** Geralmente trata-se de variáveis de tipo numérico que representam uma característica mensurável.
  - **Contínuas:** variáveis numéricas em números reais, isto é, que podem apresentar decimais. Por exemplo: preços, temperatura, peso, distâncias, etc.
  - **Cardinais:** variáveis numéricas em números inteiros, isto é, que não podem apresentar decimais e que se representam com *int*, *double*, etc. Por exemplo: n.º de produtos, de familiares, de compras realizadas, etc.
- **Qualitativas ou categóricas.** Variáveis que não são de tipo numérico, mas representam uma propriedade, etiqueta ou categoria atribuída ao exemplo.
  - **Ordinais:** categorias que podem e devem ser ordenáveis, uma vez que existe uma certa disposição ou hierarquia natural que devemos obter. Por exemplo: o nível socioeconómico ("baixo, médio, alto"), nível educativo ("primária, secundária, universidade, mestrado, doutoramento"), hierarquia ("júnior, sénior, agente, diretor", "soldado, capitão, tenente, general") etc.
  - **Nominais:** categorias não ordenáveis. Por exemplo: cores, género, marcas, os ID de produto, "louro, moreno, ruivo", etc.
  - **Binárias:** categorias não ordenáveis ou nominais que dispõem apenas de dois valores. Por exemplo: "sim/não", "homem/mulher", "cliente existente/não", "utilizador de pagamento/gratuito", "contratou o serviço/não", etc.

1.2 Regressão linear múltipla

Uma **regressão linear múltipla** constitui, tal como a regressão linear simples, um modelo estadístico que estima uma previsão numérica de uma variável objetivo ou dependente, com base na combinação linear de variáveis.

Na regressão linear simples apenas temos em conta uma única variável, enquanto numa múltipla temos em consideração mais do que uma, para verificar como essa variável afeta o objetivo.

Aplicações:

- Predição de preços de venda.
- Previsão de demanda ou nº de vendas.
- Predição de probabilidade de compra.
- Manutenção preditiva de maquinaria: predição de probabilidade de falhas, tempo até à próxima falha, nível de desgaste, etc.
- Previsões meteorológicas.

Modelação de dados

O nosso objetivo é recolher, para cada exemplo que usarmos durante o treino, todas as características que criarmos, que possam influenciar a variável objetivo e dependente (*Y*) e codificá-las na variável *X*.

A variável *X* contará, assim, com uma linha, por exemplo, com um valor por cada uma das características recolhidas nesse exemplo.

EXEMPLO: PERITO AVALIADOR DE IMÓVEIS

Para tentar treinar um modelo de ML, face a predizer o preço de venda de um imóvel, devemos planear primeiro, qual será a nossa variável objetivo.

Para este caso, será o preço de venda, mas em €? Em milhares de €, com cêntimos de € ou sem decimais? Numa escala distinta?

Estas são algumas das perguntas que podemos fazer, sobre como codificar a variável *Y*.

Da mesma forma, seguidamente, iremos planear os exemplos a considerar: estes serão imóveis cujas características e preços de venda conhecemos (procedentes, por exemplo, da base de dados de uma grande imobiliária), que formarão a variável matriz *X*.

Sobre a variável *Y* podemos alertar para: o que deveríamos usar? Os preços a que são anunciados os imóveis? Os resultados de um avaliador oficial? Apenas os preços de venda final dos imóveis? Podemos aceder-lhes?

Sobre esses exemplos, devemos considerar que características ou atributos podem afetar o preço final do imóvel e quais deles podemos obter com veracidade e precisão suficientes.

Para este exemplo, podemos planear as seguintes características e considerações:

m x *n*	$x_{...,0}$	Preço $x_{...,1}$	m^2 $x_{...,2}$	Hab. $x_{...,3}$	WC $x_{...,4}$	Tipo $x_{...,5}$	Garagem $x_{...,6}$	Ano $x_{...,7}$	Dist. $x_{...,8}$	Serv. $x_{...,9}$	Loc. $x_{...,10}$	Terr. $x_{...,11}$	Pisc. $x_{...,12}$
Casa 1 $x_{0,...}$	1	150	110	3	2	1 (piso)	1	2010	1274	8,42	1	0	0
Casa2 $x_{1,...}$	1	200	100	4	2	3 (ático)	1	2007	3265	6,76	3	2	0
Casa 3 $x_{2,...}$	1	175	130	4	2	2 (casa)	2	1988	6312	4,82	5	1	0
Casa 4 $x_{3,...}$	1	110	90	2	1	1 (piso)	0	1977	2252	5,77	2	0	1

Sendo *m* o n.º de exemplos ou imóveis que iremos considerar, para treinar, e *n* o n.º de características obtidas de cada exemplo.

- Superfície útil, em m^2: superfície útil construída ou ambas? Em m^2, em intervalos de m^2, em "pequeno, médio, amplo"? Em superfície por piso?
- Número de divisões: apenas o nº de quartos ou de divisões no total? Todas as divisões devem contar da mesma forma ou, por exemplo, o quarto de casal como dupla? E o sótão, ou a lavandaria se existirem?
- Número de casas de banho: contamos com e sem banheira, de igual forma, de forma independente ou a sua superfície?
- Tipo de imóvel: como o representamos? Com *one-hot encoding* (múltiplas características/colunas com valor 0 e apenas um 1 na correta) ou como uma categoria ordenada?
- Garagem: categoria binária ou nº de lugares de garagem? Superfície? Tipo: "fechado, aberto, individual, comunitário", etc.?
- Ano: ano de construção ou idade? Anos desde a última remodelação?
- Distância do centro: trata-se de um parâmetro cujo cálculo é muito complexo. Distância linear? Tempo? Apenas em direção ao centro ou a vários destinos habituais, tais como a praia, a saída da autoestrada, centros comerciais e de negócios, etc.?
- Serviços: outro parâmetro difícil de calcular. Podemos propor um valor numérico estimado, com base em vários critérios ou usar esses critérios base, se bem que complicaríamos bastante o modelo e, às vezes, podemos não obter todos os dados.
- Localização: devemos definir o bairro ou área onde se encontra ou um nível comparativo entre várias localizações semelhantes?
- Terraço: categoria binária, n.º de terraços, superfície total? E as varandas?

- Piscina: binário, superfície? Privada, partilhada? Zonas comuns da urbanização?

Assim, neste ponto, devemos tentar definir todas as características de um imóvel, que determinariam quanto um comprador estaria disposto a pagar por ele, o que é bastante complicado e constitui uma fonte de possíveis erros e imprecisões.

O recomendável poderia ser começar com um modelo simples para tentar explorar os seus dados e resultados e, em função desses resultados, continuar a recolher mais informação, com o objetivo de criar um modelo mais preciso e robusto, em cada versão.

Modelação de uma curva

Em várias ocasiões, necessitamos de modelar conjuntos de dados complexos com relações diferentes, entre as diversas variáveis independentes e a variável objetivo. O que fazemos se não seguirem uma relação linear simples, mas exponencial, potencial, logarítmica, etc.? Podemos resolver esses modelos com uma regressão linear múltipla comum?

Nestes casos, devemos pré-processar as variáveis. Assim, por exemplo, se queremos considerar em polinómio de grau 3, podemos converter a variável original em três variáveis do seguinte modo:

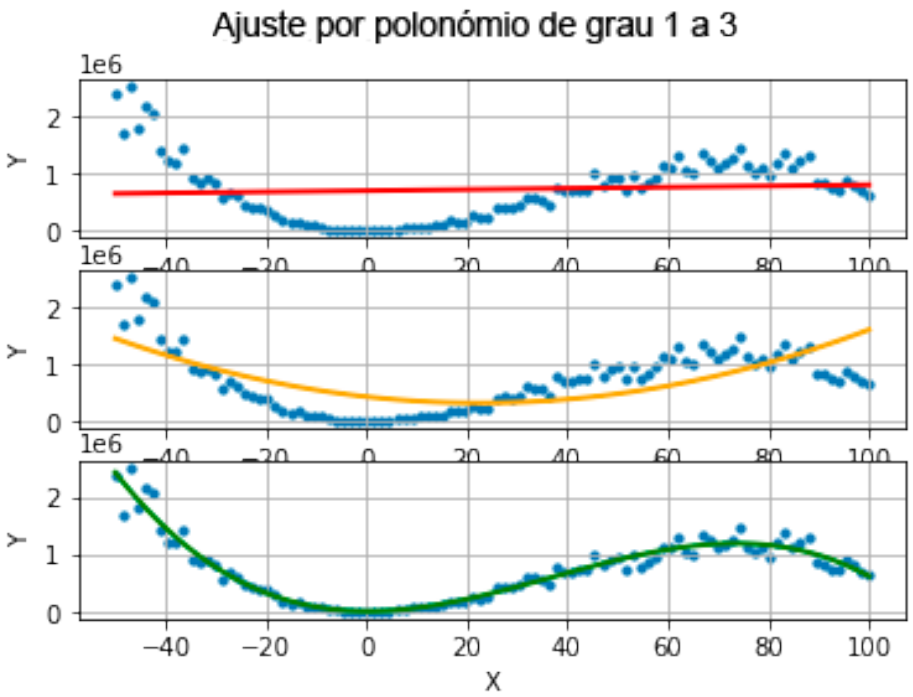


Figura 3

Podemos proceder do mesmo modo para outros tipos de ajustes, como o potencial ou logarítmico:

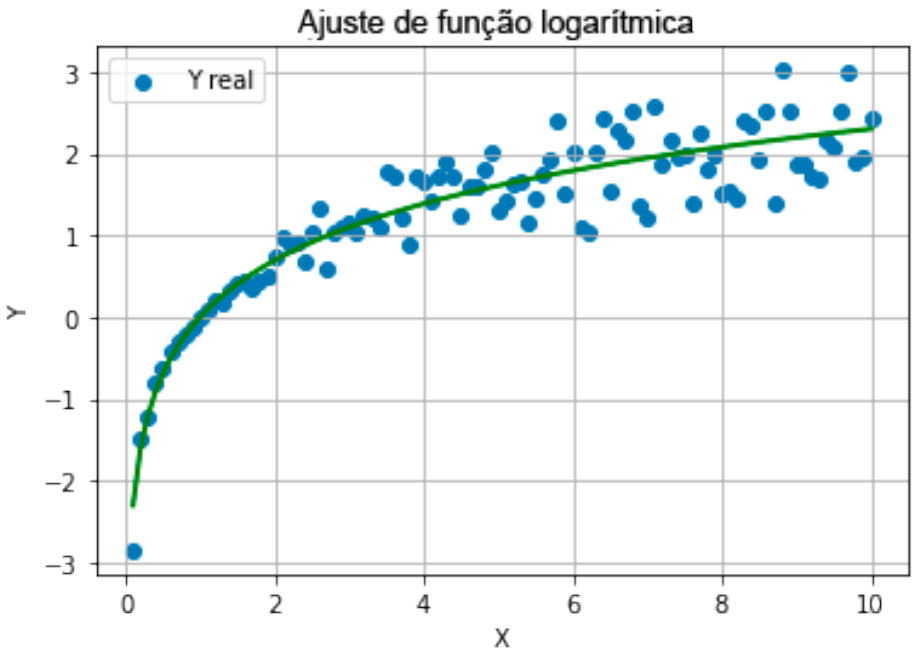


Figura 4

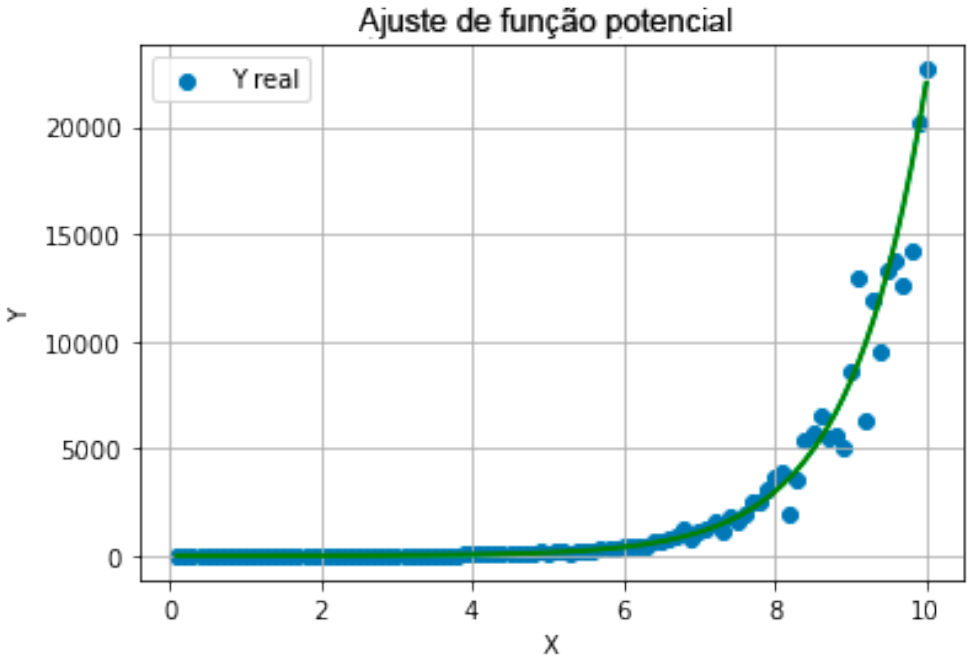


Figura 5

Desta forma, a partir de uma única variável original, consideramos várias variáveis, aparentemente independentes entre si, que são fruto do processamento dessa variável original.

Do mesmo modo, devemos recordar sempre que, segundo as propriedades de um sistema de equações linear, qualquer nova variável que seja uma combinação linear de outras variáveis. não fornecerá nova informação ao sistema ou modelo, logo deverá ser descartada ou simplificada.

Resolução analítica

Estamos a planear uma relação linear entre as variáveis *Y* e *X* que, por serem matrizes, nos permitem planear um sistema de equações lineares, que podemos resolver de forma analítica, tal como em qualquer outro problema de álgebra lineal.

Para resolver um sistema de equações lineares, resolvemos a equação seguinte:

$$Y = h_\Theta(x) = \Theta^T \times x$$
$$\Theta = (X^T \times X)^{-1} \times X^T \times Y$$

Figura 6

Código Latex:

```
Y=h_\Theta(x)=\Theta^T \times x \\
\Theta = (X^T \times X)^{-1} \times X^T \times Y
```

Contudo, surgem, com frequência, vários problemas que nos complicam a sua resolução ou a tornam impossível:

- A matriz resultante não invertível.
- O número de exemplos e características determina que a resolução, por este método, não seja viável por memória ou tempo de execução.
- Recordemos que as equações não são plenamente corretas: $Y \cong \Theta \, x \, X$
  - Estamos a tentar aproximar um modelo, mas pode não seguir realmente essa relação linear, podem faltar características, etc.
  - Os nossos dados podem não estar corretos a 100%.
  - Teremos sempre uma margem de erro ou aleatoriedade.

Função de custo

Modela o "custo" ou o erro acumulado, da hipótese proposta, para modelar este conjunto de dados sobre todos os exemplos.

O erro será a diferença entre o valor real de *Y*, para os exemplos do conjunto de aprendizagem, e a predição de *Y* para esses exemplos.

$$Y = h_\Theta(x) = \Theta^T \times x$$
$$J_\theta = \frac{1}{2m} \sum_{i=0}^{m} (h_\theta(x^i) - y^i)^2$$

Figura 7

Código Latex:

```
Y=h_\Theta(x)=\Theta^T \times x \\
J_\theta = \frac{1}{2m} \sum_{i=0}^{m} (h_\theta(x^i)-y^i)^2
```

Parâmetros:

- *m*: n.º de exemplos.
- *i*: índice do exemplo.
- *x*: vetor ou linha do exemplo.
- *y*: resultado conhecido para esse exemplo.

Assim, o objetivo de treinar o modelo é o de obter o valor de que minimize *J*, ou seja, que tenha o menor custo ou erro possível.

Resolução por métodos iterativos

Em vez de uma solução analítica, podemos resolver um sistema de equações lineares através de uma solução iterativa que, ao longo de inúmeras iterações, vai ficando cada vez mais próxima do valor ótimo.

Assim, comecemos por propor os valores iniciais de Θ aleatórios, calculamos a *J* para a Θ e continuamos a atualizar esses valores, aproximando-nos do valor ótimo, ao longo de múltiplas iterações.

Para isso, podemos utilizar qualquer algoritmo de otimização matemática.

Algoritmo de resolução

1. Compilar os exemplos, *X*.
2. Compilar os seus resultados conhecidos, *Y*.
3. Iniciar aleatoriamente os coeficientes/pesos de Θ.
4. Iterativamente, calcular o custo *J* para cada Θ e atualizar os seus valores.
5. Finalizar quando Θ convirja num valor final.