

# Inteligência Artificial

Raoni F. S. Teixeira

## Aula 7 - Regressão não-linear e Regularização

### 1 Introdução

A Figura 1 mostra um conjunto de treinamento  $\mathcal{T} = \{(x^{(1)}, y^{(1)}), \dots, (x^{(M)}, y^{(M)})\}$  com  $M = 10$  pontos extraídos de uma relação não linear entre as variáveis de entrada e saída,  $x$  e  $y$ . Esses dados foram gerados a partir da função  $\sin(2\pi x)$  com a adição de ruído aleatório.

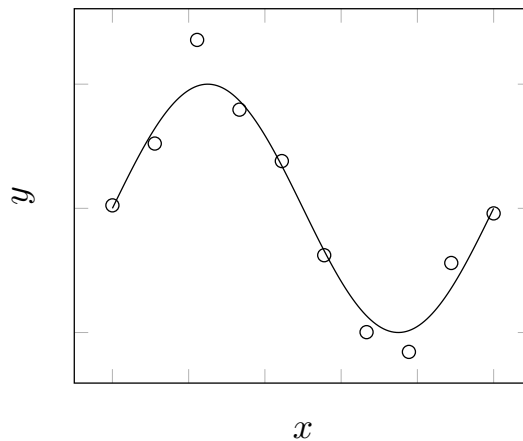


Figura 1: Conjunto de treinamento com  $M = 10$  pontos, mostrando uma relação não linear entre  $x$  e  $y$ . A curva representa a função  $\sin(2\pi x)$ , que descreve o relacionamento subjacente entre  $x$  e  $y$ .

Esta aula mostra como aplicar as ideias da anterior para prever relações não lineares.

## 2 Regressão Não Linear

Para lidar com a não linearidade nos dados, a hipótese  $h$  pode ser representada como um polinômio de grau  $K$ :

$$h_{\theta}(x) = \theta_0 + \theta_1 x + \theta_2 x^2 + \cdots + \theta_K x^K = \sum_{i=0}^K \theta_i x^i, \quad (1)$$

em que  $\theta \in \mathbb{R}^{K+1}$  é um vetor parâmetros e  $x^i$  é a entrada elevada à  $i$ -ésima potência.

Para visualizar que este polinômio é linear, definimos um mapeamento  $\phi$  como:

$$\phi(x) = \begin{bmatrix} 1 \\ x \\ x^2 \\ \vdots \\ x^K \end{bmatrix} \in \mathbb{R}^{K+1}. \quad (2)$$

No espaço transformado  $\phi(x)$ , o modelo torna-se linear:

$$h_{\theta}(x) = \underbrace{\theta^T \phi(x)}_{\text{combinação linear}}. \quad (3)$$

Embora linear nesse espaço, ele continua sendo não linear no espaço original.

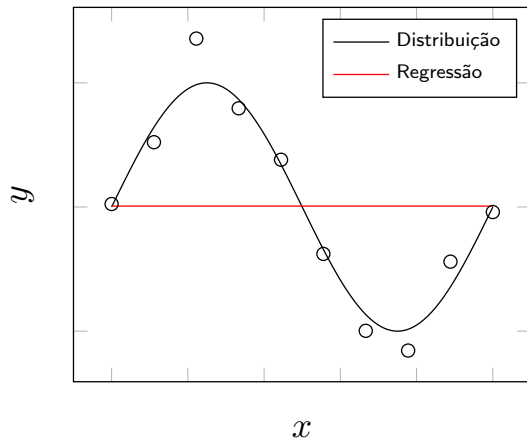
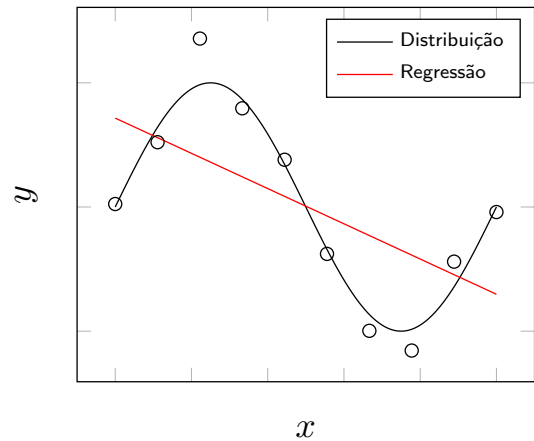
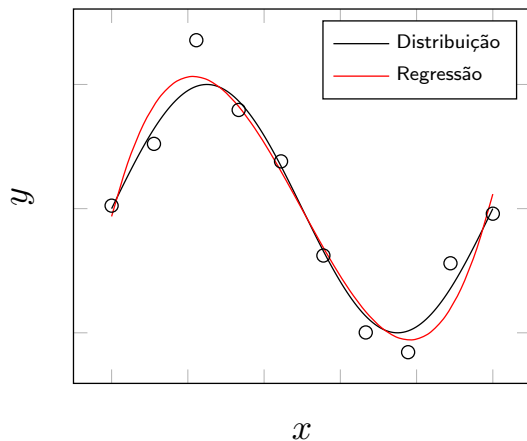
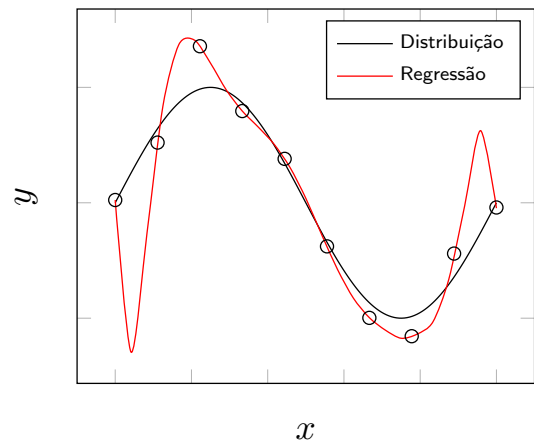
A escolha do mapeamento  $\phi$  é crítica e depende da aplicação e dos dados. Incluí-lo com polinômios ou outras combinações das variáveis permite capturar relações complexas, mas aumentar  $K$  excessivamente pode levar ao *overfitting* — quando um modelo estatístico se ajusta muito bem ao conjunto de treinamento, mas é ineficaz para prever novos resultados.

A Figura 2 ilustra regressões com polinômios de graus  $K = 0, 1, 3$  e  $9$ , baseados nos dados da Figura 1. Os polinômios constante ( $K = 0$ ) e de primeira ordem ( $K = 1$ ) apresentados nas Figuras 2a e 2b não se ajustaram bem à função  $\sin(2\pi x)$ . O polinômio de terceira ordem ( $K = 3$ ) da Figura 2c teve o melhor ajuste. Já o polinômio de ordem mais alta ( $K = 9$ ) da Figura 2d ajustou-se perfeitamente aos dados de treinamento, passando exatamente por cada ponto e resultando em custo  $J(\theta) = 0$ . No entanto, o modelo está em *overfitting* pois representa mal a função  $\sin(2\pi x)$ .

Para avaliar o desempenho do modelo em dados ainda não vistos, usamos a métrica em um conjunto de teste independente:

$$E_{\text{RMS}} = \sqrt{2 \frac{J(\theta^*)}{M}}, \quad (4)$$

em que  $J(\theta^*)$  é o custo mínimo do modelo após o treinamento e  $M$  é o número de amostras.

(a) Polinômio de grau  $K = 0$ .(b) Polinômio de grau  $K = 1$ .(c) Polinômio de grau  $K = 3$ .(d) Polinômio de grau  $K = 9$ .Figura 2: Comparação de ajustes de polinômios em diferentes graus ( $K = 0, 1, 3, 9$ ).

Esta métrica mede o erro na mesma escala dos valores previstos, tornando os modelos comparáveis independentemente do tamanho do conjunto de dados.

A Figura 3 mostra o resultado de  $E_{RMS}$  para conjuntos de treinamento e teste, considerando diferentes graus de polinômio. O erro no conjunto de teste indica o quão bem o modelo prevê novos valores de  $y$  com base em observações de  $x$ . Como vemos na Figura 3, valores pequenos de  $K$  produzem erros altos no conjunto de teste, pois esses polinômios são rígidos e não capturam as variações da função  $\sin(2\pi x)$ . Os polinômios com graus entre 3 e 8 produzem erros menores e consequentemente uma boa representação da função original.

Quando  $K = 9$ , o erro no conjunto de treinamento é zero. Isso ocorre porque o polinômio

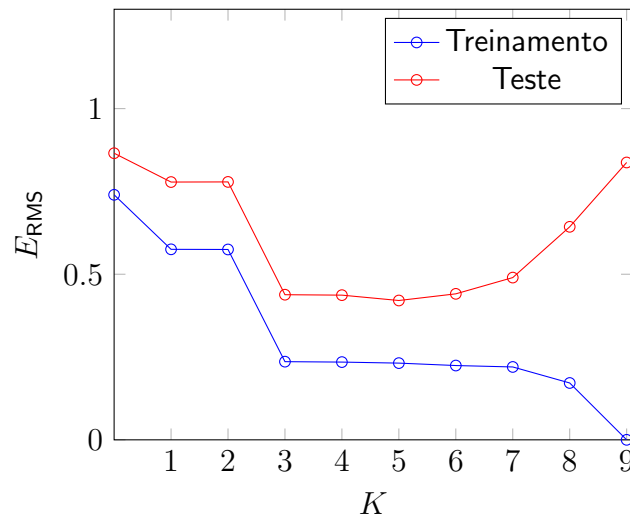


Figura 3: Raiz do erro quadrático médio para treinamento e teste e diferentes graus de polinômios. Tamanho do conjunto de treinamento  $M = 10$ .

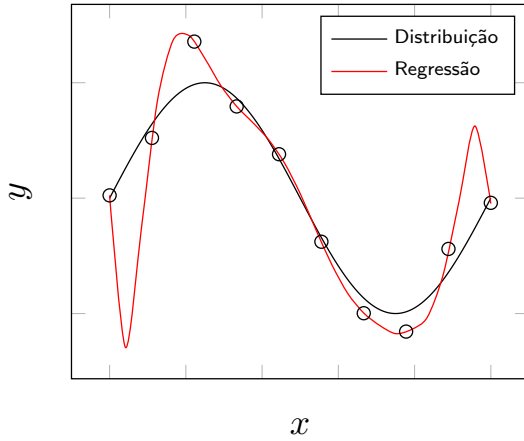
com os 10 coeficientes  $\theta_0, \dots, \theta_9$  tem graus de liberdade suficientes para ajustar-se perfeitamente aos 10 pontos do conjunto de treinamento. No entanto, o erro no conjunto de teste aumenta muito, pois a função resultante,  $h_\theta(x)$ , oscila excessivamente (veja Figura 2d).

Esse resultado pode parecer contraintuitivo: como um polinômio de ordem mais alta contém todas as ordens inferiores, era de se esperar que ele fosse pelo menos tão bom quanto um polinômio de ordem menor, como o de  $K = 3$  por exemplo. Além disso, poderíamos pensar que como a série de potências de  $\sin(2\pi x)$  inclui termos de todas as ordens, o desempenho do modelo deveria melhorar continuamente com o aumento de  $K$ .

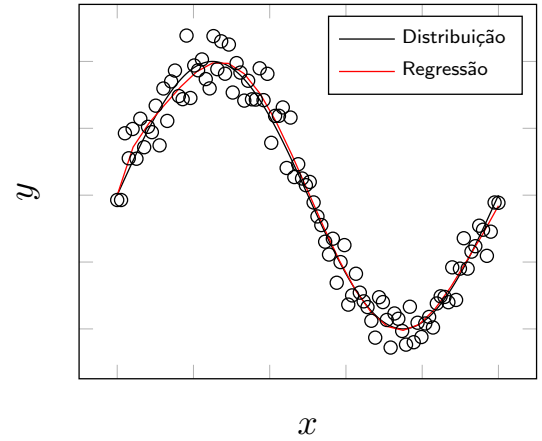
Também é útil analisar como o desempenho de um modelo muda conforme o tamanho do conjunto de dados varia, como mostra a Figura 4. Observe que quanto mais dados temos, mais complexo (ou flexível) o modelo pode ser e menos *overfitting*.

### 3 Regularização

A regularização é uma abordagem para lidar com alta dimensionalidade. Ela adiciona um termo de penalidade à função de custo, restringindo o crescimento excessivo dos coeficientes e prevenindo o *overfitting*.



(a) Polinômio de grau  $K = 9$  e quantidade de amostras  $M = 10$ .



(b) Polinômio de grau  $K = 9$  e quantidade de amostras  $M = 100$ .

Figura 4: Comparação de ajustes de polinômios com diferentes tamanhos de amostras ( $M = 10$  e  $M = 100$ ).

A regularização  $L2$  (ou *ridge regression*) modifica a função de custo da seguinte forma:

$$J(\theta) = \frac{1}{2M} \sum_{i=1}^M (h_{\theta}(x^{(i)}) - y^{(i)})^2 + \underbrace{\frac{\lambda}{2} \sum_{j=1}^n \theta_j^2}_{\text{regularização}}, \quad (5)$$

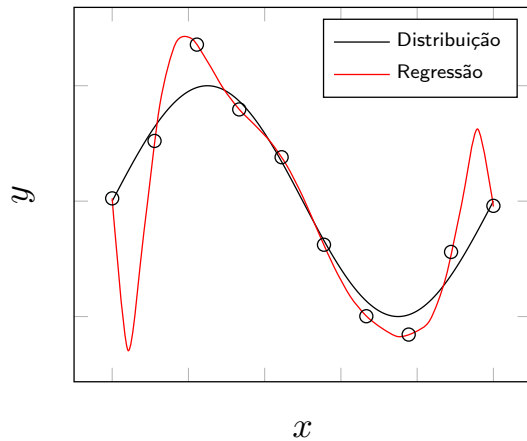
em que  $\lambda$  controla o peso da regularização. Valores maiores de  $\lambda$  aumentam a penalização dos coeficientes, favorecendo soluções mais simples. Valores muito pequenos de  $\lambda$  podem levar ao overfitting.

Outra regularização comum é o  $L1$  (ou *lasso regression*), que usa a norma  $L1$ :

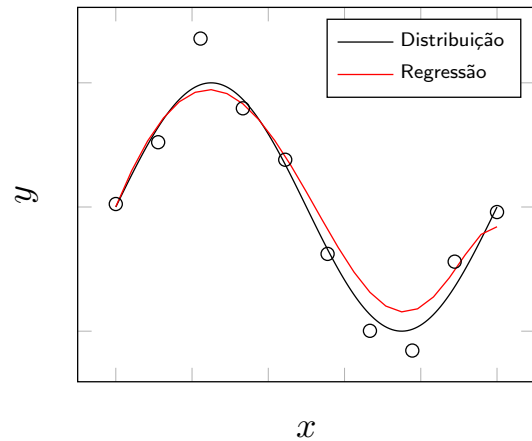
$$J(\theta) = \frac{1}{2M} \sum_{i=1}^M (h_{\theta}(x^{(i)}) - y^{(i)})^2 + \underbrace{\lambda \sum_{j=1}^n |\theta_j|}_{\text{regularização}}. \quad (6)$$

A regularização  $L1$  favorece soluções esparsas com muitos coeficientes valendo zero, o que facilita a interpretação do modelo e a seleção de características.

Essas técnicas ajudam a controlar a complexidade do modelo e melhoram sua capacidade de generalização, especialmente, em alta dimensionalidade e com poucas amostras. A Figura 5 mostra o efeito da regularização em um polinômio de grau  $K = 9$  com  $M = 10$  amostras. Ao contrário de sua contraparte sem regularização, o modelo regularizado se ajusta bem à função  $\sin(2\pi x)$ .



(a) Sem regularização



(b) Com regularização

Figura 5: Efeito da regularização no ajuste de um polinômio mantendo o grau ( $K = 9$ ) e quantidade de amostras ( $M = 10$ ).

|          | $\theta_9$ | $\theta_8$ | $\theta_7$ | $\theta_6$ | $\theta_5$ | $\theta_4$ | $\theta_3$ | $\theta_2$ | $\theta_1$ | $\theta_0$ |
|----------|------------|------------|------------|------------|------------|------------|------------|------------|------------|------------|
| Com Reg. | -5.97      | -2.17      | 2.32       | 6.30       | 7.70       | 3.96       | -5.77      | -14.09     | 7.55       | 0.00       |
| Sem Reg. | 5570.18    | -70859.78  | 136500.03  | -144864.80 | 92122.05   | -35595.19  | 8041.13    | -964.83    | 51.06      | -0.01      |

Tabela 1: Coeficientes dos polinômios de graus  $K = 9$  com e sem regularização.

A Tabela 1 evidencia o impacto da regularização nos coeficientes dos polinômios: os coeficientes de maior ordem ficam menores com a regularização. O coeficiente de grau 8,  $\theta_8$ , por exemplo, é 10 mil vezes maior no modelo não regularizado.

## 4 Exercícios

1. Como o mapeamento  $\phi$ , usado em regressão não linear, afeta a capacidade de o modelo capturar relações complexas entre as variáveis de entrada e saída?
2. Quais fatores determinam a complexidade de um modelo de regressão não linear? Como essa complexidade afeta o desempenho, especialmente em relação ao overfitting e à generalização?