



Introdução ao aprendizado de máquina 3

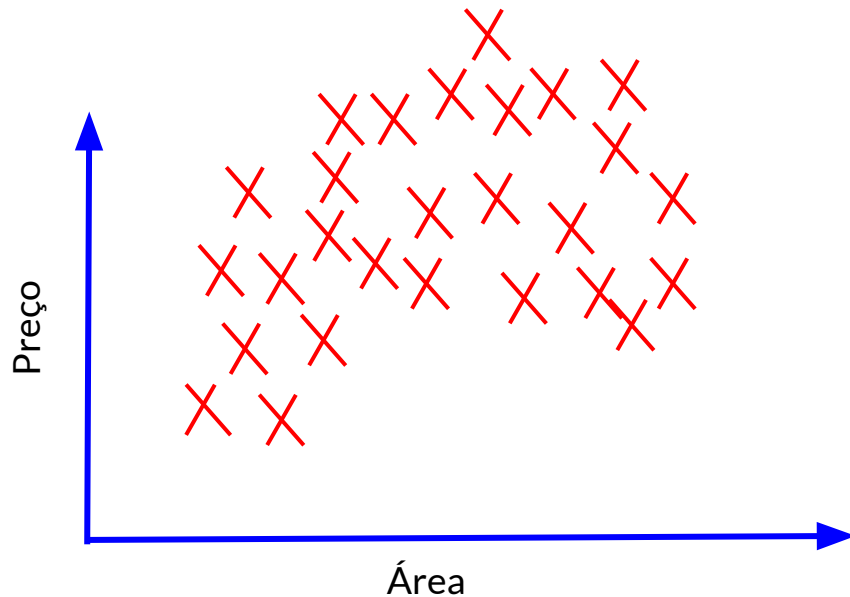
Aula 3 - Regressão linear: 1 variável, múltiplas variáveis, Lasso e Ridge, Causalidade

Como escolher o melhor modelo linear

Modelo linear pode ser escrito assim:

$$\hat{y}_i = \theta_0 + \theta_1 x_i$$

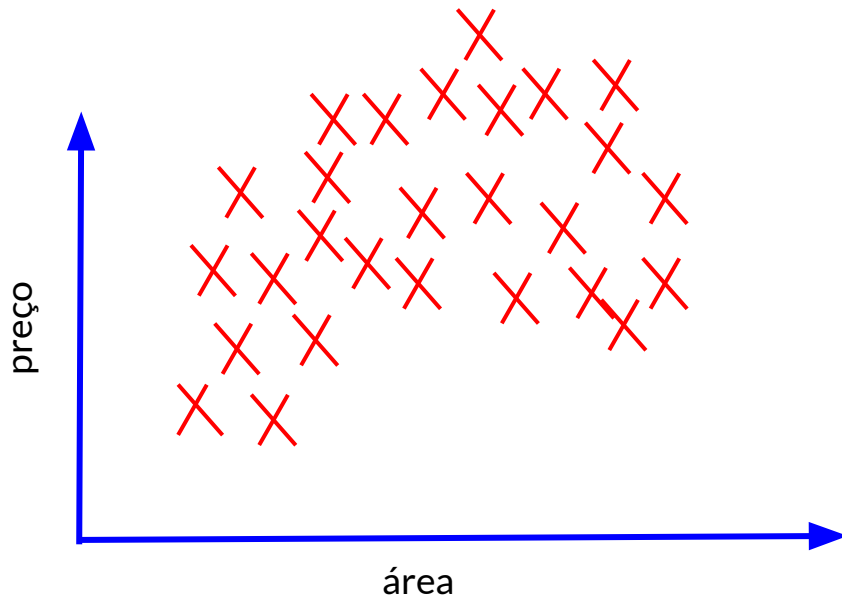
Observação	Área m ²	Preço R\$
1	80	640.000
2	160	1.200.000
...
N	100	900.000



Vamos usar o erro quadrático médio

$$\hat{y}_i = \theta_0 + \theta_1 x_i$$

$$EQM = \frac{1}{N} \sum_{i=1}^N (\hat{y}_i - y_i)^2$$



Quais são os parâmetros ótimos?



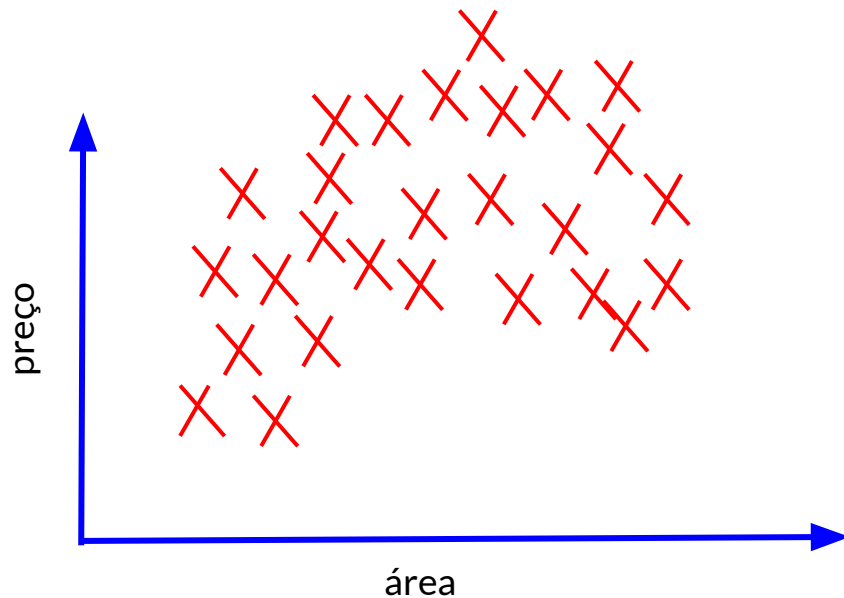
$$\min_{\theta_0, \theta_1} \frac{1}{N} \sum_{i=1}^N (\theta_0 + \theta_1 x_i - y_i)^2$$

Esse problema pode ser resolvido manualmente!!!

O resultado é intuitivo

$$\theta_0 = \bar{y} - \theta_1 \bar{x}$$

$$\theta_1 = \frac{cov(x,y)}{var(x)}$$





Matemática da regressão linear

Resolvendo o problema matematicamente




$$\min_{\theta_0, \theta_1} \frac{1}{N} \sum_{i=1}^N (\theta_0 + \theta_1 x_i - y_i)^2$$



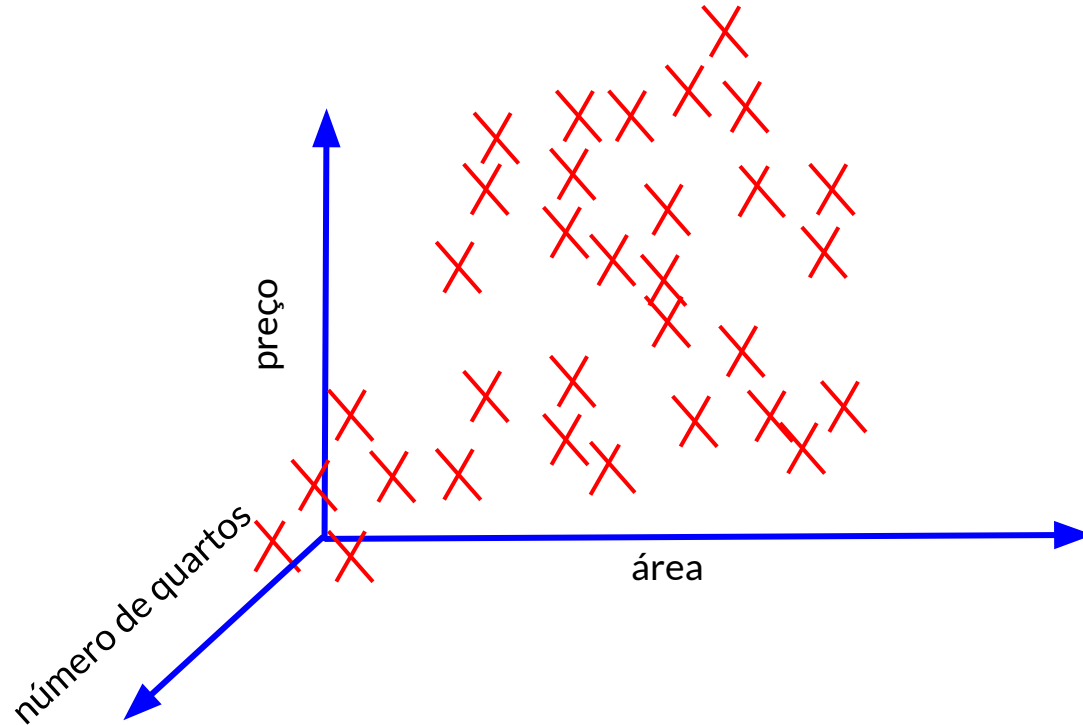
Múltiplas variáveis

Outras variáveis relevantes para o preço de casas



Observação	área x_1	Número de quartos x_2	piscina? x_3	Número de banheiros x_4	Qualidade das escolas x_5	preço y
1	80	2	0	1	6	650.000
2	160	4	1	2	9	1.600.000
...
N	100	3	0	2	8	900.000

É mais difícil de visualizar, mas a matemática ainda se aplica



Matemática do modelo linear com múltiplas variáveis



$$\min \frac{1}{N} \sum_{i=1}^N (\theta_0 + \theta_1 x_{1i} + \theta_2 x_{2i} + \dots + \theta_k x_{ki})^2$$

- Esse problema também pode ser resolvido manualmente
- Mas também podemos resolver usando o método do gradiente como vimos na aula passada.

Interpretação do modelo linear



$$\hat{\text{preço}} = 20.000 + 1.000 * \text{area}$$

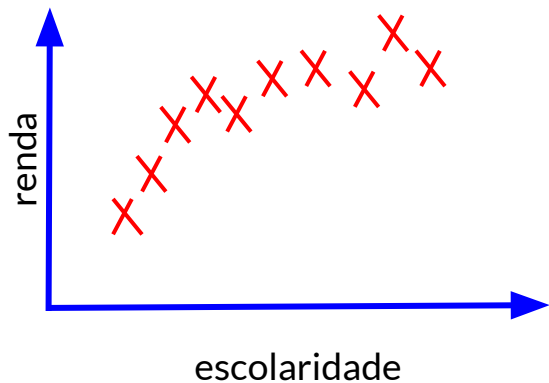
$$50.000 * \text{piscina} + 20.000 * \text{quartos}$$

$$10.000 * \text{banheiros} + 5.000 * \text{escolas}$$

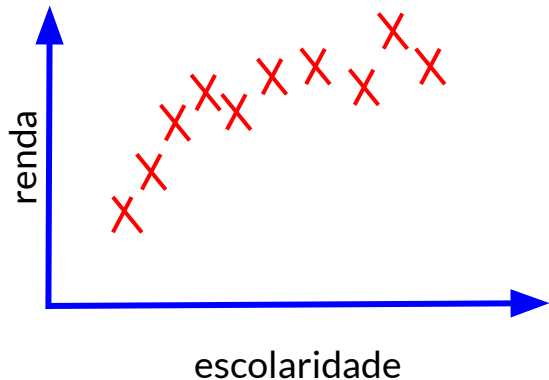
O modelo linear não é tão restritivo



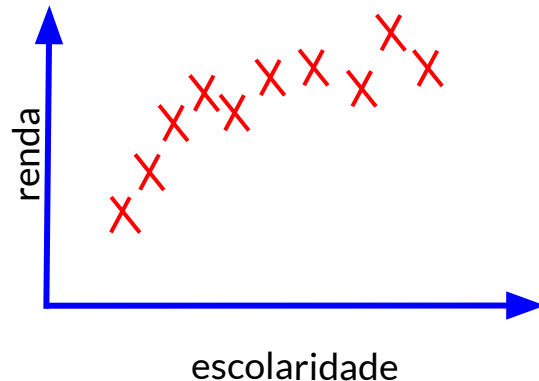
Regressão simples



Modelo logarítmico



Modelo polinomial





Métrica R2

R² é uma outra métrica de performance



Soma dos quadrados totais =

Soma dos quadrados da regressão =

Soma dos quadrados dos erros =

Prós e contras do R2



$$R^2 = 1 - \frac{\sum_{i=1}^N (y_i - \hat{y}_i)^2}{\sum_{i=1}^N (y_i - \bar{y})^2}$$

Pros:

- Intuitivo
- Fácil de calcular

Cons:

- Sempre aumenta com novas variáveis
- Ruim para selecionar modelos
- Sobreuso



Regularização e regressão Ridge

Regularização



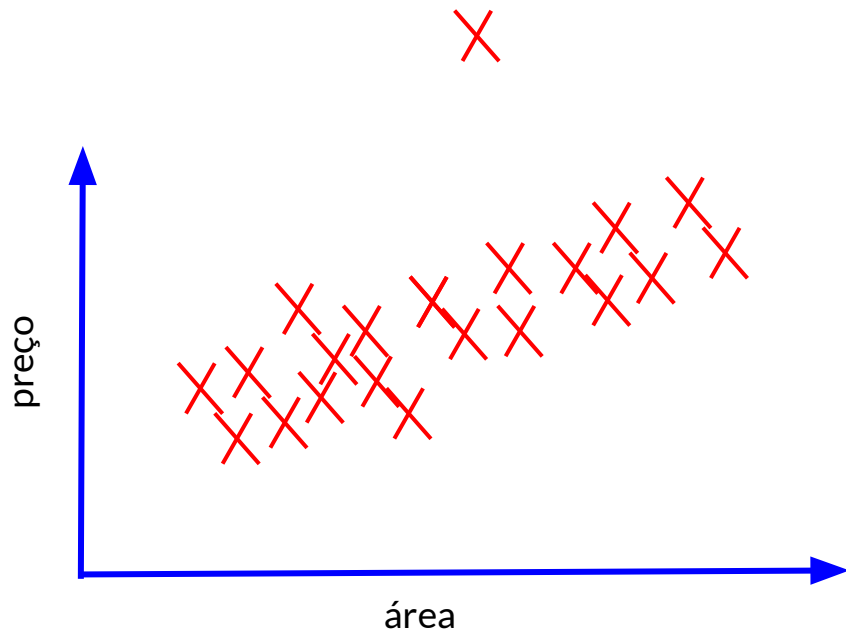
- Regularização é a técnica de restringir o modelo para torná-lo mais simples e reduzir o risco de sobre-ajuste.
- Modelos diferentes tem versões de regularização diferente, mas essa técnica é geralmente comum em vários modelos de ML.
- Na regressão linear, técnicas de regularização incluem a regressão Ridge e a regressão Lasso.

Fórmula da Regressão Ridge



Visualizando a regressão Ridge

$$\min_{\theta_0, \theta_1} \frac{1}{N} \sum_{i=1}^N (\theta_0 + \theta_1 x_i - y_i)^2 + \lambda(\theta_0^2 + \theta_1^2)$$



λ - O parâmetro de punição

- Lambda é um outro hiperparâmetro, a taxa de punição
- Quanto maior for lambda, mais o algoritmo punirá coeficientes muito grandes
- Quanto maior for lambda, menos provável que haja sobreajuste e mais provável o subajuste
- Assim como outros hiperparâmetros, nós escolhemos o lambda ótimo que leva a melhor performance no conjunto de validação



Regressão Lasso

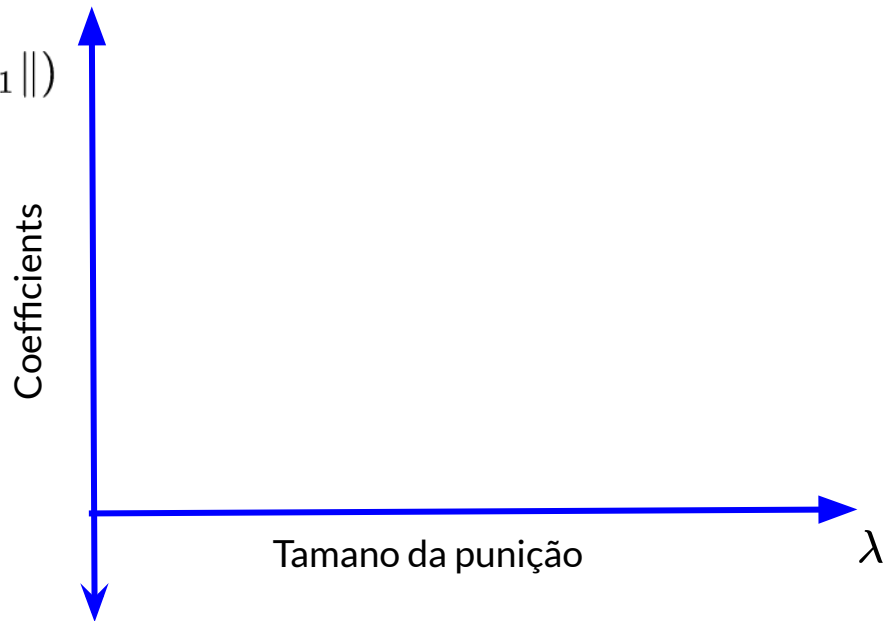
Fórmula da regressão Lasso



Visualizando a trajetória dos coeficientes

$$\min_{\theta_0, \theta_1} \frac{1}{N} \sum_{i=1}^N (\theta_0 + \theta_1 x_i - y_i)^2 + \lambda(\|\theta_0\| + \|\theta_1\|)$$

- Lasso Regression ajuda a prevenir o sobreajuste
- Também ajuda a selecionar variáveis
- Esse algoritmo força as variáveis menos relevantes para zero

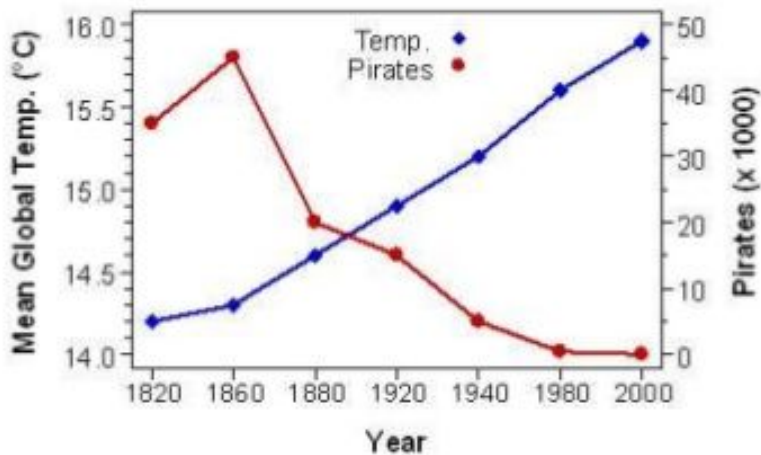




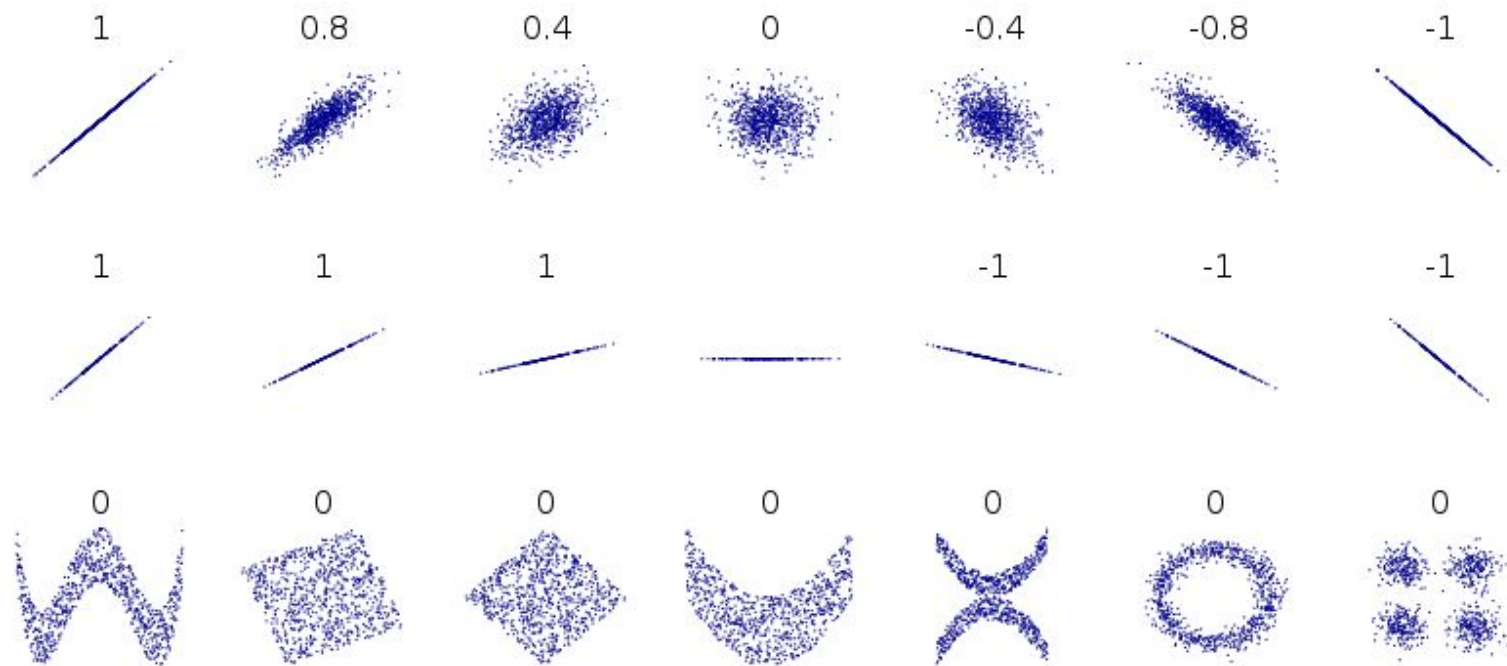
Correlação e causalidade

Correlação e causalidade

Temperatura Global X Piratas no mundo



Exemplos de correlação



Confundindo correlação e causalidade



- Casas mais caras são alugadas mais rapidamente em uma plataforma, vamos aumentar os preços das casas.
- Motoristas que dirigem mais quando ganham menos, eles são irracionais, vamos pagar menos.
- Clientes que comprem pela segunda vez tem um valor de tempo de vida, vamos dar o segundo item de graça.