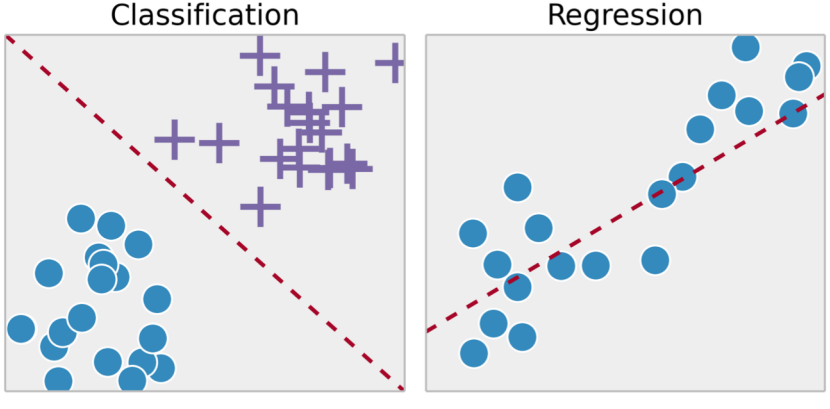
Regressão

Jones Granatyr

Regressão linear

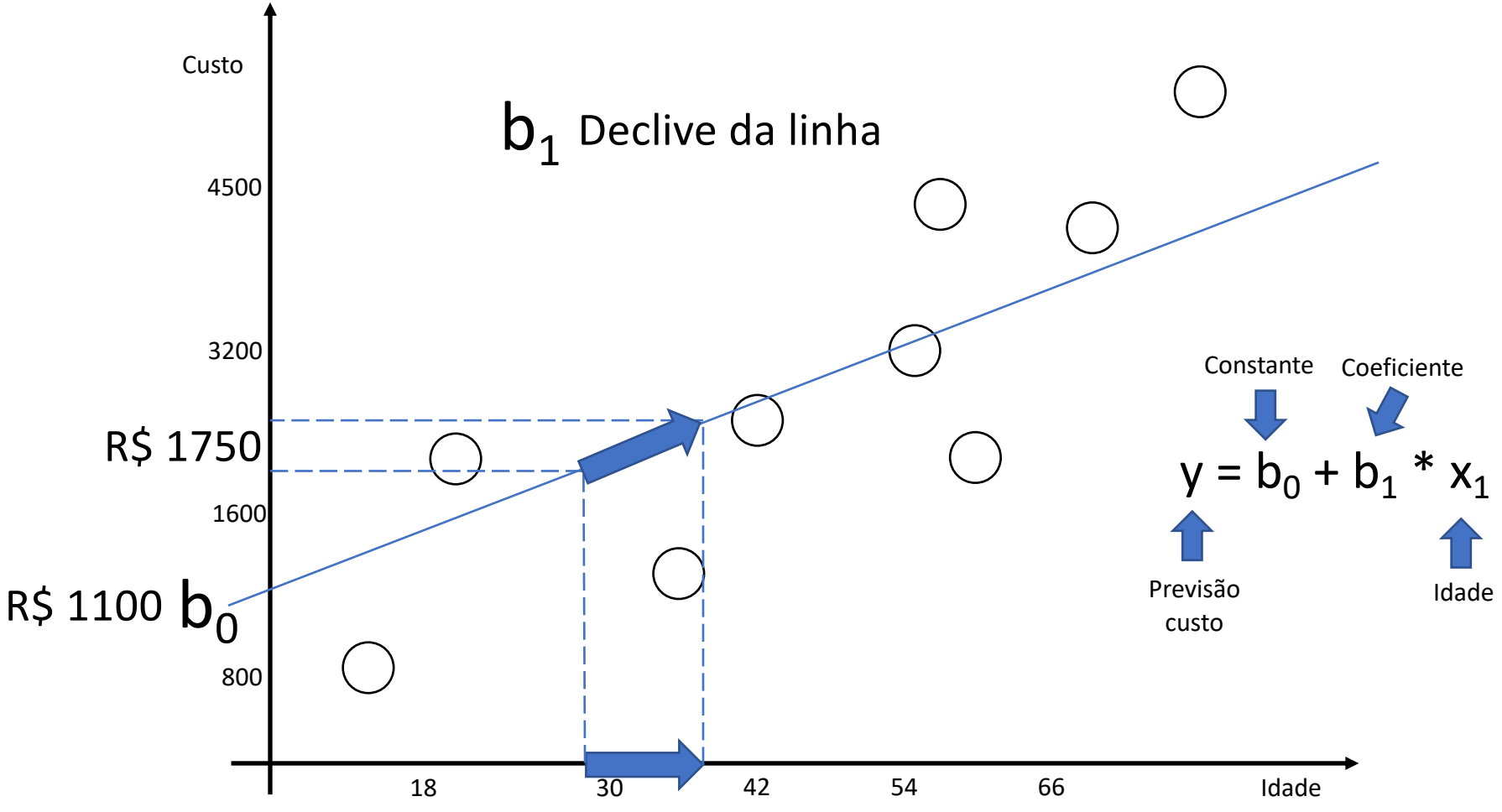
- Modelagem da relação entre variáveis numéricas (variável dependente **y** e variáveis explanatórias **x**)
- Temperatura, umidade e pressão do ar (x) → velocidade do vento (y)
- Gastos no cartão de crédito, histórico (x) → limite do cartão (y)
- Idade (x) → custo plano de saúde (y)
- Tamanho da casa (x) → preço da casa (y)



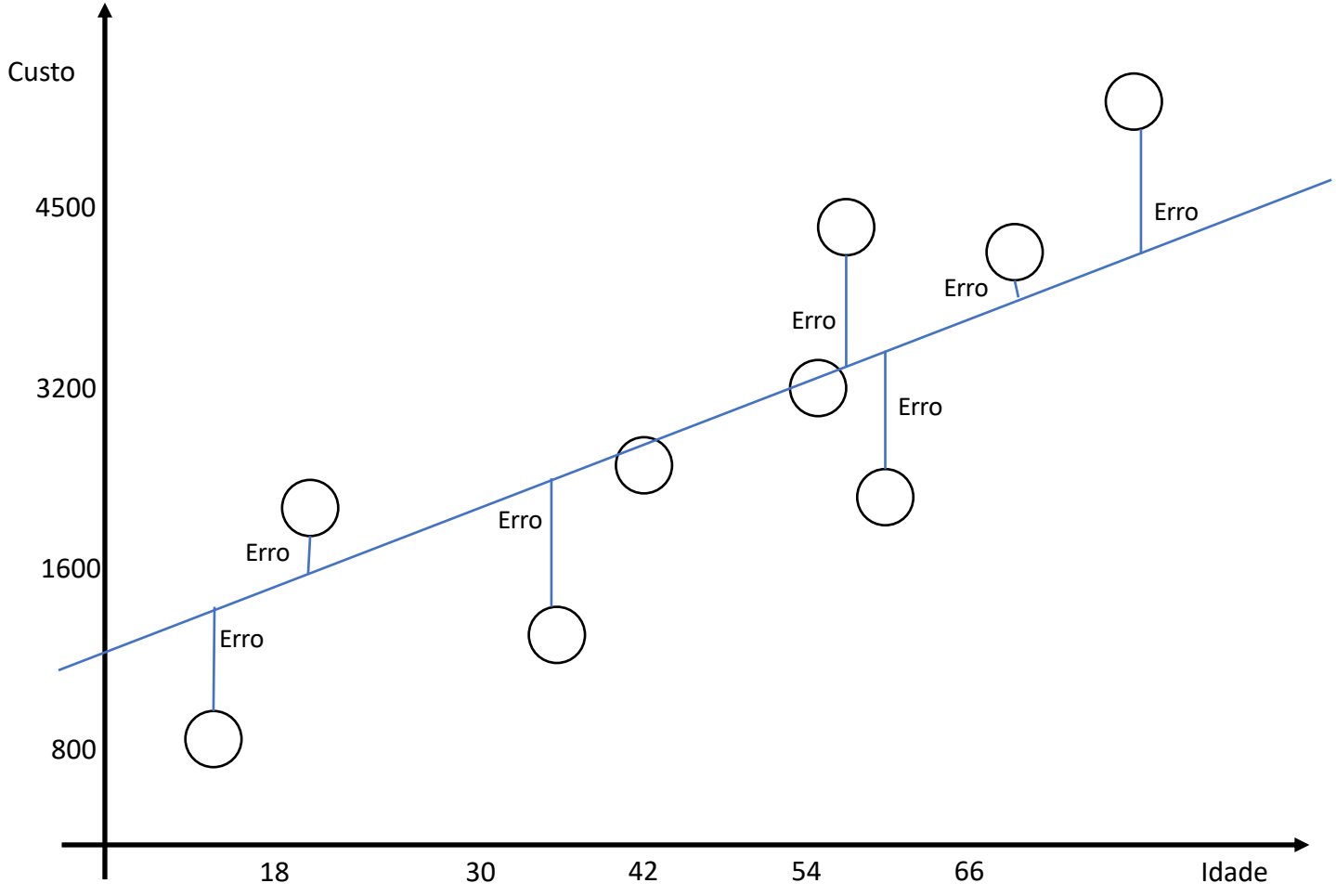
Regressão linear

Relação linear entre os atributos: quanto maior a idade, maior o custo

b_0 e b_1 definem a localização da linha (treinamento)



Regressão linear



Mean square error (MSE)

$$MSE = \frac{1}{N}\sum_{i=1}^{N}(f_i - y_i)^2$$

Preço real	Preço calculado	Erro
150	180	$(150 - 180)^2 = 900$
60	55	$(60 - 55)^2 = 25$
220	230	$(220 - 230)^2 = 100$
45	67	$(45 - 67)^2 = 484$

Soma = 1.509

MSE = 1.509 / 4 = 377,25

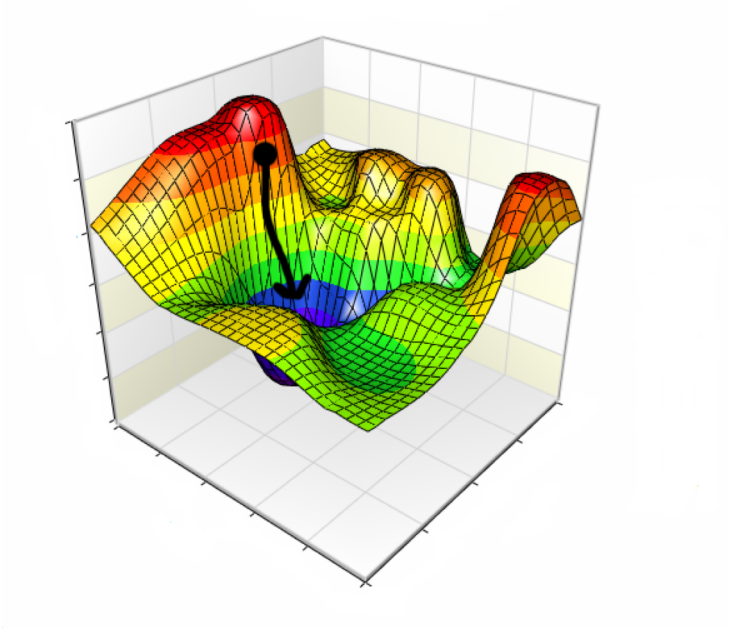
$y = b_0 + b_1 * x_1$ Objetivo: ajustar os parâmetros b_0 e b_1 para ter o menor erro!

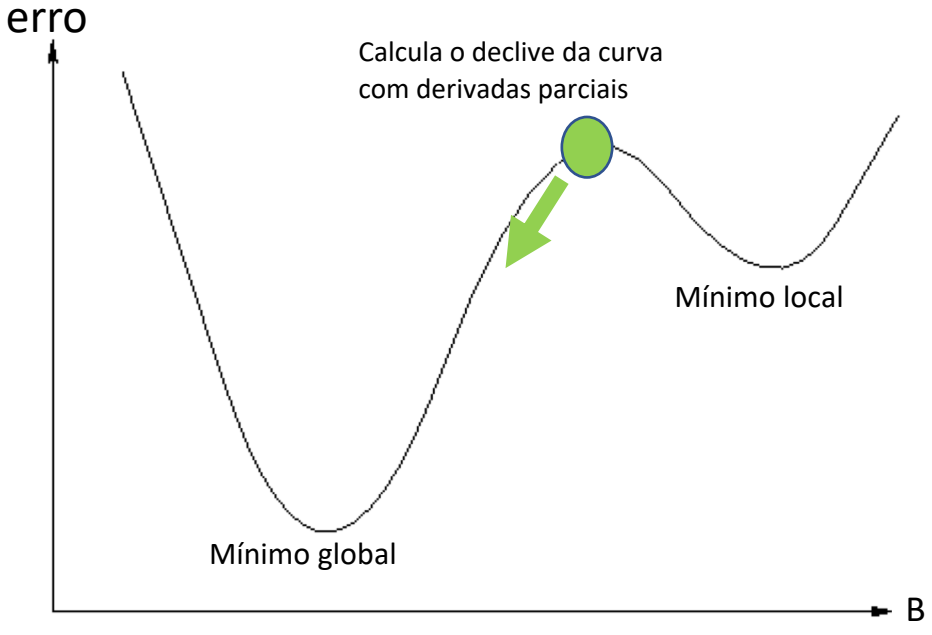
Regressão linear – ajuste dos parâmetros

- Design matrix (Álgebra Linear)
  - Bases de dados com poucos atributos
  - Inversão de matrizes que tem um custo computacional alto
- Gradient descent (descida do gradiente)
  - Desempenho melhor com muitos atributos

Descida do gradiente

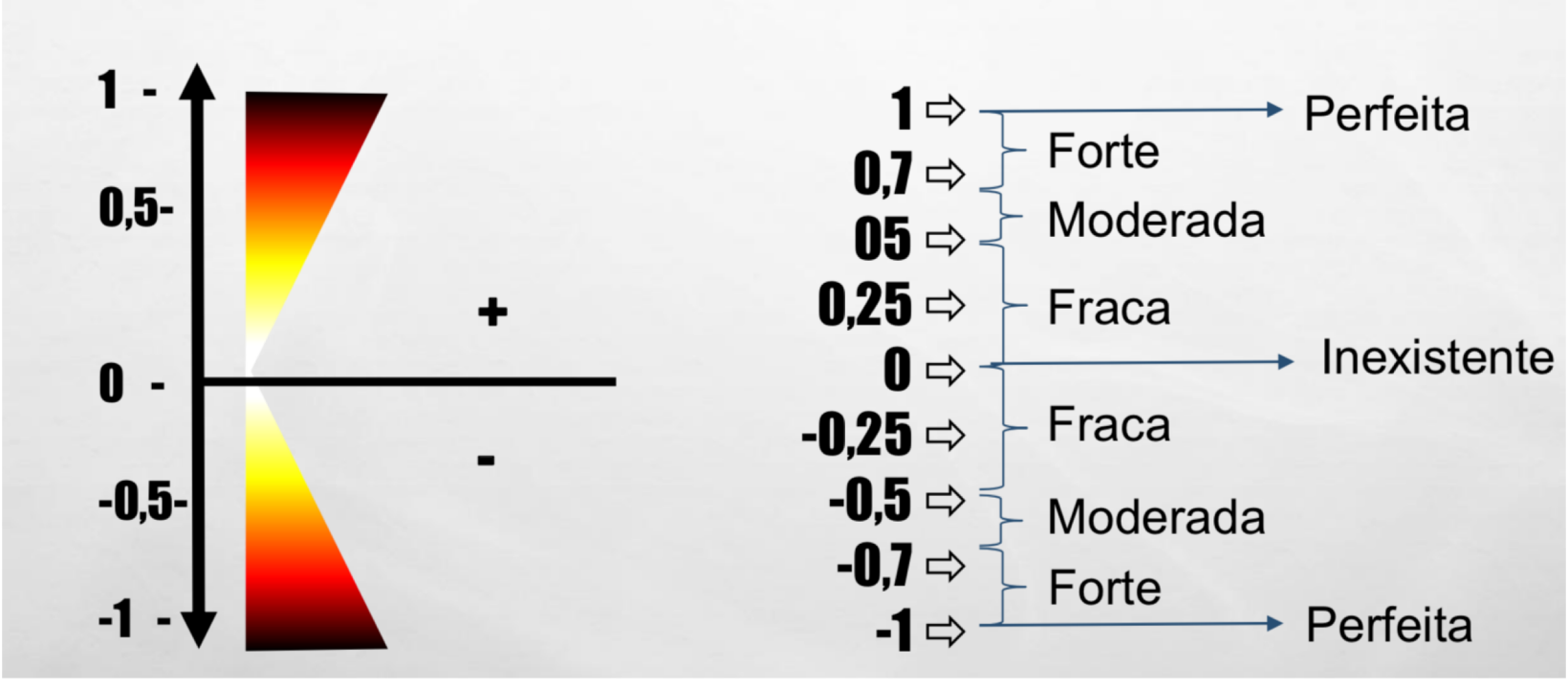
min $C(B_1, B_2 \ldots B_n)$
Taxa de aprendizagem





Correlação

- Força e direção da relação entre variáveis (valores entre -1 e 1)



Fonte: Curso Formação Cientista de Dados com R e Python

Regressão linear múltipla

$$y = b_0 + b_1 * x_1$$

$$y = b_0 + b_1 * x_1 + b_2 * x_2 + \ldots + b_n * x_n$$

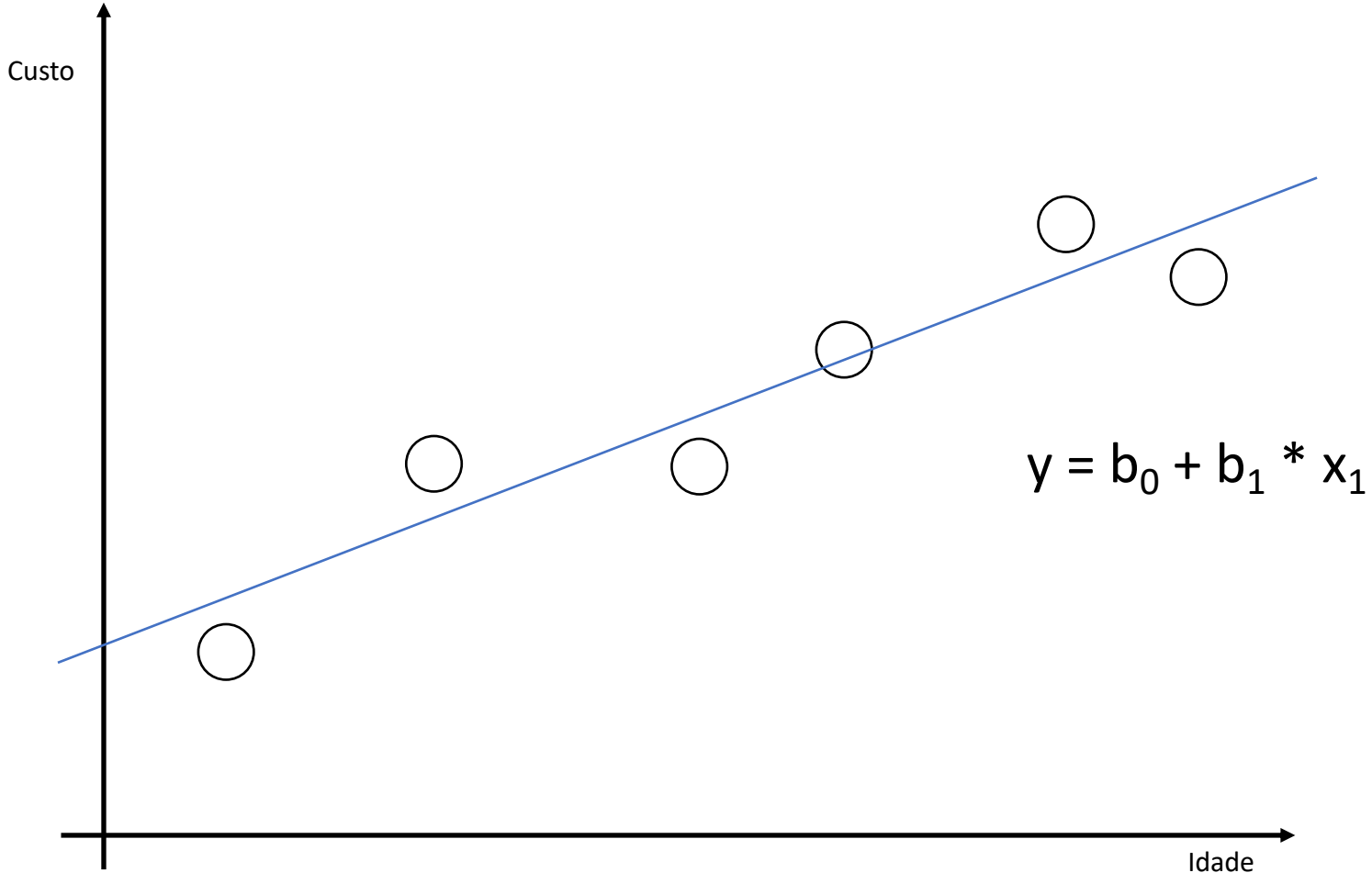
Regressão linear polinomial

$$y = b_0 + b_1 * x_1$$

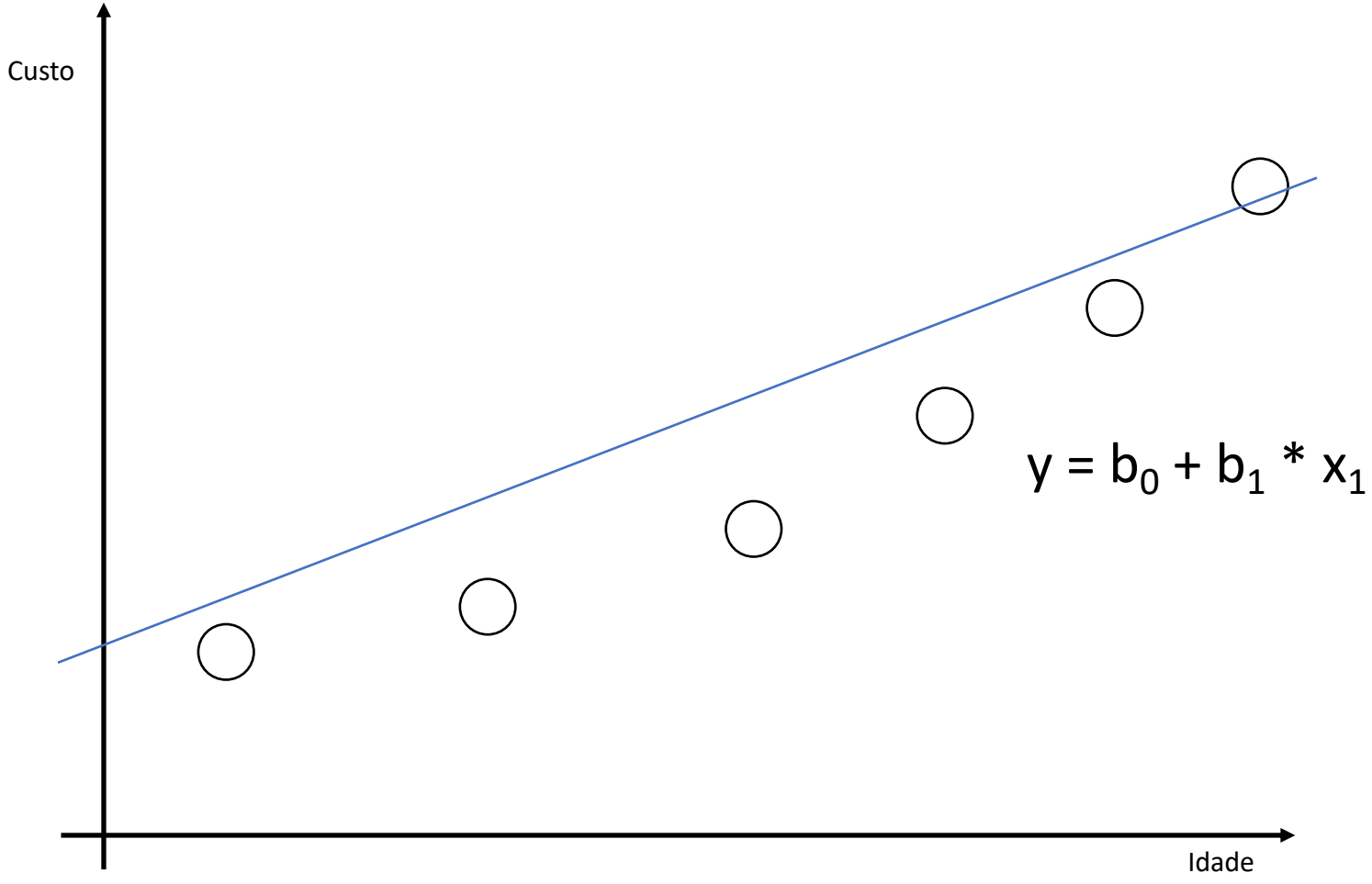
$$y = b_0 + b_1 * x_1 + b_2 * x_2 + \ldots + b_n * x_n$$

$$y = b_0 + b_1 * x_1 + b_2 * x_1^2 + \ldots + b_n * x_1^n$$

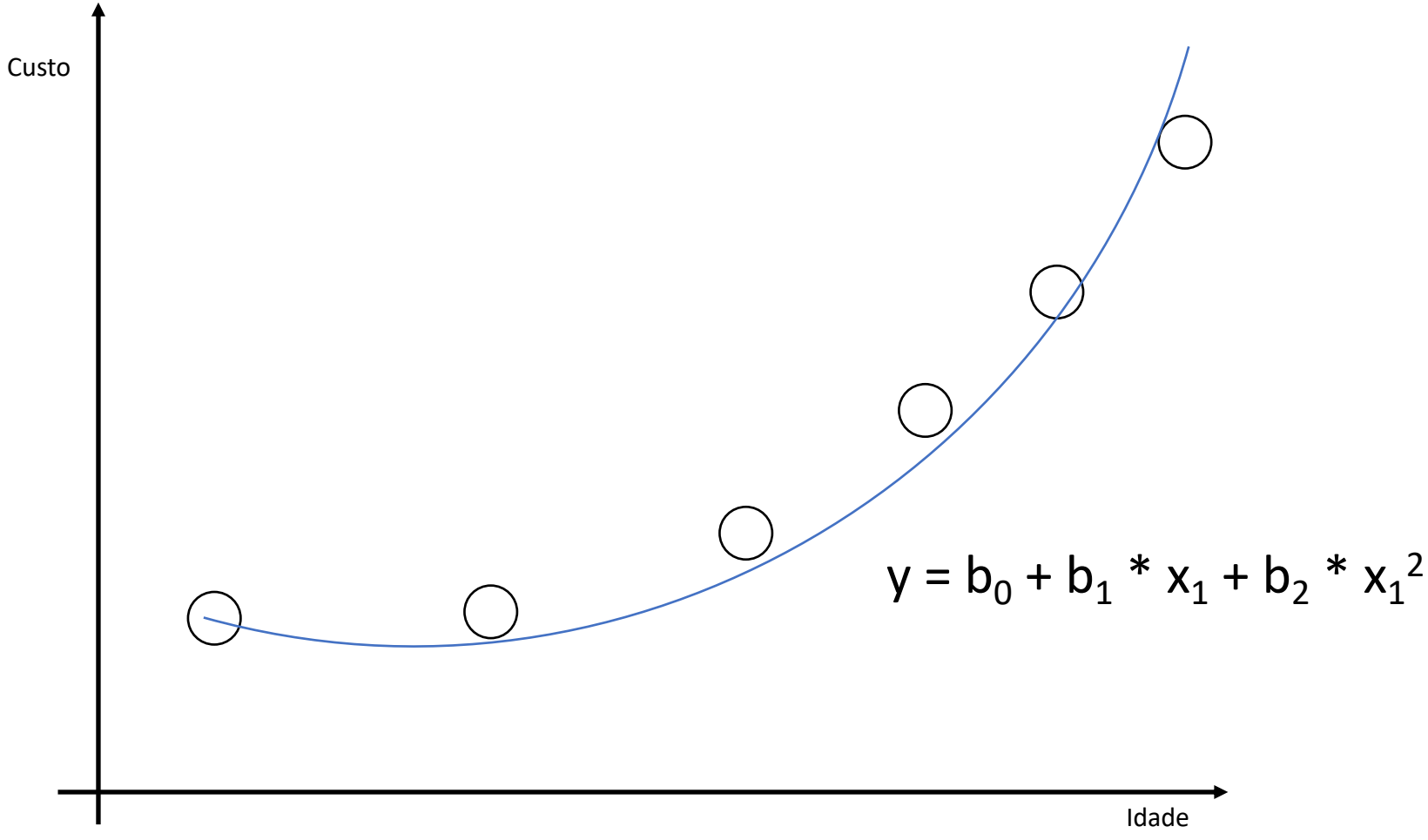
Regressão linear simples



Regressão linear simples

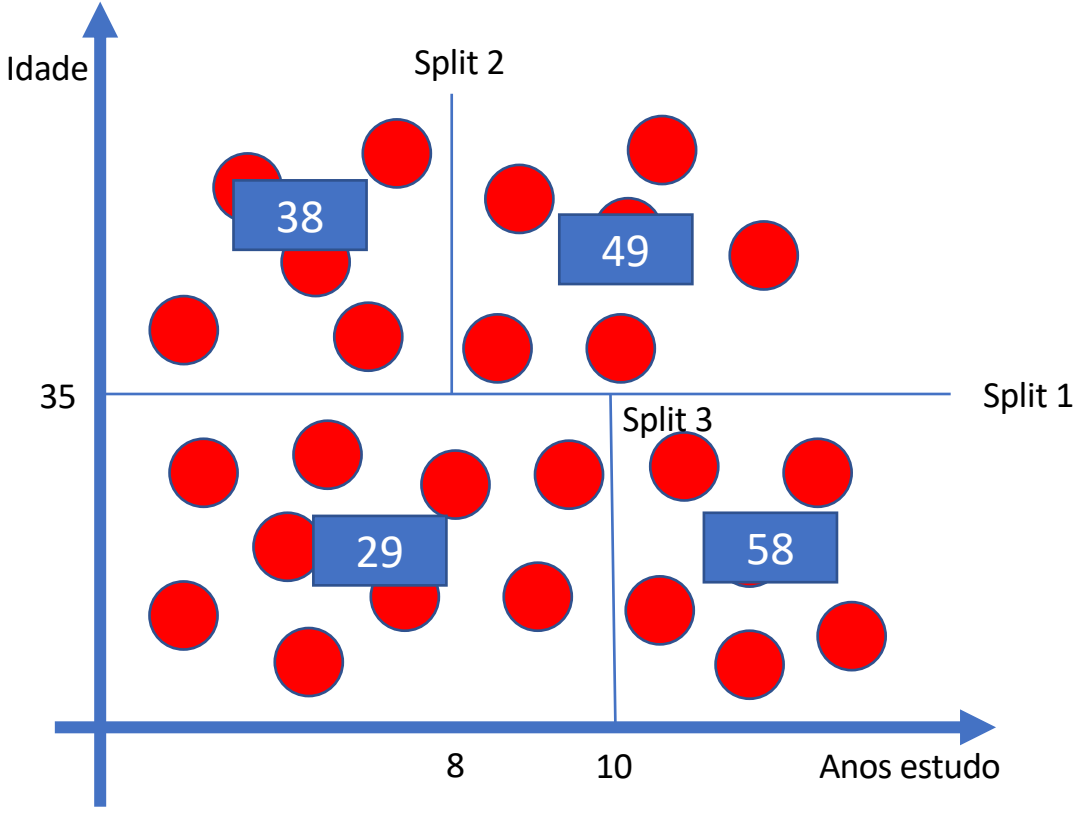


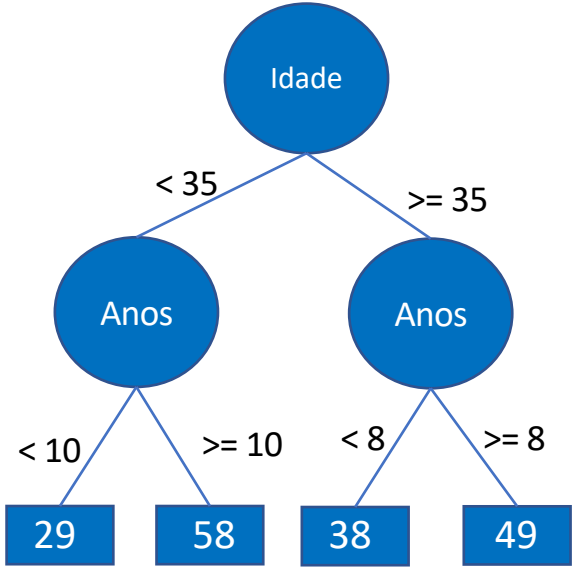
Regressão linear polinomial



Árvores com regressão

Baseado na idade e nos anos de estudo, prever o salário

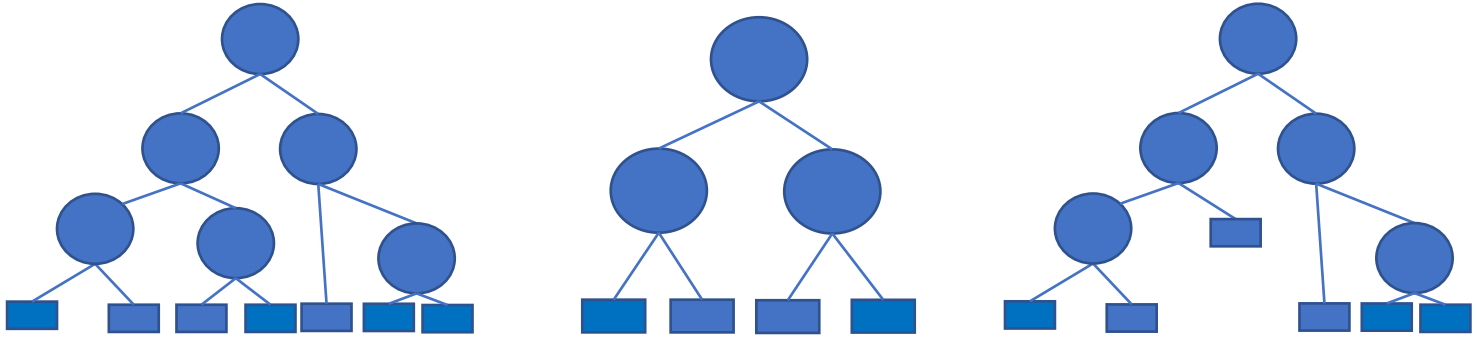




Random Forest (floresta randômica)

Random Forest

- Ensemble learning (aprendizagem em conjunto)
  - “Consultar diversos profissionais para tomar uma decisão”
  - Vários algoritmos juntos para construir um algoritmo mais “forte”
  - Usa a média (regressão) ou votos da maioria (classificação) para dar a resposta final



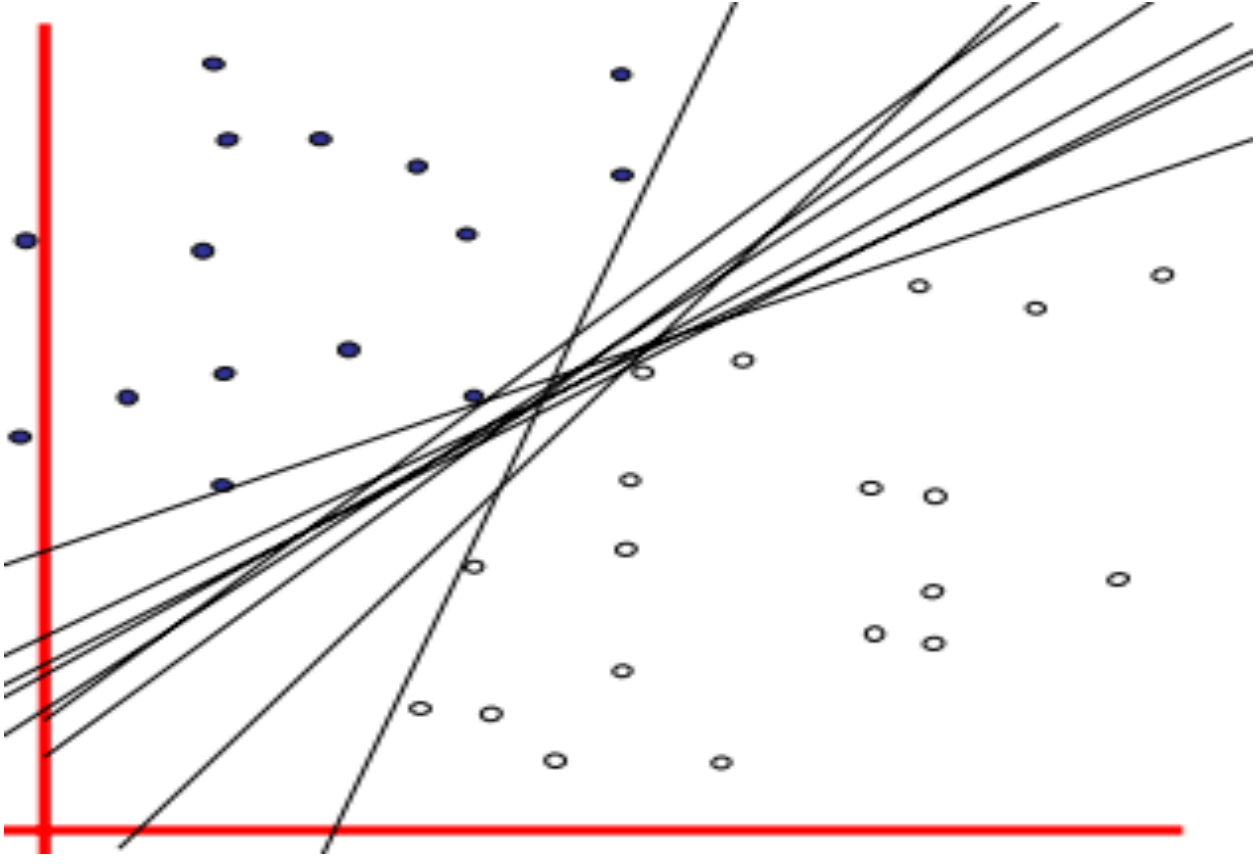
R$ 1.500 R$ 1.300 R$ 1.700

Média = R$ 1.500

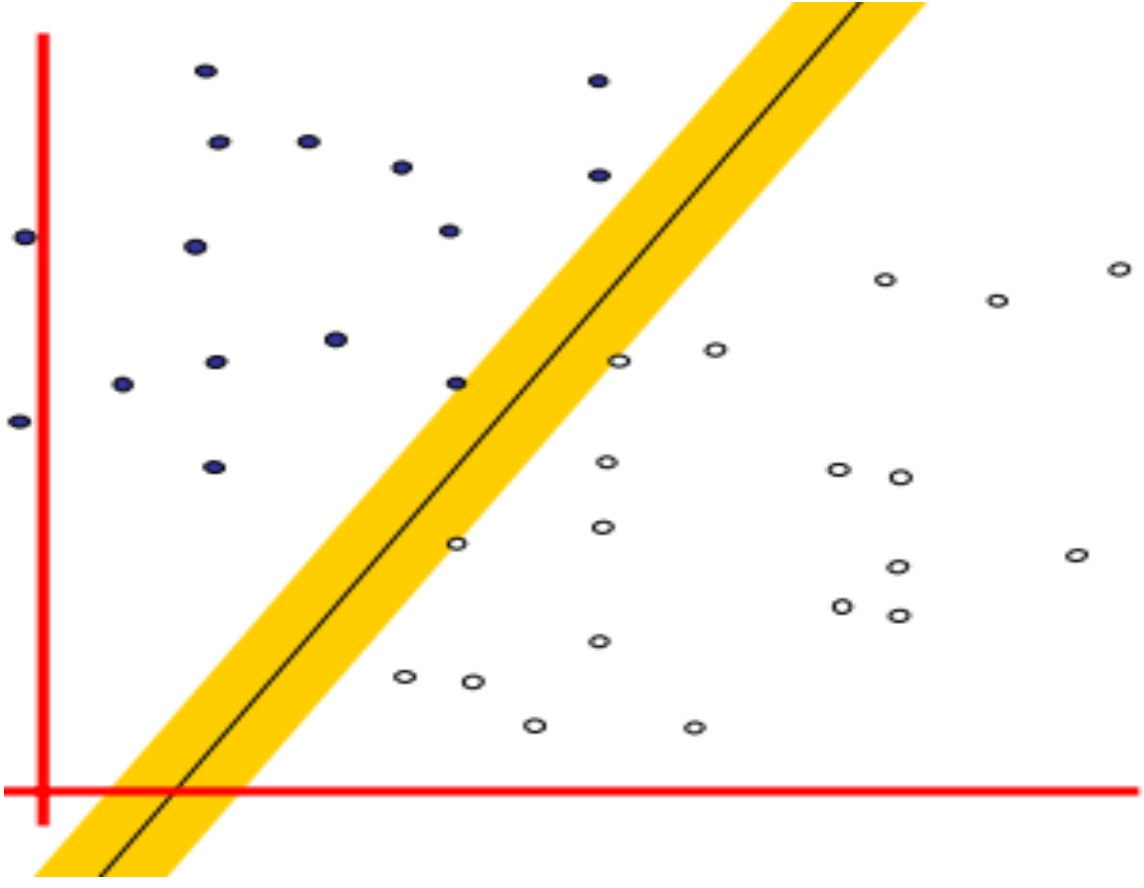
SVR (Support vector regression)

- Mantém as mesmas características das máquinas de vetores de suporte para classificação
- Mais difícil para fazer as previsões por se tratar de números (muitas possibilidades)
- Parâmetro **epsilon**
  - Penalidade do treinamento (distância para o valor real)

Qual o melhor hiperplano?



Margem máxima

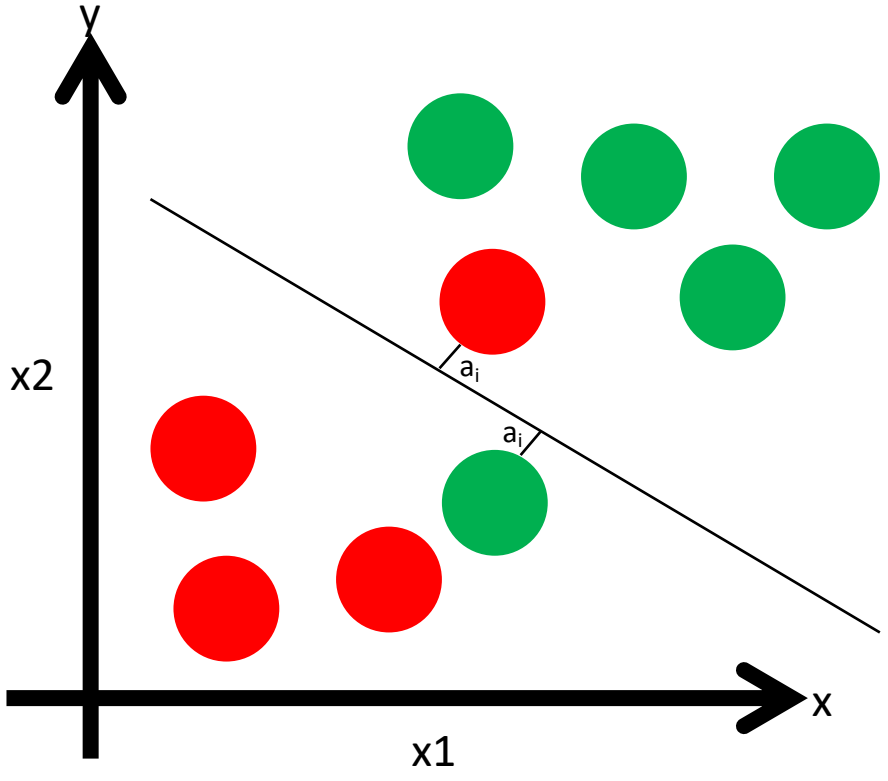


Erros e custo

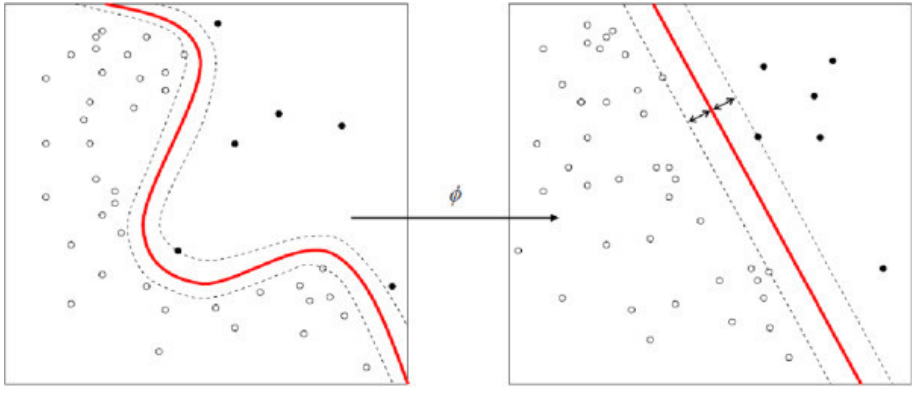
$$\frac{1}{2}\,|w|^2 + c\sum_i a_i$$

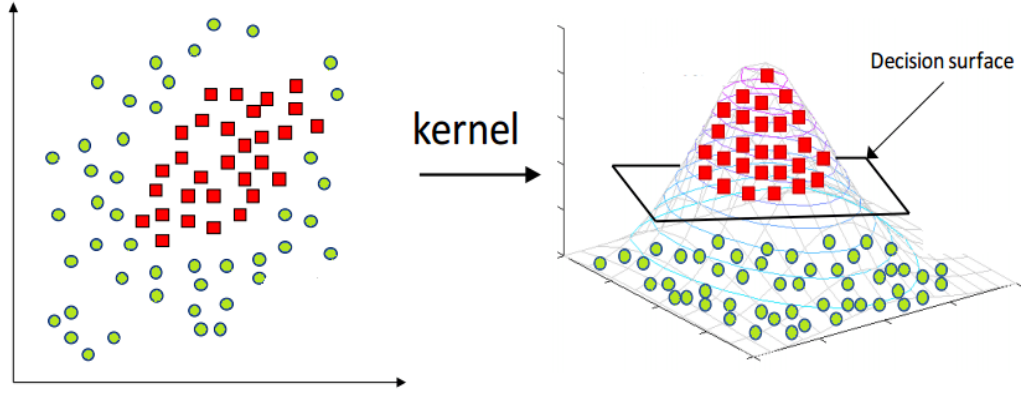
c= punição por previsão incorreta

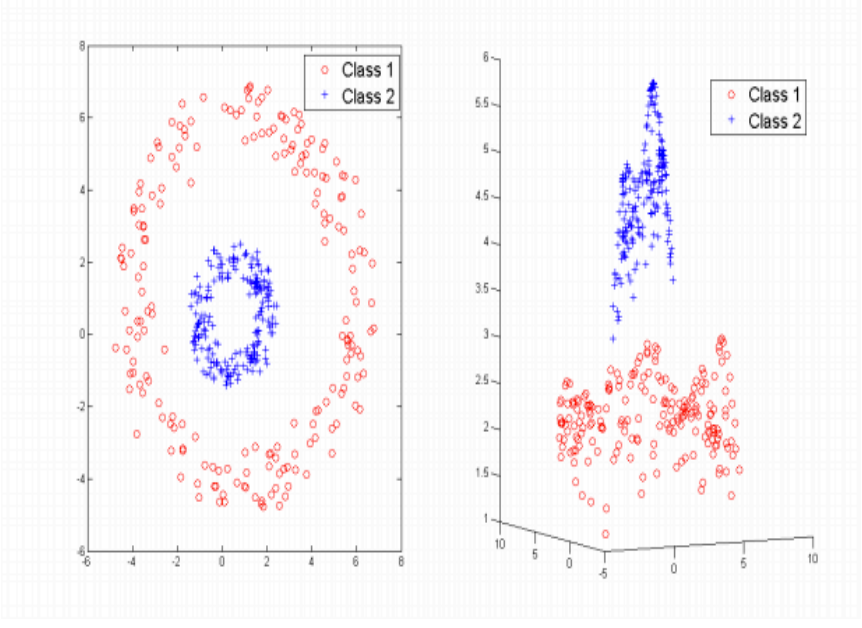
c alto = tenta 100% de separação
c baixo = permite mais erros



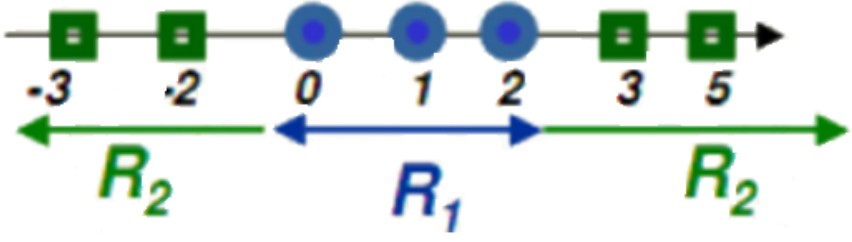
SVMs não lineares (Kernel Trick)

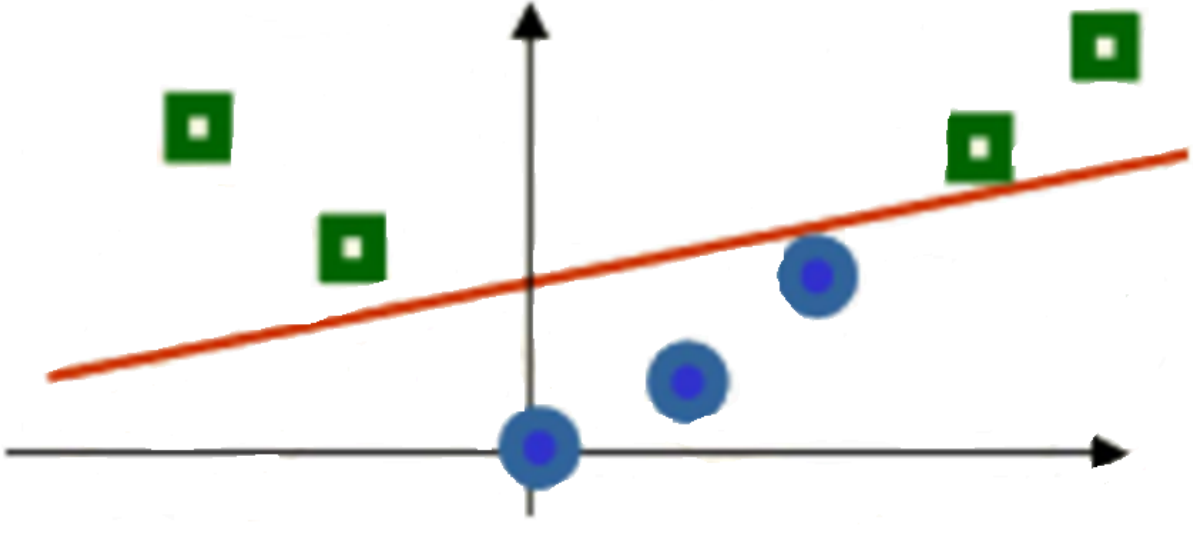






SVMs não lineares (Kernel Trick)





Conclusão