Aprendizado de Máquina

Aula 7: Algoritmos baseados em probabilidade (parte 3)

André C. P. L. F de Carvalho
ICMC/USP
andre@icmc.usp.br

Tópicos

- Regressão linear
- Regressão polinomial
- Algoritmo de ajuste de parâmetros

Sumarizando

- Regressão linear simples
- Função hipótese: $h_w(x) = w_0 + w_1 x$ ou $f(x) = w_0 + w_1 x$
- Parâmetros: w_0 , w_1
- Função custo: $J(w_0, w_1) = \frac{1}{n}\sum_{i=1}^{n}\left(y^i - f(x^i)\right)^2$
    - Para simplificar cálculos, pode ser usada a função $J(w_0, w_1) = \frac{1}{2n}\sum_{i=1}^{n}\left(y^i - f(x^i)\right)^2$
- Objetivo: $\min_{w_0 w_1} J(w_0, w_1)$

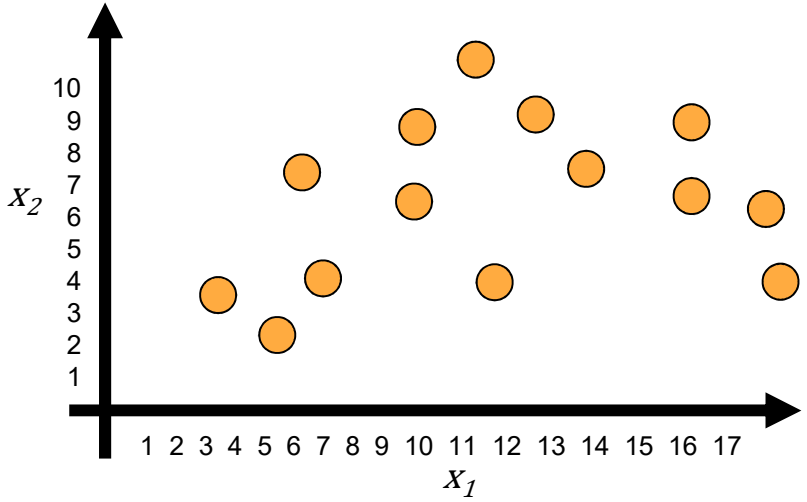
Regressão linear

- Funciona bem se relação entre variáveis independentes e variável dependente for linear
- Ex.: $f(x) = -0{,}25x + 3$
- O que ocorre se a relação for não linear?
    - Regressão linear não será capaz de encontrar um bom ajuste dos dados a uma função linear
    - Solução: usar regressão polinomial
        - Busca ajustar uma curva polinomial aos dados, para minimizar função de custo

          Ex.: $f(x) = x^4 - 4x^2 + 2x + 4$

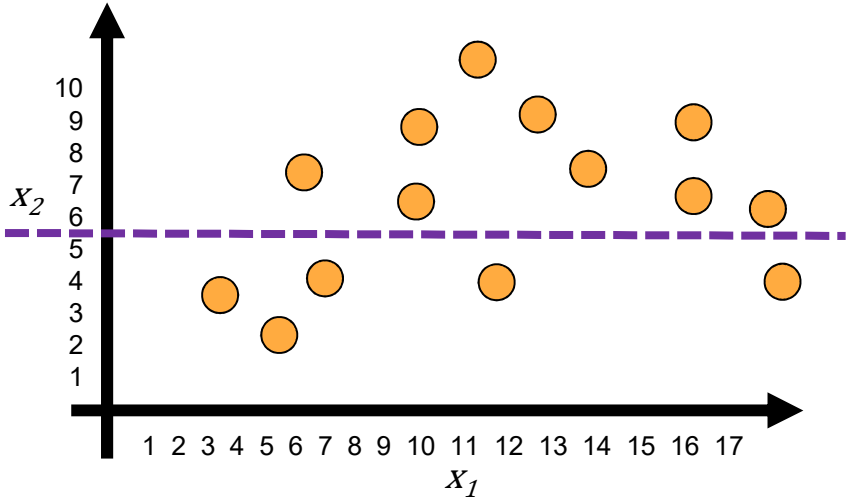
Regressão linear

- Busca uma boa função hipótese de regressão para os dados



Regressão linear

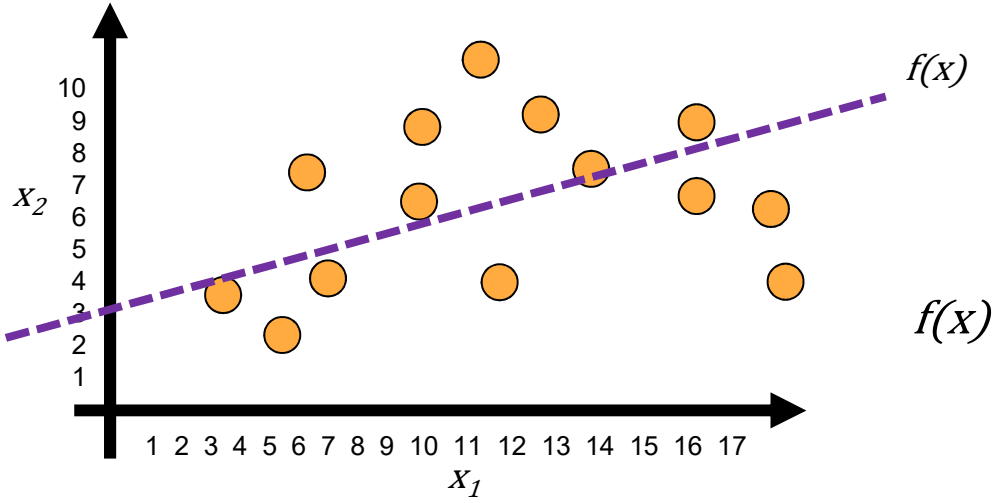
- Busca uma boa função hipótese de regressão para os dados
  - Função constante



Underfitting

Regressão linear

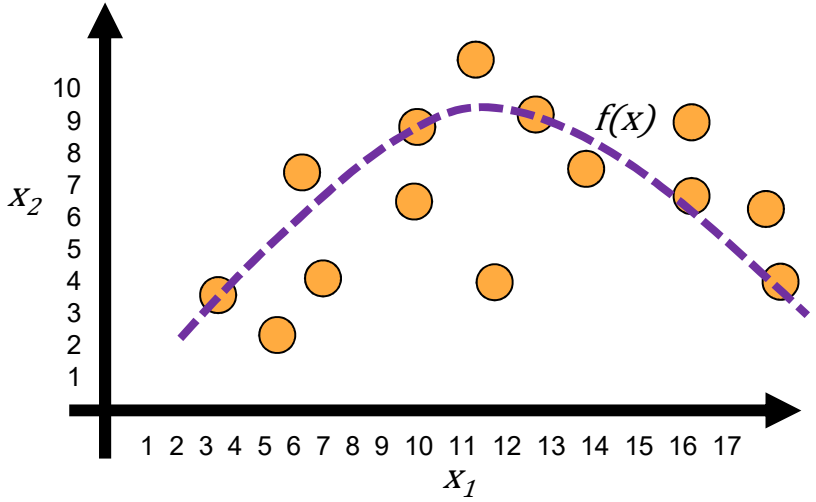
- Achar uma boa função hipótese de regressão para os dados
  - Tentar regressão linear



$$f(x) = -0{,}25x + 3$$

Regressão linear

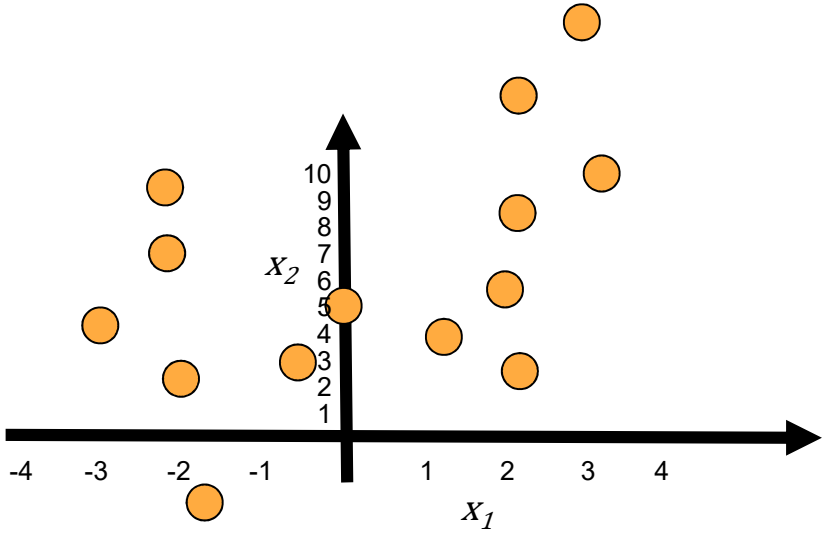
- Achar uma boa função hipótese de regressão para os dados
  - Tentar regressão polinomial



$$f(x) = -3x^2$$

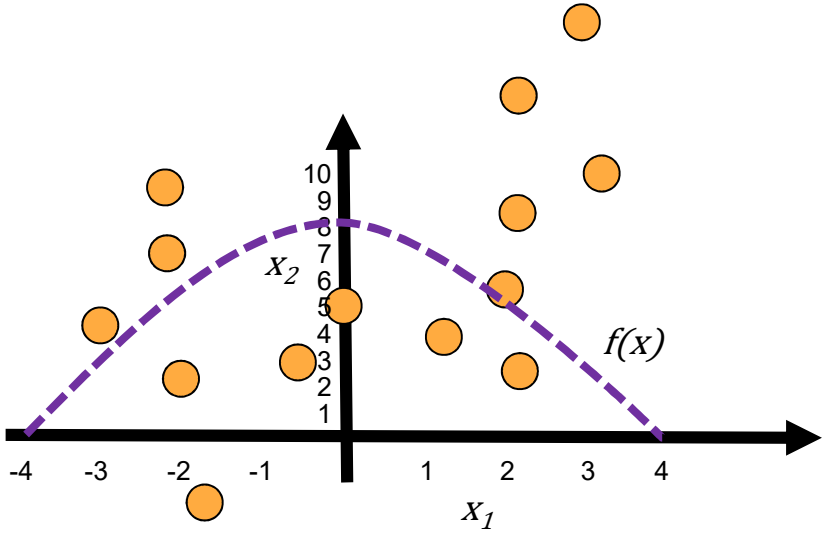
Regressão linear

- Achar uma boa função hipótese de regressão para os dados



Regressão linear

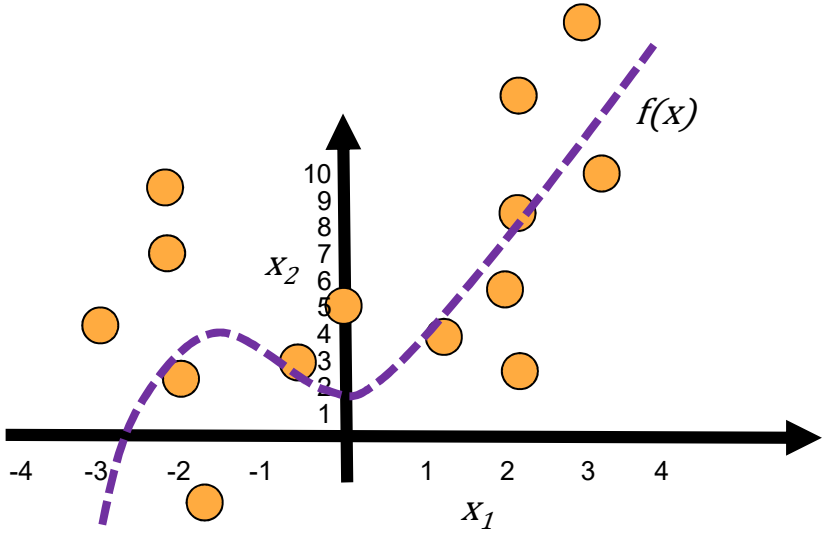
- Achar uma boa função hipótese de regressão para os dados
  - Tentar regressão polinomial de grau 2



$$f(x) = -\frac{1}{2}x^2 + 8$$

Regressão linear

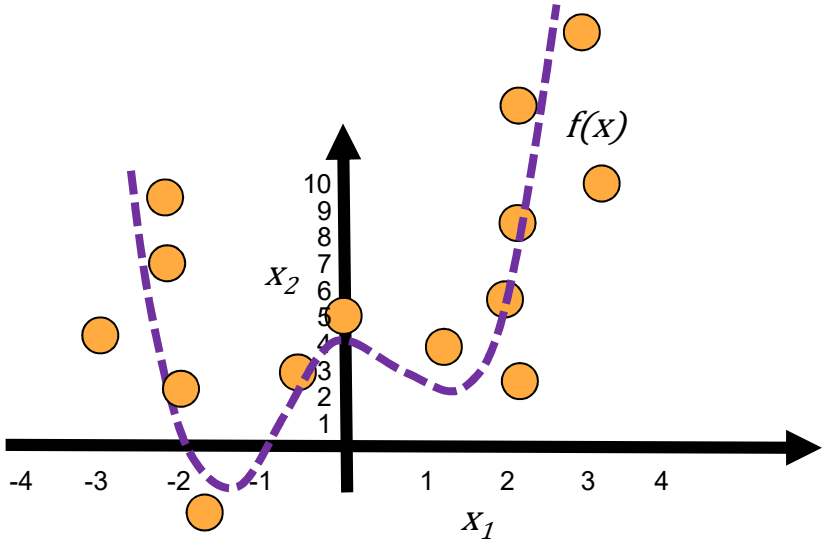
- Achar uma boa função hipótese de regressão para os dados
  - Tentar regressão polinomial de grau 3



$$f(x) = 2x^3 + 4x^2 + 2$$

Regressão linear

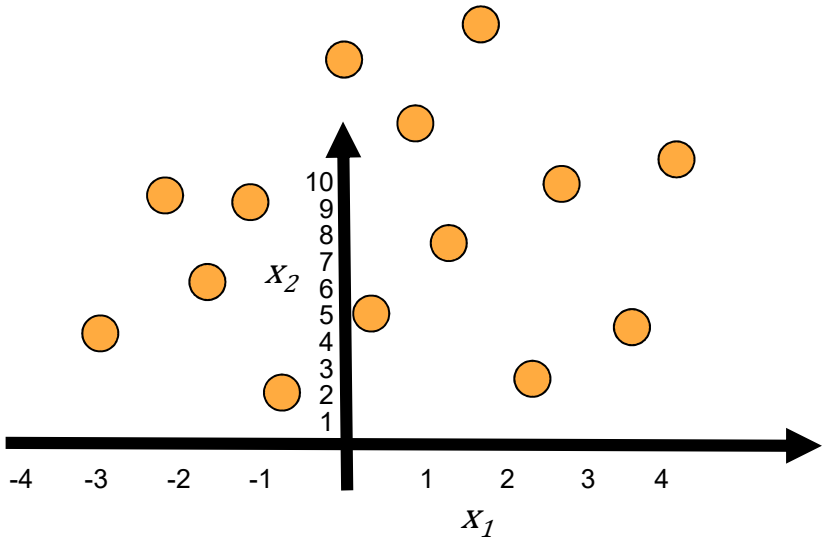
- Achar uma boa função hipótese de regressão para os dados
  - Tentar regressão polinomial de grau 4



$$f(x) = x^4 - 4x^2 + 2x + 4$$

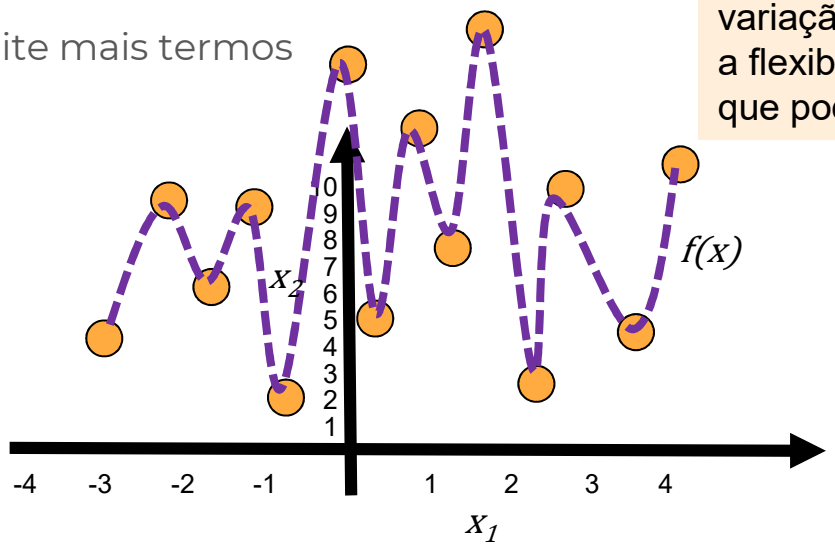
Regressão linear

- Achar uma ótima função hipótese de regressão para os dados



Regressão linear

- Achar uma ótima função hipótese de regressão para os dados
  - Polinômio com grau muito alto
    - Permite mais termos

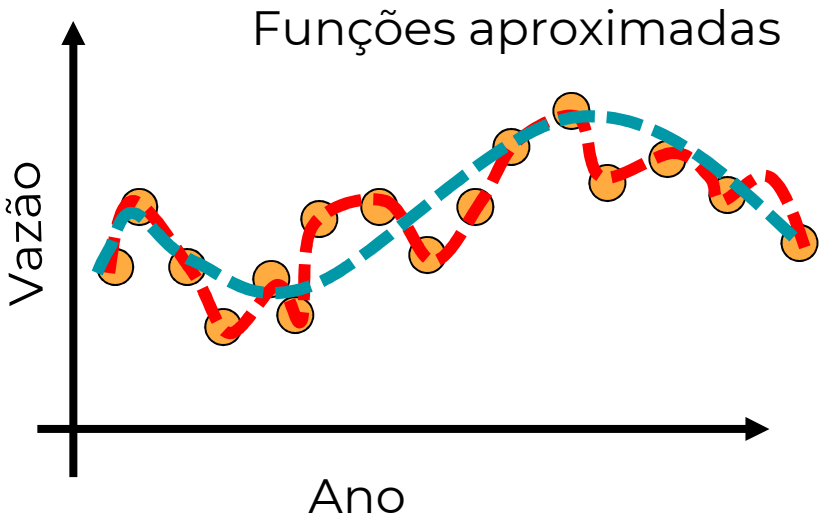


Quanto maior o número de termos, e variação dos valores dos expoentes, maior a flexibilidade (complexidade) das funções que podem ser ajustadas aos dados

Overfitting

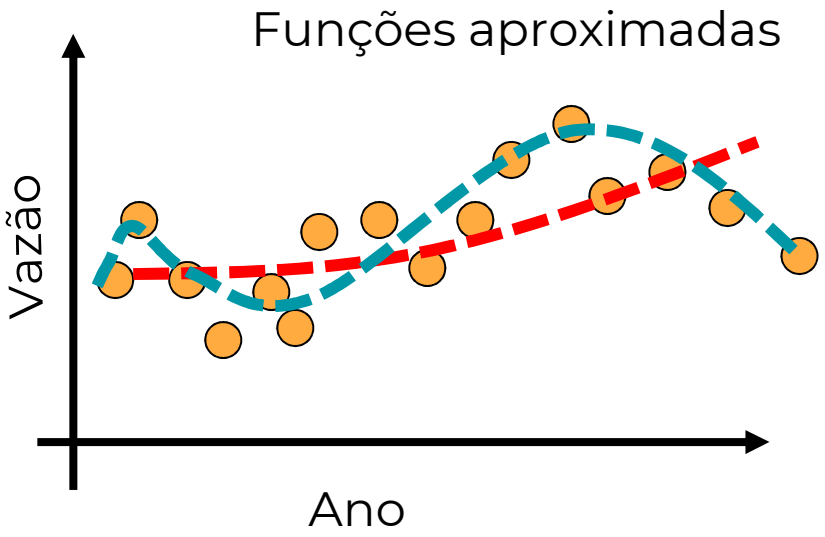
Regressão linear

- Overfitting



Tarefa não é tão simples

- Underfitting

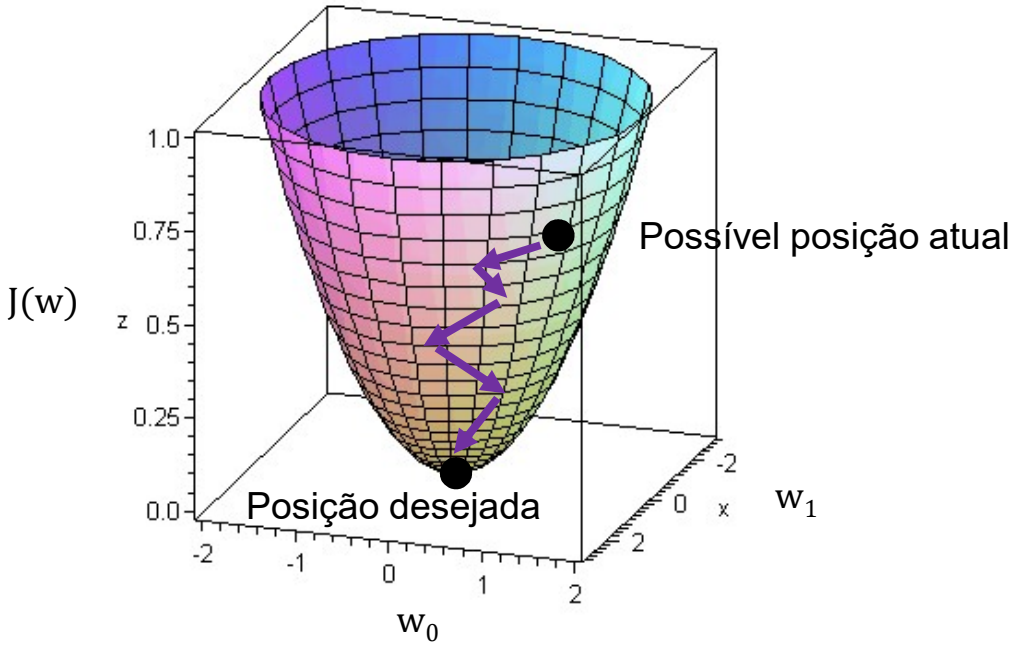


Ajuste de função de regressão

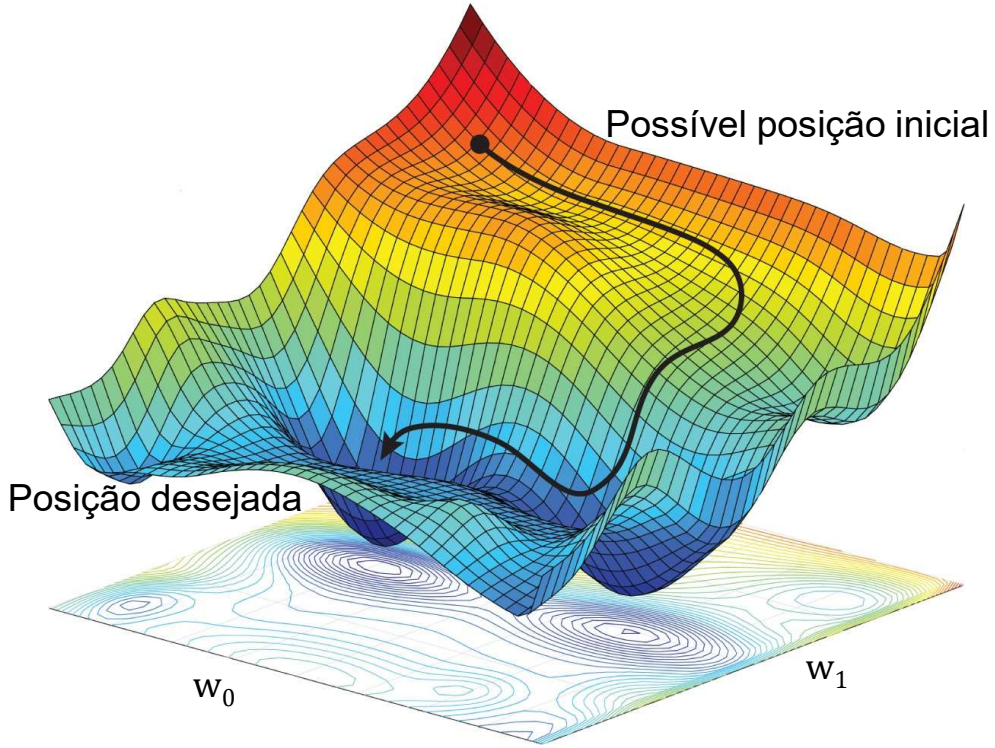
- Como ajustar os parâmetros automaticamente?
  - Usar um algoritmo de ajuste de parâmetros
    - Buscam o conjunto de valores para os parâmetros que mais reduz o custo (processo de otimização)
  - Existe vários
    - Um dos mais usados é o algoritmo **gradiente descendente**
      - Ajusta os parâmetros de forma iterativa de forma a reduzir o erro cometido com os valores de parâmetro atuais
      - Busca achar o ponto de mínimo da função de custo $J(w)$

Gradiente descendente

- As funções a terem o erro minimizado em geral têm mais de um parâmetro

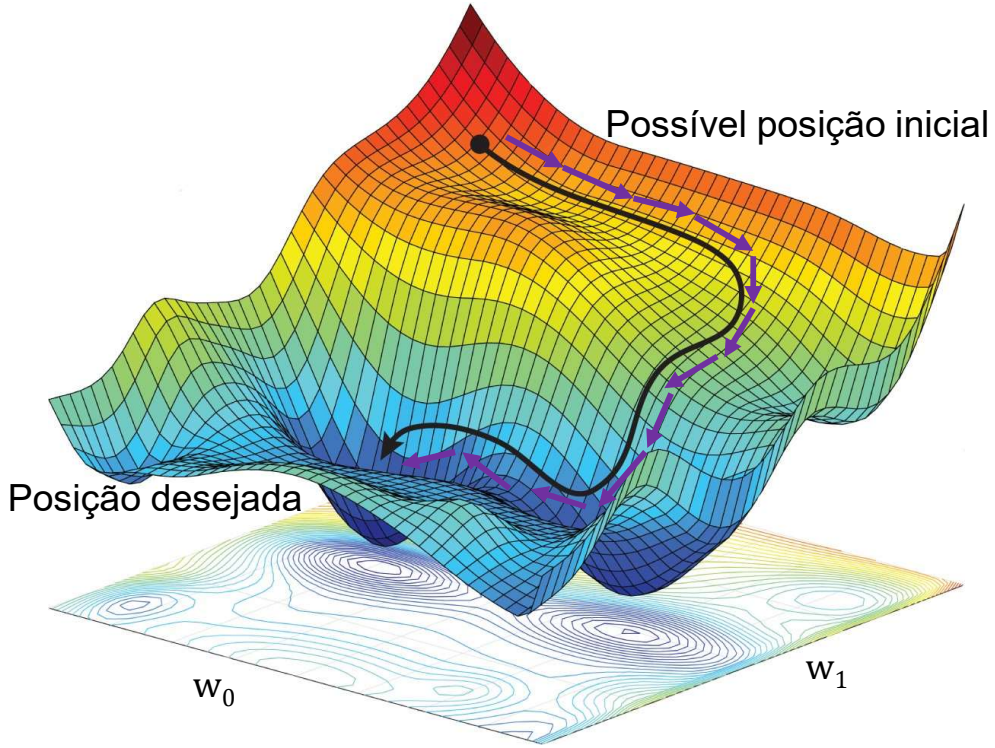


Gradiente descendente



A. Amini et al. "Spatial Uncertainty Sampling for End-to-End Control". NeurIPS Bayesian Deep Learning 2018

Gradiente descendente



Fonte: A. Amini et al. "Spatial Uncertainty Sampling for End-to-End Control". NeurIPS Bayesian Deep Learning 2018

Gradiente descendente

- Procura valores de parâmetros que levem ao ponto de mínimo da função de custo $\min_{w_0 w_1} J(w_0, w_1)$
  - Usa gradiente da função de custo $J(w)$, $\nabla J(w)$, para ajustar o valor dos parâmetros
    - Função vetorial cujos componentes são as derivadas parciais de uma função

Derivadas parciais

      - $\nabla J(w) = (\frac{\partial}{\partial w_0} J(w_0, w_1), \frac{\partial}{\partial w_1} J(w_0, w_1))$ (vetor gradiente)
      - Busca direção da função de custo que desce mais na superfície de erro
      - Deve ser possível calcular a derivada da função de custo (diferenciável)
  - Permite encontrar o valor de cada parâmetro w_i capaz de reduzir o erro
    - Pode ajustar valor de cada parâmetro em paralelo
      - $w_i = w_i - \alpha \frac{\partial}{\partial w_i} J(w_0, w_1)$ α : taxa de aprendizado

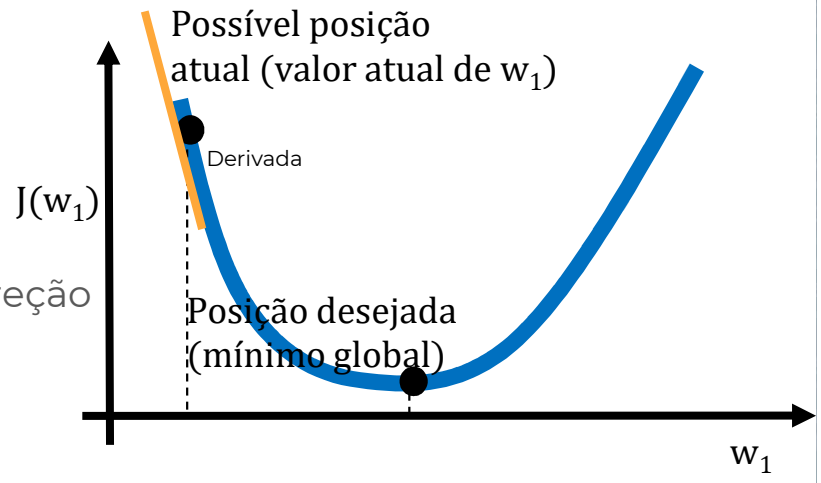
Gradiente descendente

- Ajuste dos parâmetros
  - $w_i = w_i - \alpha \frac{\partial}{\partial w_i} J(w_0, w_1)$
  - Ajuste simultâneo
    - $aux_0 = w_0 - \alpha \frac{\partial}{\partial w_0} J(w_0, w_1)$
    - $aux_1 = w_1 - \alpha \frac{\partial}{\partial w_1} J(w_0, w_1)$
    - $w_0 = aux_0$
    - $w_1 = aux_1$
  - Ajuste sequencial
    - $w_0 = w_0 - \alpha \frac{\partial}{\partial w_0} J(w_0, w_1)$
    - $w_1 = w_1 - \alpha \frac{\partial}{\partial w_1} J(w_0, w_1)$ (usa valor ajustado de w_0 neste ajuste, mais instável)

Exemplo simples de ajuste de um parâmetro 1

- Minimizar função de custo com um parâmetro
  - Supor que a função de custo $J(w)$ tem apenas um parâmetro, w_1
    - Neste caso, o gradiente da função é uma simples derivada
    - O ajuste do parâmetro w_1 é dado por:
      - $w_1 = w_1 - \alpha \frac{d}{dw_1} J(w_1)$
      - $= w_1 - \alpha J'(w_1)$
      - Valor de w_1 pode mover em apenas duas direções:
        - Esquerda ou direita
        - Usar derivada para definir direção
          - **Negativa ⇒ direita**
          - Aumenta valor de w_1

Valor do parâmetro vai para posição que mais reduz o erro



Exemplo simples de ajuste de um parâmetro 2

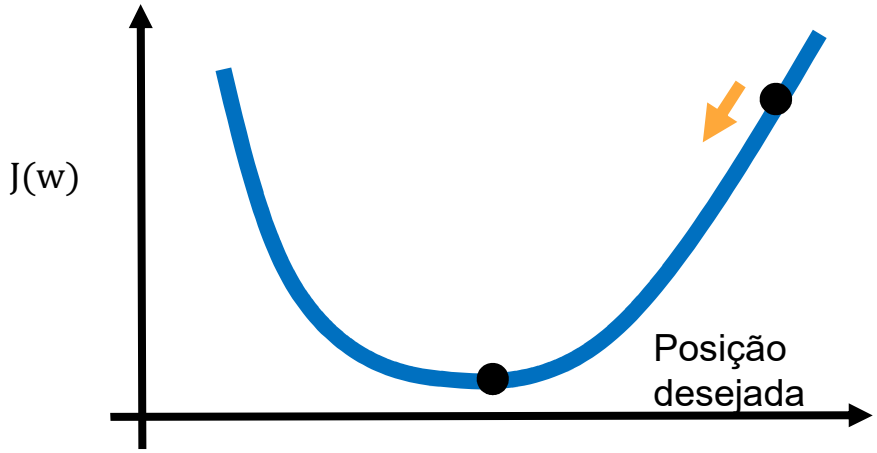
- Minimizar função de custo com um parâmetro
  - Supor que a função de custo $J(w)$ tem apenas um parâmetro, w_1
    - Neste caso, o gradiente da função é uma simples derivada
    - O ajuste do parâmetro w_1 é dado por:
      - $w_1 = w_1 - \alpha \frac{d}{dw_1} J(w_1)$
      - $= w_1 - \alpha J'(w_1)$
      - Valor de w_1 pode mover em apenas duas direções:
        - Esquerda ou direita
        - Usar derivada para definir direção
          - **Positiva ⇒ esquerda**
          - Reduz valor de w_1

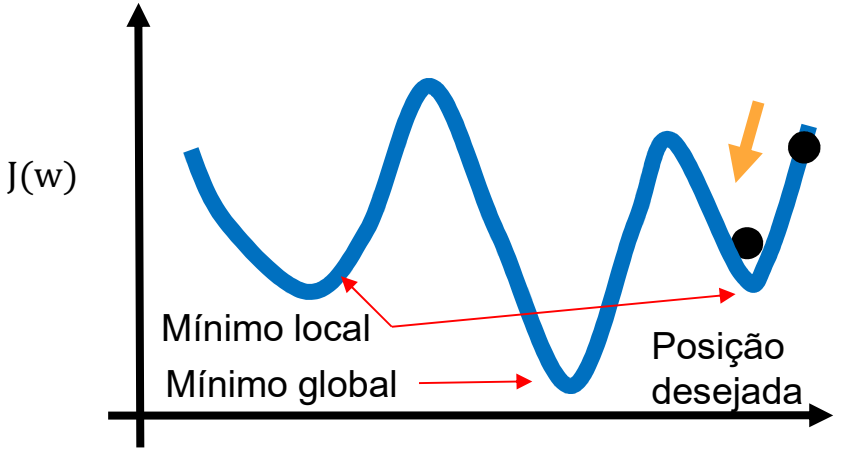
Valor do parâmetro vai para posição que mais reduz o erro

Mínimos locais

- Gradiente pode levar a mínimos locais

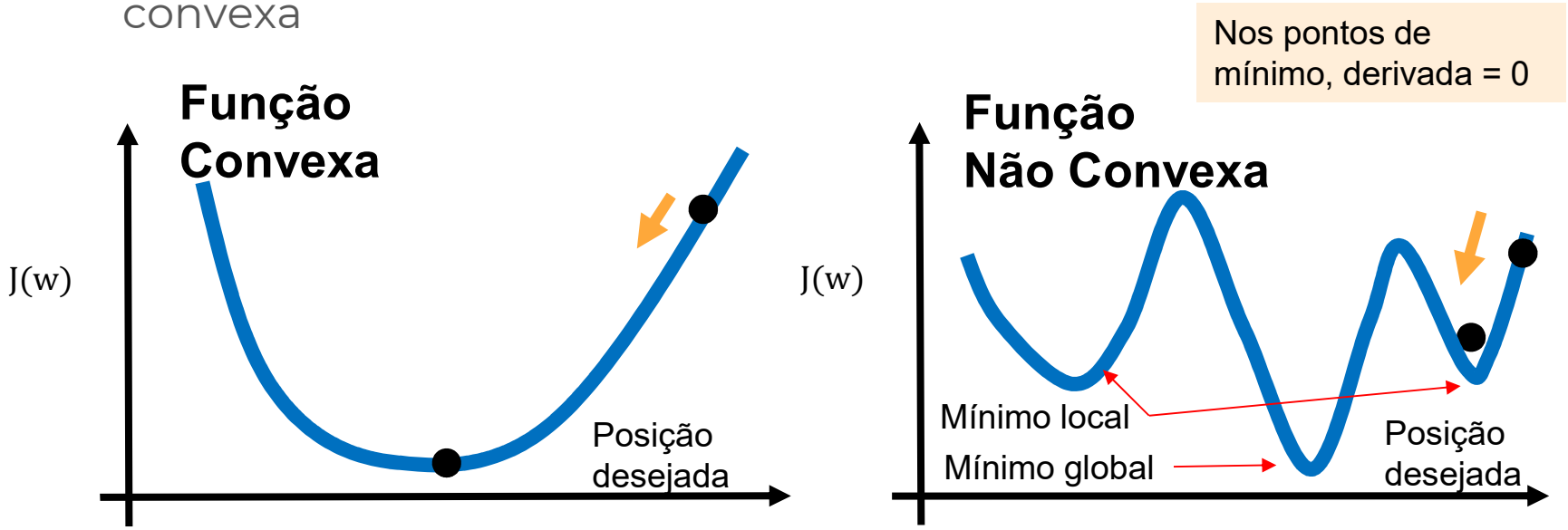
Nos pontos de mínimo, derivada = 0





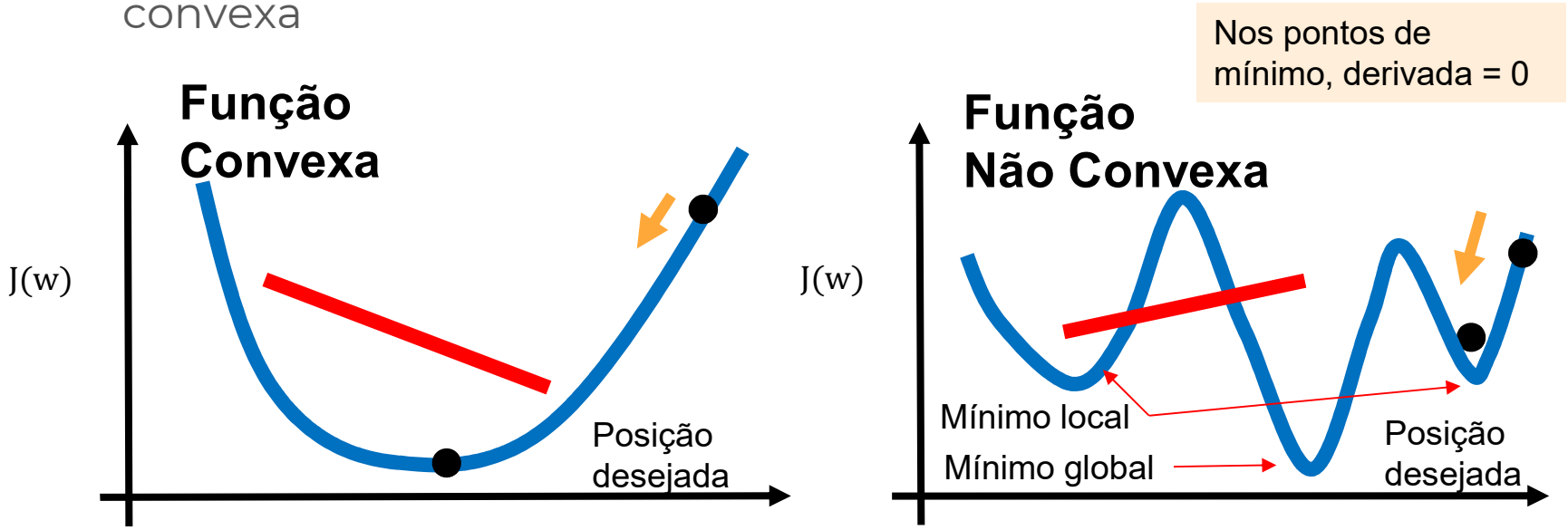
Mínimos locais

- Gradiente pode levar a mínimos locais se função J(w) for não convexa



Mínimos locais

- Gradiente pode levar a mínimos locais se função J(w) for não convexa



Gradiente descendente

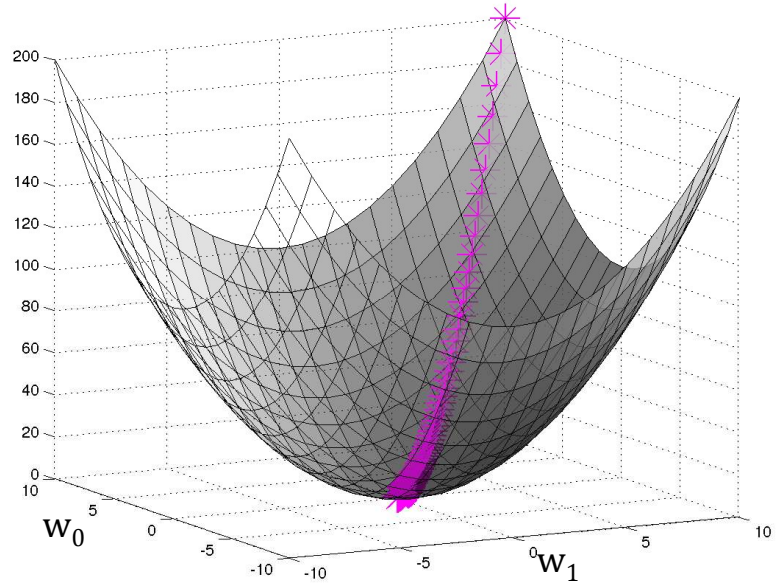
- Ajuste dos parâmetros
  - Dados que $f(x) = w_0 + w_1 x$ e que $w_i = w_i - \alpha \frac{\partial}{\partial w_i} J(w_0, w_1)$
    - $J(w_0, w_1) = \frac{1}{2n} \sum_{i=1}^{n} \left(y^i - f(x^i)\right)^2$
    - $\frac{\partial}{\partial w_0} J(w_0, w_1) = \frac{1}{n} \sum_{i=1}^{n} \left(y^i - f(x^i)\right)$
    - $\frac{\partial}{\partial w_1} J(w_0, w_1) = \frac{1}{n} \sum_{i=1}^{n} \left(y^i - f(x^i)\right) x^i$
  - Ajuste simultâneo
    - $aux_0 = w_0 - \alpha \frac{\partial}{\partial w_0} J(w_0, w_1) = w_0 - \alpha \frac{1}{n} \sum_{i=1}^{n} \left(y^i - f(x^i)\right)$
    - $aux_1 = w_1 - \alpha \frac{\partial}{\partial w_1} J(w_0, w_1) = w_1 - \alpha \frac{1}{n} \sum_{i=1}^{n} \left(y^i - f(x^i)\right) x^i$
    - $w_0 = aux_0$
    - $w_1 = aux_1$

Algoritmo gradiente descendente

1 Iniciar parâmetros com valor aleatório (ou 0)
2 Repita
ErroTotal = 0
Para cada par de treinamento (x^i, y^i)), i variando de 0 a n
Calcular a saída $f(x^i)$
Calcular $erro^i$
ErroTotal = ErroTotal + $erro^i$
Ajustar valores dos parâmetros usando ErroTotal
Até condição de parada ser atingida

Função de custo para regressão linear

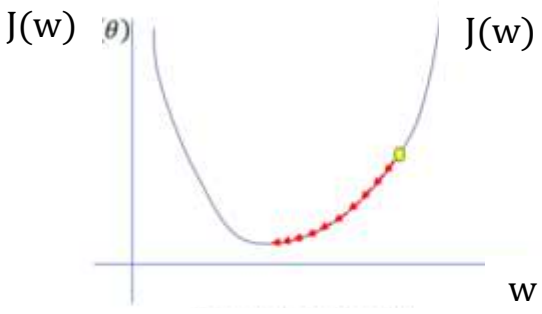
- Será sempre uma função convexa
  - Não tem mínimo local (não significa que o menor custo será = 0)



Efeito da taxa de aprendizado

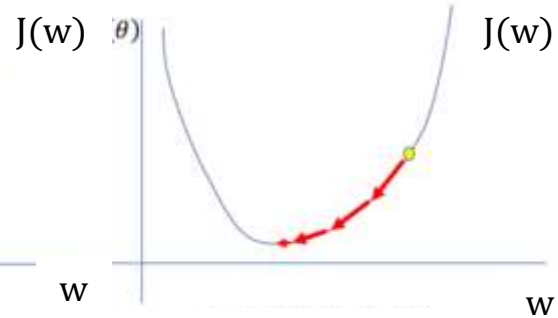
- Valor de α define convergência do gradiente descendente
  - Taxa de aprendizado (α)

Muito pequeno



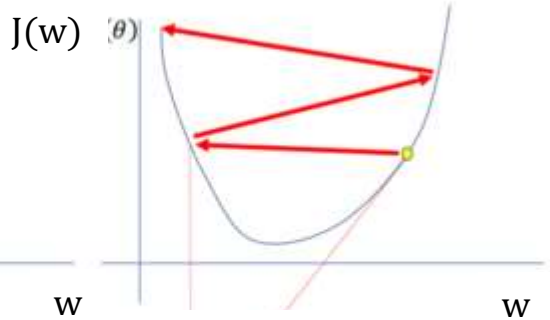
Custo deve reduzir a cada iteração, mas pode precisar de muitas iterações até convergir

Bom tamanho



Boa velocidade de convergência

Muito grande



Mudanças muito drásticas que podem passar do ponto de mínimo

Continua no próximo vídeo e conjunto de slides