UNIVERSIDADE FEDERAL DO PIAUÍ

Aprendizado

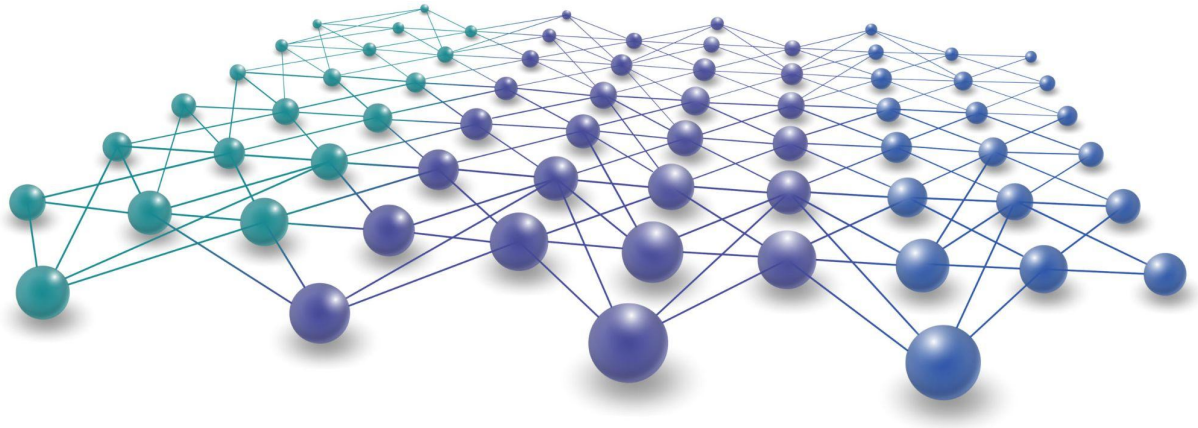
Como alcançar um bom desempenho?

Profa: Deborah Magalhães

Aprendizado

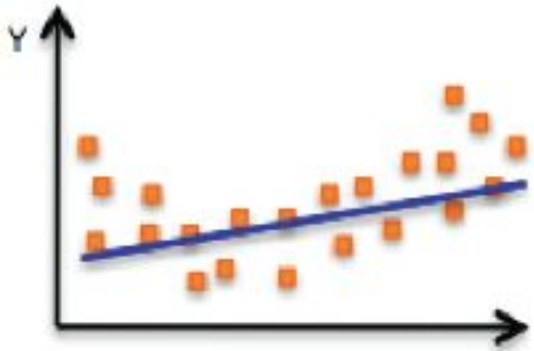
Capacidade dos nós e camadas

- A **capacidade** de uma rede neural é controlada através de:
  - **Número de nós**
  - **Número de camadas**

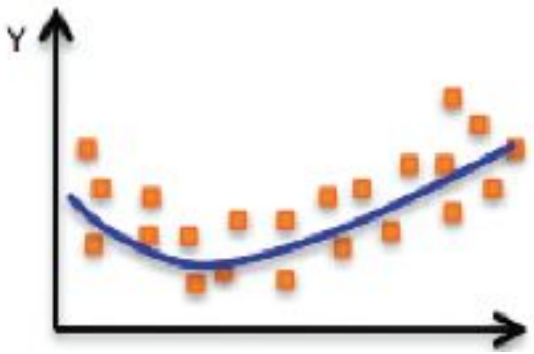


Um modelo com mais camadas (depth) e mais nós (width) por camadas possui **alta capacidade de representação** - é capaz de representar **funções mais complexas**.

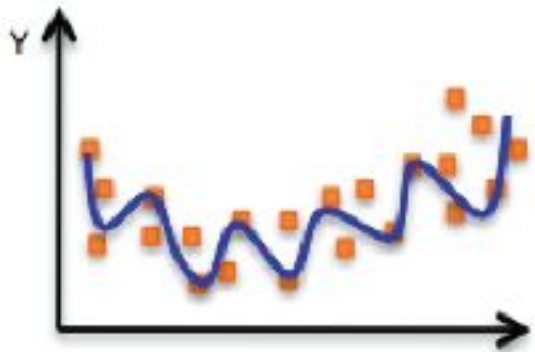
Deep Learning, 2016 - Página 428



Subajuste



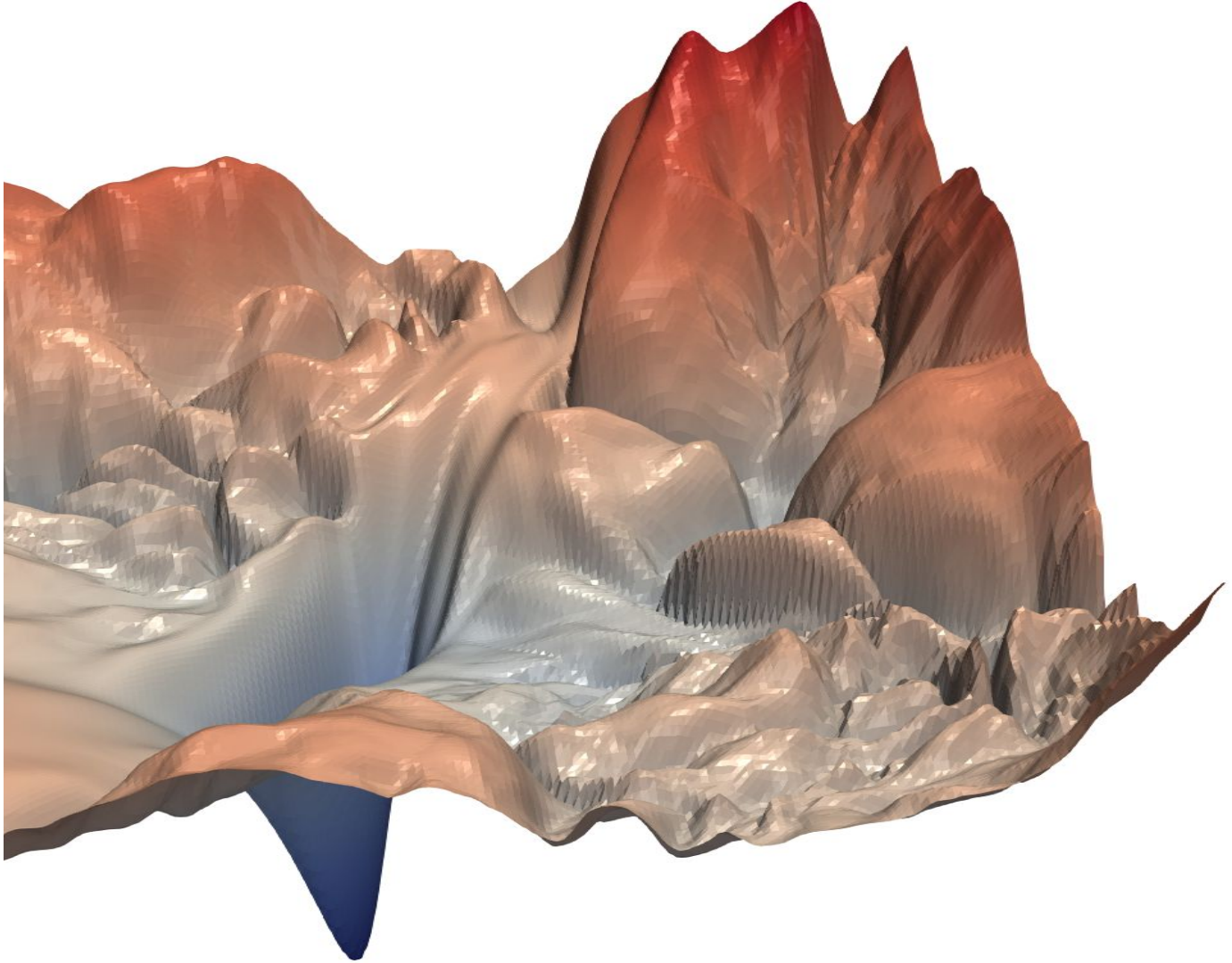
Ideal



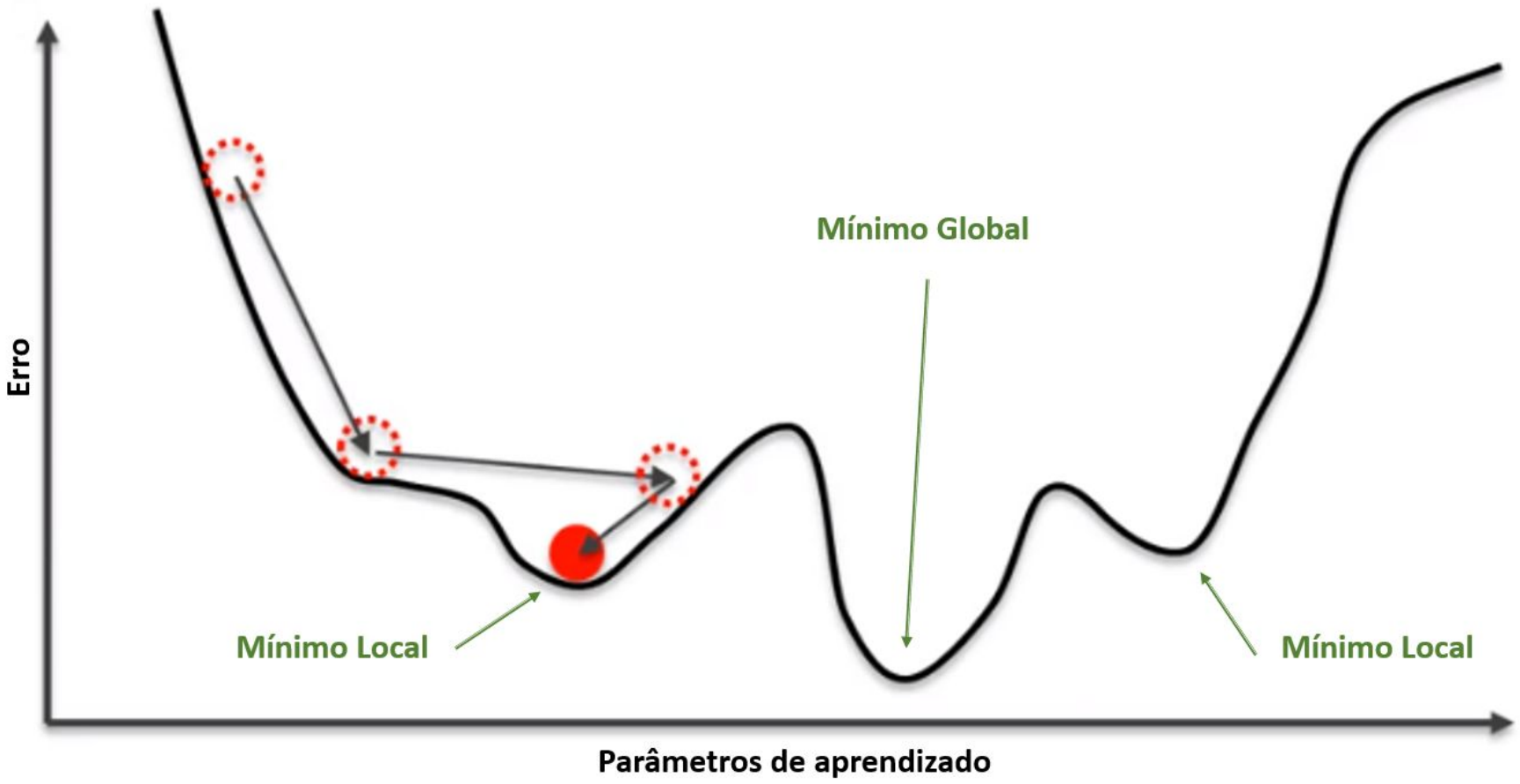
Sobreajuste

https://docs.aws.amazon.com/pt_br/machine-learning/latest/dg/model-fit-underfitting-vs-overfitting.html

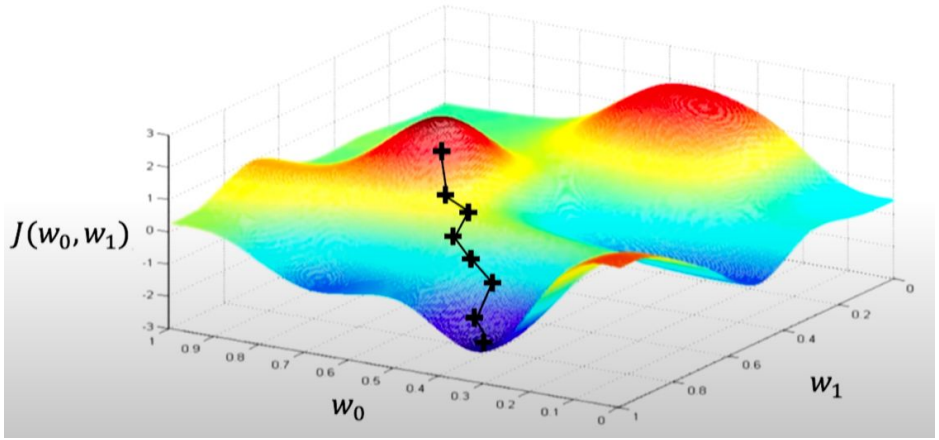
Treinar uma rede neural NÃO é trivial**!**



Li, H., Xu, Z., Taylor, G., Studer, C., & Goldstein, T. (2018). Visualizing the loss landscape of neural nets. Advances in neural information processing systems, 31.


Erro
Mínimo Global
Mínimo Local
Mínimo Local
Parâmetros de aprendizado

Gradiente Descendente



1. Inicializar os pesos randomicamente, $N(0, \sigma^2)$
2. Até convergir:
   2.1. Calcular o gradiente, $\frac{\partial J(W)}{\partial W}$
   2.2. Atualizar pesos, $W \leftarrow W - \eta \frac{\partial J(W)}{\partial W}$
3. Retornar pesos

MIT Introduction to Deep Learning - Lecture 1

A função que queremos minimizar ou maximizar é chamada de função objetivo. Quando buscamos **minimizá-la**, podemos chamá-la de função de **custo**, função de **perda** ou função de **erro**.

Deep Learning, 2016 - Página 82

Função Perda (Loss)

Objetivo: encontrar os pesos (W) da rede que resultem na menor perda (L)

$$W^* = \underset{W}{\operatorname{argmin}} \frac{1}{n} \sum_{i=1}^{n} \mathcal{L}\left(f\left(x^{(i)}; W\right), y^{(i)}\right)$$

$$W^* = \underset{W}{\operatorname{argmin}} J(W)$$

Função de Perda: maximum likelihood estimation (MLE)

- MLE minimiza a divergência entre a distribuição empírica (dados de treinamento) e a distribuição modelo (estimativas).
- Essa grau de divergência entre as duas distribuições é medido pela divergência de Kullback–Leibler (KL).

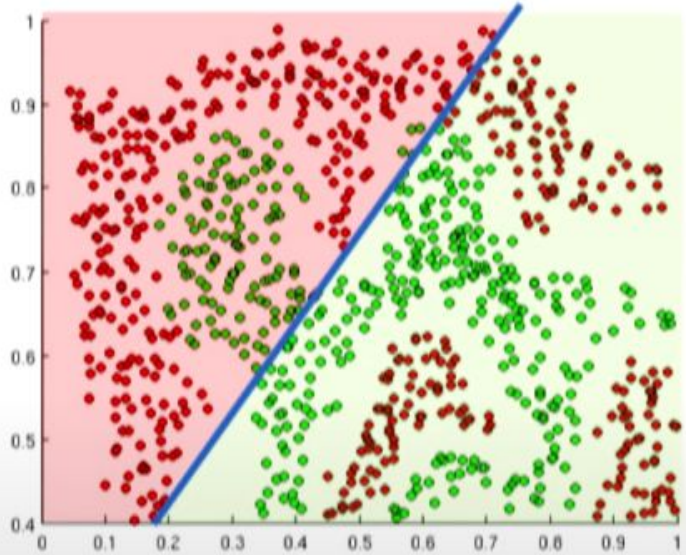
A escolha da função **custo** está intimamente **ligada** à escolha da função de ativação da camada de **saída**.

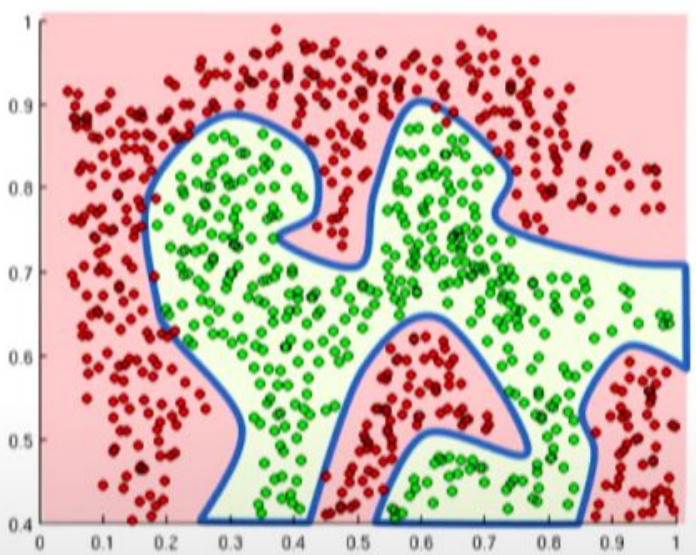
Deep Learning, 2016 - Página 181

Função de Perda

- Regressão
  - Camada de Saída: função de ativação linear;
  - Função Perda: Erro Quadrado Médio (MSE);
- Classificação Binária
  - Camada de Saída: função de ativação sigmóide;
  - Função Perda: Cross-Entropy;
- Classificação multiclasse
  - Camada de Saída: função de ativação softmax;
  - Função Perda: Cross-Entropy.

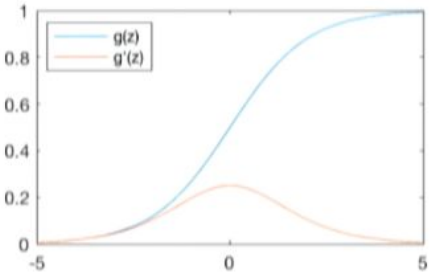
Qual o papel da função de ativação?





Exemplos de função de ativação

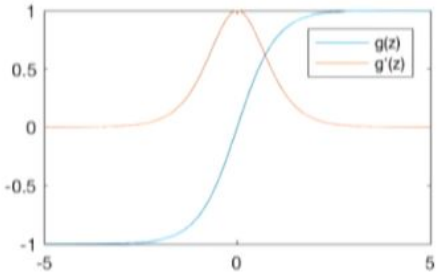
Sigmoid Function



$$g(z) = \frac{1}{1 + e^{-z}}$$

$$g'(z) = g(z)(1 - g(z))$$

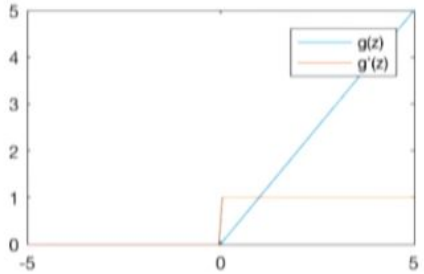
Hyperbolic Tangent



$$g(z) = \frac{e^{z} - e^{-z}}{e^{z} + e^{-z}}$$

$$g'(z) = 1 - g(z)^2$$

Rectified Linear Unit (ReLU)



$$g(z) = \max(0, z)$$

$$g'(z) = \begin{cases} 1, & z > 0 \\ 0, & \text{otherwise} \end{cases}$$

“

O **número de exemplos** do conjunto de treinamento usado na estimativa do gradiente do erro é definido pelo hiperparâmetro **batch size**

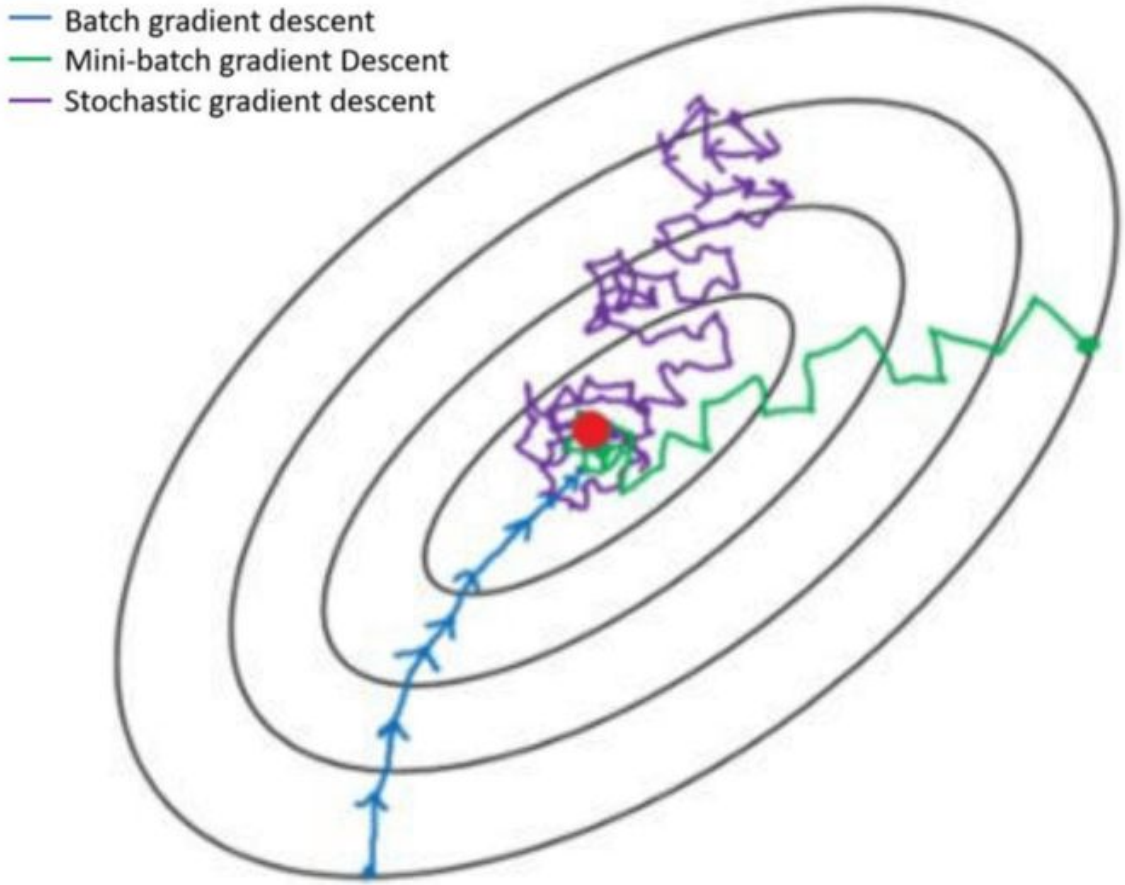
Deep Learning, 2016 - Página 278

Algoritmos de otimização que usam todo o conjunto de treinamento são chamados de métodos de gradiente determinístico ou **batch**.

Deep Learning, 2016 - Página 278

Algoritmos de otimização que usam apenas um **único** exemplo por vez são chamados de **estocásticos**.

Deep Learning, 2016 - Página 278



https://developpaper.com/difference-of-three-gradient-descent-algorithms-bgd-sgd-mbgd/

Muito Obrigada!

Se você tiver qualquer dúvida ou sugestão:

- deborah.vm@ufpi.edu.br

UNIVERSIDADE FEDERAL DO PIAUÍ