Adam

Resumo teórico

Hiperparâmetros:

$$\alpha : \text{learning rate}$$
$$\beta_1 : 0.9$$
$$\beta_2 : 0.999$$
$$\epsilon : 10^{-8}$$

Variáveis:

$$t : \text{epoca}$$
$$g : \text{vetor de gradientes}$$
$$m : \text{vetor de momentos}$$
$$v : \text{vetor de velocidades}$$
$$w : \text{vetor de pesos}$$

Algoritmo (para cada época):

$$t := t + 1$$
$$g_t := \frac{\partial E_{t-1}}{\partial w_{t-1}}$$
$$m_t := \beta_1 \cdot m_{t-1} + (1 - \beta_1) \cdot g_t$$
$$v_t := \beta_2 \cdot v_{t-1} + (1 - \beta_2) \cdot (g_t \odot g_t)$$
$$\hat{m}_t := \frac{m_t}{1 - \beta_1}$$
$$\hat{v}_t := \frac{v_t}{1 - \beta_2}$$
$$w_t := w_{t-1} - \alpha \cdot \frac{\hat{m}_t}{\sqrt{\hat{v}_t} + \epsilon}$$

O ajuste dos pesos deve ser adaptado para o caso de mini-batches.