Tema 3: Algoritmos de optimización

Aprendizaje
Profundo

Grado en Ciencia e Ingeniería de Datos (Universidad de Oviedo)

Pablo González, Pablo Pérez
{gonzalezgpablo, pabloperez}@uniovi.es
Centro de Inteligencia Artificial, Gijón

Ejercicio

Configuración del problema

- Modelo: $\hat{y} = w \times x$
- Valor inicial del peso: $w = 2.0$
- Dato de entrada: $x = 1.0$, $y = 4.0$
- Función de pérdida: $\mathcal{L} = \frac{1}{2}(y - \hat{y})^2$
- Tasa de aprendizaje: $\eta = 0.1$

Calcular, el valor del peso w actualizado después de una pasada hacia atrás utilizando:

- Descenso del gradiente
- Descenso del gradiente con momento ($\gamma = 0.9$)
- AdaGrad
- RMSProp ($\beta = 0.99$)
- Adam ($\beta_1 = 0.9$, $\beta_2 = 0.999$)

Primera iteracción

Pasada hacia adelante y pasada hacia atrás

Realizamos una pasada hacia adelante:

$$\hat{y} = w \times x$$

$$2 \times 1 = 2$$

Calculamos la loss:

$$\mathcal{L} = \frac{1}{2}(y - \hat{y})^2 = \frac{1}{2}(y - wx)^2 = \frac{1}{2}(4 - 2)^2 = 2$$

Calculamos la derivada parcial de la función de pérdida con respecto al peso:

$$\frac{\partial \mathcal{L}}{\partial w} = \frac{2}{2}(y - wx)(-x) = -(4 - w) = -(4 - 2) = -2$$

En las fórmulas de este tema, este valor suele representarse como g_t, es decir, el gradiente en esta iteracción.

Descenso del gradiente

$$w_{t+1} = w_t - update_t$$
$$update_t = \eta \frac{\partial \mathcal{L}}{\partial w} = (0.1(-2)) = -0.2$$
$$w_{t+1} = 2 - (-0.2) = 2.2$$

Descenso del gradiente con momento

$$w_{t+1} = w_t - update_t$$

$$update_t = \gamma update_{t-1} + \eta \frac{\partial \mathcal{L}}{\partial w}$$

$$update_t = 0.9 \cdot 0 + 0.1(-2) = -0.2$$

$$w_{t+1} = 2 - (-0.2) = 2.2$$

AdaGrad

$$w_{t+1} = w_t - \eta' g_t$$

$$\eta' = \frac{\eta}{\sqrt{\epsilon + G_t}}$$

$$G_t = G_{t-1} + {g_t}^2$$

$$G_t = 0 + (-2)^2 = 4$$

$$\eta' = \frac{0.1}{\sqrt{4}} = 0.05$$

$$w_{t+1} = 2 - 0.05(-2) = 2.1$$

RMSProp

$$w_{t+1} = w_t - \eta' g_t$$

$$\eta' = \frac{\eta}{\sqrt{\epsilon + G_t}}$$

$$G_t = \beta G_{t-1} + (1 - \beta){g_t}^2$$

$$G_t = 0.99 \cdot 0 + (1 - 0.99)(-2)^2 = 0.04$$

$$\eta' = \frac{0.1}{\sqrt{0.04}} = 0.5$$

$$w_{t+1} = 2 - 0.5(-2) = 3$$

Adam

Calculo de los momentos:

$$m_t = \beta_1 \cdot m_{t-1} + (1-\beta_1)g_t = 0.9 \cdot 0 + 0.1 \cdot (-2) = -0.2$$

$$v_t = \beta_2 \cdot v_{t-1} + (1-\beta_2){g_t}^2 = 0.999 \cdot 0 + (1-0.999)(-2)^2 = 0.004$$

Corrección del sesgo para los momentos:

$$\hat{m}_t = \frac{m_t}{1-\beta_1^t} = \frac{-0.2}{1-0.9^1} = -2$$

$$\hat{v}_t = \frac{v_t}{1-\beta_2^t} = \frac{0.004}{1-0.999} = 4$$

Actualización del peso:

$$w_{t+1} = w_t - \eta\frac{\hat{m}_t}{\sqrt{\hat{v}_t + \epsilon}} = 2 - 0.1\frac{-2}{\sqrt{4}} = 2.1$$

Segunda iteracción

Repaso de fórmulas necesarias

$$\hat{y} = wx = w$$

$$\mathcal{L} = \frac{1}{2}(y - wx)^2 = \frac{1}{2}(4 - w)$$

$$\frac{\partial \mathcal{L}}{\partial w} = -(y - wx) = -(4 - w)$$

Descenso del gradiente

$$w_t = 2.2$$

$$\frac{\partial \mathcal{L}}{\partial w} = -(4 - w) = -(4 - 2.2) = -1.8$$

$$w_{t+1} = w_t - update_t$$

$$update_t = \eta \frac{\partial \mathcal{L}}{\partial w} = (0.1(-1.8)) = -0.18$$

$$w_{t+1} = 2.2 - (-0.18) = 2.38$$

Descenso del gradiente con momento

$$w_t = 2.2$$

$$update_{t-1} = -0.2$$

$$\frac{\partial \mathcal{L}}{\partial w} = -(4 - w) = -(4 - 2.2) = -1.8$$

$$w_{t+1} = w_t - update_t$$

$$update_t = \gamma update_{t-1} + \eta \frac{\partial \mathcal{L}}{\partial w}$$

$$update_t = 0.9 \cdot (-0.2) + 0.1(-1.8) = -0.18 - 0.18 = -0.36$$

$$w_{t+1} = 2.2 - (-0.36) = 2.56$$

AdaGrad

$$w_t = 2.1$$

$$G_{t-1} = 4$$

$$g_t = \frac{\partial \mathcal{L}}{\partial w} = -(4 - w_t) = -(4 - 2.1) = -1.9$$

$$w_{t+1} = w_t - \eta' g_t$$

$$\eta' = \frac{\eta}{\sqrt{\epsilon + G_t}}$$

$$G_t = G_{t-1} + {g_t}^2$$

$$G_t = 4 + (-1.9)^2 = 4 + 3.61 = 7.61$$

$$\eta' = \frac{0.1}{\sqrt{7.61}} = 0.036$$

$$w_{t+1} = 2.1 - 0.036(-1.9) = 2.1684$$

RMSProp

$$w_t = 3$$

$$G_{t-1} = 0.04$$

$$g_t = \frac{\partial \mathcal{L}}{\partial w} = -(4 - w_t) = -(4 - 3) = -1$$

$$w_{t+1} = w_t - \eta' g_t$$

$$\eta' = \frac{\eta}{\sqrt{\epsilon + G_t}}$$

$$G_t = \beta G_{t-1} + (1 - \beta) {g_t}^2$$

$$G_t = 0.99 \cdot 0.04 + (1 - 0.99)(-1)^2 = 0.0496$$

$$\eta' = \frac{0.1}{\sqrt{0.0496}} = 0.449$$

$$w_{t+1} = 3 - 0.449(-1) = 3.449$$

Adam

$$w_t = 2.1, m_{t-1} = -0.2, v_{t-1} = 0.004$$

$$g_t = \frac{\partial \mathcal{L}}{\partial w} = -(4 - w_t) = -(4 - 2.1) = -1.9$$

Calculo de los momentos:

$$m_t = \beta_1 \cdot m_{t-1} + (1 - \beta_1) g_t = 0.9 \cdot (-0.2) + 0.1 \cdot (-1.9) = -0.37$$

$$v_t = \beta_2 \cdot v_{t-1} + (1 - \beta_2) {g_t}^2 = 0.999 \cdot 0.004 + (1 - 0.999)(-1.9)^2 = 0.0076$$

Corrección del sesgo para los momentos:

$$\hat{m}_t = \frac{m_t}{1 - \beta_1^t} = \frac{-0.37}{1 - 0.9^2} = -1.94$$

$$\hat{v}_t = \frac{v_t}{1 - \beta_2^t} = \frac{0.0076}{1 - 0.999^2} = 3.8$$

Actualización del peso:

$$w_{t+1} = w_t - \eta \frac{\hat{m}_t}{\sqrt{\hat{v}_t + \epsilon}} = 2.1 - 0.1 \frac{-1.94}{\sqrt{3.8}} = 2.1995$$