Input: ϵ_0, initial learning rate
Input: α, decay rate of learning rate
Input: β, momentum rate
Input: m, minibatch size
Input: k, epoch size
Input: θ, initial weights
Input: v, initial velocity
Input: *apply_nesterov*, wheather to use Nesterov method or not {True or False}
Input: $\mathbf{X}$, training dataset inputs
Input: $\mathbf{y}$, training dataset targets
$j \leftarrow 1$
while $j \leq k$ **do**
    update learning rate $\epsilon_j \leftarrow \epsilon_0 + \alpha(\epsilon_{j-1} - \epsilon_0)$
    while *stopping criteria is not satisfied* **do**
        $\{\mathbf{x}^1 ... \mathbf{x}^m\}$, $\{\mathbf{y}^1 ... \mathbf{y}^m\} \leftarrow$ get a sample from $\mathbf{X}$ and $\mathbf{y}$ randomly
        if *apply_nesterov* $= True$ **then**
            $\tilde{\theta} \leftarrow \theta + \beta v$
            calculate estimation of gradient $\hat{g} \leftarrow \frac{1}{m} \sum_{i=1}^{m} L(f(\mathbf{x}^i; \tilde{\theta}), \mathbf{y}^i)$
        else
            calculate estimation of gradient $\hat{g} \leftarrow \frac{1}{m} \sum_{i=1}^{m} L(f(\mathbf{x}^i; \theta), \mathbf{y}^i)$
        update velocity $v \leftarrow \beta v - \epsilon_j \hat{g}$
        update weights $\theta \leftarrow \theta + v$
    $j \leftarrow j + 1$ go to next epoch