RMAD для SGD без момента

Выполнил: Плетнев Никита

Выпишем алгоритм SGD без момента (эквивалентен SGD с моментом, равным 0).

1. Вход: начальное значение параметров $\mathbf{w}_0$, скорости обучения α (возможно, разные для каждого шага), функция потерь $L(\mathbf{w}, \theta, t)$.
2. Инициализировать $\mathbf{v}_1 = 0$;
3. На каждом шаге при $t = 1..T$:
   (a) Вычислить градиент $g_t \sim \nabla_{\mathbf{w}} L(\mathbf{w}_t, \theta, t)$;
   (b) Вычислить скорость $\mathbf{v}_t = -g_t$
   (c) Обновить параметры $\mathbf{w}_{t+1} = \mathbf{w}_t + \alpha_t \mathbf{v}_t$;
4. Выход: обученные параметры $\mathbf{w}_T$.

Перепишем этот алгоритм с применением обратного дифференцирования.

1. Вход: начальное значение параметров $\mathbf{w}_T$, скорости $\mathbf{v}_T$, скорость обучения α , функции потерь $L(\mathbf{w}, \theta, t)$ и $f(\mathbf{w})$.
2. Инициализировать $d\mathbf{v} = 0$, $d\theta = 0$, $d\alpha = 0$;
3. Инициализировать $d\mathbf{w} = \nabla_{\mathbf{w}} f(\mathbf{w}_T)$
4. На каждом шаге при $t = T..1$:
   (a) $d\alpha_t = d\mathbf{w}^T \mathbf{v}_t$;
   (b) $\mathbf{w}_{t-1} = \mathbf{w}_t - \alpha_t \mathbf{v}_t$;
   (c) $g_{t-1} \sim \nabla_{\mathbf{w}} L(\mathbf{w}_{t-1}, \theta, t-1)$
   (d) $v_{t-1} = -g_{t-1}$
   (e) $d\mathbf{v} = \alpha_t d\mathbf{w}$
   (f) $d\mathbf{w} = d\mathbf{w} - d\mathbf{v} \nabla_{\mathbf{w}} \nabla_{\mathbf{w}} L(\mathbf{w}_t, \theta, t)$
   (g) $d\theta = d\theta - d\mathbf{v} \nabla_{\theta} \nabla_{\mathbf{w}} L(\mathbf{w}_t, \theta, t)$
5. Выход: градиент $f(\mathbf{w}_t$ по отношению к $\mathbf{w}_1$, $\mathbf{v}_1$, α и θ.

Использование обратного дифференцирования позволяет производить вычисления со сложностью $O(T)$. Взяв в качестве θ вектор гиперпараметров $\mathbf{h}$, получаем требуемый алгоритм RMAD.