Оптимизаторы – алгоритмы, используемые для изменения пар-тов с целью ускорения работы модели.

Adagrad – оптимизатор, меняющий скорость обучения η для каждого пар-ра на каждом шаге. Работает по прин. ф-ции ошибки.

$g_t = \nabla_\theta J(\theta_t)$ – прин. ф-ция

$G_t = G_t + g_t^2$

$\theta_{t+1} = \theta_t - \frac{\eta}{\sqrt{G_t + \varepsilon}} \cdot g_t$ – обновление пар-тов

η – скорость обучения, кот. умножается для зад. пар-ра θ в данный момент времени на основе предыдущих градиентов, расч. для данного пар-ра.

Adadelta – расширение Adagrad. Ограничивает окно накопл. прошлых градиентов до нек. фикс. размера, вместо того, чтобы

хранить их все. Используется экспоненциальная средняя, а не сумма градиентов.

$$E[g^2]_t = \gamma E[g^2]_{t-1} + (1-\gamma) g_t^2$$

$$RMS[g]_t = \sqrt{E[g^2]_t + \varepsilon}$$

$$\theta_{t+1} = \theta_t - \frac{RMS[\Delta\theta]_{t-1}}{RMS[g]_t} \cdot g_t$$

Adam - работает с импульсами 1 и 2 порядков. Идея заключается в уменьшении скорости во избежание проскакивания минимума. В дополнение к хранению экспоненц. скользящих средних (как в Adadelta) сохр. экспоненц. скользящ. средние.

$$\hat{m}_t = \frac{m_t}{1-\beta_1^t} \qquad \hat{v}_t = \frac{v_t}{1-\beta_2^t}$$

$$\theta_{t+1} = \theta_t - \frac{\eta}{\sqrt{\hat{v}_t} + \varepsilon} \cdot \hat{m}_t$$

Название	Формула	Недостатки
Gradient Descent	$\theta = \theta - \alpha \nabla J(\theta)$	Можно ост в лок. мин. Вся выборка для всех наб. данных. Требуется большой объем памяти
Stochastic Gradient Descent	$\theta = \theta - \alpha \nabla J(\theta; x_i; y_i)$ $x_i; y_i$ – обуч. пример	Высок. дисперс. парам. модели. Чтобы получить сход. градиент необходимо медленно снижать значение скор. обуч.
Mini-Batch Gradient Descent	$\theta = \theta - \alpha \nabla J(\theta, B_j)$ B_j – батчи обуч. пар.	Можно ост в лок. мин. При неопт. выборе скор. обуч. потреб. много врем. для сход. град. спуск
Adagrad	$g_t = \nabla_\theta J(\theta_t)$ $\theta_{t+1} = \theta_t - \frac{\eta}{\sqrt{G_t + \varepsilon}} \cdot g_t$	Вычисл. дорого (счит. произв. второго пор-ка) Медленное обучение
Adadelta	$RMS[g]_t = \sqrt{E[g]_t^2 + \varepsilon}$ $\theta_{t+1} = \theta_t - \frac{RMS[\Delta\theta]_{t-1}}{RMS[g]_t} \cdot g_t$	Вычисл. дорого
RMSProp	$\theta_{t+1} = \theta_t - \frac{\eta}{RMS[g]_t} \cdot g_t$	Выч. дорого
Adam	$m_t = \frac{m_t}{1-\beta_1^t}; \hat{v}_t = \frac{v_t}{1-\beta_2^t}$ $\theta_{t+1} = \theta_t - \frac{\eta}{\sqrt{\hat{v}_t} + \varepsilon} \cdot \hat{m}_t$	Выч. дорого
Nesterov accelerated Gradient	$v = \gamma \cdot v + \eta \nabla J(\theta - \gamma v)$ $\theta = \theta - v$	Низкая скор-ть обучения