Лекция 2

Тема 3.1 Градиентные алгоритмы обучения нейронных сетей

Содержание:

1. Алгоритм Левенберга-Марквардта

В данном алгоритме используется квадратичное приближение целевой функции *E(w)* в окрестности полученного решения $w(t)$.

Для достижения минимуму целевой функции требуется, чтобы $\frac{dE(w(t)+p(t))}{dp(t)}=0$. При выполнении соответствующего дифференцирования можно получить условие оптимальности в виде:

$g(w(t))+H(w(t))p(t)=0$, откуда следует

$$p(t)=-[H(w(t))]^{-1}g(w(t)). \qquad (135)$$

Формула (133) однозначно указывает направление $p(t)$, которое гарантирует достижение минимального для данного шага значения целевой функции. Из него следует, что для определения этого направления необходимо в каждом цикле вычислять значение градиента g и гессиана H в точке последнего решения $w(t)$.

Формула (133) представляет собой основу ньютоновского алгоритма оптимизации и является чисто теоретическим выражением, поскольку ее применение требует положительной определенности гессиана на каждом шаге, что практически не осуществимо. Поэтому в реальных алгоритмах вместо точно определенного гессиана $H(w(t))$ используется его приближение $G(w(t))$, которое в алгоритме Левенберга-Марквардта рассчитывается на основе содержащейся в градиенте информации с учётом некоторого регуляризационного фактора [4].

Для описания данного метода представим целевую функцию в виде:

$$E(w)=\frac{1}{2}\sum_{s=1}^{M}[e_s(w)]^2, \qquad (136)$$

где $e_s = [y_s(w) - d_s(w)]$. При использовании обозначений

$$e(w) = \begin{bmatrix} e_1(w) \\ e_2(w) \\ \cdots \\ e_M(w) \end{bmatrix}, \quad J(w) = \begin{bmatrix} \frac{\partial e_1}{\partial w_1} & \frac{\partial e_1}{\partial w_2} & \cdots & \frac{\partial e_1}{\partial w_p} \\ \frac{\partial e_2}{\partial w_1} & \frac{\partial e_2}{\partial w_2} & \cdots & \frac{\partial e_2}{\partial w_p} \\ \cdots & \cdots & \cdots & \cdots \\ \frac{\partial e_M}{\partial w_1} & \frac{\partial e_M}{\partial w_2} & \cdots & \frac{\partial e_M}{\partial w_p} \end{bmatrix}, w_t = w(t). \qquad (137)$$

Вектор градиента и аппроксимированная матрица гессиана, соответствующие целевой функции (2.30) на t-ом шаге алгоритма, определяются в виде:

$$g(w(t)) = [J(w(t))]^T e(w), \qquad (138)$$

$$G(w(t)) = [J(w(t))]^T J(w(t)) + R(w(t)), \qquad (139)$$

где $R(w(t))$обозначены компоненты гессиана $H(w(t))$, содержащие высшие производные относительно $w(t)$. Аппроксимация $R(w(t))$ осуществляется с помощью регуляризационного фактора $v1(t)$, в котором переменная $v(t)$ называется параметром Левенберга-Марквардта и является скалярной величиной, изменяющейся в процессе обучения:

$$G(w(t)) = [J(w(t))]^T J(w(t)) + v1(t). \qquad (140)$$

В начале процесса обучения, когда текущее решение $w(t)$ далеко от искомого решения, следует использовать значение параметра $v(t)$ намного превышающее $[J(w(t))]^T J(w(t))$. В этом случае гессиан фактически заменяется регуляризационным фактором:

$$G(w(t)) \cong v1(t), \qquad (141)$$

а направление минимизации выбирается по методу наискорейшего спуска:

$$p(t) = -\frac{g(w(t))}{v(t)}. \qquad (142)$$

По мере приближения к искомому решению величина параметра $v(t)$ понижается и первое слагаемое в формуле (142) начинает играть более важную

роль. Таким образом, на эффективность алгоритма влияет правильный выбор величины $v(t)$. В методике, предложенной Д.Марквардтом, значение $v(t)$ изменяется по следующей схеме:

$$\text{если } E\left(\frac{v(t-1)}{r}\right) \le E(t), \text{то принять } v(t) = \frac{v(t)-1}{r};$$

$$\text{если } E\left(\frac{v(t-1)}{r}\right) > E(t) \text{ и } E(v(t-1)) < E(t), \text{то принять } v(t) = v(t-1);$$

$$\text{если } E\left(\frac{v(t-1)}{r}\right) > E(t) \text{ и } E(v(t-1)) > E(t), \text{то увеличить } m \text{ раз значение } v$$

$$\text{до достижения } E(v(t-1)r^m) \le E(t), \text{ принимая } v(t) = v(t-1)r^m,$$

где $E(t), v(t), E(t-1), v(t-1)$ обозначают значения целевой функции и параметра v на t -ом и $(t-1)$ -ом шагах алгоритма, а $r > 1$ - обозначает коэффициент уменьшения v.

Такая процедура изменения $v(t)$ применяется до момента, когда коэффициент верности отображения, рассчитываемый по формуле

$$q = \frac{E(t) - E(t-1)}{[\Delta w(t)]^T g(t) + 0{,}5[\Delta w(t)]^T G(t) \Delta w(t)}, \qquad (142)$$

достигнет значения, близкого к единице. При этом квадратичная аппроксимация целевой функции имеет высокую степень совпадения с истинными значениями, следовательно, регуляризационный фактор в формуле (142) может быть приравнен нулю, а процесс определения гессиана сводится к аппроксимации первого порядка, при этом алгоритм Левенберга-Марквардта превращается в алгоритм Гаусса-Ньютона, имеющему квадратичную сходимость.