Лекция 1

Тема 3.1 Градиентные алгоритмы обучения нейронных сетей

Содержание:

1. Алгоритм наискорейшего спуска.
2. Алгоритм переменной метрики.

1. Алгоритм наискорейшего спуска.

Если при разложении целевой функции E(w) в ряд Тейлора ограничиться ее линейным приближением, то мы получим алгоритм наискорейшего спуска. Для выполнения соотношения $E(w_{t+1}) < E(w_t)$ достаточно подобрать $g(w_t)^T p < 0$. Условию уменьшения значения целевой функции отвечает выбор вектора направления

$$p_t = -g(w_t). \tag{128}$$

В этом случае коррекция весовых коэффициентов производится по формуле:

$$w_{ij}(t+1) = w_{ij}(t) + \eta p_t \tag{129}$$

В другом виде формулу коррекции весов по методу наискорейшего спуска можно представить следующим образом:

$$w_{ij}(t+1) = w_{ij}(t) - \eta \cdot \frac{\partial E(t)}{\partial w_{ij}(t)}. \tag{130}$$

Ограничение слагаемым первого порядка при разложении функции в ряд Тейлора, не позволяет использовать информацию о ее кривизне. Это обуславливает линейную сходимость метода. Указанный недостаток, а также резкое замедление минимизации в ближайшей окрестности точки оптимального решения, когда градиент принимает очень малые значения, делают алгоритм наискорейшего спуска низкоэффективным. Тем не менее, простота, невысокие требования к объему памяти и относительно невысокая вычислительная сложность, обуславливают широкое использование

алгоритма. Повысить эффективность удается путем эвристической модификации выражения, определяющего направление градиента.

Одна из модификаций получила название алгоритма обучения с моментом. При этом подходе уточнение весов сети производится по формуле:

$$w_{ij}(t+1) = w_{ij}(t) - \eta \cdot \frac{\partial E(t)}{\partial w_{ij}(t)} + \alpha(w_{ij}(t) - w_{ij}(t-1)) \quad , \qquad (131)$$

где α - это коэффициент момента, принимающий значения в интервале [0, 1].

Первое слагаемое в формуле (3.39) соответствует алгоритму наискорейшего спуска, а второе слагаемое учитывает последнее изменение весов и не зависит от фактического значения градиента. Чем больше значение коэффициента α, тем большее значение оказывает показатель момента на подбор весов. При постоянном значении коэффициента обучения $\eta(t) = \eta$ приращение весов остается примерно одинаковым, то есть $\Delta w_{ij}(t) = \eta p(t) + \alpha \Delta w_{ij}(t)$, поэтому эффективное приращение весов можно писать формулой:

$$\Delta w_{ij}(t) = \frac{\eta}{1-\alpha} p(t). \qquad (132)$$

При значении α=0,9 это соответствует десятикратному увеличению значения коэффициента обучения и, следовательно, десятикратному ускорению процесса обучения. При малых значениях градиента показатель момента начинает доминировать, что приводит к такому приращению весов, которое соответствует увеличению значения целевой функции, позволяющему выйти из зоны локального минимума. Однако показатель момента, не должен доминировать на протяжении всего процесса обучения, поскольку это приводит к нестабильности алгоритма. На практике, увеличение целевой функции не допускается больше, чем на 4%. В противном случае, $\Delta w_{ij}(t) = 0$. При этом показатель градиента начинает доминировать над показателем момента и процесс развивается в направлении минимизации, заданном вектором градиента.

2 Алгоритм переменной метрики

В алгоритме переменной метрики 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

$$p_t = -\left[H(w_t)\right]^{-1} g(w_t) \qquad (133)$$

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), представляющая собой основу ньютоновского алгоритма оптимизации, 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 $G(w_t)$.

Основная идея метода переменной метрики заключается в том, что на каждом шаге гессиан или обратная ему величина, полученная на предыдущем шаге, модифицируются на величину некоторой поправки для обеспечения условия положительной определенности гессиана. Если прирост вектора w_t и градиента g на двух последовательных шагах итерации обозначить соответственно s_t и r_t, то есть $s_t = w_t - w_{t-1}$ и $r_t = g(w_t) - g(w_{t-1})$, а матрицу, обратную приближению гессиана $V_t = \left[G(w_t)\right]^{-1}$, $V_{t-1} = \left[G(w_{t-1})\right]^{-1}$ обозначить

V, то в соответствии с формулой Бройдена-Флетчера-Гольдфарба-Шенно процесс уточнения матрицы V можно описать рекуррентной зависимостью:

$$V_t = V_{t-1} + \left[1 + \frac{r_t^T V_{t-1} r_t}{s_t^T r_t}\right] \frac{s_t s_t^T}{s_t^T r_t} - \frac{s_t r_t^T V_{t-1} r_t s_t^T}{s_t^T r_t}. \quad (134)$$

Метод переменной метрики характеризуется более быстрой сходимостью, чем метод наискорейшего спуска. Именно этот метод считается в настоящее время одним из наиболее эффективных способов оптимизации функции нескольких переменных. Применяется для не очень больших сетей, так как требует относительно большой вычислительной сложности, связанной с необходимостью расчета в каждом цикле n^2 элементов гессиана, и значительных объемов памяти.