Лекция 5. Обзор алгоритмов обучения

Буряк Д.Ю.
к.ф.-м.н
dyb04@yandex.ru

Градиентные методы обучения

Минимизация целевой функции :

d - желаемый выход

y - реальный выход

$$E(w)=\frac{1}{2}\sum_j (y_j-d_j)^2$$

Метод обучения - градиентный спуск.

$$E(w+p)=E(w)+[g(w)]^T p+\frac{1}{2}p^T H(w)p+\ldots$$

$$g(w)=\nabla E=[\frac{\partial E}{\partial w_1},\frac{\partial E}{\partial w_2},\ldots,\frac{\partial E}{\partial w_n}]$$

H(w)-матрица вторых производных;

w - минимум *E(w),* если *g(w)=0, H(w)* – положительно определена

Обозначим w_k -решение, полученное на *k*-ом шаге.

Цель: подобрать *p* и η так, чтобы, для очередной точки w_{k+1} выполнялось условие:

$$w_{k+1}=w_k+\eta_k p_k$$

$$E(w_{k+1})<E(w_k)$$

Схема универсального алгоритма обучения

1. Проверка оптимальности текущего решения w_k.

2. Определение вектора направления оптимизации p_k для точки w_k.

3. Выбор шага η_k в направлении p_k, при котором выполняется условие
$E(w_{k+1})<E(w_k)$

4. Определение нового решения $w_{k+1} = w_k + \eta_k p_k$

а также соответствующих ему значений функции ошибки, градиента, матрицы вторых производных.

Переход на 1.

Алгоритм наискорейшего спуска

Линейное приближение функции $E(w)$.
Для выполнения условия $E(w_{k+1})<E(w_k)$, надо $g(w_k)^T p<0$

$$p_k = -g(w_k)$$

Достоинства:
- небольшая вычислительная сложность;
- невысокие требования к памяти.

Недостатки:
- не учитывает информацию о кривизне функции;
- замедление в случае маленького градиента;
- медленная сходимость.

Наискорейший спуск с моментом

$$\Delta w_k = \eta_k p_k + \alpha(w_k - w_{k-1})$$

α - коэффициент момента в интервале [0,1]

На плоских участках: $\Delta w_k = \eta p_k + \alpha \Delta w_k$

$$\Delta w_k = \frac{\eta}{1-\alpha} p_k$$

Достоинство: ускорение сходимости на плоских участках и вблизи локальных экстремумов.

Метод переменной метрики

Квадратичное приближение *E(w)*

$$E(w_k + p_k) = E(w_k) + [g(w_k)]^T p_k + \frac{1}{2} p_k^T H(w_k) p_k$$

$$\frac{dE(w_k + p_k)}{dp_k} = 0 \quad g(w_k) + H(w_k) p_k = 0$$

$$p_k = -[H(w_k)]^{-1} g(w_k)$$

Используют приближение *H(w)*: $V_k = V_{k-1} + F(w_k, w_{k-1}, g(w_k), g(w_{k-1}))$

где: V_k - приближение $[H(w_k)]^{-1}$

Достоинство: быстрая сходимость.

Недостатки:

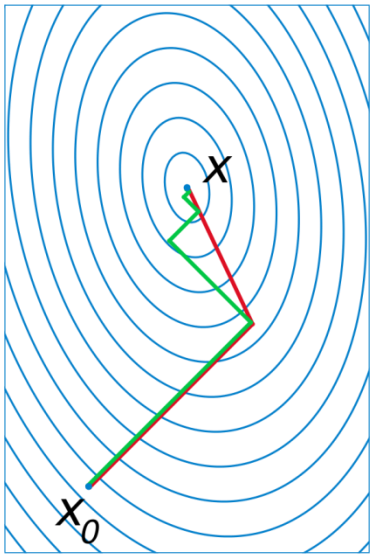
- вычислительная сложность;
- требования к памяти.

Алгоритм сопряженных градиентов

Направление поиска p_k выбирается таким образом, чтобы оно было ортогональным и сопряженным ко всем предыдущим направлениям p_0, p_1,…, p_{k-1}.

$$p_k = -g(w_k) + \beta_{k-1} p_{k-1}$$

$$\beta_{k-1} = \frac{g(w_k)^T (g(w_k) - g(w_{k-1}))}{g(w_{k-1})^T g(w_{k-1})} \quad \text{или} \quad \beta_{k-1} = \frac{g(w_k)^T (g(w_k) - g(w_{k-1}))}{-p_{k-1} g(w_{k-1})}$$



Достоинство:
- быстрее наискорейшего спуска;
- невысокие требования к памяти.

Недостаток: менее эффективен метода переменной метрики.

Алгоритм Quickprop

$$\Delta w_{ij}(k) = -\eta_k[\frac{\partial E(w)}{\partial w_{ij}} + \gamma w_{ij}(k)] + \alpha_{ij}(k)\Delta w_{ij}(k-1)$$

γ - определяется пользователем (10^{-4})

η_k =0, если *k>0* и $[\frac{\partial E(w)}{\partial w_{ij}} + \gamma w_{ij}(k)]\Delta w_{ij}(k-1) < 0$

$$\alpha_{ij} = \begin{cases} \alpha_{\max}, \beta_{ij} > \alpha_{\max}, S_{ij}(k)\Delta w_{ij}(k-1)\beta_{ij}(k) < 0 \\ \beta_{ij} \end{cases}$$

$$S_{ij}(k) = \frac{\partial E(w)}{\partial w_{ij}} + \gamma w_{ij}(k) \quad \beta_{ij}(k) = \frac{S_{ij}(k)}{S_{ij}(k-1) - S_{ij}(k)}$$

Упрощенная версия Quickprop

$$\Delta w_{ij}(k) = \begin{cases} \alpha_{ij}(k)\Delta w_{ij}(k-1), \Delta w_{ij}(k-1) \neq 0 \\ \eta_0 \frac{\partial E}{\partial w_{ij}} \end{cases}$$

$$\alpha_{ij}(k) = \min\{\frac{S_{ij}(k)}{S_{ij}(k-1) - S_{ij}(k)}, \alpha_{\max}\}$$

$$S_{ij}(k) = \frac{\partial E(w)}{\partial w_{ij}}$$

Алгоритм RPROP

$$\Delta w_{ij}(k) = -\eta_{ij}(k)\,\mathrm{sgn}\left(\frac{\partial E(w(k))}{\partial w_{ij}}\right)$$

$$\eta_{ij}(k) = \begin{cases} \min(1.2\eta_{ij}(k-1), \eta_{\max}), & S_{ij}(k)S_{ij}(k-1) > 0 \\ \max(0.5\eta_{ij}(k-1), \eta_{\min}), & S_{ij}(k)S_{ij}(k-1) < 0 \\ \eta_{ij}(k-1) \end{cases}$$

$$S_{ij}(k) = \frac{\partial E(w)}{\partial w_{ij}}$$

$$\eta_{\min} = 10^{-6}$$

$$\eta_{\max} = 50$$

Простейшие методы подбора коэффициента обучения

$$\eta = \min(\frac{1}{n_i})$$

Если $\varepsilon_i > k_w \varepsilon_{i-1}$ то $\eta_{i+1} = \eta_i \rho_d$

Если $\varepsilon_i \le k_w \varepsilon_{i-1}$ то $\eta_{i+1} = \eta_i \rho_i$

$$\varepsilon = \sqrt{\sum_j (y_j - d_j)^2}$$

Подбор на основе направленной минимизации

$$E(w) \to P_2(\eta) = a_2\eta^2 + a_1\eta + a_0$$

$$E(w) \to P_2(\eta) = a_3\eta^3 + a_2\eta^2 + a_1\eta + a_0$$

$$w_1 = w + \eta_1 p_k$$

$$w_2 = w + \eta_2 p_k$$

$$w_3 = w + \eta_3 p_k$$

$$\begin{cases} P_2(\eta_1) = E(w_1) \\ P_2(\eta_2) = E(w_2) \\ P_2(\eta_3) = E(w_3) \end{cases}$$

$$\eta = \frac{-a_1}{2a_2}$$

$$\eta = \frac{-a_2 + \sqrt{{a_1}^2 - 3a_2a_1}}{3a_3}$$

Стохастический метод обучения. Общая схема

1. Выполнить начальную инициализацию весов.

2. Вычислить выход для примера из обучающей выборки.

3. Вычислить ошибку:

4. Выбрать случайным образом вес и изменить его значение на случайную величину.
Если проведенное изменение уменьшает целевую функцию (ошибку), то сохранить его, иначе вернуться к прежнему значению веса.

5. Повторять шаги 2, 3, 4, пока сеть не обучится.

Алгоритм имитации отжига

1. Выполнить начальную инициализацию весов. Задать начальную температуру $T=T_{max}$. Вычислить значение ошибки.

2. Пока $T>0$ повторить L раз действия:
2.1. Выполнить случайную коррекцию весов $w'=w+\Delta w$.
2.2. Вычислить разницу целевых функций $c=E(w')-E(w)$.
2.3. Если $c<0$, то $w=w'$
иначе принять $w=w'$ с вероятностью $P(c)=e^{-\frac{c}{kT}}$

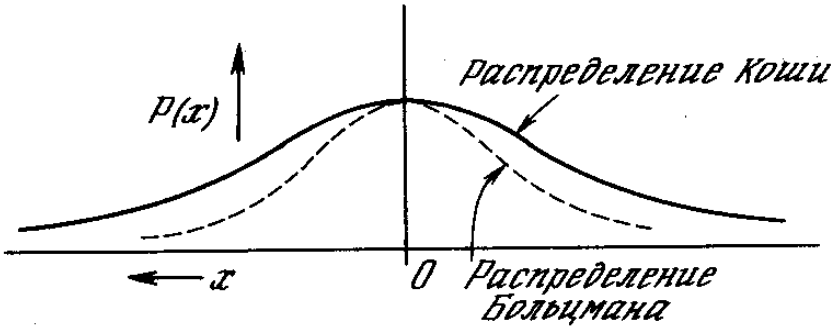
3. Уменьшить температуру T.
4. Повторять 2, 3, 4 пока $T>0$ или значение ошибки больше порога.

Приращение весов Δw: $P(\Delta w)=e^{-\frac{\Delta w^2}{T^2}}$

Алгоритм имитации отжига (Коши)

Скорость сходимости метода Больцмана: $T(t) = \frac{T_{\max}}{\log(1+t)}$

Распределение Коши: $P(x) = \frac{T(t)}{T(t)^2 + x^2}$



Скорость сходимости метода Коши: $T(t) = \frac{T_{\max}}{1+t}$

Комбинированный метод обучения

Объединение традиционного обратного распространения с обучением Коши.

$$\Delta w_k = \beta[\eta_k p_k + \alpha(w_k - w_{k-1})] + (1-\beta)\Delta w^c$$