Оценка гиперпараметров нелинейнных регрессионных моделей
Материалы лекции[1] 3 ноября 2009 г.

Часть 1.

1 Аппроксимация Лапласа

Рассмотрим непрерывную случайную переменную x и ее распределение $P(x)$,

$$P(x) = \frac{1}{Z_P} P^*(x),$$

включающее нормирующую константу Z_P ненормированной функци $P^*(x)$,

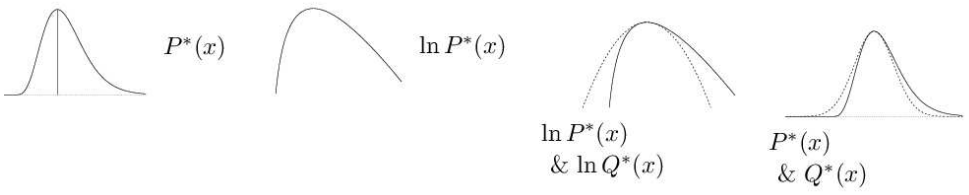
$$Z_P \equiv \int P^*(x)dx$$

Нормирующая константа Z_P неизвестна, требуется ее оценить. Для оценки используем приближение функции $P(x)$ гауссианой $Q(x)$, максимум которой совпадает с модой распределения. Найдем моду $P(x)$, а именно точку x_0, в которой $P'(x_0) = 0$, иначе

$$\left.\frac{\partial P(x)}{\partial x}\right|_{x=x_0}.$$

.

Пусть $P^*(x)$ имеет пик (локальный максимум) в точке x_0. Ряд Тейлора логариф-



ма $P^*(x)$ в окрестности этого максимума имеет вид

$$\ln P^*(x) = \ln P^*(x_0) - \frac{c}{2}(x - x_0)^2 + \cdots, \tag{1}$$

где

$$c = -\left.\frac{\partial^2}{\partial x^2} \ln P^*(x)\right|_{x=x_0}.$$

В этом разложении нет члена первого порядка, так как функция $P^*(x)$ принимает максимум при $x = x_0$.

[1]Прикладной регрессионный анализ (курс лекций, В.В.Стрижов, 2009)

Аппроксимация $P^*(x)$ ненормированным гауссианом имеет вид

$$Q^*(x) \equiv P^*(x_0) \exp\left[-\frac{c}{2}(x-x_0)^2\right],$$

Аппроксимация нормирующей константы Z_P посредством нормирующей константы этого гауссиана имеет вид

$$Z_Q = P^*(x_0)\sqrt{\frac{2\pi}{c}}$$

и приближение функцией Q имеет вид

$$Q(x) = \sqrt{\frac{c}{2\pi}} \exp\left(-\frac{c}{2}(x-x_0)^2\right).$$

Рассмотрим аппроксимацию Лапласа для многомерной случайной величины с распределением

$$P(\mathbf{x}) = \frac{P^*(\mathbf{x})}{Z_P}.$$

В точке $\mathbf{x}_0$ градиент $\nabla P(\mathbf{x}) = \mathbf{0}$. Разложение в окрестности этой точки имеет вид

$$\ln P^*(\mathbf{x}) = \ln P^*(\mathbf{x}_0) - \frac{1}{2}(\mathbf{x}-\mathbf{x}_0)^T H(\mathbf{x}-\mathbf{x}_0) + \ldots,$$

где матрица Гессе

$$H = -\nabla\nabla \ln P^*(\mathbf{x})|_{\mathbf{x}=\mathbf{x}_0}.$$

Экспоненциируя обе части разложения, получаем

$$P^*(\mathbf{x}) \approx P^*(\mathbf{x}_0) \exp\left(-\frac{1}{2}(\mathbf{x}-\mathbf{x}_0)^T H(\mathbf{x}-\mathbf{x}_0)^2\right).$$

Искомая аппроксимация $Q(\mathbf{x})$ пропорциональна $P(\mathbf{x})$ и, вместе с нормирующим коэффициентом, имеет вид

$$Q(\mathbf{x}) = \frac{|H|^{\frac{1}{2}}}{(2\pi)^{\frac{W}{2}}} \exp\left(-\frac{1}{2}(\mathbf{x}-\mathbf{x}_0)^T H(\mathbf{x}-\mathbf{x}_0)^2\right) = \mathcal{N}(\mathbf{x}|\mathbf{x}_0, H^{-1}).$$

Часть 2.

1 Оценка гиперпараметров в двухуровневом Байесовском выводе

Задана регрессионная выборка — множество пар $D = \{(\mathbf{x}_n, y_n)\}_{n=1}^N$, в котором $\mathbf{x} \in \mathbf{R}^P$ — свободная переменная и $y \in \mathbf{R}^1$ — зависимая переменная.

Задано конечное множество порождающих функций $G = \{g | g : \mathbb{R} \times \ldots \times \mathbb{R} \to \mathbb{R}\}$. Функция $g = g(\mathbf{b}, \cdot, \cdot, \ldots, \cdot)$ — гладкая параметрическая. Первый аргумент функции — вектор параметров, последующие аргументы — функции свободных переменных, принимающие значения в $\mathbb{R}^1$. Множество G индуктивно определяет набор допустимых суперпозиций $F = \{f_i\}$, $i = 1, \ldots, M$. На эти суперпозиции

Суперпозиция f_i определяет параметрическую регрессионную модель $f_i = f_i(\mathbf{w}, \mathbf{x})$. Она зависит от независимых переменных $\mathbf{x}$ и вектора параметров $\mathbf{w}$. Вектор $\mathbf{w} \in \mathbb{R}^{W_i}$ состоит из присоединенных векторов — параметров функций $g_1, \ldots, g_{r_i}$, входящих в эту суперпозицию в лексикографическом порядке, то есть $\mathbf{w} = \mathbf{b}_1 \vdots \mathbf{b}_2 \vdots \ldots \vdots \mathbf{b}_{r_i}$, где $\vdots$ — знак присоединения векторов. Требуется отыскать в множестве F модель f_i, максимизирующую заданную целевую функцию $p(\mathbf{w}|D, A, \beta, f_i)$. Функция включает гиперпараметры A, β. Число параметров модели не должно превышать заданное число W^*. Число порождающих функций, из которых она состоит не должно превышать заданное число r^*. Модель, удовлетворяющую вышеперечисленным требованиям, будем называть моделью оптимальной структуры.

2 Распределение параметров моделей

Воспользуемся двухуровневым Байесовским выводом для оценки степени предпочтения порождаемых регрессионной моделей. Рассмотрим конечное множество моделей $f_1, \ldots, f_M$, приближающих данные D, обозначим априорную вероятность i-ой модели $P(f_i)$. При появлении данных апостериорная вероятность модели $P(f_i|D)$ равна

$$P(f_i|D) = \frac{p(D|f_i)P(f_i)}{\sum_{j=1}^M p(D|f_j)P(f_j)}, \tag{1}$$

где $p(D|f_i)$ — функция правдоподобия моделей, определяющая, насколько хорошо модель f_i описывает данные D. Знаменатель дроби обеспечивает выполнение условия $\sum_{i=1}^M P(f_i|D) = 1$.

Сравним две модели с помощью апостериорных вероятностей

$$\frac{P(f_i|D)}{P(f_j|D)} = \frac{p(D|f_i)P(f_i)}{p(D|f_j)P(f_j)}. \tag{2}$$

Левая часть выражения называется отношением правдоподобия моделей. Отношение $P(f_i)/P(f_j)$ называется отношением апостериорных предпочтений моделей. Полагая априорные вероятности моделей одинаковыми, используем функции правдоподобия для выбора моделей.

Так как рассматриваемые модели f зависят от настраиваемых параметров, представим правдоподобие моделей в виде интеграла по пространству параметров

$$p(D|f) = \int p(D|\mathbf{w}, f)p(\mathbf{w}|f)d\mathbf{w}. \tag{3}$$

Априорная плотность распределения параметров $\mathbf{w}$ модели f на выборке D равна

$$p(\mathbf{w}|D, f) = \frac{p(D|\mathbf{w}, f)p(\mathbf{w}|f)}{p(D|f)}, \tag{4}$$

где $p(\mathbf{w}|f)$ — априорно заданная плотность вероятности параметров и $p(D|\mathbf{w}, f)$ — функция правдоподобия параметров. Выражения (1) и (4) называются формулами Байесовского вывода первого и второго уровня.

Рассмотрим следующую гипотезу порождения данных при восстановлении регрессии

$$y = f(\mathbf{w}, \mathbf{x}) + \nu.$$

Пусть случайная величина ν имеет нормальное распределение $\mathcal{N}(0, \sigma^2)$ с нулевым матожиданием и дисперсией σ^2, которая не зависит от свободной переменной. Обозначим $D = (X, \mathbf{y})$, где $\mathbf{y} = [y_1, \ldots, y_N]^T$ — вектор значений зависимой переменной и X — матрица

$$X = \begin{bmatrix} \mathbf{x}_1 \\ \vdots \\ \mathbf{x}_N \end{bmatrix}.$$

Полагая $\mathbf{y}$ многомерной случайной величиной, имеем $\mathbf{y} \sim \mathcal{N}(\mathbf{f}, \beta I_N)$, где I_N единичная матрица и $\beta = \sigma^{-2}$. Для фиксированной модели f плотность вероятности появления данных

$$p(y|\mathbf{x}, \mathbf{w}, \beta, f) \equiv p(D|\mathbf{w}, \beta, f) = \frac{\exp(-\beta E_D)}{Z_D(\beta)}, \tag{5}$$

где $\beta = \sigma^{-2}$, а коэффициент Z_D задан выражением, нормирующим функцию плотности в соответствии с гауссовым распределением

$$Z_D(\beta) = \left(\frac{2\pi}{\beta}\right)^{\frac{N}{2}}. \tag{6}$$

$$Z_D(\beta) = |\sigma^2 I_N|^{\frac{1}{2}}(2\pi)^{\frac{N}{2}} = \sigma^{\frac{2N}{2}}|I_N|^{\frac{1}{2}}(2\pi)^{\frac{N}{2}} = \sigma^N(2\pi)^{\frac{1}{2}} = \left(\frac{2\pi}{\beta}\right)^{\frac{N}{2}}.$$

Функция регрессионных невязок, согласно гипотезе порождения данных, равна

$$E_D = \frac{1}{2}\sum_{n=1}^{N}\left(f(x_n) - y_n\right)^2. \tag{7}$$

$$\beta E_D(D|\mathbf{w}, f) = -\frac{1}{2}(\mathbf{f}-\mathbf{y})^T(\sigma^2 I_N)^{-1}(\mathbf{f}-\mathbf{y}) == \frac{\beta}{2}\sum_{n=1}^{N}(f_n - y_n)^2,$$

$$(\sigma^2 I_N)^{-1} = \frac{I_N}{\sigma^2} = \beta I_N.$$

Рассмотрим вектор параметров модели как многомерную случайную величину $\mathbf{w}$. Пусть плотность распределения параметров имеет вид многомерного нормального распределения $\mathcal{N}(\mathbf{0}, A)$ с матрицей ковариации A,

$$p(\mathbf{w}|A, f) = \frac{\exp(-E_\mathbf{w})}{Z_\mathbf{w}(A)}, \tag{8}$$

где A — ковариационная матрица случайной величины $\mathbf{w}$. Нормирующая константа $Z_\mathbf{w}(A)$ равна

$$Z_\mathbf{w}(A) = (2\pi)^{\frac{W}{2}}|A|^{\frac{1}{2}}, \tag{9}$$

где W — число параметров модели f. Функция-штраф за большое значение параметров модели при нормальном распределении равна

$$E_\mathbf{w} = \frac{1}{2}\mathbf{w}^T A\mathbf{w}. \tag{10}$$

При заданной модели f и заданных значениях A и β выражение (4) принимает вид

$$p(\mathbf{w}|D, A, \beta, f) = \frac{p(D|\mathbf{w}, \beta, f)p(\mathbf{w}|A, f)}{p(D|A, \beta, f)}. \tag{11}$$

Записывая функцию ошибки

$$S(\mathbf{w}) = \frac{1}{2}\mathbf{w}^T A\mathbf{w} + \beta E_D, \tag{12}$$

получаем вместо (11) выражение

$$p(\mathbf{w}|D, A, \beta, f) \propto \frac{\exp(-S(\mathbf{w}))}{Z_S},$$

где Z_S — нормирующий множитель. Символ f далее будет опущен для удобства обозначений.

3 Вычисление гиперпараметров

Предлагается итеративно найти параметры и гиперпараметры модели по отдельности. На каждой итерации сначала при фиксированных гиперпараметрах отыскиваются параметры путем оптимизации функционала (12). Используется алгоритм Левенберга-Марквардта. Затем по формулам, предложенным ниже, вычисляются гиперпараметры.

Предположим, что после очередного шага итерации нам известен локальный максимум (12) и он находится в точке $\mathbf{w}_0$. Для нахождения гиперпараметров приблизим (11) методом Лапласа. Для этого построим ряд Тейлора второго порядка логарифма числителя (11) в окрестности $\mathbf{w}_0$

$$-S(\mathbf{w}) \approx -S(\mathbf{w}_0) - \frac{1}{2}\Delta\mathbf{w}^T H \Delta\mathbf{w}, \tag{13}$$

Фактически,

$$\ln\exp(-S(\mathbf{w})) = \ln\exp(S(\mathbf{w}_0) + \frac{1}{2}\Delta\mathbf{w}^T H \Delta\mathbf{w} + o(||\mathbf{w}||^3)). \tag{14}$$

где $\Delta\mathbf{w} = \mathbf{w} - \mathbf{w}_0$. В выражении (13) нет слагаемого первого порядка, так как предполагается, что $\mathbf{w}_0$ доставляет локальный минимум функции ошибки

$$\left.\frac{\partial S(\mathbf{w})}{\partial \mathbf{w}}\right|_{\mathbf{w}=\mathbf{w}_0} = \mathbf{0}.$$

Матрица H — матрица Гессе функции ошибок

$$H = -\nabla\nabla S(\mathbf{w})|_{\mathbf{w}=\mathbf{w}_0}. \tag{15}$$

Применяя экспоненту к обеим частям выражения (13) получаем требуемое приближение числителя (11)

$$\exp(-S(\mathbf{w})) \approx \exp(-S(\mathbf{w}_0))\exp(-\frac{1}{2}\Delta\mathbf{w}^T H \Delta\mathbf{w}). \tag{16}$$

Учитывая то, что интеграл выражения (11) должен равняться единице, получаем нормирующий множитель

$$Z_S = \frac{\exp(-S(\mathbf{w}_0))(2\pi)^{\frac{W}{2}}}{|H|^{\frac{1}{2}}}. \tag{17}$$

$$p(\mathbf{w}|D, A, \beta) = \frac{\exp(-S(\mathbf{w}_0))\exp(-\frac{1}{2}\Delta\mathbf{w}^T H \Delta\mathbf{w})}{Z_S(A, \beta)}. \tag{18}$$

Знаменатель (11) является числитетем (1) и определяет выбор наиболее правдоподобной модели. Для нахождения гиперпараметров максимизируем функцию $p(D|A, \beta)$ относительно A и β. Запишем ее в виде

$$p(D|A, \beta) = \int p(D|\mathbf{w}, A, \beta)p(\mathbf{w}|A)d\mathbf{w}. \tag{19}$$

Используя выражения (5) и (8) перепишем (19) в виде

$$p(D|\beta, A) = \frac{1}{Z_{\mathbf{w}}(A)} \frac{1}{Z_D(\beta)} \int \exp(-S(\mathbf{w}))d\mathbf{w}.$$

$$p(D|\beta, A) = \frac{Z_S}{Z_{\mathbf{w}}(A) Z_D(\beta)}.$$

Из (6), (9)и (17), логарифмируя (19), получим

$$p(D|A, \beta) = \frac{1}{Z_{\mathbf{w}}(A)} \frac{1}{Z_D(\beta)} \exp(-S(\mathbf{w}_0))(2\pi)^{\frac{W}{2}} |H|^{-\frac{1}{2}}.$$

$$\ln p(D|\beta, A) = -\mathbf{w}^T H \mathbf{w} - \beta E_D - \frac{1}{2} \ln |A| +$$

$$+ \frac{\mathbf{w}}{2} \ln A + \frac{N}{2} \ln \beta + \frac{N}{2} \ln(2\pi). \tag{20}$$

$$\ln p(D|A, \beta) = -\frac{1}{2} \ln |A| - \frac{N}{2} \ln 2\pi + \frac{N}{2} \ln \beta$$

$$-\beta E'_D - E'_{\mathbf{w}} - \frac{1}{2} \ln |H|. \tag{21}$$

$$\ln p(D|\beta, A) = \underbrace{-\frac{W}{2} \ln 2\pi - \frac{1}{2} \ln |A|}_{Z_{\mathbf{w}}^{-1}(A)} \underbrace{-\frac{N}{2} \ln 2\pi + \frac{N}{2} \ln \beta}_{Z_D^{-1}(\beta)} \underbrace{-S(\mathbf{w}_0) + \frac{W}{2} \ln 2\pi - \frac{1}{2} \ln |H|}_{Z_S}.$$

$$= -\frac{1}{2} \ln |A| - \frac{N}{2} \ln 2\pi + \frac{N}{2} \ln \beta \underbrace{-\beta E_D - E_{\mathbf{w}}}_{-S(\mathbf{w}_0)} - \frac{1}{2} \ln |H|.$$

Найдем максимум выражения (21) относительно гиперпараметров, приравняв его производную поочередно по A и β к нулю. Для упрощения вычислений представим $A = \mathrm{diag}(\alpha) I_W$.

$$\frac{\partial \ln p(D|\beta, A)}{\partial A} = -E_{\mathbf{w}}(\mathbf{w}_0) + \frac{\mathbf{w}}{2\alpha} + \frac{d}{d\alpha} \ln \det(H).$$

$$\frac{d \ln p(D|A, \beta)}{d\alpha} = -E'_{\mathbf{w}} + \frac{\mathbf{w}}{2\alpha} + \frac{d}{d\alpha} \ln \det(H).$$

Производная последнего слагаемого равна

$$\frac{d}{d\alpha} \ln |H| =$$

$$\frac{d}{d\alpha} \ln(\prod_{j=1}^{W} \lambda_j + \alpha) =$$

$$= \frac{d}{d\alpha} \sum_{j=1}^{W} \ln(\lambda_j + \alpha)$$

$$= \sum_{j=1}^{W} \frac{1}{\lambda_j + \alpha},$$

где λ_j — собственные значения матрицы H. Приравнивая последнее выражение к нулю и преобразовывая, получаем выражение для α

$$2\alpha E'_{\mathbf{w}} = W - \gamma, \quad \text{где} \quad \gamma = \sum_{j=1}^{W} \frac{\alpha}{\lambda_j + \alpha}. \tag{22}$$

Обозначим вычитаемое правой части через γ

$$\gamma = \sum_{j=1}^{W} \frac{\alpha}{\lambda_j + \alpha}.$$

Те компоненты суммы, в которых $\lambda_j \gg \alpha$ привносят вклад, близкий к единице, а те компоненты суммы, в которых $0 < \lambda_j \ll \alpha$, привносят вклад, близкий к нулю. Таким образом γ может быть интерпретирована как мера числа хорошо обусловленных параметров модели.

Для нахождения гиперпараметра β рассмотрим задачу оптимизации (**??**). Обозначим через μ_j собственные значение матрицы $\nabla^2 E_D$. Так как $H = \beta \nabla^2 E_D$, то $\lambda_j = \beta \mu_j$ и следовательно,

$$\frac{d\lambda_j}{d\beta} = \mu_j = \frac{\lambda_j}{\beta}.$$

Отсюда,

$$\frac{d}{d\beta} \ln |A| = \frac{d}{d\beta} \sum_{j=1}^{W} \ln(\lambda_j + \alpha) = \frac{1}{\beta} \sum_{j=1}^{W} \frac{\lambda_j}{\lambda_j + \alpha}.$$

Дифференцируя, как и в случае нахождения α, мы находим, что

Аналогично получим β

$$2\beta E'_D = N - \sum_{j=1}^{W} \frac{\lambda_j}{\lambda_j + \alpha} = N - \gamma. \tag{23}$$

Гиперпараметры α и β_i вычисляется итеративно следующим образом

$$\beta^{\text{new}} = \frac{N - \gamma}{E'_D}, \quad \alpha^{\text{new}} = \frac{W - \gamma}{E'_{\mathbf{w}}}.$$

Значения функциналов ошибок $E'_{\mathbf{w}}$ и E'_D оптимизируются после каждого вычисления новых значений гиперпараметров.

При выборе моделей выполняется следующая процедура. Экспертно задается модель-претендент. Каждому элементу модели ставится в соответствие свой гиперпараметр α. Параметры и гиперпараметры модели последовательно настраиваются. Элемент модели, имеющий наименьшее значение гиперпараметра, исключается. Модель пополняется новым элементом из множества G согласно заданному правилу. Так как на каждом шаге такой модификации модели функционал качества не ухудшается, Процедура выполняется до сходимости функционала качества (13).

Часть 3.

1 Нелинейная регрессия

Нелинейная регрессия - частный случай регрессионного анализа, в котором рассматриваемая регрессионная модель есть функция, зависящая от параметров и от одной или нескольких свободных переменных. Зависимость от параметров предполагается нелинейной.

Задана выборка из m пар $(\mathbf{x}_i, y_i)$. Задана регрессионная модель $f(\mathbf{w}, \mathbf{x})$, которая зависит от параметров $\mathbf{w} = (w_1, ..., w_n)$ и свободной переменной x. Требуется найти такие значения параметров, которые доставляли бы минимум сумме квадратов регрессионных остатков

$$E_D = \sum_{i=1}^{m} r_i^2,$$

где остатки $r_i = y_i - f(\mathbf{w}, \mathbf{x}_i)$ для $i = 1, \ldots, m$.

Для нахождения минимума функции E_D, приравняем к нулю её первые частные производные параметрам $\mathbf{w}$:

$$\frac{\partial E_D}{\partial w_j} = 2 \sum_i r_i \frac{\partial r_i}{\partial w_j} = 0 \ (j = 1, \ldots, n). \qquad (*)$$

Так как функция E_D в общем случае не имеет единственного минимума (см. Демиденко, Е. З. Оптимизация и регрессия. М.: Наука. 1989. 296 с.), то предлагается назначить начальное значение вектора параметров w_0 и приближаться к оптимальному вектору по шагам:

$$w_j \approx w_j^{k+1} = w_j^k + \Delta w_j.$$

Здесь k — номер итерации, Δw_j — вектор шага.

На каждом шаге итерации линеаризуем модель с помощью приближения рядом Тейлора относительно параметров $\mathbf{w}^k$

$$f(x_i, \mathbf{w}) \approx f(x_i, \mathbf{w}^k) + \sum_j \frac{\partial f(x_i, \mathbf{w}^k)}{\partial w_j} \left(w_j - w_j^k\right) \approx f(x_i, \mathbf{w}^k) + \sum_j J_{ij} \Delta w_j.$$

Здесь элемент матрицы Якоби J_{ij} — функция параметра w_j; значение свободной переменной $\mathbf{x}_i$ фиксировано. В терминах линеаризованной модели

$$\frac{\partial r_i}{\partial w_j} = -J_{ij}$$

и регрессионные остатки определены как

$$r_i = \Delta y_i - \sum_{j=1}^{n} J_{ij} \Delta w_j; \ \Delta y_i = y_i - f(x_i, \mathbf{w}^k).$$

Подставляя последнее выражение в выражение (*), получаем

$$-2\sum_{i=1}^{m} J_{ij}\left(\Delta y_i - \sum_{s=1}^{n} J_{is}\Delta w_s\right) = 0.$$

Преобразуя, получаем систему из n линейных уравнений, которые называются нормальным уравнением

$$\sum_{i=1}^{m}\sum_{s=1}^{n} J_{ij}J_{is}\Delta w_s = \sum_{i=1}^{m} J_{ij}\Delta y_i (j = 1, n).$$

Запишем нормальное уравнение в матричном обозначении как

$$\left(\mathbf{J^T J}\right)\mathbf{\Delta w} = \mathbf{J^T \Delta y}.$$

В том случае, когда критерий оптимальности регрессионой модели задан как взвешенная сумма квадратов остатков

$$E_D = \sum_{i=1}^{m} W_{ii} r_i^2,$$

нормальное уравнение будет иметь вид

$$\left(\mathbf{J^T W J}\right)\mathbf{\Delta w} = \mathbf{J^T W \Delta y}.$$

Для нахождения оптимальных параметров используются метод сопряженных градиентов, алгоритм Гаусса-Ньютона или алгоритм Левенберга-Марквардта.

2 Алгоритм Левенберга-Марквардта

Задана регрессионная выборка — множество пар $D = \{(\mathbf{x}_i, y_i)\}_{i=1}^m$ свободной переменной $\mathbf{x} \in \mathbb{R}^n$ и зависимой переменной $y \in \mathbb{R}$. Задана регрессионная модель — функция $f(\mathbf{w}, \mathbf{x}_i)$, непрерывно дифференцируемая в области W.

Требуется найти такое значение вектора параметров $\mathbf{w}$, которое бы доставляло локальный минимум функции ошибки

$$E_D = \sum_{i=1}^{m}\big(y_i - f(\mathbf{w}, \mathbf{x}_i)\big)^2. \tag{1}$$

Перед началом работы алгоритма задается начальный вектор параметров $\mathbf{w}$. На каждом шаге итерации этот вектор заменяется на вектор $\mathbf{w} + \Delta\mathbf{w}$. Для оценки приращения $\Delta\mathbf{w}$ используется линейное приближение функции

$$f(\mathbf{w} + \Delta\mathbf{w}, \mathbf{x}) \approx f(\mathbf{w}, \mathbf{x}) + J\Delta\mathbf{w},$$

где J — якобиан функции $f(\mathbf{w}, \mathbf{x}_i)$ в точке $\mathbf{w}$.

Приращение $\Delta\mathbf{w}$ в точке $\mathbf{w}$, доставляющее минимум E_D, равно нулю. Поэтому для нахождения последующего значения приращения $\Delta\mathbf{w}$ приравняем нулю вектор частных производных E_D по $\mathbf{w}$. Для этого представим выражение (1) в виде

$$E_D = \|\mathbf{y} - \mathbf{f}(\mathbf{w} + \Delta\mathbf{w})\|^2,$$

где $\mathbf{y} = [y_1, \ldots, y_m]^T$ и $\mathbf{f}(\mathbf{w} + \Delta\mathbf{w}) = [f(\mathbf{w} + \Delta\mathbf{w}, \mathbf{x}_1), \ldots, f(\mathbf{w} + \Delta\mathbf{w}, \mathbf{x}_m))]^T$. Преобразовывая это выражение

$$\|\mathbf{y} - \mathbf{f}(\mathbf{w} + \Delta\mathbf{w})\|^2 =$$
$$\big(\mathbf{y} - \mathbf{f}(\mathbf{w} + \Delta\mathbf{w})\big)^T \big(\mathbf{y} - \mathbf{f}(\mathbf{w} + \Delta\mathbf{w})\big) =$$
$$\mathbf{f}^T(\mathbf{w} + \Delta\mathbf{w})\mathbf{f}(\mathbf{w}) - 2\mathbf{y}^T\mathbf{f}(\mathbf{w} + \Delta\mathbf{w}) + \mathbf{y}^T\mathbf{y}$$

и дифференцируя, получим

$$\frac{\partial E_D}{\partial \mathbf{w}} = (J^T J)\Delta\mathbf{w} - J^T\big(\mathbf{y} - \mathbf{f}(\mathbf{w})\big) = 0.$$

Таким образом, чтобы найти значение $\Delta\mathbf{w}$, нужно решить систему линейных уравнений

$$\Delta\mathbf{w} = (J^T J)^{-1} J^T\big(\mathbf{y} - \mathbf{f}(\mathbf{w})\big).$$

Так как число обусловленности матрицы $J^T J$ есть квадрат числа обусловленности матрицы J то матрица $J^T J$ может оказаться существенно вырожденной. Поэтому Марквардт предложил ввести параметр регуляризации $\lambda \geq 0$,

$$\Delta\mathbf{w} = (J^T J + \lambda I)^{-1} J^T\big(\mathbf{y} - \mathbf{f}(\mathbf{w})\big),$$

где I — единичная матрица. Этот параметр назначается на каждой итерации алгоритма. Если значение ошибки E_D убывает быстро, малое значение λ сводит этот алгоритм к алгоритму Гаусса-Ньютона.

Алгоритм останавливается в том случае, если приращение $\Delta\mathbf{w}$ в последующей итерации меньше заданного значения либо если параметры $\mathbf{w}$ доставляют ошибку E_D, меньшую заданной величины. Значение вектора $\mathbf{w}$ на последней итерации считается искомым.

Недостаток алгоритма — значительное увеличение параметра λ при плохой скорости аппроксимации. При этом обращение матрицы $J^T J + \lambda I$ становится бессмысленным. Этот недостаток можно устранить, используя диагональ матрицы Гессе $J^T J$ в качестве регуляризующего слагаемого:

$$\Delta\mathbf{w} = \big(J^T J + \lambda \mathrm{diag}(J^T J)\big)^{-1} J^T\big(\mathbf{y} - \mathbf{f}(\mathbf{w})\big).$$