Zaytsev A.A., Strijov V.V., Tokmakova A.A. Estimation regression model hyperparameters using maximum likelihood // Informational Technologies, 2013, 2 : 11-15.

УДК 519.584

А. А. Зайцев, студ., *Московский физико-технический институт,* стажер-исследователь, *Datadvance*

В. В. Стрижов, к.ф.-м.н., н.с., *Вычислительный центр РАН*

А. А. Токмакова, студ., *Московский физико-технический институт*

Оценка гиперпараметров регрессионных моделей методом максимального правдоподобия[1]

Рассматривается задача выбора регрессионной модели. Предполагается, что вектор параметров модели − многомерная случайная величина с независимо распределёнными компонентами. В работе предложен способ оптимизации праметров и гиперпараметров. Приведены явные оценки гиперпараметров для случая линейных и нелинейных моделей. Показано как полученные оценки используются для отбора признаков. Предложенный подход сравнивается с подходом, использующим для лценки гиперпараметров аппроксимацию Лапласа.

Ключевые слова: регрессия, выбор признаков, распределение параметров, оценка гипертараметров, байесовский вывод.

1 Введение

В данной работе рассматривается задача выбора регрессионной модели [1] из заданного параметрического семейства регрессионных моделей. Один из возможных подходов − введение предположения о распределении параметров модели [2]. В этом случае предполагается, что функция регрессии задана оценкой вектора параметров, который считается нормлально

[1] Работа выполнена при поддержке РФФИ, грант 10-07-00422.

распределённой многомерной случайной величиной. Параметры распределения заданы вектором, в дальнейшем называемым вектором гиперпараметров модели.

Впервые этот подход к выбору признаков методом анализа распределения параметров был предложен в работе [3]. Более общий подход был предложен Маккаем в работе [2]. В этой работе Маккай ввел понятие гиперпараметров. Бишоп предложил ряд других способов оценки гиперпараметов, таких как Марковские цепи Монте-Карло и аппроксимация Лапласа [4, 5]. Подход, использующий аппроксимацию Лапласа был развит в работах [6, 7].

Предлагается для линейной регрессионной модели выписать явное выражение функции правдоподобия с учётом введенных вероятностных предположений. Максимизируя функцию правдоподобия, получаем оценки наиболее правдоподобных значений гиперпараметров модели. Такой подход позволяет получать оценки гиперпараметров регрессионных моделей. Для полученных оценок гиперпараметров явно выписываются оценки параметров модели. Они используются для отбора признаков. Предложенный подход сравнивается с подходом, использующим аппроксимацию Лапласа распределения параметров модели [6].

2 Постановка задачи

Задана выборка $D = (\mathbf{X}, \mathbf{y}) = \{(\mathbf{x}_i, y_i)\}_{i=1}^{m}$, где $\mathbf{x}_i \in \mathrm{R}^n$, $y_i \in \mathrm{R}$. Выборка D содержит m элементов. Вектор $\mathbf{x}$ состоит из n независимых переменных. Рассматривается класс регрессионных моделей вида:

$$\mathbf{y} = \mathbf{f}(\mathbf{X}, \mathbf{w}) + \boldsymbol{\varepsilon}. \qquad (1)$$

Здесь $\mathbf{X}$ – матрица плана, а $\mathbf{w}$ – вектор параметров модели $\mathbf{f}$. Предполагается, что шум $\boldsymbol{\varepsilon}$ – многомерная нормальная случайная величина с нулевым математическим ожиданием и матрицей ковариации $\mathbf{B}^{-1}$:

$$\boldsymbol{\varepsilon} : N(0, \mathbf{B}^{-1}), \tag{2}$$

вектор параметров модели $\mathbf{w}$ – многомерная нормальная случайная величина с нулевым математическим ожиданием и матрицей ковариации $\mathbf{A}^{-1}$:

$$\mathbf{w} : N(0, \mathbf{A}^{-1}). \tag{3}$$

Требуется получить оценки матриц $\mathbf{A}$, $\mathbf{B}$ согласно гипотезам порождения данных (2) и (3).

3 Функция правдоподобия линейной модели

Рассмотрим линейную регрессионную модель. Тогда выражение (1) имеет вид

$$\mathbf{y} = \mathbf{X}\mathbf{w} + \boldsymbol{\varepsilon}.$$

Плотность распределения параметров $\mathbf{w}$ согласно теореме Байеса имеет вид:

$$p(\mathbf{w} \mid \mathbf{A}, \mathbf{B}, D) = \frac{p(D \mid \mathbf{w}, \mathbf{B})\, p(\mathbf{w} \mid \mathbf{A})}{p(D \mid \mathbf{A}, \mathbf{B})}, \tag{4}$$

в котором $p(D \mid \mathbf{w}, \mathbf{B})$, $p(\mathbf{w} \mid \mathbf{A})$ – плотности многомерных нормальных случайных величин вида

$$p(D \mid \mathbf{w}, \mathbf{B}) = \frac{1}{(2\pi)^{\frac{m}{2}} \mid \mathbf{B} \mid^{-\frac{1}{2}}} \exp\left(-\frac{1}{2} (\mathbf{y} - \mathbf{X}\mathbf{w})^T \mathbf{B} (\mathbf{y} - \mathbf{X}\mathbf{w}) \right), \tag{4′}$$

согласно предположению о нормальности распределения шумов (2), и

$$p(\mathbf{w} \mid \mathbf{A}) = \frac{1}{(2\pi)^{\frac{n}{2}} \mid \mathbf{A} \mid^{-\frac{1}{2}}} \exp\left(-\frac{1}{2} \mathbf{w}^T \mathbf{A} \mathbf{w} \right), \tag{4″}$$

согласно предположению о распределении вектора параметров модели (3). Функция $p(D \mid \mathbf{A}, \mathbf{B})$ правдоподобия модели $\mathbf{f}$ имеет вид

$$p(D \mid \mathbf{A}, \mathbf{B}) = \int_{R^n} p(D \mid \mathbf{w}, \mathbf{B})\, p(\mathbf{w} \mid \mathbf{A}) d\mathbf{w}. \tag{5}$$

Для линейных моделей явно выпишем оценки гиперпараметров модели $p(D \mid \mathbf{A}, \mathbf{B})$. Отметим, что в работе [6] был предложен подход, в котором эти оценки получены с использованием аппроксимации Лапласа. Верна

следующая теорема.

Теорема. Функция правдоподобия в предположениях о распределении шума $\boldsymbol{\varepsilon}$ (2) и параметров модели $\mathbf{w}$ (3) имеет вид

$$p(D \mid \mathbf{A}, \mathbf{B}) = \frac{|\mathbf{B}|^{\frac{1}{2}} |\mathbf{A}|^{\frac{1}{2}}}{(2\pi)^{\frac{m}{2}} |\mathbf{K}|^{\frac{1}{2}}} \exp\left(\frac{1}{2}\mathbf{y}^T(\mathbf{C}^T\mathbf{K}\mathbf{C} - \mathbf{B})\mathbf{y}\right), \tag{6}$$

а его логарифм имеет вид

$$\ln p(\mathbf{D} \mid \mathbf{A}, \mathbf{B}) = -\frac{1}{2}\left(\ln |\mathbf{K}| + m \ln 2\pi - \ln |\mathbf{B}| - \ln |\mathbf{A}| - \mathbf{y}^T(\mathbf{C}^T\mathbf{K}\mathbf{C} - \mathbf{B})\mathbf{y}\right), \tag{7}$$

где

$$\mathbf{K} = \mathbf{X}^T\mathbf{B}\mathbf{X} + \mathbf{A}, \quad \mathbf{C} = \mathbf{K}^{-1}\mathbf{X}^T\mathbf{B}.$$

Доказательство. Подставляя (4′) и (4″) в (5) получим следующее выражение:

$$p(D \mid \mathbf{A}, \mathbf{B}) = \int_{R^n} \frac{1}{(2\pi)^{\frac{m}{2}} |\mathbf{B}|^{-\frac{1}{2}}} \exp\left(-\frac{1}{2}(\mathbf{y} - \mathbf{X}\mathbf{w})^T\mathbf{B}(\mathbf{y} - \mathbf{X}\mathbf{w})\right) \frac{1}{(2\pi)^{\frac{n}{2}} |\mathbf{A}|^{-\frac{1}{2}}} \exp\left(-\frac{1}{2}\mathbf{w}^T\mathbf{A}\mathbf{w}\right) d\mathbf{w}.$$

Перепишем произведение двух экспонент как экспоненту от их суммы:

$$p(D \mid \mathbf{A}, \mathbf{B}) = \int_{R^n} \frac{|\mathbf{B}|^{\frac{1}{2}} |\mathbf{A}|^{\frac{1}{2}}}{(2\pi)^{\frac{n+m}{2}}} \exp\left(-\frac{1}{2}\left((\mathbf{y} - \mathbf{X}\mathbf{w})^T\mathbf{B}(\mathbf{y} - \mathbf{X}\mathbf{w}) + \mathbf{w}^T A\mathbf{w}\right)\right) d\mathbf{w}.$$

Раскрывая скобки, получим:

$$p(D \mid \mathbf{A}, \mathbf{B}) = \int_{R^n} \frac{|\mathbf{B}|^{\frac{1}{2}} |\mathbf{A}|^{\frac{1}{2}}}{(2\pi)^{\frac{n+m}{2}}} \exp\left(-\frac{1}{2}\left(\mathbf{w}^T\mathbf{X}^T\mathbf{B}\mathbf{X}\mathbf{w} - 2\mathbf{w}^T\mathbf{X}^T\mathbf{B}\mathbf{y} + \mathbf{y}^T\mathbf{B}\mathbf{y} + \mathbf{w}^T\mathbf{A}\mathbf{w}\right)\right) d\mathbf{w}.$$

Введем обозначения $\mathbf{K} = \mathbf{A} + \mathbf{X}^T\mathbf{B}\mathbf{X}, \mathbf{C} = \mathbf{K}^{-1}\mathbf{X}^T$ и, выделяя полный квадрат по выражению $(\mathbf{w} - \mathbf{C}\mathbf{y})$, получим:

$$p(D \mid \mathbf{A}, \mathbf{B}) = \int_{R^n} \frac{|\mathbf{B}|^{\frac{1}{2}} |\mathbf{A}|^{\frac{1}{2}}}{(2\pi)^{\frac{n+m}{2}}} \exp\left(-\frac{1}{2}\left((\mathbf{w} - \mathbf{C}\mathbf{y})^T\mathbf{K}(\mathbf{w} - \mathbf{C}\mathbf{y}) - \mathbf{y}^T(\mathbf{C}^T\mathbf{K}\mathbf{C} - \mathbf{B})\mathbf{y}\right)\right) d\mathbf{w}.$$

Так как интеграл плотности многомерного нормального распределения по

вектору параметров равен единице, то

$$p(D \mid \mathbf{A},\mathbf{B}) = \frac{|\mathbf{B}|^{\frac{1}{2}}|\mathbf{A}|^{\frac{1}{2}}}{(2\pi)^{\frac{m}{2}}|\mathbf{K}|^{\frac{1}{2}}} \exp\left(\frac{1}{2}\left(\mathbf{y}^T(\mathbf{C}^T\mathbf{K}\mathbf{C}-\mathbf{B})\mathbf{y}\right)\right).$$

Следовательно, искомая функция правдоподобия модели $p(D \mid \mathbf{A},\mathbf{B})$ имеет вид (6), а его логарифм − вид (7). ■

Рассмотрим теперь случай, когда матрица $\mathbf{A}$ − диагональная, а матрица $\mathbf{B} = \beta\mathbf{I}$., где $\mathbf{I}$ − единичная матрица размера $m \times m$. Логарифм функции правдоподобия $\ln p(D \mid \mathbf{A},\mathbf{B})$ имеет вид

$$\ln p(\mathbf{D} \mid \mathbf{A},\beta) = -\frac{1}{2}\left(\ln|\mathbf{K}| + m\ln 2\pi - m\ln\beta - \ln|\mathbf{A}| - \beta\mathbf{y}^T(\beta\mathbf{X}\mathbf{K}^{-1}\mathbf{X}^T - \mathbf{I})\mathbf{y}\right),$$

где $\mathbf{K} = \mathbf{A} + \beta\mathbf{X}^T\mathbf{X}$.

3.1 Вычисление производных функции правдоподобия модели $\ln p(\mathbf{D} \mid \mathbf{A},\mathbf{B})$ по гиперпараметрам $\mathbf{A},\mathbf{B}$

Для поиска максимума функции правдоподобия будем пользоваться градиентными методами оптимизации [8], поэтому нам понадобится выражения для производных $\ln p(D \mid \mathbf{A},\mathbf{B})$ по гиперпараметрам $\mathbf{A},\mathbf{B}$.

Пусть матрица $\mathbf{A}$ имеет вид $\mathbf{A} = \{\alpha_{ij}\}, i, j = \overline{1,n}$, а матрица $\mathbf{B}$ имеет вид $\mathbf{B} = \{\beta_{ij}\}, i, j = \overline{1,m}$. Обе матрицы являются симметричными и неотрицательно определенными, так как являются матрицами ковариации.

Верны следующие два свойства производных матриц [9]. Для симметричной матрицы $\mathbf{M}$ верно, что

$$\frac{\partial \ln|\mathbf{M}|}{\partial t} = \mathrm{tr}\left(\mathbf{M}^{-1}\frac{\partial \mathbf{M}}{\partial t}\right),$$

где t − некоторый параметр, $\mathbf{M} = \mathbf{M}(t)$ и tr – след матрицы. Так же верно, что

$$\frac{\partial \mathbf{M}^{-1}}{\partial t} = -\mathbf{M}^{-1}\frac{\partial \mathbf{M}}{\partial t}\mathbf{M}^{-1}.$$

Введем обозначение $\mathbf{S}^{ij}$ – такая матрица, что для двух индексов k,l верно

$$\mathbf{S}_{kl}^{ij} = \begin{cases} 1 & k = i, l = j \text{ или } k = j, l = i, \\ 0 & \text{иначе}. \end{cases}$$

Запишем производную $\ln p(D \mid \mathbf{A}, \beta)$ по β_{ij}:

$$\frac{\partial \ln p(D \mid \mathbf{A}, \mathbf{B})}{\partial \beta_{ij}} = -\frac{1}{2}(\quad \mathrm{tr}\left(\mathbf{K}^{-1}\mathbf{X}^T\mathbf{S}^{ij}\mathbf{X}\right) - \mathrm{tr}\left(\mathbf{B}^{-1}\mathbf{S}^{ij}\right) - \mathbf{y}^T(\mathbf{S}^{ji}\mathbf{X}\mathbf{K}^{-1}\mathbf{X}^T\mathbf{B} + \mathbf{B}^T\mathbf{X}\mathbf{K}^{-1}\mathbf{X}^T\mathbf{S}^{ij} - \mathbf{B}^T\mathbf{X}\mathbf{K}^{-1}\mathbf{X}^T\mathbf{S}^{ij}\mathbf{X}\mathbf{K}^{-T}\mathbf{X}^T\mathbf{B} - \mathbf{S}^{ij})\mathbf{y}).$$

Аналогично запишем производную $\ln p(D \mid \mathbf{A}, \beta)$ по α_{ij}:

$$\frac{\partial \ln p(D \mid \mathbf{A}, \mathbf{B})}{\partial \alpha_{ij}} = -\frac{1}{2}(\quad \mathrm{tr}\left(\mathbf{K}^{-1}\mathbf{S}^{ij}\right) - \mathrm{tr}\left(\mathbf{A}^{-1}\mathbf{S}^{ij}\right) + \mathbf{y}^T\mathbf{B}^T\mathbf{X}\mathbf{K}^{-1}\mathbf{S}^{ij}\mathbf{K}^{-T}\mathbf{X}^T\mathbf{B}\mathbf{y}).$$

Запишем производные функции правдоподобия по гиперпараметрам $\mathbf{A} = \mathrm{diag}(\alpha_1, \alpha_2, \ldots, \alpha_n)$, $\mathbf{B} = \beta\mathbf{I}$:

$$\frac{\partial \ln p(D \mid \mathbf{A}, \beta)}{\partial \beta} = -\frac{1}{2}(\quad \mathrm{tr}\left(\mathbf{K}^{-1}\mathbf{X}^T\mathbf{X}\right) - \frac{m}{\beta} + \mathbf{y}^T(2\beta\mathbf{X}\mathbf{K}^{-1}\mathbf{X}^T - \mathbf{I} - \beta^2\mathbf{X}\mathbf{K}^{-1}\mathbf{X}^T\mathbf{X}\mathbf{K}^{-1}\mathbf{X}^T)\mathbf{y}),$$

$$\frac{\partial \ln p(D \mid \mathbf{A}, \beta)}{\partial \alpha_i} = -\frac{1}{2}\left(\mathrm{tr}\left(\mathbf{K}^{-1}\mathbf{I}^{ii}\right) - \frac{1}{\alpha_i} - \beta^2\mathbf{y}^T\mathbf{X}\mathbf{K}^{-1}\mathbf{I}^{ii}\mathbf{K}^{-1}\mathbf{X}^T\mathbf{y} \right).$$

Так как получены значения производных функции правдоподобия модели $\ln p(D \mid \mathbf{A}, \mathbf{B})$ по гиперпараметрам $\mathbf{A}, \mathbf{B}$, можно использовать любой градиентный метод оптимизации для оценки гиперпараметров $\mathbf{A}, \mathbf{B}$, максимизирующих функцию правдоподобия.

4 Функция правдоподобия нелинейных регрессионных моделей

Функция правдоподобия и её производные по гиперпараметрам могут не выписываться явно для нелинейных регрессионных моделей, в силу того,

что интеграл (5) может не браться аналитически. В этом случае возникает необходимость применения приближенных методов оценки гиперпараметров, таких как, например, аппроксимация Лапласа.

Оценим знаачение интеграла (5) с помощью функции максимального правдоподобия вектора параметров $\mathbf{w}$. Пусть задан вектор параметров $\mathbf{w}_0$, максимизирующих функцию правдоподобия (4), которая с точностью до коэффициента нормализации может быть выписано явно. И пусть вектору $\mathbf{w}_0$ соответствует масимум плотности распределения $p(\mathbf{w} \mid D, \mathbf{A})$. Тогда

$$p(D \mid \mathbf{A}, \mathbf{B}) \approx p_L(D \mid \mathbf{A}, \mathbf{B}) = p(D \mid \mathbf{w}_0, \mathbf{A}, \mathbf{B}) \sqrt{\frac{(2\pi)^n}{|\mathbf{H}|}}, \qquad (8)$$

где $\mathbf{H}$ − гессиан, то есть матрица элементы которой

$$\mathbf{H}_{ij} = -\frac{\partial^2}{\partial w_i \partial w_j} \ln\left(p(D \mid \mathbf{w}, \mathbf{B}) p(\mathbf{w} \mid \mathbf{A})\right).$$

Подставляя выражения для $p(D \mid \mathbf{w}_0, \mathbf{A}, \mathbf{B})$ в (8), получаем,

$$\begin{aligned} \ln p_L(D \mid \mathbf{A}, \mathbf{B}) = \; & -\frac{1}{2}\left((\mathbf{y} - f(\mathbf{w}_0, \mathbf{X}))^T \mathbf{B}(\mathbf{y} - f(\mathbf{w}_0, \mathbf{X})) - m \ln 2\pi + \ln |\mathbf{B}|\right) - \\ & -\frac{1}{2}\left(\mathbf{w}_0^T \mathbf{A} \mathbf{w}_0 + \ln |\mathbf{A}| + \ln |\mathbf{H}|\right). \end{aligned}$$

Отметим, что гессиан $\mathbf{H}$ зависит от гиперпараметров $\mathbf{A}, \mathbf{B}$. Так же как и в предыдущем разделе, явно выписываются производные логарифма функции правдоподобия по гиперпараметрам и решается задача максимизации функции правдоподобия $\ln p_L(D \mid \mathbf{A}, \mathbf{B})$.

5 Использование алгоритма Левенберга-Марквардта для оценки оптимального значения параметров

Если известны значения гиперпараметров $\mathbf{A}, \mathbf{B}$ для нелинейной регрессионной модели, то можно использовать алгоритм Левенберга-Марквардта для оценки вектора параметров $\mathbf{w}$. Пусть задано некоторое приближение для значений параметров $\mathbf{w}$. Тогда функция ошиби имеет вид:

$$S = \frac{1}{2}(\mathbf{w} + \Delta\mathbf{w})^T \mathbf{A}(\mathbf{w} + \Delta\mathbf{w}) + \frac{1}{2}(\mathbf{X}(\mathbf{w} + \Delta\mathbf{w}) - \mathbf{y})^T \mathbf{B}(\mathbf{X}(\mathbf{w} + \Delta\mathbf{w}) - \mathbf{y}). \quad (9)$$

Для минимизации функции ошибки воспользуемся алгоритмом Левенберга-Марквардта, который предназначен для оптимизации параметров нелинейных регрессионных моделей. Алгоритм заключается в последовательном приближении заданных начальных значений параметров к искомому локальному оптимуму и является обобщением метода сопряжённых градиентов и алгоритма Ньютона-Гаусса.

На нулевой итерации алгоритма задаётся начальное приближение вектора $\mathbf{w}$. Приращение $\Delta\mathbf{w}$ в точке оптимума для функции ошибки (9) равно нулю. Поэтому для нахождения экстремума приравняем вектор частных производных S по $\mathbf{w}$ к нулю. Для этого представим S в виде двух слагаемых:

$$S_1 = \frac{1}{2}(\mathbf{w} + \Delta\mathbf{w})^T \mathbf{A}(\mathbf{w} + \Delta\mathbf{w}), \quad (10)$$

$$S_2 = \frac{1}{2}(\mathbf{X}(\mathbf{w} + \Delta\mathbf{w}) - \mathbf{y})^T \mathbf{B}(X(\mathbf{w} + \Delta\mathbf{w}) - \mathbf{y}). \quad (11)$$

После дифференцирования получим следующие выражения:

$$\frac{\partial S_1}{\partial \mathbf{w}} = \frac{1}{2}(\mathbf{w} + \Delta\mathbf{w})^T (\mathbf{A} + \mathbf{A}^T),$$

$$\frac{\partial S_2}{\partial \mathbf{w}} = \frac{1}{2}[(\mathbf{X}(\mathbf{w} + \Delta\mathbf{w}) - \mathbf{y})^T \mathbf{B}^T \mathbf{X} + (\mathbf{X}(\mathbf{w} + \Delta\mathbf{w}) - \mathbf{y})^T \mathbf{B}\mathbf{X}].$$

Таким образом, чтобы найти приращение $\Delta\mathbf{w}$ необходимо решить систему линейных уравнений:

$$\nabla S = \frac{1}{2}(\mathbf{w} + \Delta\mathbf{w})^T (\mathbf{A} + \mathbf{A}^T) + \frac{1}{2}[(\mathbf{X}(\mathbf{w} + \Delta\mathbf{w}) - \mathbf{y})^T \mathbf{B}^T \mathbf{X} + (\mathbf{X}(\mathbf{w} + \Delta\mathbf{w}) - \mathbf{y})^T \mathbf{B}\mathbf{X}] = 0.$$

Раскроем скобки и приведём подобные слагаемые:

$$\mathbf{w}^T \mathbf{X}^T \mathbf{B}^T \mathbf{X} + \Delta\mathbf{w}^T \mathbf{X}^T \mathbf{B}^T \mathbf{X} - \mathbf{y}^T \mathbf{B}^T \mathbf{X} + \mathbf{w}^T \mathbf{X}^T \mathbf{B}\mathbf{X} + \Delta\mathbf{w}^T \mathbf{X}^T \mathbf{B}\mathbf{X} - \mathbf{y}^T \mathbf{B}\mathbf{X} +$$

$$+ \mathbf{w}^T \mathbf{A} + \Delta\mathbf{w}^T \mathbf{A} + \mathbf{w}^T \mathbf{A}^T + \Delta\mathbf{w}^T \mathbf{A}^T = 0.$$

Сгруппируем и перенесем в одну сторону члены, содержащие приращение параметров $\Delta\mathbf{w}$:

$$\Delta\mathbf{w}^T(\mathbf{X}^T\mathbf{B}^T\mathbf{X}+\mathbf{X}^T\mathbf{B}\mathbf{X}+\mathbf{A}+\mathbf{A}^T)=-\mathbf{w}^T\mathbf{X}^T\mathbf{B}^T\mathbf{X}+\mathbf{y}^T\mathbf{B}^T\mathbf{X}-\mathbf{w}^T\mathbf{X}^T\mathbf{B}\mathbf{X}+$$

$$+\mathbf{w}^T\mathbf{B}\mathbf{X}-\mathbf{w}^T\mathbf{A}-\mathbf{w}^T\mathbf{A}^T.$$

Выразив приращение $\Delta\mathbf{w}$, получим следующую рекуррентную формулу:

$$\Delta\mathbf{w}=[(\mathbf{A}+\mathbf{A}^T+\mathbf{X}^T(\mathbf{B}^T+\mathbf{B})\mathbf{X})^{-1}]^T(-\mathbf{w}^T(\mathbf{A}+\mathbf{A}^T)+(\mathbf{y}-\mathbf{X}\mathbf{w})^T(\mathbf{B}^T+\mathbf{B})\mathbf{X})^T.$$

Так как матрицы $\mathbf{A},\mathbf{B}$ − симметричные, положительно определенные матрицы ковариации, то приращение вектора $\Delta\mathbf{w}$:

$$\Delta\mathbf{w}=[(\mathbf{A}+\mathbf{X}^T\mathbf{B}\mathbf{X})^{-1}]^T(-\mathbf{w}^T\mathbf{A}+(\mathbf{y}-\mathbf{X}\mathbf{w})^T\mathbf{B}\mathbf{X})^T.$$

То есть, $\Delta\mathbf{w}=\left(\mathbf{X}^T\mathbf{B}\mathbf{X}+\mathbf{A}\right)^{-1}\mathbf{X}^T\mathbf{B}^T\mathbf{y}-\mathbf{w}.$

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

6 Алгоритм отбора признаков

Полученные значения гиперпараметров $\alpha_i, i=1,..,n$ для диагональной матрицы $\mathbf{A}$ могут быть использованы для отбора признаков и выбора модели линейной регрессии. Параметры w_i модели f сравниваются, используя оценки значений гиперпараметров α_i. Большие значения гиперпараметра α_i означают больший штраф на значение параметра и, следовательно, меньшую значимость данных параметров для качества модели. Малые значения α_i показывают большую значимость данного компонента модели для ее качества.

7 Вычислительный эксперимент

Результатом вычислительного эксперимента является фильтрация

шумовых и коррелирующих признаков. Тестирование алгоритма производится на временном ряде, содержащем информацию о семи компонентах, входящих в состав бетона. Исследуется два отклика: предел прочности при сжатии и морозостойкость. Ряд содержит 103 записи. Необходимо построить регрессионную модель и оценить её параметры.

При исследовании предела прочности при сжатии алгоритм приводит к следующим результатам.

На рис. 1 представлены логарифмы диагональных элементов матрицы $\mathbf{A}$. Шестой элемент почти в два раза больше всех остальных, поэтому соответствующий ему параметр модели w_6 мал, как мы видим из графика 2 и таблицы 1. Однако α_6 не настолько велик, чтобы мы могли убрать соответствующий столбец матрицы плана, так как при этом произойдёт увеличение функции ошибки на 20%.

Табл. 1. Численные значения параметров модели

w_1	w_2	w_3	w_4	w_5	w_6	w_7
0.1054	0.0317	0.0951	−0.0937	0.2790	−0.0013	0.0168

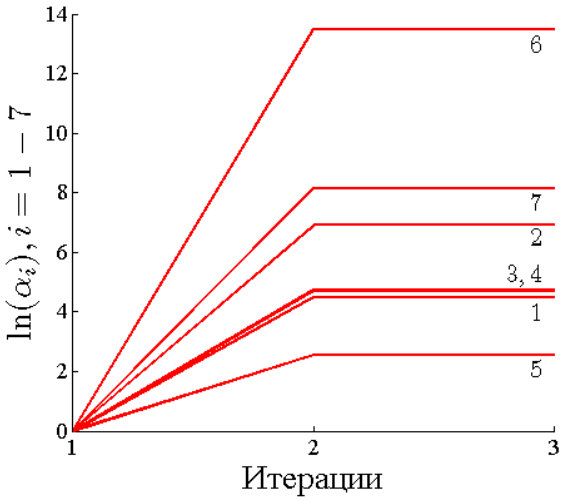

Рис. 1. Зависимость логарифмов значений диагональных элементов матрицы **A** от номера итерации

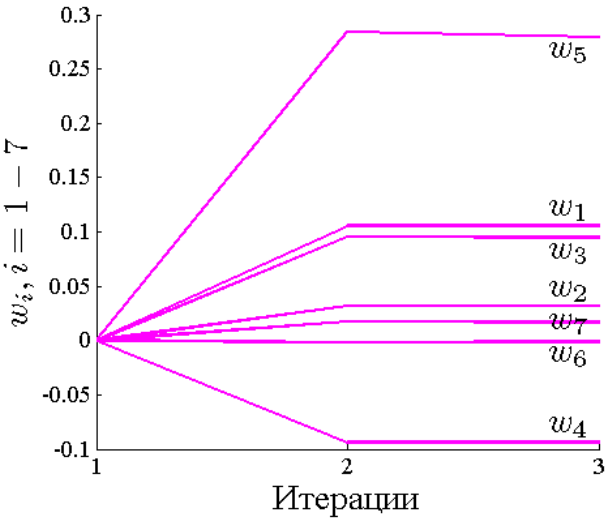

Рис. 2. Зависимость значений параметров **w** от номера итерации

При исследовании морозостойкости наблюдается вырождение матрицы **A**. Так на рисунке 3 приведён итерационный процесс для всех диагональных элементов α, кроме пятого, так как на третьей итерации $\log(\alpha_5)$ достигает значения 66, что в шесть раз превышает все остальные логарифмы элементов матрицы **A**. Рассматривая графики 4, 5 и таблицу 2, получим, что пятый признак является неинформативным и может быть исключен из матрицы плана. Функция ошибки увеличится менее, чем на 1%.

В обоих случаях использование аппроксимации Лапласа для оценки гиперпараметров приводит к увеличению функции ошибки менее, чем на 1%.

Табл. 2. Численные значения параметров модели

w_1	w_2	w_3	w_4	w_5	w_6	w_7
-0.0262	-0.1176	-0.0201	0.4801	0	-0.0238	-0.0079

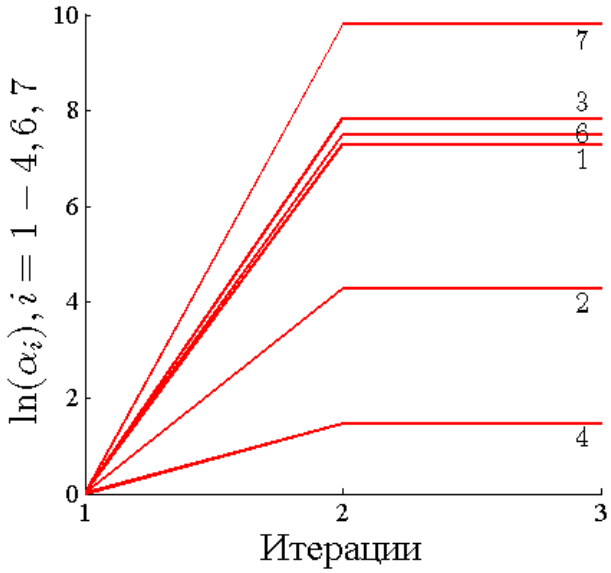

Рис. 3. Зависимость логарифмов значений диагональных элементов матрицы **A** от номера итерации

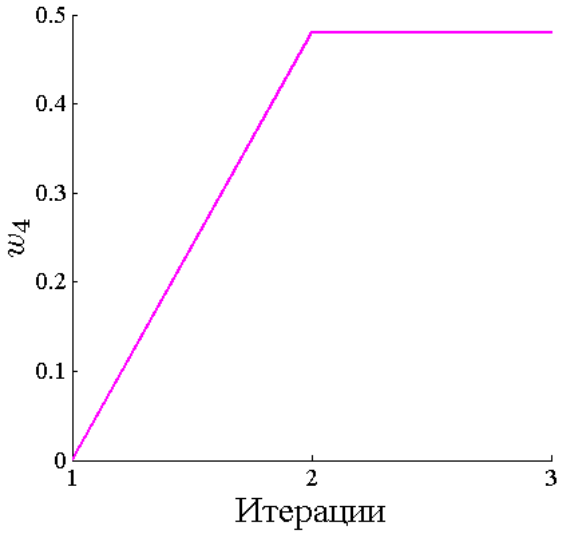

Рис. 4. Зависимость значений параметра w_4 от номера итерации

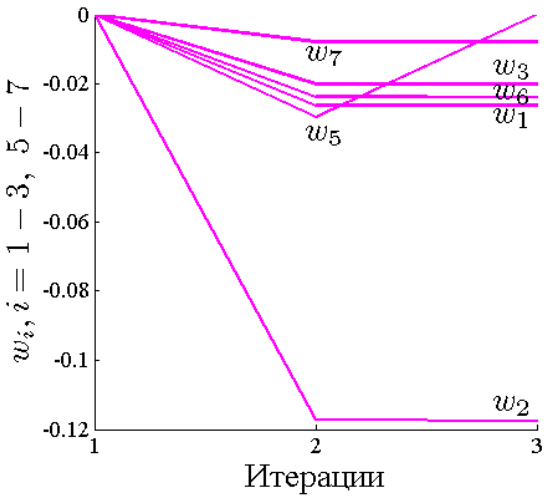

Рис. 5. Зависимость значений параметров **w** от номера итерации

8 Выводы

В работе получено точное выражение для функции правдоподобия $\ln p(D \mid \mathbf{A}, \mathbf{B})$ и предложен подход к его оптимизации. Так же проведено сравнение предложенного подхода с аппроксимацией Лапласа искомого правдоподобия. Использование точного выражения для вычисления правдоподобия позволяет получить наиболее точные оценки гиперпараметров.

Список литературы

1. Burnham K.P., Anderson D.R. Model selection and multimodel inference: a practical information-theoretic approach. Berlin: Springer. 2002.

2. MacKay D. Choice of basis for laplace approximation // Technical report, machine learning. Oxford: Oxford University. 1998.

3. LeCun Y., Denker J., Solla S., Howard R.E., Jackel L. D. Optimal brain damage //Advances in neural information processing systems II. San Mateo: Morgan Kauffman. 1990.

4. Bishop C.M., Tipping M.E. Bayesian regression and classification //Advances in learning theory: methods, models and applications. Washington: IOS Press. 2000.

P. 267−285.

5. Bishop C.M. Pattern recognition and machine learning. Berlin: Springer. 2006.

6. Стрижов В.В., Сологуб Р.А. Индуктивное порождение регрессионных моделей предполагаемой волатильности для опционных торгов // Вычислительные технологии. 2009. Т. 14, № 5. С. 102−113.

7. Стрижов В.В. Поиск параметрической регрессионной модели в индуктивно заданном множестве // Вычислительные технологии. 2007. № 1. С. 93−102.

9. Нестеров, Ю.Е. Введение в выпуклую оптимизацию. М: МЦНМО. 2010.

10. Rasmussen, C.E. Gaussian processes in machine learning // Advanced lectures on machine learning. 2004. Vol. 1. P. 63−71.

A. A. Zaytsev, Moscow Institute of Physics and Technology, Datadvance

V. V. Strijov, Computing Center of the Russian Academy of Sciences

A. A. Tokmakova, Moscow Institute of Physics and Technology

Estimation regression model hyperparameters using maximum likelihood

The papers considers the regression model selection problem. The model parameters are supposed to be a multivariate random variable with independently distributed components. A method for hyperparameters optimization is proposed. Direct way to obtain the hyperparameters estimations is shown. The papers illustrated the usage of the hyperparameters in the feature selection problem. The suggested method is compared with the Laplace approximation method.

Keywords: regression, feature selection, parameter distribution, hyperparameter estimation, Bayesian inference