Оценка гиперпараметров линейных регрессионных моделей методом максимального правдоподобия при отборе шумовых и коррелирующих признаков*

А. А. Зайцев, А. А. Токмакова
likzet@gmail.com, aleksandra-tok@yandex.ru
Московский физико-технический институт, ФУПМ, каф. «Интеллектуальные системы»

Рассматривается задача выбора регрессионной модели. Предполагается, что вектор параметров модели — многомерная случайная величина с независимо распределёнными компонентами. В работе предложен способ оптимизации параметров 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 оценки гиперпараметров аппроксимацию Лапласа.

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

Estimation regression model hyperparameters using maximum likelihood*

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

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 hyperparameter 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.*

Введение

В данной работе рассматривается задача выбора регрессионной модели [6] из заданного параметрического семейства регрессионных моделей. Один из возможных подходов — введение предположения о распределении параметров модели [8]. В этом случае предполагается, что функция регрессии задана оценкой вектора параметров, который считается нормально распределенно 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 [7]. Более общий подход был предложен Маккаем в работе [8]. В этой работе Маккай ввел понятие гиперпараметров. Бишоп предложил ряд других способов оценки гиперпараметов, таких как Марковские цепи Монте-Карло и аппроксимация Лапласа [5, 4]. Подход, использующий аппроксимацию Лапласа был развит в работах [3, 2].

Предлагается для линейной регрессионной модели выписать явное выражение функции правдоподобия с учётом введенных вероятностных предположений. Максимизируя правдоподобие, получаем оценки наиболее правдоподобных значений гиперпараметров модели. Такой подход позволяет получать оценки гиперпараметров регрессионных моделей.

Научный руководитель В. В. Стрижов

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 [3].

Во второй части работы дана постановка задачи и принятая гипотеза порождения данных. В третей части получена оценка правдоподобия гиперпараметров линейной модели и описан подход, позволяющий оценивать гиперпараметры, доставляющие максимум правдоподобия. В четвертой части описан подход, позволяющий оценивать гиперпараметры, максимизирующие правдоподобие для нелинейных регрессионных моделей с использованием аппроксимации Лапласа. В пятой части описан процедура отбора признаков, использующая полученные значения гиперпараметров. В шестой части проведено сравнение предложенного и используемых подходов на модельных и реальных данных.

Постановка задачи

Задана выборка $D = (X, \mathbf{y}) = \{(\mathbf{x}_i, y_i)\}_{i=1}^m$, где $\mathbf{x}_i \in \mathbb{R}^n$, $y_i \in \mathbb{R}$. Рассматривается класс регрессионных моделей вида:

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

Предполагается, что шум $\boldsymbol{\varepsilon}$ — многомерная нормальная случайная величина с нулевым математическим ожиданием и матрицей ковариации B^{-1}:

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

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

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

Требуется получить оценки матриц A, B согласно гипотезе порождения данных (2), (3).

Правдоподобие для линейной модели

Рассмотрим линейную регрессионную модель. Тогда (1) имеет вид

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

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

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

в котором $p(D|\mathbf{w}, B)$, $p(\mathbf{w}|A)$ — плотности многомерных нормальных случайных величин:

$$p(D|\mathbf{w}, B) = \frac{1}{(2\pi)^{\frac{m}{2}}|B|^{-\frac{1}{2}}} \exp\left(-\frac{1}{2}(\mathbf{y} - X\mathbf{w})^T B(\mathbf{y} - X\mathbf{w})\right), \tag{5}$$

согласно предположению о нормальности распределения шумов (2), и

$$p(\mathbf{w}|A) = \frac{1}{(2\pi)^{\frac{n}{2}}|A|^{-\frac{1}{2}}} \exp\left(-\frac{1}{2}\mathbf{w}^T A\mathbf{w}\right), \tag{6}$$

согласно предположению о распределении вектора параметров модели (3). Правдоподобие модели $p(D|A, B)$ имеет вид

$$p(D|A, B) = \int_{\mathbb{R}^n} p(D|\mathbf{w}, A, B)p(\mathbf{w}|A, B)d\mathbf{w}. \tag{7}$$

Для линейных моделей явно выпишем оценки гиперпараметров модели $p(D|A, B)$. Отметим, что в работе [3] был предложен подход, который оценивал эту вероятность с использованием аппроксимации Лапласа. Верна следующая теорема.

Теорема 1. *Правдоподобие в предположениях о распределении шума* $\boldsymbol{\varepsilon}$ (2) *и параметров модели* $\mathbf{w}$ (3) *имеет вид*

$$p(D|A,B) = \frac{|B|^{\frac{1}{2}}|A|^{\frac{1}{2}}}{(2\pi)^{\frac{m}{2}}|K|^{\frac{1}{2}}} \exp\left(\frac{1}{2}\mathbf{y}^T(C^TKC - B)\mathbf{y}\right), \tag{8}$$

а его логарифм имеет вид

$$\ln p(D|A,B) = -\frac{1}{2}\left(\ln|K| + m\ln 2\pi - \ln|B| - \ln|A| - \mathbf{y}^T(C^TKC - B)\mathbf{y}\right). \tag{9}$$

Здесь

$$K = X^TBX + A, \quad C = K^{-1}X^TB.$$

Доказательство.

Подставляя (5) и (6) в (7) получим следующее выражение:

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

перепишем произведение двух экспонент как экспоненту от их суммы:

$$= \int_{\mathbb{R}^n} \frac{|B|^{\frac{1}{2}}|A|^{\frac{1}{2}}}{(2\pi)^{\frac{n+m}{2}}} \exp\left(-\frac{1}{2}\left((\mathbf{y} - X\mathbf{w})^TB(\mathbf{y} - X\mathbf{w}) + \mathbf{w}^TA\mathbf{w}\right)\right) d\mathbf{w} =$$

введем обозначения $K = A + X^TBX, C = K^{-1}X^TB$ и выделим полный квадрат по $(\mathbf{w} - C\mathbf{y})$:

$$= \int_{\mathbb{R}^n} \frac{|B|^{\frac{1}{2}}|A|^{\frac{1}{2}}}{(2\pi)^{\frac{n+m}{2}}} \exp\left(-\frac{1}{2}\left((\mathbf{w} - C\mathbf{y})^TK(\mathbf{w} - Cy) - \mathbf{y}^T(C^TKC - B)\mathbf{y}\right)\right) d\mathbf{w} =$$

интеграл по плотности многомерного нормального распределения равен единице:

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

Следовательно, искомое правдоподобие модели $p(D|A,B)$ имеет вид (8), а его логарифм — вид (9). ■

Рассмотрим теперь случай, когда матрица A — диагональная, а матрица $B = \beta I$.

Следствие 1. *Если матрица* A *— диагональная, а матрица* B *имеет вид* $B = \beta I$*, то логарифм правдоподобия модели* $\ln p(D|A,B)$ *имеет вид*

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

где $K = A + \beta X^TX$.

Вычисление производных функции правдоподобия модели $\ln p(D|A,B)$ по гиперпараметрам A, B

Для поиска максимума правдоподобия будем пользоваться градиентными методами оптимизации [1], поэтому нам понадобится выражения для производных $\ln p(D|A,B)$ по гиперпараметрам A, B.

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

Верны следующие два свойства производных матриц [9]. Для симметричной матрицы M верно, что

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

где t — некоторый параметр, $M = M(t)$. Так же верно, что

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

Введем обозначение S^{ij} — такая матрица, что для двух индексов k, l выполнено, что

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

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

$$\frac{\partial \ln p(D|A,B)}{\partial \beta_{ij}} = -\frac{1}{2}\Big(\operatorname{tr}\left(K^{-1}X^T S^{ij} X\right) - \operatorname{tr}\left(B^{-1}S^{ij}\right) - \mathbf{y}^T\big(S^{ji}XK^{-1}X^TB + B^TXK^{-1}X^TS^{ij} - B^TXK^{-1}X^TS^{ij}XK^{-T}X^TB - S^{ij}\big)\mathbf{y}\Big).$$

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

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

Так же запишем производные в предположениях следствия 1, $A = \operatorname{diag}(\alpha_1, \alpha_2, \ldots, \alpha_n)$, $B = \frac{1}{\beta}I$.

$$\frac{\partial \ln p(D|A,\beta)}{\partial \beta} = -\frac{1}{2}\Big(\operatorname{tr}\left(K^{-1}X^TX\right) - \frac{m}{\beta} + \mathbf{y}^T(2\beta XK^{-1}X^T - I - \beta^2XK^{-1}X^TXK^{-1}X^T)\mathbf{y}\Big)$$

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

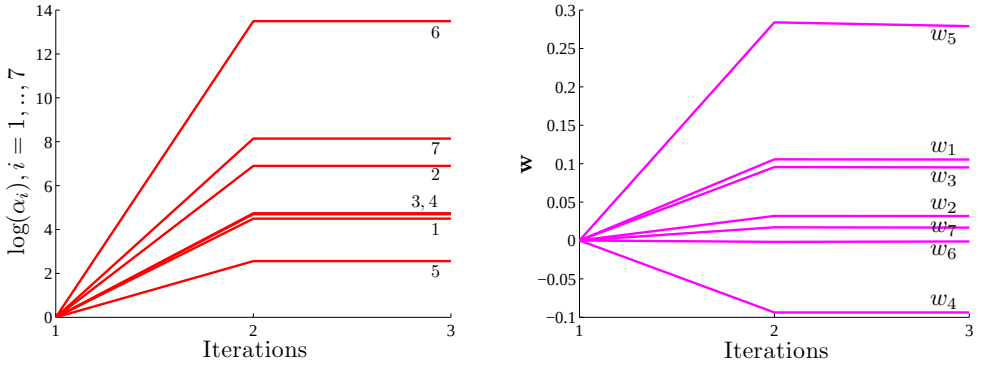
Так как получены значения производных правдоподобия модели $\ln p(D|A,B)$ по гиперпараметрам A, B, можно использовать любой градиентный метод оптимизации для поиска гиперпараметров A, B, максимизирующих правдоподобие модели.

Полученные значения гиперпараметров α_i, $i = 1, \ldots, n$ для диагональной матрицы 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 большой 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.

Модифицированный алгоритм Левенберга-Марквардта

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.

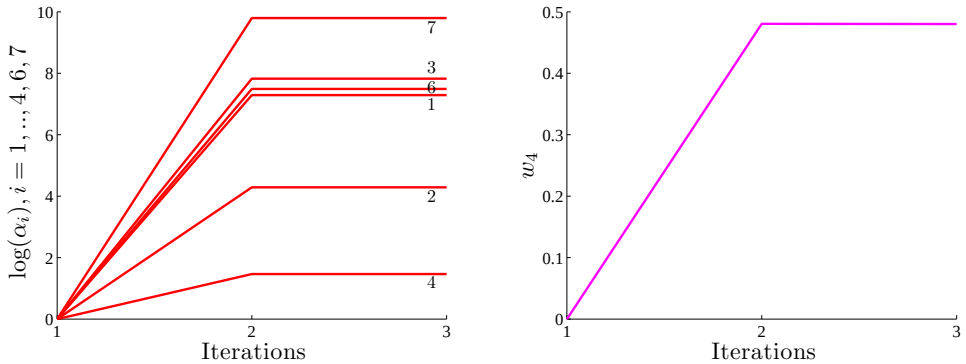


(a) Логарифм элементов матрицы A (b) Логарифм элементов матрицы A

Рис. 1. Исследование прочности при сжатии бетона

Пусть задано некоторое приближение для значений параметров модели $\mathbf{w}$. Тогда функция ошибки имеет вид:

$$S = \frac{1}{2}(\mathbf{w} + \Delta\mathbf{w})^T A(\mathbf{w} + \Delta\mathbf{w}) + \frac{1}{2}\big(X(\mathbf{w} + \Delta\mathbf{w}) - \mathbf{y}\big)^T B\big(X(\mathbf{w} + \Delta\mathbf{w}) - \mathbf{y}\big). \qquad (10)$$



(a) Логарифм элементов матрицы A (b) Параметр модели w_4

Рис. 2. Исследование морозостойкости бетона

На нулевой итерации алгоритма задаётся начальное приближение для $\mathbf{w}$. Приращение $\Delta\mathbf{w}$ в точке оптимума для функции ошибки (10) равно нулю. Поэтому для нахождения экстремума приравняем вектор частных производных S по $\mathbf{w}$ к нулю. Для этого представим S в виде двух слагаемых:

$$S_1 = \frac{1}{2}(\mathbf{w}+\Delta\mathbf{w})^T A(\mathbf{w}+\Delta\mathbf{w}),\ S_2 = \frac{1}{2}\big(X(\mathbf{w}+\Delta\mathbf{w})-\mathbf{y}\big)^T B\big(X(\mathbf{w}+\Delta\mathbf{w})-\mathbf{y}\big).$$

После дифференцирования получим следующие выражения:

$$\frac{\partial S_1}{\partial \mathbf{w}} = \frac{1}{2}(\mathbf{w}+\Delta\mathbf{w})^T(A+A^T),$$
$$\frac{\partial S_2}{\partial \mathbf{w}} = \frac{1}{2}\big[\big(X(\mathbf{w}+\Delta\mathbf{w})-\mathbf{y}\big)^T B^T X + \big(X(\mathbf{w}+\Delta\mathbf{w})-\mathbf{y}\big)^T BX\big].$$

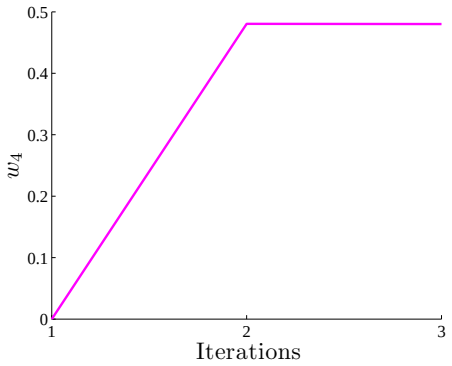


Рис. 3. Параметры модели $\mathbf{w}$

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

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

Выразив приращение $\Delta\mathbf{w}$, учитывая, что A, B симметричные матрицы, получим следующую рекуррентную формулу:

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

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.

Таблица 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

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. **??** представлены логарифмы диагональных элементов матрицы A. Шестой элемент почти в два раза больше всех остальных, поэтому соответствующий ему параметр модели w_6 мал, как мы видим из графика **??** и 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%.

Таблица 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

При исследовании морозостойкости наблюдается вырождение матрицы A. Так на рисунке 2 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 соответствующие графики 2, 3 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 вычисления правдоподобия приводит к увеличению функции ошибки менее, чем на 1%.

Выводы

В работе получено точное выражение для правдоподобия $\ln p(D|A, B)$ 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.

Литература

[1] Ю. Е. Нестеров. *Введение в выпуклую оптимизацию.* МЦНМО, 2010.

[2] В. В. Стрижов. Поиск параметрической регрессионной модели в индуктивно заданном множестве. *Журнал вычислительных технологий*, 1:93–102, 2007.

[3] В. В. Стрижов and Р. А. Сологуб. Индуктивное порождение регрессионных моделей предполагаемой волатильности для опционных торгов. *Вычислительные технологии*, 14(5):102–113, 2009.

[4] C. Bishop. *Pattern Recognition And Machine Learning.* Springer, 2006.

[5] C. M. Bishop and M. E. Tipping. Bayesian regression and classification. In *Suykens, J., Horvath, G. et al., eds. Advances in Learning Theory: Methods, Models and Applications*, volume 190, pages 267–285. IOS Press, NATO Science Series III: Computer and Systems Sciences, 2000.

[6] K.P. Burnham and D.R. Anderson. *Model selection and multimodel inference: a practical information-theoretic approach.* Springer Verlag, 2002.

[7] Y. LeCun, J. Denker, S. Solla, R. E. Howard, and L. D. Jackel. Optimal brain damage. In D. S. Touretzky, editor, *Advances in Neural Information Processing Systems II*, San Mateo, CA, 1990. Morgan Kauffman.

[8] David J.C. MacKay. Choice of basis for laplace approximation. Technical report, Machine Learning, 1998.

[9] C.E. Rasmussen. Gaussian processes in machine learning. *Advanced Lectures on Machine Learning*, 1:63–71, 2004.