Об унимодальности непрерывного расширения критерия Акаике*

Ветров Д. П., Кропотов Д. А., Пташко Н. О.
vetrovd@yandex.ru, dmitry.kropotov@gmail.com, ptashko@inbox.ru
Москва, ВМиК МГУ, Вычислительный Центр РАН

В работе рассматривается применение непрерывного расширения критерия Акаике (CAIC) к подбору параметров регуляризации в задаче обобщенной линейной регрессии. Значениями параметра регуляризации являются все симметричные неотрицательно определенные матрицы. Показывается, что на множестве всех таких матриц критерий CAIC является унимодальным. Получено явное условие вырожденности решающего правила (нулевого решения). Показано, что данное условие остается справедливым для семейства диагональных неотрицательно определенных матриц, а также для семейства матриц, пропорциональных единичной.

Для решения задач выбора моделей широко применяется информационный критерий Акаике [1], предлагающий подход, основанный на теории информации. Несмотря на то, что этот метод изначально был предложен для выбора из конечного числа моделей, он может быть расширен на случай бесконечного семейства моделей. Такое расширение позволяет заменить полный перебор на конечном множестве направленным поиском максимума непрерывного функционала, что существенно ускоряет процедуру выбора. В работах [2, 3] предложено подобное обобщение информационного критерия Акаике (CAIC) в применении для настройки параметров модели стационарной и нестационарной линейной регрессии. Одновременная настройка нескольких параметров модели с помощью критерия CAIC обычно проводится градиентным или покоординатным методами. В этой связи возникает вопрос о наличии у непрерывного критерия Акаике локальных максимумов. В данной работе рассматривается задача выбора модели с помощью CAIC в широком семействе всех симметричных неотрицательно определенных матриц. Доказано, что непрерывный критерий Акаике в этом семействе является унимодальной функцией. Выписано аналитическое решение для обобщенной линейной модели регрессии, в которой, согласно критерию Акаике, достигается наилучшая обобщающая способность. Также получено практически важное условие релевантности, допускающее непосредственную прямую проверку, которое позволяет отбросить заведомо неприемлемые модели. Рассматриваемое семейство матриц включает в себя важные подсемейства: семейство всех диагональных матриц и семейство матриц, пропорциональных единичной. Полученные здесь результаты для широкого семейства могут быть частично перенесены для данных подсемейств. В частности, показано, что условие релевантности остается справедливым и для рассматриваемых подсемейств.

*Работа выполнена при финансовой поддержке РФФИ, проекты №08-01-00405, №08-01-90016, №08-01-90427, №07-01-00211.

Непрерывное расширение критерия Акаике

Рассмотрим классическую задачу обобщенной линейной регрессии. Пусть $(X, \boldsymbol{t}) = \{(\boldsymbol{x}_i, t_i)\}_{i=1}^n$ — обучающая выборка, где $\boldsymbol{x}_i = (x_i^1, \ldots, x_i^d) \in \mathbb{R}^d$ — вектор наблюдаемых признаков объекта, $t_i \in \mathbb{R}$ — значение зависимой переменной. Зафиксируем некоторое множество базисных функций $\{\varphi_i(\boldsymbol{x})\}_{i=1}^m$, $\varphi_j : \mathbb{R}^d \to \mathbb{R}$. Требуется найти вектор весов $\boldsymbol{w} \in \mathbb{R}$ такой, что функция

$$y(\boldsymbol{x}) = \boldsymbol{w}^\intercal \boldsymbol{\varphi}(\boldsymbol{x}) = \sum_{j=1}^m w_j \varphi_j(\boldsymbol{x})$$

приближала бы значения переменной t на объектах обучающей выборки X. Пусть $\Phi = (\varphi_{ij})_{n\times m} = (\varphi_j(\boldsymbol{x}_i))_{n\times m}$ — матрица базисных функций, вычисленных для каждого объекта обучающей выборки. Классический подход к обучению линейной регрессии состоит в оптимизации регуляризованного правдоподобия

$$\boldsymbol{w}_{\text{MP}} = \arg\max_{\boldsymbol{w}} p(\boldsymbol{t} \mid X, \boldsymbol{w}) p(\boldsymbol{w} \mid \alpha), \qquad (1)$$

где

$$p(\boldsymbol{t} \mid X, \boldsymbol{w}) = \frac{1}{\sqrt{(2\pi)^n}\sigma^n} \exp\left(-\frac{1}{2\sigma^2}\|\Phi\boldsymbol{w} - \boldsymbol{t}\|^2\right) \qquad (2)$$

— функция правдоподобия.

В качестве регуляризатора $p(\boldsymbol{w} \mid \alpha)$ часто рассматривается квадратичный функционал с некоторым коэффициентом регуляризации $\alpha \geqslant 0$:

$$p(\boldsymbol{w} \mid \alpha) = \left(\frac{\alpha}{2\pi}\right)^{m/2} \exp\left(-\frac{\alpha}{2}\sum_{j=1}^m w_j^2\right). \qquad (3)$$

В методе релевантных векторов RVM [4] для автоматического отбора релевантных базисных функций семейство регуляризаторов (3) предлагается расширить, и для каждого веса w_j ввести свой

коэффициент регуляризации α_j:

$$p(\boldsymbol{w}\,|\,\boldsymbol{\alpha}) = \prod_{j=1}^{m} \sqrt{\frac{\alpha_j}{2\pi}} \exp\left(-\frac{\alpha_j}{2} w_j^2\right) = \\ = \sqrt{\frac{\det(R)}{(2\pi)^m}} \exp\left(-\frac{1}{2}\boldsymbol{w}^\intercal R \boldsymbol{w}\right), \quad (4)$$

где $R = \operatorname{diag}(\alpha_1, \ldots, \alpha_m)$ — матрица регуляризации, $\alpha_j \geqslant 0$. Расширим используемое в RVM семейство регуляризаторов на случай всех (необязательно диагональных) симметричных неотрицательно определенных матриц $R = R^\intercal \succeq 0$. Такое семейство матриц позволяет не только находить релевантное подмножество базисных функций, но и одновременно выделять релевантные линейные комбинации исходных базисных функций.

Подбор матрицы регуляризации будем осуществлять, используя непрерывное расширение критерия Акаике, CAIC [2]:

$$R = \arg\max f(R);$$
$$f(R) = \log p(\boldsymbol{t}\,|\,X, \boldsymbol{w}_{\text{MP}}) - \operatorname{tr}\big(H(H{+}R)^{-1}\big). \quad (5)$$

Здесь $H = -\nabla\nabla \log p(\boldsymbol{t}\,|\,X, \boldsymbol{w}) = \sigma^{-2}\Phi^\intercal\Phi$.

Параметр σ также может быть найден путем максимизации CAIC, что приводит к следующему итеративному процессу его вычисления на основе текущего значения R:

$$(\sigma^2)^{\text{new}} = \frac{\|\boldsymbol{t} - \Phi\boldsymbol{w}_{\text{MP}}\|^2}{n - \operatorname{tr} H(H{+}R)^{-1}R(H{+}R)^{-1}}. \quad (6)$$

Решение задачи оптимизации

Обозначим через $\boldsymbol{w}_{\text{ML}}$ оценку максимального правдоподобия на выборке X. Можно показать, что при условии (2) и (4) максимизация CAIC (5) эквивалентна следующей оптимизационной задаче:

$$\begin{cases} -\frac{1}{2}\boldsymbol{w}_{\text{ML}}^\intercal H(H{+}R)^{-1}H(H{+}R)^{-1}H\boldsymbol{w}_{\text{ML}} + \\ + \boldsymbol{w}_{\text{ML}}^\intercal H(H{+}R)^{-1}H\boldsymbol{w}_{\text{ML}} - \\ \qquad - \operatorname{tr}(H(H{+}R)^{-1}) \to \max\limits_{R}; \\ R = R^\intercal \succeq 0. \end{cases} \quad (7)$$

Обозначим $\boldsymbol{v} = H^{\frac{1}{2}}\boldsymbol{w}_{\text{ML}}$,

$$A = H^{\frac{1}{2}}(H{+}R)^{-1}H^{\frac{1}{2}} = (I + H^{-\frac{1}{2}}RH^{-\frac{1}{2}})^{-1}.$$

Тогда задача (7) может быть переписана в следующем виде:

$$\begin{cases} -\frac{1}{2}\boldsymbol{v}^\intercal AA\boldsymbol{v} + \boldsymbol{v}^\intercal A\boldsymbol{v} - \operatorname{tr} A \to \max\limits_{A}; \\ A^{-1} - I \succeq 0; \\ A^\intercal = A. \end{cases} \quad (8)$$

Используя условие симметричности матрицы A, представим ее в виде $A = Q\Lambda Q^\intercal$, где $Q^\intercal = Q^{-1}$,

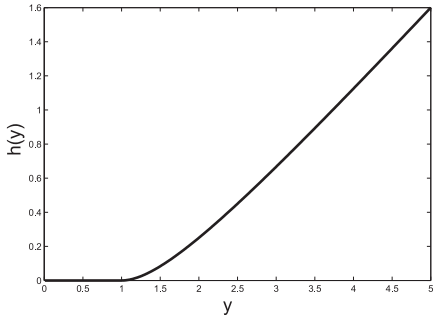

Рис. 1.

$\Lambda = \operatorname{diag}(\lambda_1, \ldots, \lambda_m)$ и $(\lambda_1, \ldots, \lambda_m)$ — набор собственных чисел матрицы A. Заметим, что такое разложение единственно. Обозначим $\boldsymbol{x} = Q^\intercal\boldsymbol{v}$. Тогда задача (8) переформулируется следующим образом:

$$\begin{cases} g(\boldsymbol{x}, \boldsymbol{\lambda}) = \sum\limits_{j=1}^{m}\big(-\frac{1}{2}x_j^2\lambda_j^2 + x_j^2\lambda_j - \lambda_j\big) \to \max\limits_{\boldsymbol{x},\boldsymbol{\lambda}}; \\ \sum\limits_{j=1}^{m} x_j^2 = \|\boldsymbol{v}\|^2; \\ 0 \leqslant \lambda_j \leqslant 1, \quad j = 1, \ldots, m. \end{cases} \quad (9)$$

Максимальные значения функции $g(\boldsymbol{x}, \boldsymbol{\lambda})$ при фиксированных $\boldsymbol{x}$ достигаются при следующих значениях $\boldsymbol{\lambda}$:

$$\lambda_j^*(\boldsymbol{x}) = \max\left(0, 1 - \frac{1}{x_j^2}\right). \quad (10)$$

Введем функцию $h(y)$ (см. рис. 1):

$$h(y) = \begin{cases} \frac{y}{2} + \frac{1}{2y} - 1, & y \geqslant 1; \\ 0, & 0 \leqslant y \leqslant 1. \end{cases} \quad (11)$$

Заметим, что функция $h(y)$ выпукла. Обозначим $y_j = x_j^2$. Тогда, подставив выражение (10) в систему (9), получим следующую задачу распределения ресурсов:

$$\begin{cases} \sum\limits_{j=1}^{m} h(y_j) \to \max\limits_{\boldsymbol{y}}; \\ \sum\limits_{j=1}^{m} y_j = \|\boldsymbol{v}\|^2. \end{cases} \quad (12)$$

В случае $\|\boldsymbol{v}\| \leqslant 1$ критерий $\sum\limits_{j=1}^{m} h(y_j)$ тождественно равен 0 (см. рис. 1), и компоненты y_j принимают любые значения от 0 до 1 при условии, что $\sum\limits_{j=1}^{m} y_j = \|\boldsymbol{v}\|^2$. Таким образом, все $x_j^2 < 1$ и, следовательно, все $\lambda_j^* = 0$, то есть $A = 0$, и матрица регуляризации $R^{-1} = 0$. В этом случае регуляризатор штрафует любую попытку настройки на данные, и решающее правило становится вырожденным с $\boldsymbol{w}_{\text{MP}} = \boldsymbol{0}$. На практике такой случай соответствует ситуации, когда выбранное семейство

базисных функций $\{\varphi_i(\boldsymbol{x})\}_{i=1}^m$ не позволяет восстановить зависимость скрытой переменной от признаков. Заметим, что данный результат остается справедливым в подсемействах моделей, отвечающих диагональным и скалярным (пропорциональным единичной) матрицам регуляризации, т. к. вырожденная матрица регуляризации $R^{-1} = 0$ входит в эти подсемейства. Процесс поиска наилучшей модели в этих подсемействах является итеративным, поэтому возможность отсечь заведомо неадекватное подсемейство по аналитически проверяемому условию

$$\boldsymbol{w}_{ML}^{\intercal} H \boldsymbol{w}_{ML} \leqslant 1 \tag{13}$$

позволяет значительно сократить время поиска наилучшей модели.

Если $\|\boldsymbol{v}\| > 1$, тогда, вследствие строгой выпуклости $h(y)$, максимум достигается в том случае, когда все ресурсы сосредоточены в одной из компонент y_i. Без ограничения общности выберем в качестве такой компоненты y_1, т. е. $y_1 = \|\boldsymbol{v}\|^2$, $y_2 = \cdots = y_m = 0$. Тогда первый собственный вектор матрицы A сонаправлен вектору $\boldsymbol{v}$. Остальные собственные векторы имеют нулевые собственные значения и могут быть выбраны произвольно при условии сохранения тождества $Q^{\intercal} = Q^{-1}$. Отсюда, используя определение матрицы A, получим выражение для матрицы R^{-1}:

$$\begin{aligned} R^{-1} &= H^{-\frac{1}{2}}(A^{-1} - I)^{-1} H^{-\frac{1}{2}} = \\ &= H^{-\frac{1}{2}} Q \operatorname{diag}\big(\|\boldsymbol{v}\|^2 - 1, 0, \ldots, 0\big) Q^{\intercal} H^{-\frac{1}{2}} = \\ &= \frac{\|\boldsymbol{v}\|^2 - 1}{\|\boldsymbol{v}\|^2} H^{-\frac{1}{2}} \boldsymbol{v}\boldsymbol{v}^{\intercal} H^{-\frac{1}{2}} = \\ &= \frac{\boldsymbol{w}_{\mathrm{ML}}^{\intercal} H \boldsymbol{w}_{\mathrm{ML}} - 1}{\boldsymbol{w}_{\mathrm{ML}}^{\intercal} H \boldsymbol{w}_{\mathrm{ML}}} \boldsymbol{w}_{\mathrm{ML}} \boldsymbol{w}_{\mathrm{ML}}^{\intercal}. \end{aligned}$$

Таким образом, конечное выражение для матрицы R не зависит от выбора компоненты y_i при решении задачи распределения ресурсов (12), следовательно, исходная задача (7) имеет единственную точку максимума. Итак доказана следующая

Теорема 1. *Функционал*

$$f(R) = \log p(\boldsymbol{t} \,|\, X, \boldsymbol{w}_{\mathrm{MP}}) - \operatorname{tr}\big(H(H+R)^{-1}\big)$$

является унимодальным на множестве всех матриц $R = R^{\intercal} \succeq 0$, *если* $\|H^{\frac{1}{2}} \boldsymbol{w}_{\mathrm{ML}}\| > 1$.

Выводы

Доказанная теорема позволяет использовать точку максимума R критерия CAIC для построения конечного решения задачи восстановления регрессии, гарантируя, что выбранная модель будет наилучшей среди всех допустимых моделей $R = R^{\intercal} \succeq 0$.

Кроме того, полученный результат важен для выбора модели в задаче линейной регрессии, рассмотренной в [5]. В указанной работе подбор коэффициентов регуляризации, связанных индивидуально с каждым весом, производится путем максимизации непрерывного критерия Акаике (CAIC). В этом случае критерий (5) оптимизируется на множестве всех диагональных матриц $R \succeq 0$. Унимодальность критерия на данном множестве остается открытой проблемой. Доказанная в данной работе теорема косвенно подтверждает предположение об унимодальности критерия и в этом случае. Полученное в работе условие релевантности (13) позволяет эффективно отсекать заведомо неадекватные наблюдаемым данным модели до начала итеративной настройки параметров модели. Аналитическое решение, максимизирующее непрерывный критерий Акаике, может быть использовано для построения решающих правил с лучшей обобщающей способностью.

Литература

[1] *Akaike H.* A new look at statistical model identification // IEEE Trans. Automatic Control. 1974. V. 25. P. 461–464.

[2] *Kropotov D. A., Vetrov D. P.* General solutions for information-based and bayesian approaches to model selection in linear regression and their equivalence // Pattern Recognition and Image Analysis. 2009. V. 3. P. 447–455.

[3] *Ezhova E., Mottl V., Krasotkina O.* Estimation of time-varying linear regression with unknown time-volatility via continuous generalization of the Akaike information criterion // World Academy of Sciences, Engineering and Technology. 2009. V. 51.

[4] *Tipping M. E.* The relevance vector machine // Advances Neural Information Processing Systems. 2000. V. 12. P. 652–658.

[5] *Kropotov D. A., Ptashko N. O., Vetrov D. P.* Relevant regressors selection by continuous AIC // Pattern Recognition and Image Analysis. 2009. V. 3. Pp. 456–464.