ЖУРНАЛ ВЫЧИСЛИТЕЛЬНОЙ МАТЕМАТИКИ И МАТЕМАТИЧЕСКОЙ ФИЗИКИ, 2009, том 49, № 11, с. 1–14

УДК 519.71

ЭФФЕКТИВНЫЙ МЕТОД ОТБОРА ПРИЗНАКОВ В ЛИНЕЙНОЙ РЕГРЕССИИ С ПОМОЩЬЮ ОБОБЩЕНИЯ ИНФОРМАЦИОННОГО КРИТЕРИЯ АКАИКЕ[1)]

© 2009 г. Д. П. Ветров*, Д. А. Кропотов, Н. О. Пташко***

*(*119992 Москва, Ленинские горы, МГУ ВМиК;*
***119333 Москва, ул.Вавилова, 40, ВЦ РАН)*
e-mail: vetrovd@yandex.ru; dkropotov yandex.ru; ptashko inbox.ru

Поступила в редакцию 12.05.2009 г.

Предлагается метод отбора признаков для линейной регрессии с помощью обобщения информационного критерия Акаике. Использование классического информационного критерия Акаике (ИКА) для отбора признаков связано с полным перебором по всем подмножествам признаков, что приводит к неоправданно большим вычислительным и временным затратам. Предлагается новый информационный критерий, который является непрерывным обобщением ИКА. В результате задача отбора признаков сводится к задаче гладкой оптимизации. Выводится эффективная процедура решения полученной задачи оптимизации. Экспериментальные исследования показывают, что разработанный метод действительно позволяет быстро и эффективно отбирать признаки в линейной регрессии. В экспериментах новая процедура также сравнивается с методом релевантных векторов, который является методом отбора признаков на основе байесовского подхода. Показано, что обе процедуры близки по результатам. Основное отличие нового метода состоит в том, что некоторые коэффициенты регуляризации становятся тождественно равными нулю. Это позволяет избежать эффекта переупрощения модели, который характерен для метода релевантных векторов. Также рассматривается специальный случай (так называемая недиагональная регуляризация), в котором оба метода оказываются идентичными. Библ. 18. Фиг. 4. Табл. 2.

Ключевые слова: распознавание образов, линейная регрессия, отбор признаков, информационный критерий Акаике.

1. ВВЕДЕНИЕ

Байесовские методы широко используются для построения процедур автоматического выбора модели. Метод релевантных векторов (МРВ), предложенный в [1], является одним из примеров применения байесовской парадигмы в задаче линейной регрессии. В МРВ с каждым весом регрессора в линейном решающем правиле связывается индивидуальный коэффициент регуляризации (L2-регуляризация). Коэффициенты регуляризации подбираются автоматически путем максимизации правдоподобия модели (обоснованности). В результате такой процедуры, известной также как АОР (автоматическое определение релевантности (см. [2])), большинство коэффициентов регуляризации становятся равными бесконечности, что соответствует обнулению весов регрессоров и удалению соответствующих им регрессоров из модели. L1-регуляризация, использующая общий коэффициент регуляризации (см. [3]) либо использующая несобственное априорное распределения Джеффри с последующим интегрированием по коэффициентам регуляризации, также показывают высокую разреженность (число нулевых весов), сохраняя при этом хорошую обобщающую способность (см. [4]–[6]).

Известный информационный критерий Байеса–Шварца (см. [7]) может быть рассмотрен как грубая аппроксимация логарифма маргинального правдоподобия (обоснованности, см. [8]). Также для решения задач выбора моделей широко применяется информационный критерий Акаике (ИКА, см. [9]), предлагающий альтернативный подход, основанный на теории информации. Несмотря на то, что этот метод был изначально предложен для выбора из конечного числа моделей, он может быть расширен на случай континуального семейства моделей. В работе пред-

[1)] Работа выполнена при финансовой поддержке РФФИ (коды проектов 08-01-00405, 08-01-90016, 08-01-90427, 07-01-00211).

ложено подобное обобщение информационного критерия Акаике для выбора модели с дальнейшим применением в задаче линейной регрессии. Подбор коэффициентов регуляризации, связанных индивидуально с каждым весом, производится путем максимизации непрерывного аналога критерия Акаике (ОИКА). Особый интерес представляет разреженность решений, получаемых с помощью нового метода, и сравнение нового метода с классическим МРВ.

В разд. 2 приведен вывод непрерывного аналога критерия Акаике (ОИКА – обобщенный информационный критерий Акаике). В разд. 3 показано применение критерия ОИКА к задаче обобщенной линейной регрессии и выведены итеративные формулы пересчета коэффициентов регуляризации. В разд. 4 представлены результаты экспериментов и проведено сравнение ОИКА с классическим МРВ.

2. ОБОБЩЕНИЕ ИНФОРМАЦИОННОГО КРИТЕРИЯ АКАИКЕ

Опишем расширение ИКА на непрерывный случай. Пусть задана обучающая выборка $Z = (\mathbf{z}_1, ..., \mathbf{z}_n)$, $\mathbf{z} \in \mathbb{R}^d$. Требуется восстановить неизвестную плотность распределения $p(\mathbf{x})$ на элементах множества X, где Z и X – выборки длины n из одного вероятностного распределения.

Для описания общей схемы поиска $p(\mathbf{x})$ введем понятие модели.

Определение 1. *Вероятностной моделью алгоритмов восстановления плотностей* назовем тройку $\langle \Omega, p(X|\mathbf{w}), p(\mathbf{w})\rangle$, где $\Omega = \{\mathbf{w}\}$ – значения параметров плотностей распределения, $p(X|\mathbf{w}) = \prod_{i=1}^{n} p(\mathbf{x}_i|\mathbf{w})$ – функция правдоподобия выборки X при фиксированном значении $\mathbf{w}$ и $p(\mathbf{w})$ – априорное распределение на $\mathbf{w}$.

Предположим, что априорное распределение зависит от некоторого параметра A, т.е. может быть записано в виде $p(\mathbf{w}|A)$. Тогда, варьируя A, получаем параметрическое семейство моделей алгоритмов восстановления плотностей $\{\langle \Omega, p(X|\mathbf{w}), p(\mathbf{w}|A)\rangle, A \in \mathcal{A}\}$. В этом случае A называется параметром вероятностной модели.

Определение 2. Назовем *байесовской оценкой параметра* $\mathbf{w}$ значение $\mathbf{w}_{MP}(Z, A)$, максимизирующее величину регуляризованного правдоподобия, т.е.

$$\mathbf{w}_{MP}(Z, A) \stackrel{\Delta}{=} \arg\max_{\mathbf{w}} p(Z|\mathbf{w})p(\mathbf{w}|A).$$

Заметим, что байесовская оценка зависит как от обучающей выборки Z, так и от параметра вероятностной модели A. Пусть

$$\mathbf{w}_n^*(A) \stackrel{\Delta}{=} \mathbb{E}_Z \mathbf{w}_{MP}(Z, A),$$

$$C_n(A) \stackrel{\Delta}{=} \mathbb{E}_Z[\mathbf{w}_{MP}(Z, A) - \mathbf{w}_n^*(A)][\mathbf{w}_{MP}(Z, A) - \mathbf{w}_n^*(A)]^{\mathrm{T}},$$

где математические ожидания берутся по всем выборкам длины n из данного распределения $p(\mathbf{x})$.

Определение 3. *Матрицей Фишера* назовем выражение вида

$$F \stackrel{\Delta}{=} -\int \nabla_{\mathbf{w}} \nabla_{\mathbf{w}} \log p(\mathbf{x}|\mathbf{w}) p(\mathbf{x}) d\mathbf{x}.$$

Пусть $F_n = nF$. Заметим, что $F_n = \mathbb{E}_X \nabla_{\mathbf{w}} \nabla_{\mathbf{w}} \log p(X|\mathbf{w}) = \nabla_{\mathbf{w}} \nabla_{\mathbf{w}} \mathbb{E}_X \log p(X|\mathbf{w})$. В дальнейшем под символом ∇ будем понимать $\nabla_{\mathbf{w}}$.

В случае фиксированного значения параметра A (т.е. фиксированной вероятностной модели) в качестве оценки $p(\mathbf{x})$ будем использовать $p(\mathbf{x}|\mathbf{w}_{MP}(Z, A))$. Параметр модели может быть подобран, следуя идее Акаике, путем максимизации информации Кульбака по A:

$$\mathbb{E}_X \mathbb{E}_Z \log p(X|\mathbf{w}_{MP}(Z, A)) = \int\int p(Z)p(X)\log p(X|\mathbf{w}_{MP}(Z, A))dXdZ \longrightarrow \max_A. \tag{2.1}$$

Итак, задача обучения состоит в нахождении значения параметра вероятностной модели A, оптимального в смысле критерия (2.1). Аналитически вычислить данный интеграл не удается. Сформулируем условия, накладываемые на семейство вероятностных моделей, при выполнении которых данное выражение можно упростить.

Теорема 1. *Пусть A – симметричная неотрицательно-определенная квадратная матрица действительных чисел. Пусть при любом A для вероятностной модели $\Omega(A)$ справедливо следующее*:

1) $p(X|\mathbf{w}) = \prod_{i=1}^{n} p(\mathbf{x}_i|\mathbf{w})$, *т.е. объекты обучающей выборки – это независимые, одинаково распределенные случайные величины;*

2) $\log p(X|\mathbf{w})$ – *квадратичная функция по* $\mathbf{w}$;

3) $p(\mathbf{w}|A) = \mathcal{N}(\mathbf{w}|\mathbf{0}, A^{-1})$, *т.е* $\mathbf{w}$ *распределен нормально с центром в нуле и матрицей ковариации* A^{-1};

4) *случайные величины* $\mathbf{w}_{MP}(X, A)$ *и* $\nabla\nabla\log p(X|\mathbf{w}_{MP}(X, A))$ *независимы.*

Тогда верно соотношение

$$\mathbb{E}_X\mathbb{E}_Z\log p(X|\mathbf{w}_{MP}(Z, A)) = \mathbb{E}_X\log p(X|\mathbf{w}_{MP}(X, A)) - \operatorname{tr}(F_n + A)C_n(A). \tag{2.2}$$

Для упрощения записи будем опускать зависимость $\mathbf{w}_{MP}$, $\mathbf{w}_n^*$ и C_n от A (подразумевая ее). Используя теорему о замене переменных под знаком интеграла Лебега (см. [10, гл. II, п. 6]), можем переписать исходное выражение в виде

$$\mathbb{E}_X\mathbb{E}_Z\log p(X|\mathbf{w}_{MP}(Z)) = \mathbb{E}_{w_{MP}}\mathbb{E}_X\log p(X|\mathbf{w}_{MP}). \tag{2.3}$$

Раскладывая внутреннее математическое ожидание в ряд Тейлора в точке $\mathbf{w}_n^*$, получаем

$$\begin{aligned}\mathbb{E}_{\mathbf{w}_{MP}}\mathbb{E}_X\log p(X|\mathbf{w}_{MP}) &= \mathbb{E}_{\mathbf{w}_{MP}}\mathbb{E}_X\log p(X|\mathbf{w}_n^*) + \mathbb{E}_{\mathbf{w}_{MP}}\mathbb{E}_X[\nabla\log p(X|\mathbf{w}_n^*)]^{\mathrm{T}}(\mathbf{w}_{MP} - \mathbf{w}_n^*) + \\ &+ \frac{1}{2}\mathbb{E}_{\mathbf{w}_{MP}}(\mathbf{w}_{MP} - \mathbf{w}_n^*)^{\mathrm{T}}[\nabla\nabla\mathbb{E}_X\log p(X|\mathbf{w}_n^*)](\mathbf{w}_{MP} - \mathbf{w}_n^*) = \mathbb{E}_X\log p(X|\mathbf{w}_n^*) - \frac{1}{2}\operatorname{tr}F_nC_n.\end{aligned} \tag{2.4}$$

Для оценки первого слагаемого в (2.4) разложим в ряд Тейлора $\log p(X|\mathbf{w}_n^*)$ в точке $\mathbf{w}_{MP}(X)$. Тогда имеем

$$\begin{aligned}\mathbb{E}_X\log p(X|\mathbf{w}_n^*) &= \mathbb{E}_X\log p(X|\mathbf{w}_{MP}(X)) + \mathbb{E}_X\{[\nabla\log p(X|\mathbf{w}_{MP}(X))]^{\mathrm{T}}[\mathbf{w}_n^* - \mathbf{w}_{MP}(X)]\} + \\ &+ \frac{1}{2}\mathbb{E}_X\{[\mathbf{w}_n^* - \mathbf{w}_{MP}(X)]^{\mathrm{T}}\nabla\nabla\log p(X|\mathbf{w}_{MP}(X))[\mathbf{w}_n^* - \mathbf{w}_{MP}(X)]\}.\end{aligned} \tag{2.5}$$

Так как $\log p(X|\mathbf{w})$ квадратична по $\mathbf{w}$, то легко показать, что

$$\nabla\log p(X|\mathbf{w}_{MP}) = A\mathbf{w}_{MP}.$$

Используя данный факт и условие независимости случайных величин $\mathbf{w}_{MP}(X)$ и $\nabla\nabla\log p(X|\mathbf{w}_{MP}(X))$, можно упростить два последних слагаемых в (2.5):

$$\begin{aligned}&\mathbb{E}_X[\nabla\log p(X|\mathbf{w}_{MP}(X))]^{\mathrm{T}}[\mathbf{w}_n^* - \mathbf{w}_{MP}(X)] + \\ &+ \frac{1}{2}\mathbb{E}_X[\mathbf{w}_n^* - \mathbf{w}_{MP}(X)]^{\mathrm{T}}\nabla\nabla\log p(X|\mathbf{w}_{MP}(X))[\mathbf{w}_n^* - \mathbf{w}_{MP}(X)] = \\ &= \mathbb{E}_X\mathbf{w}_{MP}^{\mathrm{T}}(X)A[\mathbf{w}_n^* - \mathbf{w}_{MP}(X)] + \\ &+ \frac{1}{2}\operatorname{tr}\{\mathbb{E}_X\nabla\nabla\log p(X|\mathbf{w}_{MP}(X))\mathbb{E}_X[\mathbf{w}_n^* - \mathbf{w}_{MP}(X)]^{\mathrm{T}}[\mathbf{w}_n^* - \mathbf{w}_{MP}(X)]\} = -\operatorname{tr}AC_n - \frac{1}{2}\operatorname{tr}F_nC_n.\end{aligned}$$

Подставляя полученное выражение в (2.5) и объединяя результат с (2.4), получаем

$$\begin{aligned}&\mathbb{E}_Z\mathbb{E}_X\log p(X|\mathbf{w}_{MP}(Z, A)) = \\ &= \mathbb{E}_X\log p(X|\mathbf{w}_{MP}(X, A)) - \operatorname{tr}AC_n - \frac{1}{2}\operatorname{tr}F_nC_n - \frac{1}{2}\operatorname{tr}F_nC_n = \\ &= \mathbb{E}_X\log p(X|\mathbf{w}_{MP}(X, A)) - \operatorname{tr}(F_n + A)C_n.\end{aligned} \tag{2.6}$$

Теорема доказана.

Следствие 1. При использовании в методах распознавания критерий (2.1) может быть приближенно вычислен по формуле

$$\mathbb{E}_X\mathbb{E}_Z\log p(X|\mathbf{w}_{MP}(Z,A)) \approx \log p(Z|\mathbf{w}_{MP}(Z,A)) - \operatorname{tr}(H(Z)+A)^{-1}H(Z), \tag{2.7}$$

где $H(Z) = \nabla\nabla\log p(Z|\mathbf{w}) = \Sigma_{i=1}^{n}\nabla\nabla\log p(\mathbf{z}_i|\mathbf{w})$ – гессиан логарифма правдоподобия.

Покажем, что $C_n \approx (F_n + A)^{-1}F_n(F_n + A)^{-1}$. Обозначим через $\mathbf{w}_{ML}(X)$ оценку максимального правдоподобия на выборке X. Известно (см. [11]), что $\sim\mathcal{N}(\mathbf{w}_{ML}|\mathbf{w}_*,\ F_n^{-1})$, где $\mathbf{w}_* = \operatorname{argmax}\int p(\mathbf{x}|\mathbf{w})d\mathbf{x}$. При условии квадратичности $\log p(\mathbf{x}|\mathbf{w})$ по $\mathbf{w}$ легко показать, что

$$\mathbf{w}_{MP}(X,A) = [H(X)+A]^{-1}H(X)\mathbf{w}_{ML}(X), \tag{2.8}$$

где $H(X) = \Sigma_{i=1}^{n}\nabla\nabla\log p(\mathbf{x}_i|\mathbf{w})$.

С учетом $\mathbb{E}\nabla\nabla\log p(\mathbf{x}_i|\mathbf{w}) = F$, используя закон больших чисел (см. [10, гл. III, п. 3]), записываем

$$\forall\varepsilon > 0 \quad P\left(\frac{|H(X) - F_n|}{n} \geq \varepsilon\right) \longrightarrow 0 \quad \text{при} \quad n \longrightarrow \infty \tag{2.9}$$

Рассмотрим множество $I_n(\varepsilon) = \left\{X\middle|\frac{|H(X) - F_n|}{n} \geq \varepsilon\right\}$. Из (2.9) следует, что $P(I_n(\varepsilon)) \longrightarrow 0$ при $n \longrightarrow \infty$; при этом на множестве $\mathbb{R}^{n\times d}\setminus I_n(\varepsilon)$ справедливо представление

$$\frac{H(X)}{n} = \frac{F_n}{n} + \delta_1(n), \quad \text{где} \quad \|\delta_1(n)\| \longrightarrow 0 \quad \text{при} \quad n \longrightarrow \infty.$$

При фиксированных $\varepsilon > 0$ и $n > 0$ имеем

$$\mathbb{E}\mathbf{w}_{MP}(X) = \int\limits_{\mathbb{R}^{n\times d}\setminus I_n(\varepsilon)} \mathbf{w}_{MP}(X)p(X)dX + \int\limits_{I_n(\varepsilon)} \mathbf{w}_{MP}(X)p(X)dX. \tag{2.10}$$

Предполагая ограниченность $\mathbf{w}_{MP}(X)$, применяем теорему о среднем ко второму слагаемому в (2.10):

$$\mathbb{E}\mathbf{w}_{MP}(X) = \int\limits_{\mathbb{R}^{n\times d}\setminus I_n(\varepsilon)} [H(X)+A]^{-1}H(X)\mathbf{w}_{ML}(X)p(X)dX + L_{\mathbf{w}_{MP}}P(I_n(\varepsilon)), \tag{2.11}$$

где $L_{\mathbf{w}_{MP}}$ – некоторая положительная константа. Заметим, что при достаточно больших значениях n справедлива оценка $\|\delta_1(n)\| \leq \|F_n/n + A/n\|^{-1}$; тогда верно (см. [12, гл. 5, п. 6]) следующее разложение: $[F_n/n + A/n + \delta_1(n)]^{-1} = (F_n/n + A/n)^{-1} + \delta_2(n)$, где $\|\delta_2(n)\| \longrightarrow 0$ при $n \longrightarrow \infty$. Далее,

$$\int\limits_{\mathbb{R}^{n\times d}\setminus I_n(\varepsilon)} [H(X)+A]^{-1}H(X)\mathbf{w}_{ML}(X)p(X)dX =$$

$$= \int\limits_{\mathbb{R}^{n\times d}\setminus I_n(\varepsilon)} \left[\frac{H(X)}{n} + \frac{A}{n}\right]^{-1}\frac{H(X)}{n}\mathbf{w}_{ML}(X)p(X)dX =$$

$$= \int\limits_{\mathbb{R}^{n\times d}\setminus I_n(\varepsilon)} \left(\frac{F_n}{n} + \frac{A}{n} + \delta_1(n)\right)^{-1}\left(\frac{F_n}{n} + \delta_1(n)\right)\mathbf{w}_{ML}(X)p(X)dX = \tag{2.12}$$

$$= \int\limits_{\mathbb{R}^{n\times d}\setminus I_n(\varepsilon)} \left[\left(\frac{F_n}{n}+\frac{A}{n}\right)^{-1}+\delta_2(n)\right]\left(\frac{F_n}{n}+\delta_1(n)\right)\mathbf{w}_{ML}(X)p(X)dX =$$

$$= \int\limits_{\mathbb{R}^{n\times d}\setminus I_n(\varepsilon)} (F_n+A)^{-1}F_n\mathbf{w}_{ML}(X)p(X)dX + \int\limits_{\mathbb{R}^{n\times d}\setminus I_n(\varepsilon)} \delta_3(n)p(X)dX,$$

где $\|\delta_3(n)\| \longrightarrow 0$ при $n \longrightarrow \infty$. Итак, получаем

$$\mathbb{E}\mathbf{w}_{MP}(X) = (F_n+A)^{-1}F_n \int\limits_{\mathbb{R}^{n\times d}\setminus I_n(\varepsilon)} \mathbf{w}_{ML}(X)p(X)dX + \int\limits_{\mathbb{R}^{n\times d}\setminus I_n(\varepsilon)} \delta_3(n)p(X)dX + L_{\mathbf{w}_{MP}}P(I_n(\varepsilon)), \qquad (2.13)$$

В силу того что $L_{\mathbf{w}_{MP}}P(I_n(\varepsilon))$ – положительная константа, $P(I_n(\varepsilon)) \longrightarrow 0$ и $\|\delta_3(n)\| \longrightarrow 0$ при $n \longrightarrow \infty$, при увеличении объема выборки множество $\mathbb{R}^{n\times d}\setminus I_n(\varepsilon)$ будет стремиться к множеству $\mathbb{R}^{n\times d}$, а два последних слагаемых в (2.13) – к нулю. Отбрасывая два последних слагаемых, получаем следующее приближение к $\mathbb{E}\mathbf{w}_{MP}(X)$:

$$\mathbb{E}\mathbf{w}_{MP}(X) \approx (F_n+A)^{-1}F_n\mathbb{E}\mathbf{w}_{ML}(X).$$

Аналогично можем провести приближение и для $\mathbb{E}\mathbf{w}_{MP}(X)\mathbf{w}_{MP}^{\mathrm{T}}(X)$:

$$\mathbb{E}\mathbf{w}_{MP}(X)\mathbf{w}_{MP}^{\mathrm{T}}(X) \approx (F_n+A)^{-1}F_n\mathbb{E}\mathbf{w}_{ML}(X)\mathbf{w}_{ML}^{\mathrm{T}}(X)F_n(F_n+A)^{-1}.$$

Учитывая, что $F_n^{-1} = [\mathbb{E}\mathbf{w}_{ML}(X)\mathbf{w}_{ML}^{\mathrm{T}}(X) - (\mathbb{E}\mathbf{w}_{ML}(X))(\mathbb{E}\mathbf{w}_{ML}(X))^{\mathrm{T}}]$, получаем

$$\begin{aligned} C_n &= \mathbb{E}\mathbf{w}_{MP}(X)\mathbf{w}_{MP}^{\mathrm{T}}(X) - (\mathbb{E}\mathbf{w}_{MP}(X))(\mathbb{E}\mathbf{w}_{MP}(X))^{\mathrm{T}} \approx \\ &\approx (F_n+A)^{-1}F_n[\mathbb{E}\mathbf{w}_{ML}(X)\mathbf{w}_{ML}^{\mathrm{T}}(X) - (\mathbb{E}\mathbf{w}_{ML}(X))(\mathbb{E}\mathbf{w}_{ML}(X))^{\mathrm{T}}]F_n(F_n+A)^{-1} = \\ &= (F_n+A)^{-1}F_nF_n^{-1}F_n(F_n+A)^{-1} = (F_n+A)^{-1}F_n(F_n+A)^{-1}. \end{aligned} \qquad (2.14)$$

Подставив приближенное значение C_n во второе слагаемое (2.2), получим

$$\mathbb{E}_X\mathbb{E}_Z\log p(X|\mathbf{w}_{MP}(Z,A)) = \mathbb{E}_X\log p(X|\mathbf{w}_{MP}(X,A)) - \mathrm{tr}(F_n+A)^{-1}F_n$$

Подставляя в полученное выражение вместо X обучающую выборку Z и вместо матрицы F_n ее несмещенную оценку $H(Z)$, получаем утверждение следствия.

Заметим, что если A – диагональная матрица с элементами, равными либо 0, либо $+\infty$, то $\mathrm{tr}(F_n(F_n+A)^{-1}) = k$ – количество нулевых диагональных элементов A; при этом (2.7) становится с точностью до бесконечно малой константы эквивалентно классическому информацинному критерию Акаике. Подобное непрерывное расширение критерия (ОИКА) может также быть рассмотрено, как частный случай девиантного информационного критерия, описанного в [13].

3. ПРИМЕНЕНИЕ ОИКА К ЗАДАЧЕ ОБОБЩЕННОЙ ЛИНЕЙНОЙ РЕГРЕССИИ

3.1. Постановка задачи оптимизации

Рассмотрим классическую задачу обобщенной линейной регрессии. Пусть $(X, \mathbf{t}) = \{(\mathbf{x}_1, t_1), \ldots (\mathbf{x}_n, t_n)\}$ – обучающая выборка, где $\mathbf{x}_i = (x_i^1, \ldots, x_i^d) \in \mathbb{R}^d$ – вектор наблюдаемых признаков объекта, а $t_i \in \mathbb{R}$ – значение зависимой переменной. Зафиксируем некоторое множество базисных функций $\{\phi_1(\mathbf{x}), \ldots, \phi_m(\mathbf{x})\}$, $\phi_j : \mathbb{R}^d \longrightarrow \mathbb{R}^d$. Требуется найти вектор весов $\mathbf{w} \in \mathbb{R}$ такой, чтобы функция

$$y(\mathbf{x}) = \mathbf{w}^{\mathrm{T}}\boldsymbol{\phi}(\mathbf{x}) = \sum_{j=1}^{m} w_j\phi_j(\mathbf{x})$$

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

$$\mathbf{w}_{MP} = \arg\max_{\mathbf{w}} p(\mathbf{t}|X, \mathbf{w})p(\mathbf{w}|\alpha), \tag{3.1}$$

где

$$p(\mathbf{t}|X, \mathbf{w}) = \frac{1}{\sqrt{(2\pi)^n\sigma^n}}\exp\left(-\frac{1}{2\sigma^2}\|\Phi\mathbf{w} - \mathbf{t}\|^2\right) \tag{3.2}$$

есть функция правдоподобия,

$$p(\mathbf{w}|\alpha) = \sqrt{\left(\frac{\alpha}{2\pi}\right)^m}\exp\left(-\frac{\alpha}{2}\|\mathbf{w}\|^2\right)$$

есть априорное распределение на веса. Априорное распределение имеет смысл регуляризатора, штрафующего большие значения $\mathbf{w}$. Более общий случай семейства регуляризаторов рассмотрен в методе релевантных векторов (MPB, см. [1]), где для каждого веса w_j вводится свой коэффициент регуляризации, а априорное распределение имеет вид

$$p(\mathbf{w}|\alpha) = \prod_{j=1}^{m}\sqrt{\frac{\alpha_j}{2\pi}}\exp\left(-\frac{\alpha_j}{2}w_j^2\right) = \frac{\det(A)}{(2\pi)^{m/2}}\exp\left(-\frac{1}{2}\mathbf{w}^{\mathrm{T}}A\mathbf{w}\right), \tag{3.3}$$

где $A = \mathrm{diag}(\alpha_1, \ldots, \alpha_m)$ – матрица регуляризации, $\alpha_j \geq 0$. Такое априорное распределение позволяет проводить выбор базисных функций. В случае если $\alpha_j = 0$, то никаких дополнительных ограничений на значение веса $w_{MP,j}$ не накладывается и его значение совпадает с точкой максимума правдоподобия $w_{ML,j}$. Если параметр регуляризации α_j стремится к плюс бесконечности, то соответствующая базисная функция $\phi_j(\cdot)$ исключается из модели, так как ее вес $w_{MP,j} = 0$. Таким образом, априорное распределение (3.3) вместе с функцией правдоподобия (3.2) и методом байесовского оценивания (3.1) позволяет решать задачу селекции базисных функций. Данная задача переходит в задачу отбора признаков, если в качестве базисных функций выбираются исходные признаки $\phi_j(\mathbf{x}) = x^j$. Если в качестве базисных функций выбираются ядровые или потенциальные функции с центром в объектах обучающей выборки $\phi_j(\mathbf{x}) = K(\mathbf{x}_j, \mathbf{x})$, то данный подход позволяет отбирать релевантные объекты.

Заметим, что задачу восстановления регрессии в статистической постановке можно рассматривать как частный случай задачи восстановления плотностей. Используя введенные функции, сформулируем процедуру обучения (подбора $\mathbf{w}$ и α) в терминах вероятностных моделей алгоритмов восстановления плотностей. Параметрическое семейство вероятностных моделей может быть записано в виде

$$\{\langle \mathbb{R}^m, P(\mathbf{t}|X, \mathbf{w}), p(\mathbf{w}|\alpha)\rangle, \alpha \in \mathbb{R}^m\}. \tag{3.4}$$

При фиксированной вероятностной модели α в качестве решения задачи выбираем $\mathbf{w}_{MP}(X, \alpha)$ – байесовскую оценку вектора весов $\mathbf{w}$. Рассмотрим далее способ подбора α.

Условия теоремы 1 для семейства вероятностных моделей (3.4) выполнены. Поэтому для выбора наилучшей модели α воспользуемся следствием теоремы 1.

$$\begin{aligned}\alpha = \arg\max f(\alpha) &= \arg\max\{\log p(\mathbf{t}|X, \mathbf{w}_{MP}) - \mathrm{tr}[H(H + A)^{-1}]\} = \\ &= \arg\max\{\mathscr{L}(\mathbf{w}_{MP}) - \mathrm{tr}[H(H + A)^{-1}]\}.\end{aligned} \tag{3.5}$$

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

3.2. Процедура оптимизации

Поиск решения задачи оптимизации (3.5) будем проводить с помощью покоординатного спуска – отдельно по каждой компоненте α_j. Для вывода итеративных уравнений пересчета α_j воспользуемся тождеством блочного матричного обращения:

$$\begin{pmatrix} P & Q \\ R & S \end{pmatrix}^{-1} = \begin{pmatrix} P^{-1} + P^{-1}QBRP^{-1} & -P^{-1}QB \\ -BRP^{-1} & B \end{pmatrix}. \tag{3.6}$$

Здесь $P \in \mathbb{R}^{p\times p}$, $Q \in \mathbb{R}^{p\times q}$, $R \in \mathbb{R}^{q\times p}$, $S \in \mathbb{R}^{q\times q}$ – некоторые матрицы, а $B = (S - RP^{-1}Q)^{-1}$ – дополнение Шура.

Далее применим это тождество к матрице $(H + A)$, представленной в следующем виде:

$$(H + A) = \begin{pmatrix} P & \mathbf{q} \\ \mathbf{q}^{\mathrm{T}} & h_{mm} + \alpha_m \end{pmatrix}.$$

Для простоты изложения, не ограничивая общности, будем выводить итеративные уравнения для α_m. Пересчет остальных компонент вектора $\boldsymbol{\alpha}$ производится аналогичным образом. Используя (3.6), получаем

$$(H + A)^{-1} = \begin{pmatrix} P^{-1} + \beta_m P^{-1}\mathbf{q}\mathbf{q}^{\mathrm{T}}P^{-1} & -\beta_m P^{-1}\mathbf{q} \\ -\beta_m \mathbf{q}^{\mathrm{T}}P^{-1} & \beta_m \end{pmatrix},$$

где $\beta_m = (h_{mm} + \alpha_m - \mathbf{q}^{\mathrm{T}}P^{-1}\mathbf{q})^{-1}$ – скалярное дополнение Шура. Тогда $\mathrm{tr}[H(H + A)^{-1}]$ может быть выражено как функция от α_m, при условии, что остальные α_j фиксированы:

$$\mathrm{tr}[H(H + A)^{-1}] = \mathrm{tr}(H_{(m)}P^{-1}) + \beta_m[\mathbf{q}^{\mathrm{T}}P^{-1}(H_{(m)}P^{-1}\mathbf{q} - 2\mathbf{q}^{\mathrm{T}}P^{-1}\mathbf{q} + h_{mm}].$$

Здесь нижний индекс $_{(m)}$ означает вектор (или матрицу), у которого удалена m-ая строка (и столбец).

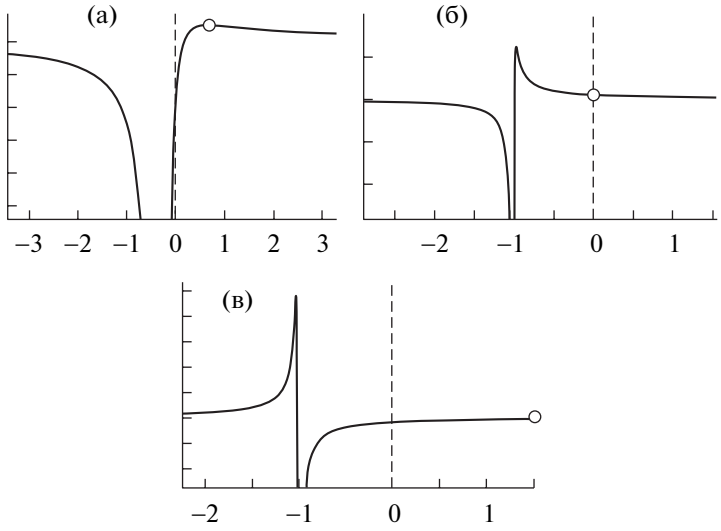
Заметим, что значение $\mathrm{tr}(H_{(m)}P^{-1})$ в точности равно значению $\mathrm{tr}[H(H + A)^{-1}]$ при $\alpha_m = +\infty$, т.е. при удалении из модели m-й базисной функции.

Рассмотрим разницу между точкой максимума апостериорного распределения $\mathbf{w}_{MP}$ и точкой максимума апостериорного распределения при бесконечно большом значении коэффициента регуляризации для m-й базисной функции (при этом значения остальных компонент вектора $\boldsymbol{\alpha}$ остаются неизменными) $\mathbf{w}^*_{MP} = \mathbf{w}_{MP}\big|_{\alpha_m = +\infty} \in \mathbb{R}^m$. Пусть $\boldsymbol{\psi} = H\mathbf{w}_{ML}$. Используя соотношение $\mathbf{w}_{MP} = (H + A)^{-1}H\mathbf{w}_{ML}$, получаем

$$\begin{aligned} \mathbf{w}_{MP} - \mathbf{w}^*_{MP} &= [(H + A)^{-1} - (H + A)^{-1}\big|_{\alpha_m = +\infty}]\boldsymbol{\psi} = \\ &= \left[\begin{pmatrix} P^{-1} + \beta_m P^{-1}\mathbf{q}\mathbf{q}^{\mathrm{T}}P^{-1} & -\beta_m P^{-1}\mathbf{q} \\ -\beta_m \mathbf{q}^{\mathrm{T}}P^{-1} & \beta_m \end{pmatrix} - \begin{pmatrix} P^{-1} & \mathbf{0} \\ \mathbf{0}^{\mathrm{T}} & 0 \end{pmatrix}\right]\begin{pmatrix} \boldsymbol{\psi}_{(m)} \\ \psi_m \end{pmatrix} = \\ &= \beta_m \begin{pmatrix} P^{-1}\mathbf{q}\mathbf{q}^{\mathrm{T}}P\boldsymbol{\psi}_{(m)} - \psi_m P^{-1}\mathbf{q} \\ -\mathbf{q}^{\mathrm{T}}P^{-1}\boldsymbol{\psi}_{(m)} + \psi_m \end{pmatrix} = \beta_m \boldsymbol{\xi}_m. \end{aligned} \tag{3.7}$$

Рассмотрим разность между значениями логарифма правдоподобия в точках $\mathbf{w}_{MP}$ и $\mathbf{w}^*_{MP}$:

$$\mathscr{L}(\mathbf{w}_{MP}) - \mathscr{L}(\mathbf{w}^*_{MP}) = \beta_m \nabla\mathscr{L}(\mathbf{w}^*_{MP})^{\mathrm{T}}\boldsymbol{\xi}_m - \frac{\beta_m^2}{2}\boldsymbol{\xi}_m^{\mathrm{T}}H\boldsymbol{\xi}_m = \beta_m \boldsymbol{\zeta}_m^{\mathrm{T}}\boldsymbol{\xi}_m - \frac{\beta_m^2}{2}\boldsymbol{\xi}_m^{\mathrm{T}}H\boldsymbol{\xi}_m.$$



Фиг. 1.

Используя (2.8), записываем градиент в виде

$$\zeta_m = \nabla\mathcal{L}(\mathbf{w}^*_{MP}) = -H(\mathbf{w}^*_{MP} - \mathbf{w}_{ML}) =$$

$$= -[H(H+A)^{-1}\big|_{\alpha_m=+\infty} - I]\psi = -\begin{pmatrix} (H_{(m)}P^{-1} - I)\psi_{(m)} \\ \mathbf{q}^{\mathrm{T}}P^{-1}\psi_{(m)} - \psi_m \end{pmatrix}. \tag{3.8}$$

В результате значение критерия ОИКА как функции от β_m при фиксированных $\alpha_j, j \neq m$ представляется в следующем виде:

$$(\beta_m) = (0) - \frac{1}{2}\beta_m^2 \xi_m^{\mathrm{T}} H \xi_m + \beta_m \zeta_m^{\mathrm{T}} \xi_m - \beta_m \mathbf{q}^{\mathrm{T}} P^{-1} H_{(m)} P^{-1}\mathbf{q} + 2\beta_m \mathbf{q}^{\mathrm{T}} P^{-1}\mathbf{q} - \beta_m h_{mm}. \tag{3.9}$$

Критерий квадратичен по β_m и, следовательно, имеет единственный максимум, который вычисляется аналитически по формуле

$$\beta_m^* = \frac{\zeta_m^{\mathrm{T}}\xi_m - \mathbf{q}^{\mathrm{T}}P^{-1}H_{(m)}P^{-1}\mathbf{q} + 2\mathbf{q}^{\mathrm{T}}P^{-1}\mathbf{q} - h_{mm}}{\xi_m^{\mathrm{T}} H \xi_m} = \frac{b}{a}.$$

Используя выражения для ξ_m (3.7) и ζ_m (3.8), значения a и b вычисляем следующим образом:

$$a = (\mathbf{q}^{\mathrm{T}}P^{-1}\psi_{(m)} - \psi_m)^2(\mathbf{q}^{\mathrm{T}}P^{-1}H_{(m)}P^{-1}\mathbf{q} - 2\mathbf{q}^{\mathrm{T}}P^{-1}\mathbf{q} + h_{mm}), \tag{3.10}$$

$$\begin{aligned} b = &-(\psi_{(m)}^{\mathrm{T}}P^{-1}H_{(m)}P^{-1}\mathbf{q} - 2\psi_{(m)}^{\mathrm{T}}P^{-1}\mathbf{q} + \psi_m)(\mathbf{q}^{\mathrm{T}}P^{-1}\psi_{(m)} - \psi_m) - \\ &- \mathbf{q}^{\mathrm{T}}P^{-1}H_{(m)}P^{-1}\mathbf{q} + 2\mathbf{q}^{\mathrm{T}}P^{-1}\mathbf{q} - h_{mm} \end{aligned} \tag{3.11}$$

Перейдем от вспомогательной переменой β_m к исходной α_m:

$$\beta_m = (h_{mm} + \alpha_m - \mathbf{q}^{\mathrm{T}}P^{-1}\mathbf{q})^{-1};$$

следовательно,

$$\alpha_m^* = \mathbf{q}^{\mathrm{T}}P^{-1}\mathbf{q} - h_{mm} + \frac{1}{\beta_m^*}.$$

При использовании последнего выражения необходимо учитывать также, что $\alpha_m \geq 0$. Зависимость ОИКА от α_m имеет характерную форму ириса, показанную на фиг. 1. В случае (а) максимум

достигается для положительного α_j. В случае (б) критерий монотонно убывает в области $\alpha_j \geq 0$ и, следовательно, оптимальное значение $\alpha_j = 0$. В случае (в) оптимальное неотрицательное значение α_j равно $+\infty$. Критерий равен минус бесконечности, когда матрица $H + A$ вырождена, т.е. $\alpha_m = \mathbf{q}^{\mathrm{T}} P^{-1}\mathbf{q} - h_{mm}$. Согласно свойству дополнения Шура, $\mathbf{q}^{\mathrm{T}} P^{-1}\mathbf{q} - h_{mm} \leq 0$, т.е. "стебель" всегда соответствует неположительным α_m. В зависимости от взаиморасположения точки максимума и "стебля" значение α_m пересчитывается разными способами:

$$\alpha_m^{(\mathrm{new})} = \begin{cases} \alpha_m^*, & \alpha_m^* \geq 0, \\ 0, & \mathbf{q}^{\mathrm{T}} P^{-1}\mathbf{q} - h_{mm} < \alpha_m^* < 0, \\ +\infty, & \alpha_m^* < \mathbf{q}^{\mathrm{T}} P^{-1}\mathbf{q} - h_{mm}. \end{cases} \tag{3.12}$$

Выражения для α_j при $j \neq m$ аналогичны.

Для подбора значения параметра σ^2 продифференцируем ОИКА по σ^{-2}. Приравнивая производную к нулю, получаем следующую формулу пересчета:

$$\sigma^{2(\mathrm{new})} = \frac{\left\|\mathbf{t} - \boldsymbol{\Phi}\mathbf{w}_{MP}\right\|^2}{n - 2\mathrm{tr}(A(H+A)^{-1}H(H+A)^{-1})}. \tag{3.13}$$

Итак, доказана

Теорема 2. *Справедливы соотношения*

$$\underset{\alpha_j \geq 0}{\mathrm{argmax}}(\alpha_j) = \alpha_j^{(\mathrm{new})}, \quad j = 1, 2, \ldots, m,$$

$$\underset{\sigma^2 \geq 0}{\mathrm{argmax}}(\sigma^2) = \sigma^{2(\mathrm{new})},$$

где $\alpha_j^{(\mathrm{new})}$ и $\sigma^{2(\mathrm{new})}$ *рассчитываются по формулам* (3.12) *и* (3.13) *соответственно. Формула* (3.12) *подразумевает, что оптимизация критерия производится поочередно по каждой из* $\alpha_j, j = 1, 2, \ldots, m$, *при фиксированных остальных компонентах* $\boldsymbol{\alpha}$.

Используя полученный результат, можно построить итерационный процесс оптимизации критерия. На каждом шаге оптимизируется тот параметр α_j, который обеспечивает максимальный прирост критерия (см. Алгоритм 1). Такой алгоритм схож с методом обучения MPB, предложенным в [14].

Алгоритм 1 ОИКА MPB

вход Обучающая выборка $(X, \mathbf{t}) = \{\mathbf{x}_i, t_i\}_{i=1}^n$, $\mathbf{x}_i \in \mathbb{R}^d$, $t_i \in \mathbb{R}$, множество базистных функций $\{\phi_j(\mathbf{x})\}_{j=1}^m$.

Инициализировать $\alpha_j = +\inf\ \forall j = 1, 2, \ldots, m$, $\sigma^2 = \sigma_0^2$, $\Phi = \{\phi_j(\mathbf{x})_i\}_{i,j=1}^{n,m}$ и $A = \mathrm{diag}(\alpha_1, \ldots, \alpha_m)$.

Найти максимум логарифма правдоподобия $\mathbf{w}_{ML} = (\Phi^{\mathrm{T}}\Phi)^{-1}\Phi^{\mathrm{T}}\mathbf{t}$.

повторять

Вычислить $H = \sigma^{-2}\Phi^{\mathrm{T}}\Phi$ и $\Psi = H\mathbf{w}_{ML}$.

для $j = 1, 2, \ldots, m$ **цикл**

Вычислить $H_{(j)}$, $P^{-1} = (H_{(j)} + A_{(j)})^{-1}$ и $\mathbf{q}$ т.е. j-й столбец H без j-го элемента.

Вычислить a и b, используя выражения (3.10) и (3.11).

Вычислить оптимальное значение $\alpha_j^* = \mathbf{q}^{\mathrm{T}} P^{-1}\mathbf{q} - h_{jj} + a/b$ и текущее приращение ОИКА $\Delta_j = b^2/a$.

если $\alpha_j^* < 0$ **тогда**

если $\alpha_j^* > \mathbf{q}^{\mathrm{T}} P^{-1}\mathbf{q} - h_{jj}$ **тогда**

$\alpha_j^* = 0$, $\beta_0 = 1/(h_{jj} - \mathbf{q}^{\mathrm{T}} P^{-1}\mathbf{q})$, $\Delta_j = -\beta_0^2/(2a) + \beta_0 b$.

иначе

$\alpha_j^* = +\infty$, $\Delta_j = 0$.

конец если

конец если

если $\alpha_j \neq +\infty$ **тогда**

$\beta_{\mathrm{old}} = 1/(h_{jj} - \mathbf{q}^{\mathrm{T}} P^{-1}\mathbf{q} + \alpha_j)$

$\Delta_j^{\mathrm{old}} = -\beta_j^{\mathrm{old}}/(2a) + \beta_{\mathrm{old}} b$, $\Delta_j = \Delta_j - \Delta_j^{\mathrm{old}}$.

конец если

конец если

Найти $j^* = \arg\max_j \Delta_j$ и установить $\alpha_{j^*} = \alpha_{j^*}^*$.

Вычислить $A = \mathrm{diag}(\alpha_1, .., \alpha_m)$, $\mathbf{w}_{MP} = (H + A)^{-1} H \mathbf{w}_{ML}$ и пересчитать σ^2, используя (3.13).

пока процесс не сошелся

выход Решающее правило для нового объекта $\mathbf{x}$: $f(\mathbf{x}) = \sum_{j=1}^{m} w_{MP,j}\phi_j(\mathbf{x})$

3.3. Недиагональная регуляризация

Альтернативнй подход для определения коэффициентов регуляризации α предложен в методе релевантных векторов (MPB, см. [15]). В этом методе используется байесовская парадигма для оценивания параметров и коэффициенты регуляризации находятся с помощью максимизации правдоподобия модели (обоснованности):

$$EV(\alpha) = \int p(\mathbf{t}|X, \mathbf{w}) p(\mathbf{w}|\alpha) d\mathbf{w} \longrightarrow \max_{\alpha} \tag{3.14}$$

Здесь функция правдоподобия и априорное распределение выбираются, как и раньше, по формулам (3.2) и (3.3). Обоснованность максимизируется с помощью итерационной процедуры, в которой на каждом шаге подынтегральная функция аппроксимируется нормальным распределением.

Другой способ оптимизации обоснованности предложен в рамках подхода недиагональной регуляризации (см. [16]). В этом случае предполагается, что матрица регуляризации A является диагональной в базисе из собственных векторов гессиана логарифма правдоподобия. Тогда можно перейти к новым переменным $\mathbf{u}$, являющимся линейными комбинациями $\mathbf{w}$, таким, что в новом базисе матрицы H, A и $H + A$ станут диагональными. Этот переход существенно упрощает процесс оптимизации обоснованности, так как многомерный интеграл (3.14) переходит в произведение одномерных интегралов, каждый из которых зависит от своего параметра регуляризации α_j. В результате оптимизационный процесс сходится за одну итерацию, так как оптимальные значения коэффициентов регуляризации не зависят друг от друга. Более того, в отличие от MPB и ридж-регрессии, данная процедура инвариантна относительно линейных преобразований базисных функций (т.е. при любом невырожденном линейном преобразовании $\boldsymbol{\phi}(\mathbf{x})$ результат обучения регрессии остается неизменным).

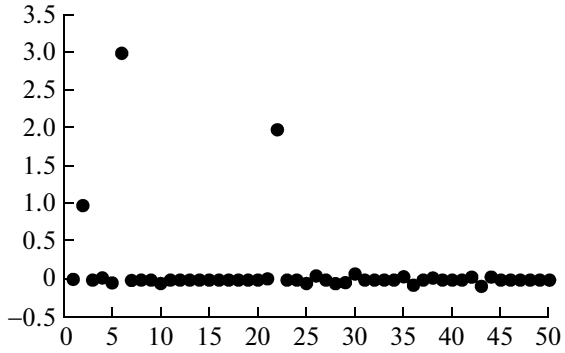
Экспериментальное сравнение классического метода релевантных векторов и предлагаемого в работе метода на основе обобщения информационного критерия Акаике ниже. Показано, что в случае недиагональной регуляризации оба подхода оказываются эквивалентными.

Предполагая, что матрица $H + A$ диагональна и, следовательно, $\mathbf{q} = \mathbf{0}$, получаем

$$\beta_m = \frac{1}{h_{mm} + \alpha_m},$$

$$\xi_m = \begin{pmatrix} \mathbf{0} \\ h_{mm} u_{ML,m} \end{pmatrix},$$

$$\zeta_m^{\mathrm{T}} \xi_m = (h_{mm} u_{ML,m})^2.$$



Фиг. 2.

Оптимальное значение β определяется следующим выражением:

$$\beta_m^* = \frac{(h_{mm}u_{ML,m})^2 - h_{mm}}{h_{mm}^3 u_{ML,m}^2}.$$

Отсюда имеем

$$\alpha_m = \begin{cases} \dfrac{h_{mm}}{h_{mm}u_{ML,m}^2 - 1}, & h_{mm}u_{ML,m}^2 > 1, \\ +\infty, & 0 < h_{mm}u_{ML,m}^2 < 1, \end{cases} \tag{3.15}$$

Случай, соответствующий $\alpha_j = 0$ (см. фиг. 1б), невозможен, так как $h_{mm}u_{ML,m}^2$ всегда неотрицательно. Таким образом, получены те же выражения для α_j, как и в случае оптимизации обоснованности при настройке коэффициентов регуляризации, связанных с собственными векторами гессиана (см. [16]). Следовательно, недиагональная регуляризация с помощью критерия Акаике эквивалентна недиагональной байесовской регуляризации для задачи восстановления регрессии.

4. ЭКСПЕРИМЕНТЫ И ОБСУЖДЕНИЕ

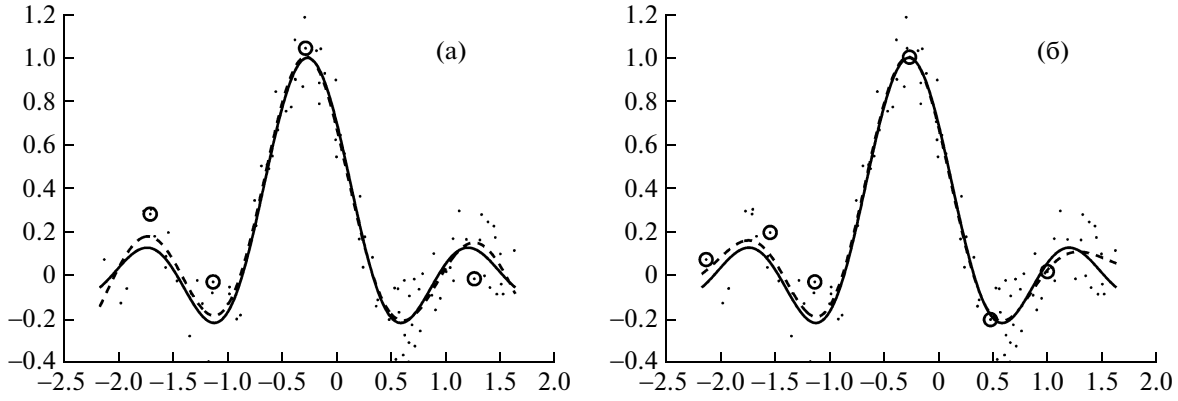
4.1. Отбор признаков

Информационный критерий Акаике широко используется для отбора регрессоров в классической линейной регрессии. Тем не менее, эта задача является чрезвычайно трудоемкой из-за необходимости перебирать всевозможные подмножества регрессоров. Предлагаемый метод (ОИКА) значительно облегчает этот процесс, так как при этом задача дискретной оптимизации сводится к задаче гладкой оптимизации, для которой удается построить эффективную итерационную процедуру.

Рассмотрим модельную задачу регрессии с 49 признаками, имеющими стандартные гауссовские распределения, 100 объектами и значением целевой переменной

$$t = x_2 + 3x_6 + 2x_{22} + \varepsilon,$$

где $\varepsilon \sim \mathcal{N}(\varepsilon|0, 0.5)$. Запустив ОИКА с $\phi_j(\mathbf{x}) = x^j$, получим 15 релевантных признаков, из которых 12 имеют коэффициенты регуляризации $\alpha_j > 100$, среди остальных $\alpha_2 = 0.93$, $\alpha_6 = 0.10$, and $\alpha_{22} = 0$. Соответствующие веса показаны на фиг. 2. Легко видеть, что только три веса значительно отличаются от нуля и близки к истинным значениям.



Фиг. 3.

4.2. Функция Sinc

ОИКА обладает свойством разреженности и может рассматриваться как альтернатива МРВ, где коэффициенты регуляризации подбираются с использованием принципа максимальной обоснованности:

$$\alpha = \arg\max \int p(\mathbf{t}|X, \mathbf{w}) p(\mathbf{w}|\alpha) d\mathbf{w}.$$

Различие между двумя методами можно проследить на модельной задаче – зашумленной функции $\frac{\sin(x)}{x}$ с равномерным шумом на отрезке [–0.2, 0.2] и 100 объектами в выборке. На фиг. 3 представлены регрессии, полученные методами МРВ (график а) и ОИКА (график б). Базисные функции $\phi_j(\mathbf{x}) = \exp\left(-\frac{\|\mathbf{x}_j - \mathbf{x}\|^2}{2\sigma^2}\right), j = 1, 2, ..., n$, где $\mathbf{x}_j$ – объекты обучающей выборки; параметр ширины гауссианы $\sigma = 0.4$. Истинная зависимость показана сплошной линией, пунктирная линия соответствует прогнозируемым значениям, релевантные объекты – кружочки. МРВ и ОИКА выделяют 6 и 4 релевантных объектов соответственно. Из графика видно, что регрессия, полученная с помощью МРВ, в целом больше прижимается к нулю, особенно на концах отрезка. Это можно объяснить тем фактом, что все коэффициенты регуляризации отличны от нуля, поэтому даже релевантные базисные функции подвергаются небольшой регуляризации. С другой стороны, решение ОИКА получается более разреженным. При этом двум из четырех базисных функций соответствуют строго нулевые коэффициенты регуляризации. Таким образом, по сравнению с ОИКА, в МРВ наблюдается эффект “недообучения” (переупрощения модели), который часто отмечается для данного метода (см. [17]).

4.3. Сравнительная оценка

Было проведено сравнение методов МРВ, ОИКА и линейной ридж-регрессии (ЛР) на 11 задачах, взятых из хранилища UCI[2)] и Regression Toolbox by Heikki Hyotyniemi[3)].

Во всех методах были установлены следующие параметры. Количество базисных функций $m = n + 1$, $\phi_j(\mathbf{x}) = \exp\left(-\frac{\|\mathbf{x} - \mathbf{x}_j\|^2}{2\sigma}\right)$ и $\phi_{n+1}(\mathbf{x}) = 1$. Параметр ширины σ подбирался с использованием кросс-валидации на основе пятикратного разбиения обучающей выборки на два подмножества (5x2-fold cross-validation) (см. [18]). Для каждой обучающей выборки СКО также оценивалось с использованием 5x2-fold кросс-валидации.

[2)] http://archive.ics.uci.edu/ml/
[3)] http://www.control.hut.fi/Hyotyniemi/publications/01_ report125/RegrToolbox

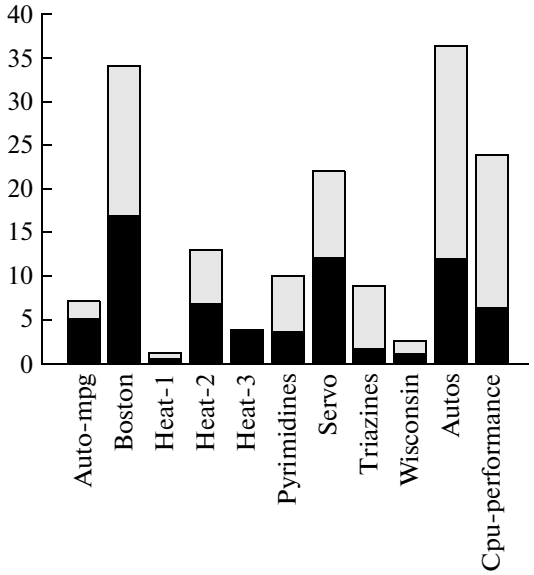
Таблица 1. Корень из среднего квадрата отклонения для различных алгоритмов

Задача	ОИКА	MPB	ЛР
Auto-mpg	2.95 ± 0.06	*2.93 ± 0.04*	**2.92 ± 0.03**
Boston	*3.78 ± 0.22*	**3.75 ± 0.19**	3.86 ± 0.11
HeatExchange-1	*7.90 ± 0.09*	**7.88 ± 0.10**	9.16 ± 0.68
HeatExchange-2	*8.75 ± 0.97*	9.27 ± 1.07	**8.35 ± 1.17**
HeatExchange-3	**0.80 ± 0.07**	*0.82 ± 0.04*	0.84 ± 0.05
Pyrimidines	**0.10 ± 0.01**	**0.10 ± 0.01**	0.11 ± 0.01
Servo	*0.91 ± 0.07*	0.95 ± 0.06	**0.90 ± 0.02**
Triazines	**0.16 ± 0.01**	*0.17 ± 0.00*	*0.17 ± 0.01*
Wisconsin (wdbc)	*25.80 ± 2.31*	*25.27 ± 1.60*	29.18 ± 4.52
Autos	**0.33 ± 0.06**	**0.33 ± 0.02**	0.46 ± 0.04
Cpu-performance	**0.36 ± 0.04**	*0.40 ± 0.04*	0.48 ± 0.20
Ранг	19.00	20.50	26.50
Шрифтовая легенда	**Место1**	*Место2*	Место3

Таблица 2. Разреженность различных алгоритмов (число релевантных объектов)

Задача	ОИКА	MPB	ЛР
Auto-mpg	7.10 ± 4.94	8.60 ± 2.99	199.00 ± 0.00
Boston	34.00 ± 5.24	24.50 ± 3.02	253.00 ± 0.00
HeatExchange-1	1.20 ± 0.45	5.60 ± 5.47	45.00 ± 0.00
HeatExchange-2	13.10 ± 10.17	10.10 ± 4.39	45.00 ± 0.00
HeatExchange-3	3.90 ± 2.86	2.60 ± 0.22	45.00 ± 0.00
Pyrimidines	10.10 ± 2.97	9.80 ± 5.53	37.00 ± 0.00
Servo	22.00 ± 11.80	15.80 ± 3.47	83.50 ± 0.00
Triazines	8.90 ± 5.48	31.30 ± 19.84	93.00 ± 0.00
Wisconsin (wdbc)	2.60 ± 1.39	8.30 ± 4.96	23.50 ± 0.00
Autos	36.30 ± 6.23	23.80 ± 3.27	100.50 ± 0.00
Cpu-performance	23.90 ± 0.89	26.70 ± 3.03	104.50 ± 0.00

Для ридж-регрессии во всех задачах коэффициенты регуляризации были установлены равными 10^{-6}. Для ОИКА и MPB дополнительно вычислялась разреженность (количество ненулевых весов). В табл. 1 и 2 отражены результаты экспериментов. На фиг. 4. проиллюстрировано число



Фиг. 4.

релевантных объектов для ОИКА в различных задачах. Черная часть столбца соответствует числу релевантных объектов с нулевыми коэффициентами регуляризации.

ЗАКЛЮЧЕНИЕ

Заметим, что, как и в МРВ, большинство α_j в ОИКА стремятся к бесконечности, обеспечивая, таким образом, разреженность получаемого решения. Более того, во многих случаях методы дают близкие результаты. Основным выводом данной работы является тот факт, что информационный критерий Акаике может быть использован для проведения процедуры автоматического определения релевантности наравне с байесовскими методами.

В отличие от МРВ, в случае ОИКА некоторые коэффициенты регуляризации становятся тождественно равными нулю. Подход, основанный на использовании ОИКА, перспективен для решения задачи отбора признаков в линейной регрессии, традиционно решаемой с помощью информационного критерия Акаике. Вместо проведения вычислительно сложной процедуры полного перебора при решении дискретной задачи отбора признаков становится возможным переход к непрерывной задаче гладкой оптимизации и использование ОИКА.

Результаты экспериментов позволяют сделать вывод, что байесовское обучение и информационный подход имеют много общего и, возможно, являются двумя косвенными характеристиками одного и того же явления.

Одним из направлений будущих исследований является применение ОИКА к задаче классификации. Одним из возможных путей здесь является сведение задачи классификации к задаче регрессии (см. [15]).

Авторы выражают признательность В.В. Моттлю за ценные замечания и обсуждение работы.

СПИСОК ЛИТЕРАТУРЫ

1. *Tipping M.E.* The relevance vector machine // Advances Neural Information Processing Systems. 2000. V. 12. P. 652–658.
2. *MacKay D.J.C.* The evidence framework applied to classification networks // Neural Comput. 1992. V. 4. P. 720–736.
3. *Tibshirani R.* Regression shrinkage and selection via the lasso // J. Roy. Stat. Soc. 1996. V. 58. P. 267–288.
4. *Figueiredo M.* Adaptive sparseness for supervised learning // IEEE Trans. Pattern Analys. Mach. Intelligence. 2003. V. 25. P. 1150–1159.
5. *Williams P.M.* Bayesian regularization and pruning using a laplace prior // Neural Comput. 1995. V. 7. P. 117–143.
6. *Cawley G.C., Talbot N.L.C., Girolami M.* Sparse multinomial logistic regression via bayesian l1 regularisation // Advances Neural Informat. Processing Systems. 2007. V. 19. P. 209–216
7. *Schwarz G.* Estimating the dimension of a model // Ann. Statistics. 1978. V. 6. P. 461–464.
8. *Bishop C.M.* Pattern recognition and machine learning. New York: Springer, 2006.
9. *Akaike H.* A new look at statistical model identification // IEEE Trans. Automatic Control. 1974. V. 25. P 461–464.
10. *Ширяев А.Н.* Вероятность. М.: Наука, 1979.
11. *Боровков А.А.* Математическая статистика. М.: Физматлит, 2007.
12. *Хорн Р., Джонсон Ч.* Матричный анализ. М.: Мир, 1989.
13. *Spiegelhalter D., Best N., Carlin B., van der Linde A.* Bayesian measures of model complexity and fit // J. Roy. Statist. Soc. 2002. V. 64. P. 583–640.
14. *Faul A.C., Tipping M.E.* Analysis of sparse bayesian learning // Advances Neural Informat. Processing Systems. 2002. V. 14. P. 383–389.
15. *Tipping M.E.* Sparse bayesian learning and the relevance vector machines // J. Mach. Learning Res. 2001. V. 1. P. 211–244.
16. *Kropotov D.A., Vetrov D.P.* On one method of non-diagonal regularization in sparse bayesian learning // Proc. 24th Internat. Conf. Mach. Learning. Corvalis: Omnipress, 2007. P. 457–464.
17. *Qi Y., Minka T., Picard R., Ghahramani Z.* Predictive automatic relevance determination by expectation propagation // Proc. 21st Internat. Conf. Mach. Learning. Banff: Omnipress, 2004. P. 671–678.
18. *Dietterich T.G.* Approximate statistical tests for comparing supervised classification learning algorithms // Neural Comput. 1998. V. 10. P. 1895–1924.