Tokmakova A.A., Strijov V.V. Estimation of linear model hyperparameters for noisy or correlated feature selection problem // Informatics and applications, 2012, 6(4) : 66-75.

УДК 519.584

ОЦЕНИВАНИЕ ГИПЕРПАРАМЕТРОВ ЛИНЕЙНЫХ РЕГРЕССИОННЫХ МОДЕЛЕЙ ПРИ ОТБОРЕ ШУМОВЫХ И КОРРЕЛИРУЮЩИХ ПРИЗНАКОВ[1]

А. А. ТОКМАКОВА[2], В. В. СТРИЖОВ[3]

А. А. ТОКМАКОВА (р. 1991) — студентка Московского физико-технического института, 119991, ГСП-1, Москва, Вавилова д. 42, оф. 151
Тел. служебный: 8 (495) 135-4163
Факс: 8 (495) 137-2848
E-mail: aleksandra-tok@yandex.ru

В. В. СТРИЖОВ (р. 1967) — кандидат физико-математических наук, научный сотрудник Отдела вычислительных методов прогнозирования Вычислительного Центра РАН, доцент кафедры Интеллектуального анализа данных Московского Физико-технического института.
119991, ГСП-1, Москва, Вавилова д. 42, оф. 151
Тел. служебный: 8 (495) 135-4163
Факс: 8 (495) 137-2848
E-mail: strijov@ccas.ru

Аннотация

В работе решается задача отбора признаков при восстановлении линейной регрессии. Принята гипотеза о нормальном распределении вектора зависимой переменной и параметров модели. Для оценки ковариационной матрицы параметров используется аппроксимация Лапласа: логарифм функции ошибки приближается функцией плотности нормального распределения. Исследуется проблема присутствия в выборке шумовых и коррелирующих признаков, так как при их наличии матрица ковариаций параметров модели становится вырожденной. Предлагается алгоритм, производящий отбор информативных признаков. В вычислительном эксперименте приводятся результаты исследования на временном ряде.

Ключевые слова: *байесовский вывод, ковариационная матрица, гиперпараметры модели, отбор признаков, регрессия.*

1 Введение

Часто при анализе временных рядов требуется рассмотрение большого количества признаков. В связи с этим возникают проблемы, связанные с наличием в выборке большого количества мультикоррелирующих признаков или с высокой зашумлённостью выборки. В работе выдвинута гипотеза о нормальном распределении вектора зависимой переменной и вектора параметров модели [1,2]. Необходимо оценить ковариационные матрицы этих распределений и установить связь между пространством данных и пространством параметров, что позволит произвести отбор шумовых и коррелирующих признаков.

Развитие методов отбора признаков имеет богатую историю. Начиная с 1960г., активно развивались шаговые методы (Stepwise Regression) [3]. Главная идея этих методов состоит в отборе признаков, вносящих наибольший вклад в зависимую переменную. Вводится критерий, на основании которого алгоритм добавляет или удаляет признаки. Широкое применение получили частные случаи шаговой регрессии — алгоритмы LARS(Least Angle Regression) [4] и LASSO(Least Absolute Shrinkage and Selection Operator) [5]. Алгоритм LARS заключается в последовательном добавлении признаков. На каждом шаге веса признаков меняются

[1]Работа выполнена при поддержке РФФИ, грант №10-07-00422.
[2]Московский физико-технический институт, aleksandra-tok@yandex.ru
[3]Вычислительный центр РАН, strijov@ccas.ru

таким образом, чтобы доставить наибольшую корреляцию восстановленного вектора зависимых переменных с вектором регрессионных остатков. Алгоритм позволяет сократить количество свободных переменных и избежать проблемы неустойчивой оценки весов. Метод LASSO вводит ограничения на норму вектора коэффициентов модели, что приводит к обращению в ноль некоторых коэффициентов модели. Метод приводит к повышению устойчивости модели и позволяет отбирать признаки, оказывающие наибольшее влияние на вектор ответов.

Одной из причин возникновения задачи отбора признаков является их мультиколлинеарность. Первые шаги по решению этой проблемы были сделаны А. И. Тихоновым в 1963г., который ввел понятие регуляризации — дополнительного ограничения на задачу [6]. В работе [7] введено понятие регуляризации и описан общий метод решения задач. Так как работы А. И. Тихонова были опубликованы на западе только лишь в 1977г., в 1970г. A. E. Hoerl и R. W. Kennard предложили метод гребневой регрессии [8]. В минимизируемую функцию вводилось дополнительное слагаемое, что повышало устойчивость решения [9], однако не позволяло производить отбор признаков. Позднее стали появляться методы, использующие качественно иной подход для решения проблемы мультиколлинеарности. Например, Belsley предложил метод для удаления признаков [10], использующий сингулярное разложение матрицы плана. Алгоритм находит коэффициент, характеризующий степень зависимости признаков друг от друга. Позднее появился метод фактора инфляции дисперсии (Variance Inflation Factor) [11], оценивающий увеличение дисперсии заданного коэффициента регрессии, что свидетельствует о высокой корреляции данных.

В данной работе для отбора признаков линейной регрессионной модели предлагается выполнить анализ пространства параметров. Вектор параметров рассматривается как многомерная случайная величина. Оценивается наиболее вероятное значение параметров. При оценке используется связный байесовский вывод [12,13].

Основываясь на гипотезе о нормальном распределении параметров модели [1], оценивается ковариационная матрица распределения параметров [2, 14]. На её главной диагонали стоят дисперсии случайных величин, что позволяет установить степень значимости данного конкретного параметра в модели. При таком подходе к отбору признаков не возникает необходимости разбиения выборки на обучение и контроль. Для оценки ковариационной матрицы в работе используется аппроксимация Лапласа [15]. Логарифм функции ошибки приближается функцией плотности нормального распределения, и появляется возможность вычисления нормировочной константы.

2 Постановка задачи

Дана регрессионная выборка: $D = \{\mathbf{x}_i, y_i\}_{i=1}^m = (X, \mathbf{y})$, где $\mathbf{x}_i \in \mathbb{R}^n, i = 1, \ldots, m$ — векторы независимой переменной, а $y_i \in \mathbb{R}, i = 1, \ldots, m$ — значения зависимой переменной. Решается задача восстановления регрессии

$$\mathbf{y} = \mathbf{f}(\mathbf{w}, X), \tag{1}$$

где $\mathbf{f}(\mathbf{w}, X)$ — некоторая параметрическая вектор-функция. Пусть многомерная случайная величина $\mathbf{y}$ имеет нормальное распределение:

$$\mathbf{y} \sim \mathcal{N}(\mathbf{f}, \sigma^2 I_m),$$

где $\mathbf{f}$ — вектор-функция, σ^2 — дисперсия распределения, I_m — единичная матрица размерности m.

Требуется приблизить функцию $\mathbf{f}(\mathbf{w}, X)$ параметрической функцией $\widehat{\mathbf{f}}(X, \mathbf{w})$ из заданного класса $\mathcal{F}$ (линейные функции), причем $|\mathcal{F}|$ конечно. Отображение $\mathbf{f} : \mathbb{R}^m \times \mathbb{W}^n \to \mathbb{R}^m$ будем называть моделью. Здесь $\mathbb{R}^m$ — пространство данных, а $\mathbb{W}^n \subseteq \mathbb{R}^n$ — пространство параметров. В задаче линейной регрессии задача приближения функции $\mathbf{f}(\mathbf{w}, X)$ эквивалентна задаче отбора признаков. В данном случае модель определяется параметрами, которые соответствуют множеству индексов активных признаков $\mathcal{A} \subseteq \mathcal{J} = \{1, 2, ..., n\}$. Таким образом, при выборе модели требуется найти такое множество индексов $\mathcal{A}^*$, которое бы доставляло минимум функции:

$$\mathcal{A}^* = \arg\min_{\mathcal{A} \subseteq \mathcal{J}} S(\mathbf{f}_\mathcal{A} | \mathbf{w}^*, D),$$

где $S(\mathbf{f}|\mathbf{w}, D)$ — функция ошибки, $\mathbf{f}_\mathcal{A}$ — параметрическая вектор-функция, вычисляемая только на множестве активных признаков, заданном индексами из множества $\mathcal{A}$. При этом параметры $\mathbf{w}^*$ модели должны доставлять минимум функции:

$$\mathbf{w}^* = \arg\min_{\mathbf{w}\in\mathbb{W}} S(\mathbf{w}|\mathbf{f}_\mathcal{A}, D).$$

3 Вид функции ошибки

Пользуясь предположением о том, что вектор зависимой переменной — многомерная случайная величина с нормальным распределением, запишем конкретный вид функции ошибки $S(\mathbf{w})$ для поставленной задачи.

Пусть многомерная случайная величина $\mathbf{y}$ имеет нормальное распределение. Обозначим $\beta^{-1} = \sigma^2$. Тогда распределение зависимой переменной $\mathbf{y}$ можно представить в виде:

$$p(\mathbf{y}) = (2\pi\beta^{-1})^{-\frac{m}{2}} \exp\left(-\frac{1}{2}(\mathbf{y}-\mathbf{f})^T\beta I(\mathbf{y}-\mathbf{f})\right). \tag{2}$$

Рассмотрим функцию правдоподобия данных, которая имеет вид

$$p(\mathbf{y}|X, \mathbf{w}, \beta, \mathbf{f}) \stackrel{\text{def}}{=} p(D|\mathbf{w}, \beta, \mathbf{f}) = \frac{\exp(-\beta E_D)}{Z_D(\beta)}. \tag{3}$$

Здесь E_D — функция ошибки. Из выражений (2) и (3), определим её как:

$$E_D = \frac{1}{2}(\mathbf{y}-\mathbf{f})^T(\mathbf{y}-\mathbf{f}).$$

Коэффициент Z_D нормирует плотность нормального распределения и равен

$$Z_D(\beta) = (2\pi\beta^{-1})^{\frac{m}{2}}. \tag{4}$$

Рассмотрим равенство (1). Слева стоит многомерная случайная величина $\mathbf{y}$, имеющая нормальное распределение. Матрица X не является случайной величиной, поэтому предположим, что $\mathbf{w} \in \mathbb{W}^n$ также является многомерной случайной величиной с нормальным распределением. Параметрами этого распределения будут математическое ожидание $\mathbf{w}_0$ и матрица ковариаций A^{-1}:

$$p(\mathbf{w}|A, \mathbf{f}) = \frac{\exp(-E_\mathbf{w})}{Z_\mathbf{w}(A)}. \tag{5}$$

Определим функцию-штраф за большое значение параметров модели для принятого распределения как $E_\mathbf{w} = \frac{1}{2}(\mathbf{w}-\mathbf{w}_0)^T A(\mathbf{w}-\mathbf{w}_0)$. Нормирующая константа $Z_\mathbf{w}$ в этом случае равна:

$$Z_\mathbf{w}(A) = (2\pi)^{\frac{n}{2}}|A^{-1}|^{\frac{1}{2}}. \tag{6}$$

Апостериорное распределение параметров модели для заданных A и β имеет вид:

$$p(\mathbf{w}|D, A, \beta, \mathbf{f}) = \frac{p(D|\mathbf{w}, \beta, \mathbf{f})p(\mathbf{w}|A, \mathbf{f})}{p(D|A, \beta, \mathbf{f})}, \tag{7}$$

$$\frac{p(D|\mathbf{w}, \beta, \mathbf{f})p(\mathbf{w}|A, \mathbf{f})}{p(D|A, \beta, \mathbf{f})} = \frac{\exp(-\beta E_D)\exp(-E_\mathbf{w})}{Z_D(\beta)Z_\mathbf{w}(A)} = \frac{\exp\big(-(\beta E_D + E_\mathbf{w})\big)}{Z_D(\beta)Z_\mathbf{w}(A)}. \tag{8}$$

В выражении (7) приняты следующие обозначения:

$p(\mathbf{w}|D, A, \beta, \mathbf{f})$ — апостериорное распределение параметров;

$p(D|\mathbf{w}, \beta, \mathbf{f})$ — функция правдоподобия данных;

$p(\mathbf{w}|A, \mathbf{f})$ — априорное распределение параметров;

$p(D|A,\beta,\mathbf{f})$ — функция правдоподобия модели.

Записывая функцию ошибки как

$$S = E_{\mathbf{w}} + \beta E_D = \frac{1}{2}(\mathbf{w}-\mathbf{w}_0)^T A(\mathbf{w}-\mathbf{w}_0) + \frac{1}{2}(\mathbf{y}-\mathbf{f})^T \beta I(\mathbf{y}-\mathbf{f}), \tag{9}$$

получим следующее выражение для апостериорного распределения параметров:

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

где $Z_S = Z_S(A,\beta)$ — нормирующий коэффициент. Оценка нормировочного коэффициента производится с помощью аппроксимации Лапласа.

4 Аппроксимация Лапласа

Аппроксимация Лапласа позволяет оценить нормировочный коэффициент для ненормированной плотности вероятности. Пусть задано ненормированное распределение $p^*(\mathbf{w})$. Требуется найти нормировочную константу:

$$Z = \int p^*(\mathbf{w}) d\mathbf{w},$$

при которой распределение $p(\mathbf{w}) = Z^{-1}p^*(\mathbf{w})$. Предположим, что $p^*(\mathbf{w})$ имеет максимум в точке $\mathbf{w}_0$, то есть

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

Прологарифмируем и разложим $p^*(\mathbf{w})$ по Тейлору в окрестности $\mathbf{w}_0$:

$$\ln p^*(\mathbf{w}) = \ln p^*(\mathbf{w}_0) + 0 - \frac{1}{2}(\mathbf{w}-\mathbf{w}_0)^T A(\mathbf{w}-\mathbf{w}_0) + \ldots, \tag{10}$$

где матрица Гессе $A = [\alpha_{ij}]$ определена как:

$$\alpha_{ij} = -\left.\frac{\partial \ln p^*(\mathbf{w})}{\partial w_i \partial w_j}\right|_{\mathbf{w}=\mathbf{w}_0},$$

то есть $A = -\nabla\nabla \ln p^*(\mathbf{w})|_{\mathbf{w}=\mathbf{w}_0}$, где ∇ — градиент функции.

Отбрасывая все члены выше квадратичного в разложении и беря экспоненту обеих частей выражения (10), получим:

$$p^*(\mathbf{w}) \approx p^*(\mathbf{w}_0)\exp\left(-\frac{1}{2}(\mathbf{w}-\mathbf{w}_0)^T A(\mathbf{w}-\mathbf{w}_0)\right).$$

Тогда нормальное распределение $\widehat{p}(\mathbf{w})$, приближающее нормированное распределение $p(\mathbf{w})$ имеет вид

$$\widehat{p}(\mathbf{w}) = \mathcal{N}(\mathbf{w}_0, A^{-1}) = \frac{1}{(2\pi)^{\frac{n}{2}}|A^{-1}|^{\frac{1}{2}}}\exp\left(-\frac{1}{2}(\mathbf{w}-\mathbf{w}_0)^T A(\mathbf{w}-\mathbf{w}_0)\right).$$

Следовательно, нормировочная константа имеет вид

$$Z = p^*(\mathbf{w}_0)\frac{(2\pi)^{\frac{n}{2}}}{|A|^{\frac{1}{2}}}. \tag{11}$$

5 Оценка ковариационных матриц

Анализируя функцию ошибки $S(\mathbf{w})$, построим алгоритм, позволяющий выявлять шумовые и коррелирующие признаки.

Пусть нам известен локальный минимум $S(\mathbf{w})$, и он находится в точке $\mathbf{w}_0$. Рассмотрим матрицу Гессе функции ошибок $H = -\nabla\nabla S(\mathbf{w})|_{\mathbf{w}=\mathbf{w}_0}$, где ∇ — градиент функции. При появлении в выборке шумовых или коррелирующих признаков, происходит резкое возрастание некоторых элементов матрицы H. Необходимо установить связь между компонентами матрицы Гессе и ковариационной матрицей параметров, для того чтобы произвести отбор активных параметров $\mathcal{A}$ и повысить устойчивость решения.

Рассмотрим ряд Тейлора второго порядка логарифма числителя (7):

$$-S(\mathbf{w}) \approx -S(\mathbf{w}_0) - \frac{1}{2}\Delta\mathbf{w}^T H \Delta\mathbf{w}, \tag{12}$$

где $\Delta\mathbf{w} = \mathbf{w} - \mathbf{w}_0$. В выражении (12) нет слагаемого первого порядка, так как предполагается, что точка $\mathbf{w}_0$ доставляет локальный минимум функции $S(\mathbf{w})$. Следовательно:

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

Применяя экспоненту к обеим частям выражения (12) получим необходимое приближение:

$$\exp\big(-S(\mathbf{w})\big) \approx \exp\big(-S(\mathbf{w}_0)\big) \exp\left(-\frac{1}{2}\Delta\mathbf{w}^T H \Delta\mathbf{w}\right). \tag{13}$$

При полученном приближении выражение (13) будет выглядеть следующим образом:

$$p(\mathbf{w}|D, A, \beta) \approx \frac{\exp\big(-S(\mathbf{w}_0)\big) \exp\left(-\frac{1}{2}\Delta\mathbf{w}^T H \Delta\mathbf{w}\right)}{Z_S(A, \beta)}, \tag{14}$$

где $Z_S(A, \beta)$ выступает в роли нормировочного коэффициента плотности вероятностного распределения. Оценка для коэффициента Z_S получена с помощью аппроксимации Лапласа (пояснения см. в главе 4):

$$Z_S = \frac{\exp\big(-S(\mathbf{w}_0)\big)(2\pi)^{\frac{n}{2}}}{|H|^{\frac{1}{2}}}. \tag{15}$$

Подставив (15) в (14), получим оценку правдоподобия модели, на основании которой будем производить отбор оптимальных гиперпараметров модели

$$p(\mathbf{w}|D, A, \beta) = \frac{|H|^{\frac{1}{2}} \exp\left(-\frac{1}{2}\Delta\mathbf{w}^T H \Delta\mathbf{w}\right)}{(2\pi)^{\frac{n}{2}}}. \tag{16}$$

Выражение (14) определяет выбор наиболее правдоподобной модели. Для нахождения гиперпараметров воспользуемся принципом максимума правдоподобия $p(D|A, \beta)$ относительно A и β. Запишем $p(D|A, \beta)$ в следующем виде:

$$p(D|A, \beta) = \int p(D|\mathbf{w}, A, \beta) p(\mathbf{w}|A) d\mathbf{w}. \tag{17}$$

Используя выражения (5) и (3) перепишем функцию правдоподобия в виде:

$$p(D|A, \beta) = \frac{1}{Z_{\mathbf{w}}(A)} \frac{1}{Z_D(\beta)} \int \exp\big(-S(\mathbf{w})\big) d\mathbf{w}. \tag{18}$$

Из соображений нормировки интеграл выражения (7) равен единице, то есть:

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

Следовательно интеграл в правой части (18) в точности равен Z_S. Поэтому:

$$p(D|A,\beta) = \frac{1}{Z_{\mathbf{w}}(A)} \frac{1}{Z_D(\beta)} \exp\bigl(-S(\mathbf{w}_0)\bigr)(2\pi)^{\frac{n}{2}} |H|^{-\frac{1}{2}}. \tag{19}$$

Подставив значение $Z_{\mathbf{w}}$ из (6) и Z_D из (4) в (19), получим:

$$p(D|A,\beta) = (2\pi)^{-\frac{n}{2}} |A^{-1}|^{-\frac{1}{2}} (2\pi)^{-\frac{m}{2}} (\beta^{-1})^{\frac{m}{2}} \exp\bigl(-S(\mathbf{w}_0)\bigr)(2\pi)^{\frac{n}{2}} |H|^{-\frac{1}{2}}. \tag{20}$$

Получим оценку логарифма правдоподобия:

$$\ln p(D|A,\beta,\mathbf{f}) = -\frac{1}{2}\ln|A^{-1}| - \frac{m}{2}\ln 2\pi + \frac{m}{2}\ln\beta^{-1} - S(\mathbf{w}_0) - \frac{1}{2}\ln|H|. \tag{21}$$

Поочерёдно приравнивая частные производные по A и β выражения (21) к нулю, найдём максимум (21) по гиперпараметрам.

Пусть матрица A диагональна. Введем обозначение $\boldsymbol{\alpha} = [\alpha_1, ..., \alpha_n]^T$ для вектора, состоящего из элементов диагонали матрицы A. Представим гессиан в виде:

$$H = -\nabla\nabla S(\mathbf{w}) = -\nabla\nabla(\beta E_D + E_{\mathbf{w}}) = -\beta\nabla\nabla E_D - \nabla\nabla E_{\mathbf{w}} = H_D + H_{\mathbf{w}},$$

где H_D зависит от β, а $H_{\mathbf{w}}$ зависит от A. Так как $\nabla\nabla E_{w_i} = \frac{\partial^2}{\partial w_i}\left(\frac{1}{2}\alpha_i(w_i - w_{0i})^2\right) = \alpha_i$, то часть гессиана $H_{\mathbf{w}}$ диагональна. Покажем, что при некоторых допущениях H_D также будет диагональной матрицей. Для этого рассмотрим два случая:

1) если все признаки независимы, то матрица H_D будет диагональной, так как недиагональные элементы матрицы Гессе отражают степень зависимости измеряемых величин;

2) при наличии в выборке шумовых или коррелирующих признаков будет наблюдаться возрастание диагональных элементов матрицы (дисперсий признаков), в сравнении с которыми недиагональными элементами можно пренебречь. Таким образом получим, что и в этом случае матрицу H_D можно считать диагональной (на диагонали собственные числа).

Таким образом представим H_D в следующем виде: $H_D = \mathrm{diag}(h_1, ..., h_n)$. Для выявления связи между параметрами и гиперпараметрами модели рассмотрим выражение (21). Воспользуемся необходимым условием минимума и приравняем к нулю первые производные выражения (21) по α_i:

$$\frac{1}{\alpha_i} - (w_i - w_0)^2 - \frac{1}{\beta h_i + \alpha_i} = 0. \tag{22}$$

Данное уравнение имеет два корня. Однако один из них не имеет смысла, так как A^{-1} — диагональная ковариационная (положительно определённая) матрица, следовательно по критерию Сильвестра (симметричная квадратная матрица является положительно определенной тогда и только тогда, когда все её главные миноры положительны) не имеет отрицательных компонент:

$$\alpha_i = \frac{1}{2}\lambda_i\left(\sqrt{1 + \frac{4}{(w_i - w_0)^2\lambda_i}} - 1\right), \tag{23}$$

где $\lambda_i = \beta h_i$.

Приравняв производную по β выражения (21), найдём оптимальное значение β:

$$\frac{m}{2\beta} - E_D - \frac{1}{2\beta}\gamma = 0,$$

где

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

Таким образом

$$\beta = \frac{m - \gamma}{2E_D}. \tag{24}$$

Выражения (23) и (24) не позволяют явно вычислить значения $\boldsymbol{\alpha}$ и β. Поэтому итерационный процесс организуется следующим образом. На каждом шаге вычисляем $\mathbf{w}$ (минимизируя функцию ошибки из выражения (9)), далее, используя полученное приближение, находим вектор гиперпараметров $\boldsymbol{\alpha}$, затем значение гиперпараметра β. Процедура продолжается до сходимости как параметров, так и гиперпараметров, то есть до сходимости функции правдоподобия модели $p(D|A, \beta, \mathbf{w})$.

При появлении шумовых или коррелирующих признаков происходит возрастание диагональных элементов (большое значение дисперсии свидетельствует о неинформативности признака). В следствие этого недиагональные элементы становятся настолько малы, что можно считать матрицу H_D диагональной. Поэтому необходимо принудительно занижать возрастающие диагональные элементы, тем самым производя отсеивание шумовых и коррелирующих признаков.

6 Псевдокод алгоритма оценки гиперпараметров регрессионной модели

Вход: вектор зависимой переменной $\mathbf{y}$, модель $\mathrm{mdl}(\mathbf{w}, X)$
$\mathbf{w}_0 = 0$;
$\mathbf{w} = 0$;
$A = \mathrm{diag}(n, 1)$;
$\beta = 1$;
для $k = 2, \ldots, MaxIterations$
вычислить $A, \beta, \mathbf{w}$:
$\mathbf{w} = \mathbf{FindParameters}(S(\mathbf{w}), A, \beta, \mathbf{w}, \mathbf{w}_0, \mathbf{y})$;
для $j = 2, \ldots, MaxIterations$
добиться сходимости A и β при данном векторе $\mathbf{w}$:
$H = \mathbf{CalcHessian}(S(\mathbf{w}), A, \beta, \mathbf{w}, \mathbf{w}_0, \mathbf{y})$;
если $\frac{\max(H)}{\min(H)} > 10^6$ **то**
$idx = \mathrm{find}(\max(H))$; {индекс строки/столбца (диагональный элемент) с max элементом}
занулить строку и столбец Гессиана, содержащие максимальный элемент;
выход;
$\lambda = \beta * \mathrm{diag}(H)$;
$A = \frac{1}{2}\lambda(\sqrt{1 + \frac{4}{(\mathbf{w}-\mathbf{w}_0)^2\lambda}} - 1)$;
если $idx \neq 0$ **то**
занулить соответствующие диагональные элементы матрицы A (необходимо для сходимости гиперпараметра α);
выход;
$\gamma = \sum \frac{\lambda_j}{\lambda_j + \alpha_j}$;
$\mathbf{f} = \mathrm{mdl}(\mathbf{w}, X)$;
$E_D = \frac{1}{2}(\mathbf{y} - \mathbf{f})^T(\mathbf{y} - \mathbf{f})$;
$\beta = \frac{(m-\gamma)}{2E_D}$;
если $\sum(\alpha_k - \alpha_{k-1})^2 < Convergency$ and $(\beta_k - \beta_{k-1})^2 < Convergency$; **то**
закончить выполнение цикла на текущей итерации;
выход;
если $j = MaxIterations$ **то**
вывести сообщение о величине ошибки и закончить выполнение программы;
выход;
если $\sum(w_k - w_{k-1})^2 < Convergency$ **то**
закончить выполнение программы;
конец

ПРОЦЕДУРА **FindParameters**$(S(\mathbf{w}), A, \beta, \mathbf{w}, \mathbf{w}_0, \mathbf{y})$
пока не найден минимум функции $S(\mathbf{w})$ по $\mathbf{w}$
  $\mathbf{f} = \text{mdl}(\mathbf{w}, X)$;
  $S(\mathbf{w}) = \frac{1}{2}(\mathbf{w} - \mathbf{w}_0)^T A(\mathbf{w} - \mathbf{w}_0) + \frac{1}{2}(\mathbf{y} - \mathbf{f})^T \beta I(\mathbf{y} - \mathbf{f})$;
выход;
вернуть $\mathbf{w}$;

ПРОЦЕДУРА **CalcHessian**$(S(\mathbf{w}), A, \beta, \mathbf{w}, \mathbf{w}_0, \mathbf{y})$
$h = 10^{-6}$; {шаг разностной схемы}
для $i = 1, ..., l$
  для $j = 1, ..., l$
    посчитать элемент матрицы Гессе:
    $\mathbf{e}_i = 0$; {вектор приращения}
    $e_i(i) = 1$;
    $\mathbf{e}_j = 0$;
    $e_j(j) = 1$;
    $H(i,j) = \frac{S(\mathbf{w}+(\mathbf{e}_i+\mathbf{e}_j)h) - S(\mathbf{w}+\mathbf{e}_i h) - S(\mathbf{w}+\mathbf{e}_j h) + S(\mathbf{w})}{h^2}$;
  выход;
выход;
вернуть H;

7 Алгоритмы отбора признаков

Для того, чтобы подчеркнуть особенности описанного в работе алгоритма, приведем примеры ранее предложенных методов регуляризации, приводящих к повышению устойчивости решения и отбору признаков в задаче линейной регрессии [5, 8].

7.1 Гребневая регрессия

Запишем функцию ошибки для линейной модели вида (1):

$$Q(\mathbf{w}) = ||X\mathbf{w} - \mathbf{y}||^2.$$

Для минимизации функции воспользуемся необходимым условием минимума:

$$\frac{\partial Q}{\partial \mathbf{w}} = 2X^T(X\mathbf{w} - \mathbf{y}) = 0,$$

откуда следует, что $X^T X\mathbf{w} = X^T\mathbf{y}$. Если матрица $X^T X$ невырождена, то решением системы является вектор:

$$\mathbf{w}^* = (X^T X)^{-1} X^T \mathbf{y}.$$

Если ковариационная матрица $X^T X$ имеет неполный ранг, то её обращение невозможно. Также выделяют случай мультиколлинеарности: матрица $X^T X$ имеет полный ранг, но близка к некоторой матрице неполного ранга. В этом случае увеличивается разброс коэффициентов $\mathbf{w}^*$, появляются большие по абсолютной величине коэффициенты. Решение становится неустойчивым (небольшие изменения матрицы X ведут к большим изменениям величины $\mathbf{w}^*$).

Для решения проблемы мультиколлинеарности к функционалу Q добавляют регуляризатор, штрафующий большие значения нормы вектора $\mathbf{w}$: $Q_\tau = ||X\mathbf{w} - \mathbf{y}||^2 + \tau||\mathbf{w}||^2$. Решением полученной задачи является вектор:

$$\mathbf{w}^* = (X^T X + \tau I_m)^{-1} X^T \mathbf{y}.$$

Увеличение τ приводит к уменьшению нормы вектора $\mathbf{w}$, однако при этом ни один из параметров не обращается в ноль. То есть повышая устойчивость модели, гребневая регрессия не производит отбор признаков.

7.2 Лассо Тибширани

В данном методе вместо добавления штрафного слагаемого к функционалу качества вводится ограничение-неравенство, запрещающее большие абсолютные значения коэффициентов:

$$\begin{cases} Q(\mathbf{w}) = ||X\mathbf{w} - \mathbf{y}||^2 \to \min\limits_{\mathbf{w}\in\mathbb{W}}, \\ \sum\limits_{j=1}^{W} |w_j| < \theta. \end{cases}$$

Чем меньше значение θ, тем больше коэффициентов w_j обнуляется, таким образом происходит исключение j-го признака. Недостатком этого метода относительно алгоритма, представленного в работе, является необходимость в разделении выборки на две части: для обучения и контроля.

Также при использовании методов регуляризации возникает проблема выбора константы регуляризации. Для её вычисления обычно используют скользящий контроль, что значительно повышает трудоёмкость всей задачи в целом.

8 Вычислительный эксперимент

Результатом вычислительного эксперимента является отбор шумовых и коррелирующих признаков. Тестирование алгоритма производится на временном ряде продаж нарезного хлеба в зависимости от времени. Ряд содержит 195 записей. Модель, аппроксимирующая ряд: $\mathbf{y} = 0.2256 + 0.1996\boldsymbol{\xi} + 0.0496\sin(10\boldsymbol{\xi})$, где $\boldsymbol{\xi} \in \mathbb{R}^n$ — регрессионная выборка. Введем следующие обозначения: $\boldsymbol{\xi}^0$, $\boldsymbol{\xi}^1$ — значение каждого элемента выборки в нулевой и первой степени соответственно, $\sin(10\boldsymbol{\xi})$ — поэлементное применение элементарной функции к вектору $\boldsymbol{\xi}$. На рис. 1 представлена выборка и аппроксимирующая её модель.

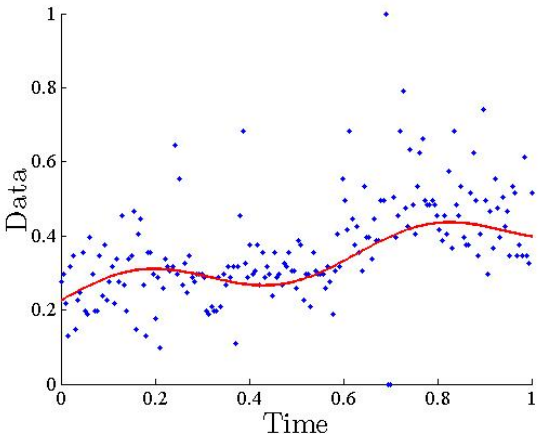

Рис. 1: Данные и аппроксимирующая модель

Пусть матрица плана X представлена в следующем виде $X = [\boldsymbol{\chi}_1, ..., \boldsymbol{\chi}_n]$, где $\boldsymbol{\chi} \in \mathbb{R}^m$. В данном случае она состоит из трёх столбцов: $\boldsymbol{\xi}^0$, $\boldsymbol{\xi}^1$, $\sin(10\boldsymbol{\xi})$.

8.1 Отбор шумовых признаков

Шумовая выборка сформирована при помощи добавления столбца случайных чисел с нормальным распределением. Модель, аппроксимирующая данные в эксперименте: $\mathbf{y} = w_1\boldsymbol{\chi}_1 + w_2\boldsymbol{\chi}_2 + w_3\boldsymbol{\chi}_3 + w_4\boldsymbol{\chi}_4$, где $\boldsymbol{\chi}_1 = \boldsymbol{\xi}^0$, $\boldsymbol{\chi}_2 \sim \mathcal{N}(0, 2)$, $\boldsymbol{\chi}_3 = \boldsymbol{\xi}^1$, $\boldsymbol{\chi}_4 = \sin(10\boldsymbol{\xi})$. При наличии в выборке шумового элемента процедура сходится за восемь итераций. Ниже на рис. 2 проиллюстрированы изменения матрицы Гессе H на каждом шаге процедуры.

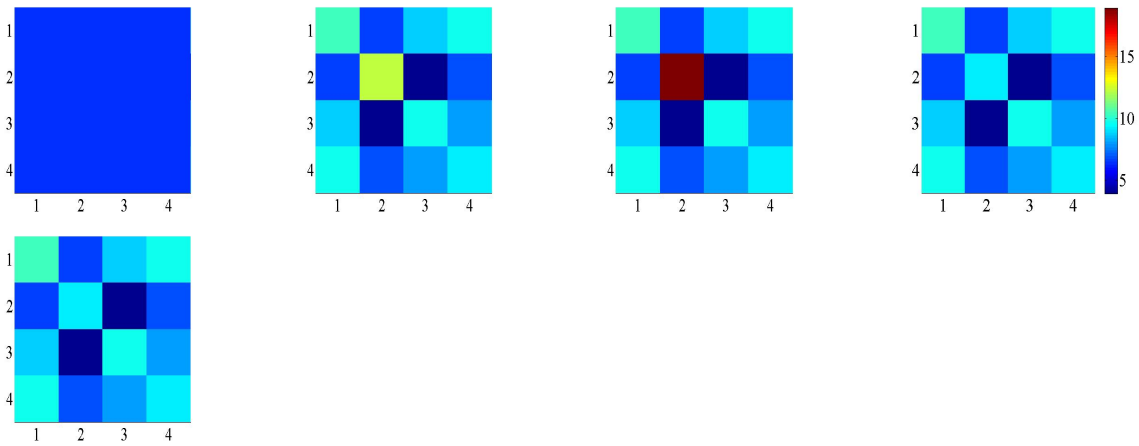

Рис. 2: Итерационный процесс для матрицы Гессе (случай шумового параметра)

На 2-ой итерации наблюдается резкое отличие диагонального элемента $(2,2)$. В течение итераций 2 и 3 он продолжает возрастать, пока не достигает критической относительной величины (принята эмпирическая оценка отношения максимального элемента матрицы к минимальному 10^6). Далее на 4-ой итерации выполняется его зануление. Таким образом происходит выявление шумового признака.

На рис. 3 и 4 представлены диагональные элементы матрицы A. Первый график иллюстрирует изменения второго диагонального элемента α_2, который соответствует шумовому параметру модели. Резкий скачок объясняется тем, что на данной итерации алгоритм находится вблизи локального минимума $\mathbf{w}_0$, и, несмотря на возрастание диагональных элементов матрицы H, знаменатель формулы (23) мал. Далее происходит зануление элементов матрице Гессе, и соответствующий гиперпараметр α становится равным нулю.

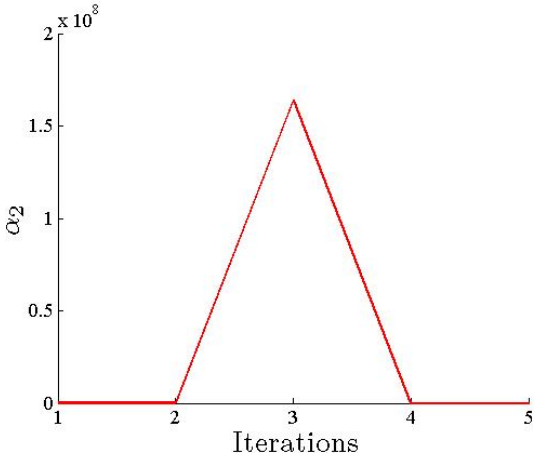

Рис. 3: Элемент матрицы A, соответствующий шумовому параметру модели

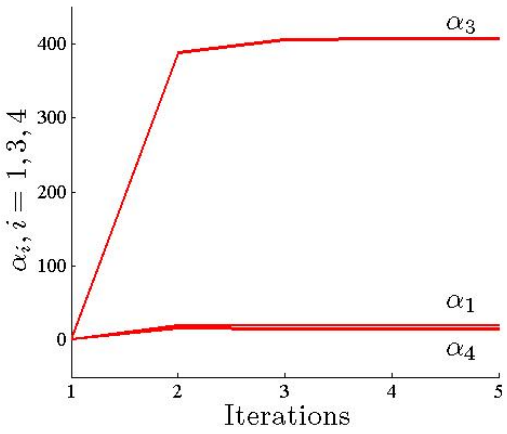

Рис. 4: Элементы матрицы A, соответствующие нешумовым параметрам модели

На графиках 5 и 6 представлен скалярный гиперпараметр β и процесс изменения параметров модели w_i соответственно.

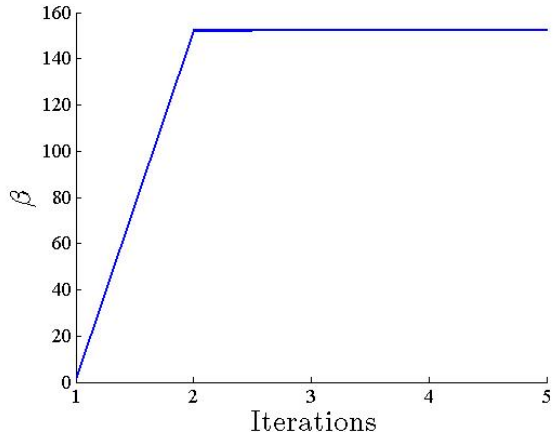

Рис. 5: Скалярный гиперпараметр β (случай шумового параметра)

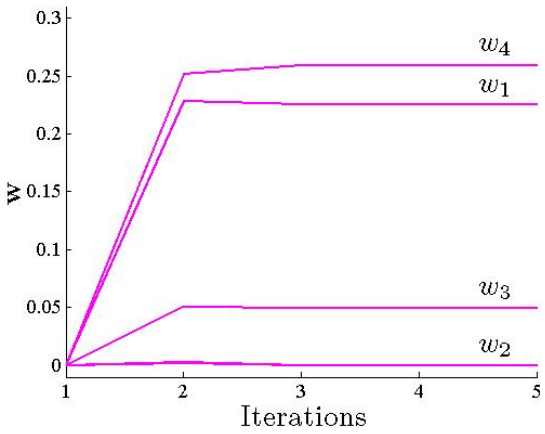

Рис. 6: Параметры модели $\mathbf{w}$ (случай шумового параметра)

8.2 Отбор коррелирующих признаков

Выборка с коррелирующими признаками сформирована при помощи добавления в матрицу плана столбца $1.3\boldsymbol{\chi}_2$. Таким образом, модель, аппроксимирующая данные в эксперименте: $y = w_1\boldsymbol{\chi}_1 + w_2\boldsymbol{\chi}_2 + w_3\boldsymbol{\chi}_3 + w_4\boldsymbol{\chi}_4$, где $\boldsymbol{\chi}_1 = \boldsymbol{\xi}^0$, $\boldsymbol{\chi}_2 = \boldsymbol{\xi}^1$, $\boldsymbol{\chi}_3 = 1.3\boldsymbol{\xi}^1$, $\boldsymbol{\chi}_4 = \sin(10\boldsymbol{\xi})$. Ниже на рис. 7 поэлементно проиллюстрирована матрица Гессе H.

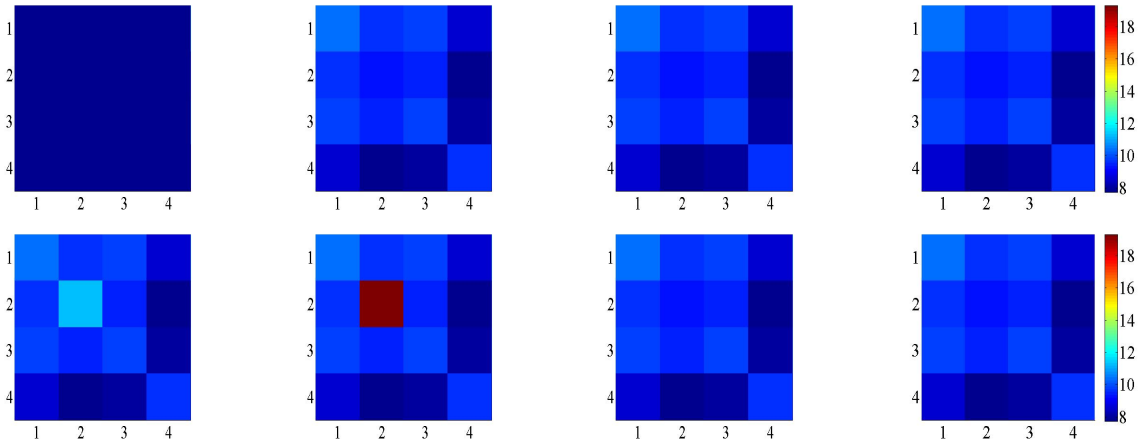

Рис. 7: Итерационный процесс для матрицы Гессе (случай коррелирующих параметров модели)

При наличии коррелирующих признаков также наблюдается возрастание диагональных элементов. Это происходит из-за того, что алгоритм выбирает ближайший вектор $\boldsymbol{\chi}$ к вектору $\mathbf{y}$ (в пространстве векторов матрицы X), а коррелирующий с ним считает шумовым.

На графиках 8 и 9 представлены диагональные элементы матрицы A.

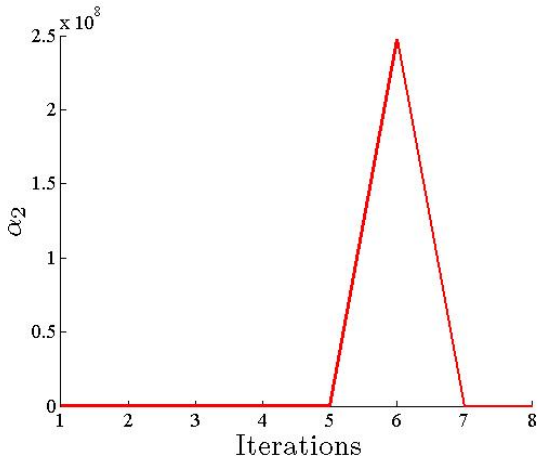

Рис. 8: Элементы матрицы A, соответствующие независимым параметрам модели

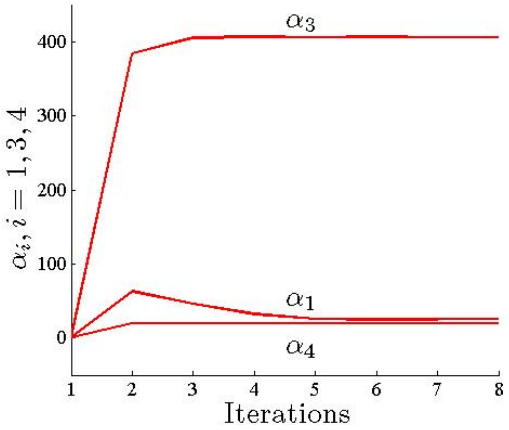

Рис. 9: Элемент матрицы A, соответствующий коррелирующему параметру модели

На рис. 10 представлены изменения скалярного гиперпараметры β.

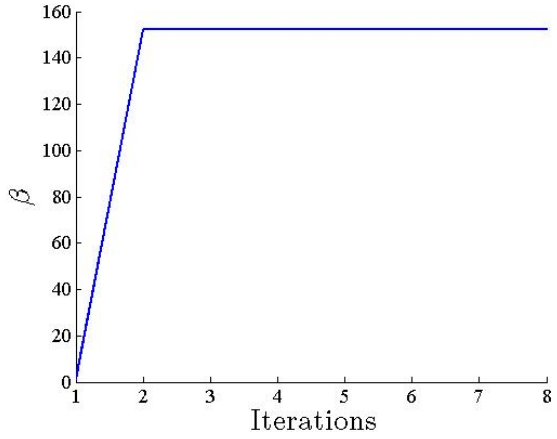

Рис. 10: Скалярный гиперпараметр β (случай зависимых параметров)

На рис. 11 представлены изменения параметров модели w_i в течении итерационного процесса. Коррелирующий параметр w_2 сначала возрастает, а затем стремится к нулю. Это происходит из-за того, что пространство параметров модели многоэкстремально.

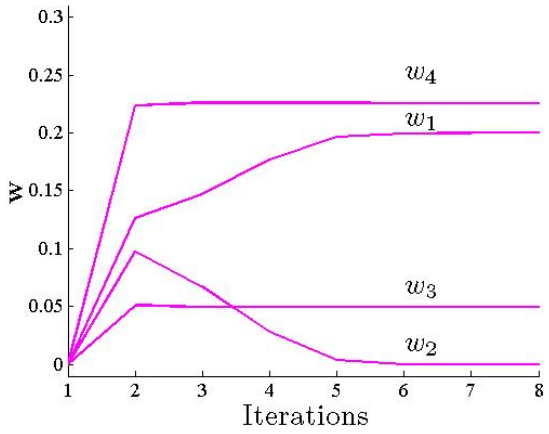

Рис. 11: Вектор параметров модели **w** (случай зависимых параметров)

9 Заключение

В работе предложен способ отсеивания шумовых и коррелирующих признаков, а также алгоритм оценки ковариационной матрицы параметров модели. Преимуществами данного алгоритма перед методами, описанными во введении, являются: 1) нет необходимости разделения данных на обучающую и контрольную выборку; 2) алгоритм не содержит никаких параметров, которые необходимо оценивать или задавать дополнительно (как, например, в методах регуляризации); 3) добиваясь сходимости как параметров, так и гиперпараметров, предложенный алгоритм повышает устойчивость выбранной регрессионной модели.

Список литературы

[1] *Strijov V. V., Weber G.-W.* Nonlinear regression model generation using hyperparameter optimization // Computers and Mathematics with Applications, 2010, vol. 60, no. 4, pp. 981-988.

[2] *Стрижов В. В.* Поиск параметрической регрессионной модели в индуктивно заданном множестве // Вычислительные технологии, 2007, vol. 1, pp. 93-102.

[3] *Efroymson M. A.* Multiple regression analysis. – New York: Ralston, Wiley, 1960.

[4] *Efron B., Hastie T., Johnstone J., Tibshirani R.* Least Angle Regression // Annals of Statistics, 2004, vol. 32, no. 3, pp. 407-499.

[5] *Tibshirani R.* Regression shrinkage and Selection via the Lasso // Journal of the Royal Statistical Society, 1996, vol. 32, no. 1, pp. 267-288.

[6] *Ильин В. А.* О работах А. Н. Тихонова по методам решения некорректно поставленных задач // Успехи математичексих наук, 1997, vol. 1, pp. 168-175.

[7] *Тихонов А. Н.* Решение некорректно поставленных задач и метод регуляризации. – М.: ДАН, 1963, vol. 151, pp. 501-504.

[8] *Hoerl A. E., Kennard R. W.* Ridge regression: Biased estimation for nonortthogonal problems // Technometrics, 1970, vol. 3, no. 12, pp. 55-67.

[9] *Bjorkstrom A.* Ridge regression and inverse problems. Tech. rep.: Stockholm University. – Stockholm, 2001.

[10] *Belsley D. A.* Conditioning Diagnostics: Collinearity and Weak Data in Regression. New York: John Wiley and Sons, 1991.

[11] *Marquardt D. W.* Generalized inverses, ridge regression, biased linear estimation, and nonlinear estimation // Technometrics, 1996, vol. 12, no. 3, pp. 605-607.

[12] *MacKay D.* Laplace's Method /В кн.: Information Theory, Inference, and Learning Algotirhms. – Cambridge: Cambridge University Press, 2005, pp. 341-351.

[13] *Nabney I.* Bayesian Techniques /В кн.: Netlab: Algorithms for Pattern Recognition. – New York: Springer, 2002, pp. 325-366.

[14] *Стрижов В. В.* Методы выбора регрессионных моделей. – М.: ВЦ РАН, 2010.

[15] *Bishop C. M.* Linear models for classification / В кн.: Pattern Recognition and Machine Learning. Под ред.: M. Jordan, J. Kleinberg, B. Scholkopf. – New York: Springer Science+Business Media, 1960, pp. 213-216.

ESTIMATION OF LINEAR MODEL HYPERPARAMETERS FOR NOISE OR CORRELATED FEATURE SELECTION PROBLEM

A. A. TOKMAKOVA[4], V. V. STRIJOV[5]

A. A. TOKMAKOVA (b. 1991) — student, Moscow Institute of Physics and Technology.
119991, Moscow, Vavilova 42, of. 151
Telephone: 8 (495) 135-4163
Fax: 8 (495) 137-2848
E-mail: aleksandra-tok@yandex.ru

V. V. STRIJOV (b. 1967) — Ph.D. in physics and mathematics, Research Fellow at the Computing Centre of the Russian Academy of Sciences.
119991, Moscow, Vavilova 42, of. 151
Telephone: 8 (495) 135-4163
Fax: 8 (495) 137-2848
E-mail: strijov@ccas.ru

Аннотация

This paper deals with the problem of feature selection in linear regression models. To select features authors estimate the covariance matrix of the model parameters. Dependent variable and model parameters are assumed to be normally distributed vectors. Laplace approximation is used for estimation of the covariance matrix: logarithm of the error function is approximated by the normal distribution function. The problem of noise or correlated features is also examined, since in this case the model parameters covariance matrix becomes singular. An algorithm for feature selection is proposed. The results of the study for a time series are given in the computational experiment.

Keywords: *feature selection, regression, coherent Bayesian inference, covariance matrix, model parameters.*

[4]Moscow Institute of Physics and Technology, aleksandra-tok@yandex.ru
[5]Computing Center of the Russian Academy of Sciences, strijov@ccas.ru