РОССИЙСКАЯ АКАДЕМИЯ НАУК
ВЫЧИСЛИТЕЛЬНЫЙ ЦЕНТР

СООБЩЕНИЯ ПО ПРИКЛАДНОЙ МАТЕМАТИКЕ

В. В. СТРИЖОВ, Е. А. КРЫМОВА

МЕТОДЫ ВЫБОРА РЕГРЕССИОННЫХ МОДЕЛЕЙ

ВЫЧИСЛИТЕЛЬНЫЙ ЦЕНТР РАН
МОСКВА, 2010

УДК 519.584

Ответственный редактор
доктор физ.-матем. наук К. В. Воронцов

При решении практических задач восстановления регрессии, для отыскания адекватной модели предлагается использовать порожденные признаки, полученные с помощью измеряемых исходных признаков. Это влечет существенное повышение размерности признакового пространства и, как следствие, необходимость использования алгоритмов выбора признаков. Ниже рассматриваются и сравниваются классические и новые алгоритмы выбора признаков. Работа алгоритмов проиллюстрирована прикладными задачами.
Основу работы составляет курс лекций, читаемый автором в Московском физико-техническом институте.

Ключевые слова: *регрессия, выбор моделей, выбор признаков, регуляризация, совместный байесовский вывод.*

Рецензенты: А. Г. Дьяконов,
Ю. В. Чехович

Научное издание

1. Введение

Решение практических задач восстановления регрессионной зависимости требует рассмотрения большого числа порождаемых признаков. Процедура построения регрессионных моделей состоит из двух шагов. На первом шаге, на основе свободных переменных — результатов измерений — порождается набор признаков. На втором шаге производится выбор признаков. При выборе признаков выполняется настройка параметров модели и оценивается ее качество. Модель с настроенными параметрами, доставляющая минимум заданному функционалу качества, называется моделью оптимальной структуры.

Развитие методов отбора признаков в регрессионном анализе имеет насыщенную историю. Среди методов отбора признаков широкое распространение получил шаговый метод, впервые предложенный в 1960 г. М. А. Эфроимсоном [1] На каждом шаге признаки проверяются на возможность добавления признака в модель или возможность удаления из модели. При этом используется F-статистика или критерий Акаике, Байесовский критерий, критерий Маллоуза [2–4].

В 1963 г. А. И. Тихонов ввел понятие регуляризации — дополнительного ограничения на задачу [5]. В работах [6, 7] введено понятие класса регуляризуемых некорректно поставленных задач и обоснован общий метод решения таких задач, названный методом регуляризации.

Работы А. И. Тихонова были опубликованы на западе только в 1977 г. В 1970 г. А. Хоэрл и Р. Кеннард предложили метод гребневой регрессии, в котором использовалась регуляризация [8]. Было введено дополнительное регуляризующее слагаемое в минимизируемый функционал, стало возможным улучшить устойчивость решения [9].

В первом издании книги Н. Дрейпера 1966 г. [10] приведен ступенчатый алгоритм выбора признаков. На каждой итерации выбирается признак, имеющий наибольшую проекцию на вектор

ответов, после этого делается небольшой шаг в направлении решения [11]. Среди полученных на каждом шаге моделей находится оптимальная, тем самым производится отбор признаков.

В 1971 г. А. Г. Ивахненко начал разрабатывать семейство методов группового учета аргументов МГУА [12]. Согласно его подходу на каждом шаге происходит отбор моделей и построение на их основе более сложных моделей [13–15]. Метод позволяет сократить перебор и осуществить отбор признаков.

В книге Д. Холланда 1975 г. [16] рассматривался общий подход к построению адаптивных систем. Любая адаптивная задача может быть представлена в терминах теории эволюции. Задача, сформулированная таким образом, может быть решена с помощью генетического алгоритма, воспроизводящего процессы эволюции. В основе использования генетического алгоритма лежит теорема шим (scheme), из доказательства которой следует, что при определенных условиях алгоритм дает экспоненциально быструю сходимость решения к локальному оптимуму.

В работах С. Шена 1980 г. и 1991 г. [17, 18] рассмотрен алгоритм последовательного добавления признаков с ортогонализацией (Forward Orthogonal search). Отбор признаков происходит автоматически при выборе оптимальной модели [19, 20].

Для упрощения структуры модели также используется метод оптимального прореживания, согласно которому элементы модели, оказывающие малое влияние на ошибку аппроксимации, можно исключить из модели без значительного ухудшения качества аппроксимации [21–23]. Алгоритм предложен в 1990 г. А. Н. Горбанем и основан на анализе первых производных в ходе обучения градиентными методами.

Еще один метод регуляризации, Лассо, был предложен Р. Тибширани в 1996 г. [24]. В нем вводится ограничение на L_1-норму вектора параметров модели, что приводит к обнулению части параметров модели и улучшению устойчивости решения.

В 2002 г. Б. Эфрон, Т. Хасти, И. Джонстон и Р. Тибширани предложили метод наименьших углов (Least Angle

Regression) [25]. Алгоритм заключается в последовательном добавлении признаков. На каждом шаге признак выбирается таким образом, что вектор регрессионных остатков равноуголен уже добавленным в модель признакам [26].

Перечислим рассмотренные ниже методы отбора признаков:

1) полный перебор моделей [14];
2) генетический алгоритм [27];
3) метод группового учета аргументов [13, 14, 28];
4) шаговая регрессия [10, 1, 29];
5) гребневая регрессия [10];
6) алгоритм Лассо [24];
7) ступенчатая регрессия [10];
8) последовательное добавление признаков с ортогонализацией [17, 18];
9) метод наименьших углов LARS [25];
10) оптимальное прореживание в линейной регрессии [21, 22];
11) метод моделей наибольшей обоснованности.

Основная проблема, возникающая при порождении признаков — проблема мультиколлинеарности. Термин «мультиколлинеарность» введен Р. Фришем при рассмотрении линейных зависимостей между признаками [30]. Мультиколлинеарность проявляется в сильной корреляции между двумя или более признаками, что затрудняет оценивание параметров модели. Мультиколлинеарность называют полной, если существует функциональная зависимость между всеми признаками. При этом становится

невозможно однозначно оценить параметры модели. На практике встречаются случаи частичной мультиколлинеарности, когда имеется высокая степень корреляции между некоторыми признаками. Тогда решение получить можно, однако оценки параметров модели и их дисперсии могут быть неустойчивы. Увеличиваются дисперсии оценок и абсолютные значения регрессионных параметров, что усложняет их интерпретацию.

Перечислим некоторые признаки наличия мультиколлинеарности: значительные изменения в оценках параметров при добавлении или удалении признака, превышения некоторого порога абсолютным значением корреляции между признаками, близость к нулю определителя матрицы попарных корреляций признаков.

Рассмотрим основные способы обнаружения мультиколлинеарности: проверка корреляции между признаками [10], исследование факторов инфляции дисперсии (Variance Inflation Factor) [31], метод Белсли [32].

Пусть X — матрица признаков, столбец в которой соответствует значениям признака при различных измерениях. Корреляционной матрицей называется матрица, элементами которой являются выборочные корреляции между столбцами:

$$\rho_{ij} = \frac{(\mathbf{x}_i - \mathbf{1}\bar{x}_i)^T(\mathbf{x}_j - \mathbf{1}\bar{x}_j)}{\|\mathbf{x}_i - \mathbf{1}\bar{x}_i\|\|\mathbf{x}_j - \mathbf{1}\bar{x}_j\|},$$

где $\mathbf{x}_i, \mathbf{x}_j$ — столбцы X, $\bar{x}_i, \bar{x}_j$ — среднее значения соответствующих признаков, $\mathbf{1}$ — столбец из единиц, число которых равно числу признаков, см. [33]. В случае центрированных и нормированных признаков

$$\rho_{ij} = \mathbf{x}_i^T \mathbf{x}_j.$$

Допустимым уровнем значимости называется минимальный уровень, вычисленный для данного значения статистики (значения коллинеарности преобразуются к t-статистике с $m-2$ степенями свободы), в случае выполнения гипотезы неколлинеарности признаков.

Если задан некоторый уровень значимости α и вычисленные допустимые уровни значимости p_{ij} меньше заданного α, то считается, что мультиколлинеарность велика. При этом следует помнить, что такие значения не обязательно являются следствием мультиколлинеарности.

Для оценки мультиколлинеарности используются факторы инфляции дисперсии (Variance Inflation Factor). При этом строится линейная регрессия всех признаков, кроме i-го, на i-й признак. Коэффициенты регрессии вычисляются с помощью метода наименьших квадратов.

Обозначим $\boldsymbol{\beta}$ вектор параметров линейной регрессии и σ^2 дисперсию регрессионных остатков при условии их гомоскедастичности. Значение фактора VIF определим как

$$\mathrm{VIF}_i = \frac{1}{1 - R_i^2},$$

где R_i^2 — коэффициент детерминации, вычисленный для i-го признака.

$$R_i^2 = 1 - \frac{\|\mathbf{x}_i - \mathbf{f}(\mathbf{x}_1, \ldots, \mathbf{x}_{i-1}, \mathbf{x}_{i+1}, \ldots, \mathbf{x}_n)\|^2}{\|\mathbf{x}_i - \mathbf{1}\bar{x}_i\|^2},$$

где $\mathbf{f}(\mathbf{x}_1, \ldots, \mathbf{x}_{i-1}, \mathbf{x}_{i+1}, \ldots, \mathbf{x}_n)$ — приближение i-го признака, полученное регрессией всех признаков, кроме i-го на i-й признак, $\bar{x}_i$ — средние значения i-го признака. Дисперсия параметров

$$\mathcal{V}(\beta_i) = \mathrm{VIF}\frac{\sigma^2}{\|\mathbf{x}_i - \mathbf{1}\bar{x}_i\|^2}.$$

Наличие мультиколлинеарности определяется по значениям VIF. Если $\mathrm{VIF}_i > 10$, то считается [34], что мультиколлинеарность велика.

В данной работе при исследовании мультиколлинеарости авторы используют метод Белсли. Пусть матрица признаков X имеет размерность $m \times n$, центрирована и нормирована. Проводится

сингулярное разложение [35–39] матрицы X,

$$X = U\Lambda V^T,$$

где U,V — ортогональные матрицы размерностью соответственно $m \times m$ и $n \times n$ и Λ — диагональная матрица с элементами (сингулярными числами) на диагонали, такими что

$$\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_{r+1},$$

где r — ранг X. Столбцы матрицы V являются собственными векторами, а квадраты сингулярных чисел — собственными значениями матрицы $X^T X$.

$$X^T X = V\Lambda^T U^T U\Lambda V^T = V\Lambda^2 V^T,$$

$$X^T XV = V\Lambda^2.$$

Индекс обусловленности с номером i — отношение максимального сингулярного числа к i-му сингулярному числу

$$\eta_i = \frac{\lambda_{max}}{\lambda_i}.$$

Для обнаружения мультиколлинеарности признаков построим таблицу, в которой каждому индексу обусловленности η_i соответствуют значения q_{ij} — долевые коэффициенты, которые в сумме по индексу j дают единицу. Они представляют собой части от общей величины $\sigma^{-2}\mathcal{V}(\beta_i)$.

Представим дисперсию оценок параметров в виде произведения. Если матрица $X^T X$ является обратимой, то можно записать

$$(X^T X)^{-1} = (V^T)^{-1}\Lambda^{-2}V^{-1} = V\Lambda^{-2}V^T.$$

Тогда

$$\sigma^{-2}\mathcal{V}(\boldsymbol{\beta}) = (X^T X)^{-1} = V\Lambda^{-2}V^T.$$

Таблица 1. Разложение $\mathcal{V}(\beta_i)$

Индекс обусловленности	$\mathcal{V}(\beta_1)$	$\mathcal{V}(\beta_2)$	$\dots$	$\mathcal{V}(\beta_n)$
η_1	q_{11}	q_{21}	$\cdots$	q_{n1}
η_2	q_{12}	q_{22}	$\cdots$	q_{n2}
$\vdots$	$\vdots$	$\vdots$		$\vdots$
η_n	q_{1n}	q_{2n}	$\cdots$	q_{nn}
$\sum_j q_{ij}$	1	1	$\dots$	1

Обозначим $V = \big(v_{ij}\big)$ и перепишем предыдущее выражение

$$\sigma^{-2}\mathcal{V}(\beta_i) = \frac{v_{i1}^2}{\lambda_{i1}^2} + \frac{v_{i2}^2}{\lambda_{i2}^2} + \cdots + \frac{v_{in}^2}{\lambda_{in}^2} = \\ = (q_{i1} + q_{i2} + \cdots + q_{in}) \sum_{j=1}^{n} \frac{v_{ij}^2}{\lambda_{ij}^2},$$

где q_{ij} — отношение соответствующего слагаемого в разложении $\sigma^{-2}\mathcal{V}(\beta_i)$ ко всей сумме. Наличие мультиколлинеарности определяется по таблице: большие величины η_i означают, что возможно есть зависимость между признаками. Большие значения q_{ij} в соответствующих строках относятся к признакам, между которыми эта зависимость существует. Маленькие величины η_i также исследуются: между признаками, соответствующими большим значениям q_{ij}, зависимости не существует.

Основными методами устранения мультиколлинеарности являются либо выбор признаков, либо введение ограничений на параметры модели [40, 41].

В данной работе предлагается компромиссный вариант алгоритма выбора регрессионных моделей. Его целью является получение наиболее адекватной и одновременно с этим наименее мультиколлинеарной модели. Он заключается в последовательном порождении моделей наибольшей обоснованности и основан

на работах по связаному Байесовскому выводу и порождающему походу к оценке параметров моделей [42–46]. При этом значения обоснованности различных моделей сравниваются. Выставляется порог интересующего нас значения правдоподобия. В процессе порождения модели модифицируются таким образом, что при добавлении признаков уменьшается сумма квадратов регрессионных остатков, а при удалении признаков уменьшается мультиколлинеарность модели.

Результаты сравнения алгоритмов приведены в таблице. Сравнение выполнялось на задаче поиска модели волатильности опционов. Использовались исторические данные торгов опционом Brent Crude Oil [47]. В таблицу входят значения функционала качества на обучающей и контрольной выборке, информационный критерий Акаике, число переменных модели. Исходя из значений критериев делается вывод об эффективности работы алгоритмов.

2. Постановка задачи

Задана выборка $D = \left\{(\mathbf{x}^i, y^i)\right\}_{i=1}^m$ — множество m пар, состоящих из вектора значений n свободных переменных $\mathbf{x}^i = (x_j^i)_{j=1}^n \in \mathbb{R}^n$, и значения одной зависимой переменной $y^i \in \mathbb{R}^1$. Индекс i объектов и индекс j свободных переменных далее будем рассматривать как элементы множеств $i \in I = \{1, \ldots, m\}$ и $j \in J = \{1, \ldots, n\}$.

Дополнительно задано разбиение выборки $I = \mathcal{L} \sqcup \mathcal{C}$ на обучающую и контрольную. Для каждого набора данных, рассматриваемого в вычислительном эксперименте, наборы индексов $\mathcal{L}$, $\mathcal{C}$ определены до начала эксперимента.

Задан класс регрессионных моделей $\mathfrak{F} = \{f_s\}$ — параметрических функций, линейных относительно параметров,

$$y^i = f_s(\boldsymbol{\beta}_s, \mathbf{x}^i) = \sum_{j \in J_s} \beta_j x_j^i, \tag{1}$$

в которой $s \in \{1, \ldots, 2^n\}$ является индексом модели, $\boldsymbol{\beta}_s = (\beta_j)_{j \in J_s}$ — вектор параметров, заданный индексом модели f_s. Под сложностью модели (1) понимается число линейно входящих параметров. Введено ограничение на число элементов линейной комбинации (1): в множество $\mathfrak{F}$ могут входить только модели сложностью не больше R.

Алгоритм выбора модели задает метод оптимизации, доставляющий оптимальное значение параметрам $\tilde{\boldsymbol{\beta}}$ модели f на обучающей выборке $\{(\mathbf{x}^i, y^i) \colon i \in \mathcal{L}\}$.

Пусть случайная аддитивная переменная ν регрессионной модели

$$y = f(\boldsymbol{\beta}, \mathbf{x}) + \nu \tag{2}$$

имеет нормальное распределение $\nu \sim \mathcal{N}(0, \sigma^2)$.

При условии гомоскедастичности регрессионных остатков распределение зависимой переменной имеет вид

$$p(y|x, \boldsymbol{\beta}, \sigma^2, f) = \frac{\exp\left(-\frac{1}{\sigma^2} S\right)}{(2\pi\sigma^2)^{\frac{n}{2}}},$$

где S — сумма квадратов невязок $y^i - f(\boldsymbol{\beta}, \mathbf{x}^i)$. Это распределение задает критерий качества модели, равный сумме квадратов регрессионных остатков.

$$S = \sum_{i \in \mathcal{X}} \left(y^i - f(\boldsymbol{\beta}, \mathbf{x}^i)\right)^2, \tag{3}$$

где $\mathcal{X}$ — некоторое множество индексов. Параметры модели находятся из минимизации критерия (3) на обучающей выборке, то есть при $\mathcal{X} \stackrel{\text{def}}{=} \mathcal{L}$. Требуется найти такую модель оптимальной структуры $f_s \in \mathfrak{F}$, которая доставляет наименьшее значение функционалу качества (3) на контрольной выборке $\mathcal{X} \stackrel{\text{def}}{=} \mathcal{C}$.

2.1. Порождение признаков

Множества измеряемых признаков зачастую бывает недостаточно для построения модели удовлетворительного качества. Тре-

буется расширить множество признаков с помощью функциональных преобразований исходных признаков с целью уменьшения недоопределенности линейной модели.

Предлагается следующий способ порождения признаков выборки D. Задано множество свободных переменных (измеряемых признаков)

$$\Xi = \{\xi_u\}_{u=1}^{U}$$

и конечное множество функций

$$G = \{g_v\}_{v=1}^{V}.$$

Эти функции называются порождающими функциями. Обозначим $a_i = g_v(\xi_u)$, где индекс $i = (v-1)U + u$. Рассмотрим декартово произведение $G \times \Xi$, элементу (g_v, ξ_u) которого поставлена в соответствие суперпозиция $g_v(\xi_u)$, однозначно определяемая индексами v, u.

Назначена базовая модель порождения признаков. В качестве модели, описывающей отношение между зависимой переменной y и свободными переменными a_i, используется полином Колмогорова-Габора

$$y = \beta_0 + \sum_{\iota=1}^{UV} \beta_\iota a_\iota + \sum_{\iota=1}^{UV} \sum_{\zeta=1}^{UV} \beta_{\iota\zeta} a_\iota a_\zeta + \ldots,$$

где вектор коэффициентов

$$\boldsymbol{\beta} = (\beta_0, \beta_\iota, \beta_{\iota\zeta}, \ldots)_{\iota,\zeta,\ldots=1,2,\ldots}.$$

Запишем этот полином в виде линейной комбинации порожденных переменных, где индекс j —номер члена линейной комбинации:

$$y = \sum_{j \in J} \beta_j x_j, \ |J| = n. \tag{4}$$

Переменные x_j поставлены в однозначное соответствие мономам полинома. Выражение (1) являются частным случаем выражения (4) для фиксированной модели с индексом s.

Для удобства описания алгоритмов выбора признаков обозначим вектор-столбец $\mathbf{x}_j = \left(x_j^1, \ldots, x_j^m\right)$ и $\mathbf{y} = \left(y^1, \ldots, y^m\right)$ Тогда

$$\mathbf{y} = \beta_1 \mathbf{x}_1 + \ldots + \beta_n \mathbf{x}_n$$

или в матричном представлении

$$\mathbf{y} = X\boldsymbol{\beta},$$

где X — матрица признаков со столбцами $\mathbf{x}_1, \ldots, \mathbf{x}_n$, $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_n)$ — вектор параметров.

2.2. Допустимые суперпозиции

Методы индуктивного построения регрессионных моделей используют в качестве моделей-претендентов различные суперпозиции свободных переменных. Рассмотрим условия при которых суперпозиции являются допустимыми. Функции $g_v \in G$ проиндексированы числами $v \in \mathcal{V} = \{1, ..., V\}$. Задано отображение $\iota : \mathcal{V}^R \to \mathcal{A}$. Элементы $A_\iota \in \mathcal{A}$ — всевозможные сочетания с повторениями из V по K, где $K = 1, \ldots, R$. Мощность множества $\mathcal{A}$ равна

$$|\mathcal{A}| = \sum_{K=1}^{R} \bar{C}_K^V = \sum_{K=1}^{R} \frac{(K+V-1)!}{(K-1)!V!},$$

где $\bar{C}$ — число сочетаний с повторениями.

Элементы набора $A_\iota = \{a_\iota(k)\}$ проиндексированы числами $k = 1, \ldots, K_\iota$. Так как $a \in \mathcal{V}$, элементы $a_\iota(k)$ однозначно соответствуют функциям g_v из G. Каждому набору A_ι поставим в соответствие набор матриц инцидентности $\{\rho_i(A_\iota)\}, i \in \mathbb{N}$. Индекс i матрицы ρ задает уникальную суперпозицию f_i функций g из G;

обозначим $\rho_i = \rho_i(A_\iota)$. Число элементов этой суперпозиции равно K_ι. Матрица инцидентности

$$\rho_i : \{1, \ldots, K_\iota\} \times \{1, \ldots, K_\iota\} \to \{0, 1\}$$

задает орграф и суперпозицию функций f_i нескольких аргументов. Суперпозиция f_i называется *допустимой*, если выполнены следующие условия.

1. Орграф ρ_i является ациклическим.

2. Орграф является односвязным без изолированных вершин, то есть справедливо равенство

$$\sum_{k=1}^{K_\iota} \sum_{l=1}^{K_\iota} \rho_i(l, k) = \sum_{k=1}^{K_\iota} s(a_\iota(k)),$$

   где $s = s(v)$ — число аргументов функции g_v. Число единиц в орграфе ρ_i равно суммарному числу аргументов в суперпозиции f_i.

3. Число аргументов каждого элемента суперпозиции должно совпадать с числом аргументов соответствующей порождающей функции

$$\sum_{l=1}^{K_\iota} \rho_i(l, k) = s(a_\iota(k)), \quad \text{для всех} \quad k = 1, \ldots, K_\iota.$$

   Число вершин орграфа, смежных вершине с номером k, есть число $s(a_\iota(k))$ аргументов функции g_v при $v = a_\iota(k)$.

2.3. Стандартизация данных

Выборка D стандартизируется таким образом, чтобы для $j \in J$ выполнялись условия нормированности и центрированности столбцов матрицы признаков X, центрированности

вектора $\mathbf{y}$:

$$\sum_{i=1}^{m} x_j^i = 0, \quad \sum_{i=1}^{m} (x_j^i)^2 = 1, \quad \sum_{i=1}^{m} y^i = 0. \tag{5}$$

Предполагается, что векторы $\mathbf{x}_j = \left(x_j^1, \dots x_j^m\right)$ и $\mathbf{x}_k = \left(x_k^1, \dots x_k^m\right)$ для всех $j, k \in J$, $j \neq k$ линейно независимы. Линейно зависимые векторы исключаются из дальнейшего рассмотрения.

3. Алгоритмы выбора моделей

3.1. Алгоритм полного перебора

Алгоритм полного перебора порождает все возможные подмножества признаков $\{\mathbf{x}_j\}_{j \in J}$.

Алгоритм последовательно строит модели-претенденты неубывающей сложности. Перебирается $\sum_{i=1}^{R} C_i^n$ моделей (при $R = n$ нужно перебрать 2^n моделей). Например,

$$\begin{array}{lcl} f_1 & = & \beta_1 \mathbf{x}_1, \\ \dots & \dots & \dots, \\ f_3 & = & \beta_{31} \mathbf{x}_1 + \beta_{32} \mathbf{x}_2, \\ \dots & \dots & \dots, \\ f_7 & = & \beta_{71} \mathbf{x}_1 + \beta_{72} \mathbf{x}_2 + \beta_{73} \mathbf{x}_3. \end{array} \tag{6}$$

Параметры каждой модели настраиваются методом наименьших квадратов по обучающей выборке. Наилучшая модель выбирается исходя из минимума ошибки на контрольной выборке. Для упрощения описания множества признаков введем переменную выбора монома — вектор $\mathbf{c} = (c_1, \dots, c_n)$. Его элемент $c_j \in \{0, 1\}$ принимает значение 1, если $j \in J_s$, в противном случае 0. Тогда любая модель из (6) будет иметь вид $y = \sum\limits_{j \in J_s} c_j \beta_j x_j$. При больших R время работы алгоритма недопустимо велико.

3.2. Генетический алгоритм

Алгортим состоит из итеративно повторяемых шагов. Из текущего множества (популяции) отбирается заданное число лучших моделей (особей). С помощью операций скрещивания и мутации происходит порождение новых особей. Процесс повторяется, пока не выполнится условие останова.

Используем переменную выбора признака — вектор $\mathbf{c} = (c_1, \ldots, c_n)$. Алгоритм содержит следующие параметры для отбора моделей: F — число лучших моделей в популяции, F_1 — число моделей для скрещивания, P_2 — вероятность выбора модели для мутации. Начальный набор моделей выбирается случайным образом. Итеративно выполняются следующие операции.

1. Отбор: согласно критерию (3) при $\mathcal{X} \stackrel{\text{def}}{=} \mathcal{C}$ выбирается F лучших моделей.

2. Случайным образом выбираются F_1 моделей для скрещивания и мутации.

3. Скрещивание: операция, при которой из двух моделей порождается две новые. Выбранные модели случайным образом разбиваются на пары. В каждой паре переменные выбора $\mathbf{c}^q = (c_1^q, \ldots, c_n^q)$ и $\mathbf{c}^p = (c_1^p, \ldots, c_n^p)$ разбиваются точкой кроссинговера, выбираемой случайно из множества $\{1, \ldots, n\}$, на две части. Происходит обмен элементов векторов $\mathbf{c}^p$ и $\mathbf{c}^p$:

$$\left\{ \begin{array}{l} (c_1^q, \ldots, c_k^q, c_{k+1}^q, \ldots, c_n^q) \\ (c_1^p, \ldots, c_k^p, c_{k+1}^p, \ldots, c_n^p) \end{array} \right. \mapsto$$

$$\mapsto \left\{ \begin{array}{l} (c_1^q, \ldots, c_k^q, c_{k+1}^p, \ldots, c_n^p) \\ (c_1^p, \ldots, c_k^p, c_{k+1}^q, \ldots, c_n^q) \end{array} \right. .$$

4. Каждая модель из полученного множества с вероятностью P_2 подвергается мутации: случайным образом из множества $\{1, \ldots, n\}$ выбирается индекс переменной выбора j.

В результате операции значение компоненты c_j меняется на противоположное (если был выбран элемент $c_j = 0$, то после операции он меняет свое значение на 1 и наоборот).

После операций 3 и 4 новые модели настраиваются исходя из условия минимизации критерия (3) при $\mathcal{X} \stackrel{\text{def}}{=} \mathcal{L}$. Операция производятся заданное число раз.

3.3. Метод группового учета аргументов

Алгоритмы МГУА воспроизводят схему массовой селекции, согласно которой последовательно порождаются и выбираются модели возрастающей сложности. При этом используются критерии, предложенные в рамках МГУА. Внутренним называется критерий (3), использующий только обучающую выборку $X_{\mathcal{L}}$. Внешним называется критерий, использующий тестовую выборку $X_{\mathcal{C}}$, при этом предполагается, что параметры модели настроены на обучающей выборке [13].

Предложены следующие внешние критерии.

1. Критерий регулярности — сумма квадратов ошибок на контрольной выборке:

$$\Delta^2(\mathcal{C}) = \|\mathbf{y}_{\mathcal{C}} - X_{\mathcal{C}}\boldsymbol{\beta}_{\mathcal{L}}\|_2^2.$$

   Модификация критерия регулярности с нормировкой на зависимую переменную:

$$\Delta^2(\mathcal{C}) = \frac{\|\mathbf{y}_C - X_{\mathcal{C}}\boldsymbol{\beta}_{\mathcal{L}}\|_2^2}{\|\mathbf{y}_{\mathcal{C}}\|_2^2}.$$

2. Критерий минимального смещения:

$$\nu^2 = \|X\boldsymbol{\beta}_{\mathcal{L}} - X\boldsymbol{\beta}_{\mathcal{C}}\|_2^2,$$

   в котором $\boldsymbol{\beta}_{\mathcal{L}}$ — параметры модели, находятся с помощью метода наименьших квадратов на обучающей выборке. Критерий требует, чтобы оценки коэффициентов в оптимальной

модели, вычисленные на обучающей и контрольной выборках, различались минимально.

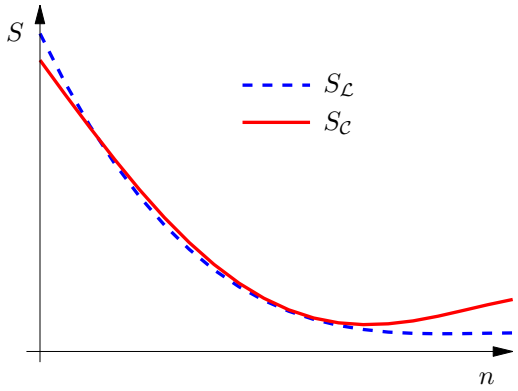

Рис. 1. Вид графиков функций внешнего и внутреннего критериев $S_{\mathcal{C}}$ и $S_{\mathcal{L}}$ в зависимости от числа параметров модели n

На каждом шаге алгоритма происходит порождение моделей. Модели настраиваются из условий минимизации внутреннего критерия. Затем выбираются P лучших моделей согласно внешнему критерию. Обозначим линейную комбинацию признаков как $f(x_1, x_2) = \beta_1 x_1 + \beta_2 x_2$, где $\beta_1,\ \beta_2$ — коэффициенты модели. Верхним индексом зависимой переменной обозначается номер шага. Порождение моделей происходит следующим образом.

Первый шаг селекции состоит из моделей, являющихся линей-

ной комбинацией двух признаков

$$\begin{array}{lcl} y_1^1 & = & f(x_1, x_2), \\ y_2^1 & = & f(x_1, x_3), \\ \dots & \dots & \dots, \\ y_{Q_1}^1 & = & f(x_{n-1}, x_n), \end{array}$$

где $Q_1 = C_n^2$. P_1 лучших моделей согласно внешнему критерию обозначим $y_1^1, \dots, y_{P_1}^1$. Второй шаг селекции:

$$\begin{array}{lcl} y_1^2 & = & f(y_1^1, y_2^1), \\ y_2^2 & = & f(y_1^1, y_3^1), \\ \dots & \dots & \dots, \\ y_{Q_2}^2 & = & f(y_{P_1-1}^1, y_{P_1}^1). \end{array}$$

После отыскания параметров, P_2 лучших моделей обозначим $y_1^2, \dots, y_{P_2}^2$. Рассмотрим k-й шаг селекции:

$$\begin{array}{lcl} y_1^k & = & f(y_1^{k-1}, y_2^{k-1}), \\ y_2^k & = & f(y_1^{k-1}, y_3^{k-1}), \\ \dots & \dots & \dots, \\ y_{Q_k}^k & = & f(y_{P_{k-1}-1}^{k-1}, y_{P_{k-1}}^{k-1}). \end{array}$$

Процесс порождения моделей прекращается, когда на новом шаге происходит увеличение внешнего критерия.

3.4. Шаговая регрессия

Шаговыми методами называются методы, заключающиеся в последовательном удалении или добавлении признаков согласно определенному критерию. Обычно используется критерий Фишера (F-критерий)

$$F = \frac{S_1 - S_2}{S_2} \frac{m - n_2}{n_1 - n_2},$$

где индекс 2 соответствует второй линейной модели, индекс 1 соответствует первой линейной модели, которая является модификацией второй модели; $n_1,\ n_2$ — соответствующие числа параметров в моделях. Если значение критерия больше заданного, то вторая модель считается лучше первой.

Шаговая регрессия включает два основных шага: шаг Add (последовательное добавление признаков) и шаг Del (последовательное удаление признаков).

Алгоритм последовательного добавления признаков. Рассмотрим текущий набор индексов признаков $\mathcal{A}$. Начальным набором является пустой набор $\mathcal{A} = \emptyset$. К текущему набору $\mathcal{A}$ присоединяется по одному признаку, который доставляет максимум F-критерию:

$$j^* = \arg\max_{j \in J} F_{\text{Add}} \propto \arg\max_{j \in J} \frac{S(\mathcal{A}) - S(\mathcal{A} \cup \{j\})}{S(\mathcal{A} \cup \{j\})}. \tag{7}$$

Алгоритм последовательного удаления признаков. Начальный набор состоит из всех признаков. На каждом шаге происходит удаление признака так, чтобы значение F-критерию было минимально:

$$j^* = \arg\min_{j \in J} F_{\text{Del}} \propto \arg\min_{j \in J} \frac{S(A \backslash \mathbf{x}^j) - S(A)}{S(A)}. \tag{8}$$

Алгоритм шаговой регрессии заключается в последовательном добавлении и удалении признаков при проверке значимости уже добавленных ранее признаков. Начальным набором является пустой набор индексов признаков $\mathcal{A} = \emptyset$. Добавляется несколько признаков, пока значение критерия на шаге не станет меньше заданного F_1, затем из набора $\mathcal{A}$ удаляются признаки, значение частного F-критерия (8) для которых не превосходит заданного значения F_2.

Останов алгоритма производится при достижении минимума, заданного критерием Маллоуза C_p:

$$C_p = \frac{S}{MSE} + 2k - m,$$

где $MSE = \frac{S}{n}$ — среднеквадратичная ошибка, вычисленная для модели, настроенной с помощью метода наименьших квадратов на всем множестве признаков, k —сложность модели. Критерий штрафует модели с большим числом признаков. Минимизация критерия позволяет найти множество, состоящее из значимых признаков.

Существует несколько недостатков метода, например, важная переменная может никогда не включаться в модель, а второстепенные признаки будут включены. Несмотря на это, шаговая регрессия применима в ситуации, когда число объектов выборки существенно превышает количество признаков.

3.5. Гребневая регрессия

Метод заключается во введении дополнительного регуляризующего слагаемого в минимизируемый функционал (строго говоря, это метод не является методом выбора признаков, так как не указывает, какие признаки следует исключить из модели). Плохая обусловленность матрицы $X^TX = \Sigma$ приводит к неустойчивости решения нормального уравнения линейной регрессии. Регуляризация позволяет уменьшить число обусловленности матрицы Σ и получить более устойчивое решение.

При регуляризации параметры модели находятся из минимизации функционала

$$\boldsymbol{\beta}^* = \arg\min_{\boldsymbol{\beta}} \left(\sum_{i=1}^{m} \left(y_i - \sum_{j=1}^{n} \beta^j x_j^i \right)^2 + \tau \|\boldsymbol{\beta}\|^2 \right)$$

или, в эквивалентной записи,

$$\boldsymbol{\beta}^* = \arg\min_{\boldsymbol{\beta}} \|\mathbf{y} - X\boldsymbol{\beta}\|^2, \quad \text{при } \|\boldsymbol{\beta}\|^2 \leq s.$$

Решением задачи минимизации является вектор

$$\boldsymbol{\beta}^* = (X^TX + \tau I_n)^{-1} X^T\mathbf{y}.$$

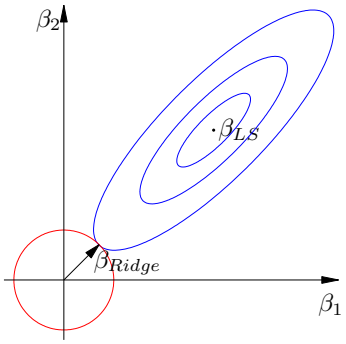

Рис. 2. Пространство параметров и ограничения, которые задача регуляризации накладывает на параметры

Увеличение параметра τ приводит к уменьшению нормы вектора параметров модели и повышению эффективной размерности признакового пространства [48]. Действительно, рассмотрим сингулярное разложение $\Sigma = U\Lambda V^T$. Числом обусловленности матрицы называется отношение максимального сингулярного числа к минимальному:

$$\kappa = \frac{\lambda_1}{\lambda_n}.$$

Рассмотрим число обусловленности κ регуляризованной матрицы Σ нормального уравнения:

$$\kappa\big(X^TX + \tau I\big) = \frac{\lambda_1 + \tau}{\lambda_n + \tau}$$

где λ_i — собственные числа матрицы X^TX. С увеличением параметра регуляризации τ уменьшается число обусловленности матрицы X^TX.

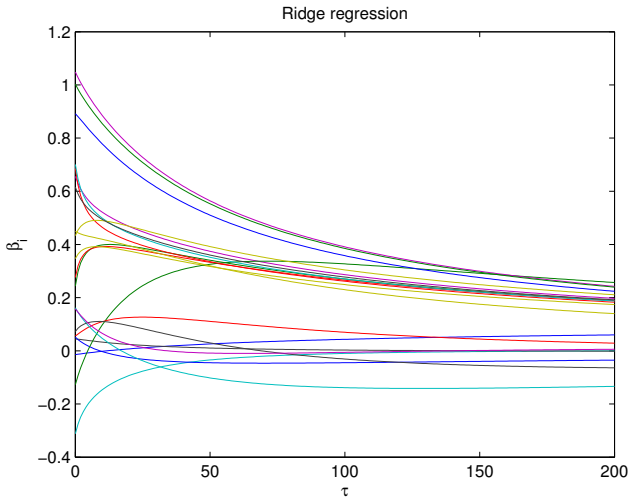

Рис. 3. Оценки коэффициентов регрессии, полученные с помощью алгоритма гребневой регрессии в зависимости от коэффициента регуляризации τ

В [49] приводится другой способ регуляризации $\Sigma_\tau = (1-\tau)\Sigma + \tau(D_\Sigma)$, где D_Σ — диагональная матрица, значения в которой — диагональные элементы матрицы Σ.

На рис. 2 показано пространство параметров модели. Критерий S — квадратичная функция относительно параметров $\boldsymbol{\beta}$, поэтому кривая $S = \text{const}$ является эллипсоидом. Регуляризирующий параметр, отличный от нуля, задает сферу в этом пространстве. Точка касания эллипсода сферой является решением нормального уравнения при фиксированном τ. При этом касание эллипсоида в нулевой точке исключено и обнуления параметров модели не происходит. Метод улучшает устойчивость параметров регрессионной модели, но не приводит к обращению в ноль ни одного из них.

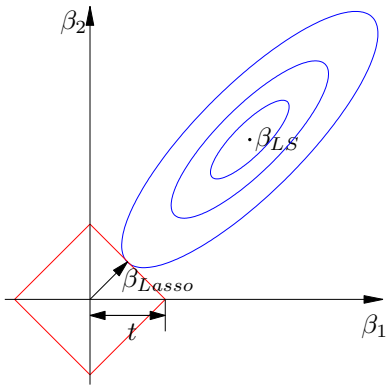

Рис. 4. Пространство параметров и ограничения, которые задача регуляризации накладывает на параметры

3.6. Лассо

Лассо (Least absolute shrinkage and selection operator) — метод оценивания параметров линейной модели. При оценивании вводится ограничение на сумму абсолютных значений параметров модели. В отличие от гребневой регрессии в Лассо некоторые параметры становятся равными нулю, а значит, происходит отбор признаков.

Рассматривается сумма модулей параметров модели,

$$T(\boldsymbol{\beta}) = \sum_{j=1}^{n} |\beta_j|.$$

Регрессионные параметры выбираются из условия минимизации критерия (3), $\min\{S(\boldsymbol{\beta})\}$, при ограничении

$$T(\boldsymbol{\beta}) \leq t, \tag{9}$$

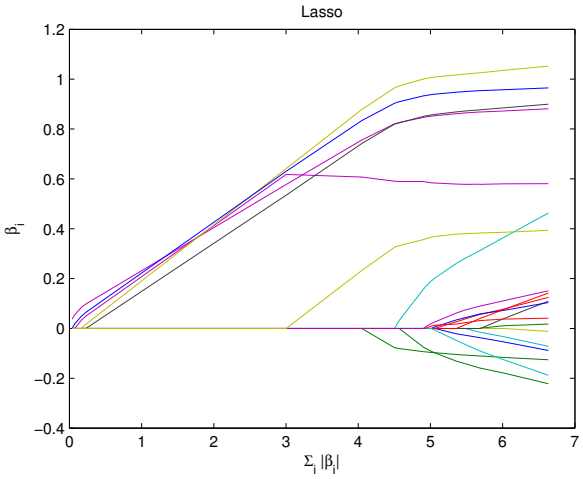

Рис. 5. Оценки коэффициентов регрессии, полученные с помощью алгоритма Лассо в зависимости от L_1-нормы вектора коэффициентов

где t — параметр регуляризации. Для решения используется метод квадратичного программирования с линейным ограничением-неравенством.

При больших t решение совпадает с решением, полученным методом наименьших квадратов. Чем меньше t, тем больше коэффициентов β_j принимают нулевое значение. Таким образом, Лассо осуществляет отбор признаков.

Задача может быть решена методом наименьших квадратов с 2^n ограничениями-неравенствами, соответствующими 2^n возможным знакам параметров β_j. Найдем решение при фиксированном $t \geq 0$. Введем $\boldsymbol{\delta}_i$, $i = 1, 2, \ldots, 2^n$ — m-мерные векторы вида $(\pm 1, \pm 1, \ldots, \pm 1)$. Тогда условия (9) эквивалентны $\boldsymbol{\delta}_i^T \boldsymbol{\beta} \leq t$ для всех i. Для заданного вектора $\boldsymbol{\beta}$ пусть $E = \{i : \boldsymbol{\delta}_i^T \boldsymbol{\beta} = t\}$ и $D = \{i : \boldsymbol{\delta}_i^T \boldsymbol{\beta} < t\}$, где E — набор индексов, соответствующих равенствам, D — набор индексов, для которых неравенство

не выполняется. Введем матрицу G_E, строками которой являются $\boldsymbol{\delta}_i,\ i \in E$, и $\mathbf{1}$ — вектор из единиц длиной, равной числу строк в G_E.

Начальное приближение для алгоритма: $E = \{i_0\}$, где $\boldsymbol{\delta}_{i_0} = \operatorname{sign}(\boldsymbol{\beta}^0)$, $\boldsymbol{\beta}^0$ — оценка вектора параметров методом наименьших квадратов без ограничений-неравенств. Пока $\sum |\beta_j| > t$,

1) найти $\tilde{\boldsymbol{\beta}}$, минимизирующий (3) при $G_E\boldsymbol{\beta} \le \mathbf{1}t$,

2) добавить i в набор E, где $\boldsymbol{\delta}_i = \operatorname{sign}(\boldsymbol{\beta})$. Найти $\tilde{\beta}$, минимизирующий $g(\boldsymbol{\beta})$ при $G_E\boldsymbol{\beta} \le t\mathbf{1}$.

Эта процедура сходится за конечное число шагов, так как на каждом шаге добавляется по одному элементу и число добавляемых элементов конечно. Решение, получаемое на последнем шаге, является решением всей задачи.

Альтернативный метод решения. Каждый параметр β_j задачи оптимизации записывается в виде $\beta_j^+ - \beta_j^-$, где β_j^+ и β_j^- неотрицательны. Тогда ограничения-неравенства принимают вид

$$\begin{array}{l} \beta_j^+ \ge 0, \beta_j^- \ge 0, \\ \sum_{j=1}^n (\beta_j^+ + \beta_j^-) \le t. \end{array}$$

Таким образом, начальная задача с n переменными и 2^n ограничениями может быть преобразована в новую задачу с $2n$ переменными, но меньшим числом ограничений $(2n + 1)$.

На рис. 2 показано пространство параметров модели. Кривая $S = \text{const}$ является эллипсоидом. Параметр t, отличный от нуля, задает многомерный октаэдр в этом пространстве. Точка касания эллипсода и октаэдра является решением нормального уравнения при фиксированном t. При касании сферы и ребра октаэдра происходит обнуление коэффициента.

3.7. Ступенчатая регрессия

Алгоритм ступенчатой регрессии состоит в последовательном добавлении признаков, наиболее коррелирущих с вектором ре-

грессионных остатков.

Начальный набор признаков пуст, вектор остатков $\mathbf{r}_0 = \mathbf{y}$.

Рассмотрим k-й шаг алгоритма. Сначала находится признак, корреляция которого с вектором остатков максимальна

$$j = \arg\max_i \mathbf{r}_k^T \mathbf{x}_i.$$

Обновляется вектор остатков

$$\mathbf{r}_{k+1} = \mathbf{r}_k + \epsilon \text{sign}\big(\mathbf{r}_k^T \mathbf{x}_i\big)\mathbf{x}_j,$$

где ϵ — достаточно маленькое число. Выбор большого ϵ, то есть $\epsilon = \frac{|\mathbf{r}_k^T \mathbf{x}_i|}{\|\mathbf{x}_i\|^2}$, приводит к обычному алгоритму последовательного добавления признаков Add.

Процесс продолжается, пока значение критерия (3) не будет незначительно изменяться за шаг.

В алгоритме ступенчатой регрессии направление для шага находится так же, как в Add, но шаг берется достаточно маленьким. Это приводит к большим вычислительным затратам, но и к более тщательному отбору признаков.

3.8. Добавление признаков с ортогонализацией

Метод последовательного добавления признаков с ортогонализацией (Forward Orthogonal Search) основан на ортогонализации признаков-столбцов матрицы X. Ортогонализация делает возможным вычисление индивидуального вклада каждого признака в вектор значений зависимой переменной. Ортогонализация матрицы X может быть произведена с помощью процедур ортогонализации Грамма-Шмидта или Хаусхолдера. Метод состоит в последовательном добавлении признаков с процедурой ортогонализации.

Запишем ортогональное разложение матрицы

$$X = QR,$$

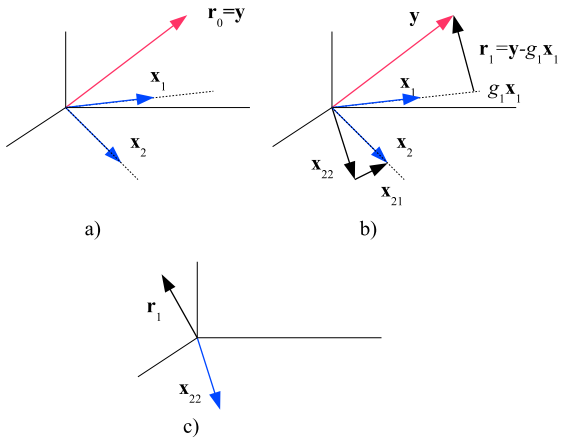

Рис. 6. Шаги алгоритма последовательного добавления признаков с ортогонализацией

где Q — матрица, столбцы которой являются ортогональным базисом матрицы признаков X, а R — верхняя треугольная матрица. Тогда

$$y = X\boldsymbol{\beta} = Q\mathbf{g},$$

где $\mathbf{g} = R\boldsymbol{\beta}$. Пусть на k-м шаге получен вектор остатков $\mathbf{r}_k$. Обозначим $\mathcal{A}_k$ — текущий набор признаков, $\bar{\mathcal{A}}_k$ — остальные признаки. Начальное значение вектора остатков $\mathbf{r}_0 = \mathbf{y}$ (рис. 6, а), текущий набор признаков $\mathcal{A}_0$ пуст.

Рассмотрим k-й шаг алгоритма.

1. Находим признак j, который составляет наименьший угол с вектором остатков $\mathbf{r}_k$:

$$j = \max_{i \in \bar{\mathcal{A}}_{k-1}} \frac{\mathbf{x}_i^T \mathbf{r}_k}{\|\mathbf{x}_i\| \|\mathbf{r}_k\|}.$$

Признак j добавляется в текущий набор $\mathcal{A}_k$ и удаляется из $\bar{\mathcal{A}}_k$.

2. Находим $\mathbf{r}_k^{\mathrm{Pr}}$ — проекцию вектора остатков $\mathbf{r}_k$ на $\mathbf{x}_j$ (рис. 6, б):
$$\mathbf{r}_k^{\mathrm{Pr}} = \frac{\mathbf{x}_j^T \mathbf{r}_k}{\|\mathbf{x}_j\|^2} \mathbf{x}_j.$$

3. Находим параметр, соответствующий добавленному признаку:
$$g_j = \frac{\|\mathbf{r}_k^{\mathrm{Pr}}\|}{\|\mathbf{x}^j\|}. \tag{10}$$

4. Обновляется вектор остатков (рис. 6, б):
$$\mathbf{r}_{k+1} = \mathbf{r}_k - \mathbf{r}_k^{\mathrm{Pr}}.$$

5. Признаки, не входящие в текущий набор, проецируются на подпространство, ортогональное пространству признаков из $\mathcal{A}_k$ (рис. 6, в). Происходит ортогонализация векторов признаков:
$$X_{\bar{\mathcal{A}}_k} = X_{\bar{\mathcal{A}}_{k-1}} - \mathbf{x}_j \frac{(\mathbf{x}_j)^T X_{\mathcal{A}_{k-1}}}{\|\mathbf{x}_j\|^2}.$$

Вектор параметров $\boldsymbol{\beta}$ находится из (10):
$$\boldsymbol{\beta} = R^{-1}\mathbf{g}.$$

Алгоритм последовательно добавляет признаки за n шагов.

Алгоритм FOS за счет нахождения ортонормированного базиса в пространстве признаков позволяет отбирать наименее коррелированные признаки. FOS, как и Add, делает максимальные шаги в выбираемом направлении.

3.9. Метод наименьших углов

На каждом шаге алгоритма LARS (Least Angle Regression) происходит изменение вектора параметров модели так, чтобы доставить добавляемому признаку наибольшую корреляцию с вектором регрессионных остатков.

LARS последовательными шагами строит оценку коэффициентов $\boldsymbol{\beta}$. На k-м шаге только k элементов вектора $\boldsymbol{\beta}$ отличны от нуля. Алгоритм последовательно вычисляет приближение зависимой переменной

$$\boldsymbol{\mu} = X\boldsymbol{\beta}.$$

Для приближений используется вектор корреляций столбцов матрицы X с вектором остатков $\mathbf{y} - \boldsymbol{\mu}$:

$$\mathbf{c}(\boldsymbol{\mu}) = X^T(\mathbf{y} - \boldsymbol{\mu}).$$

На k-м шаге новое значение приближения вектора зависимой переменной $\mathbf{y}$ вычисляется как

$$\boldsymbol{\mu}_k = \boldsymbol{\mu}_{k-1} + \gamma_k \mathbf{u}_k.$$

Здесь $\mathbf{u}_k$ — вектор единичной длины, вычисляемый следующим образом. Пусть $\mathcal{A}$ — подмножество индексов $\{1, \ldots, n\}$ столбцов $\mathbf{x}_j$ матрицы X. Это подмножество задает матрицу

$$X_{\mathcal{A}} = [s_{j_1}\mathbf{x}_{j_1}, \ldots, s_{j_{|\mathcal{A}|}}\mathbf{x}_{j_{|\mathcal{A}|}}], j \in \mathcal{A},$$

где множитель $s \in \{+1, -1\}$ и $|\mathcal{A}|$ — мощность множества $\mathcal{A}$. Обозначим ковариационную матрицу

$$\mathcal{G} = X_{\mathcal{A}}^T X_{\mathcal{A}}$$

и $A_{\mathcal{A}} = (\mathbf{1}_{\mathcal{A}}^T \mathcal{G}_{\mathcal{A}}^{-1} \mathbf{1}_{\mathcal{A}})^{-\frac{1}{2}}$, где $\mathbf{1}_{\mathcal{A}}$ — вектор, состоящий из $|\mathcal{A}|$ единиц.

Вычислим единичный вектор

$$\mathbf{u}_{\mathcal{A}} = X_{\mathcal{A}}\mathbf{w}_{\mathcal{A}},$$

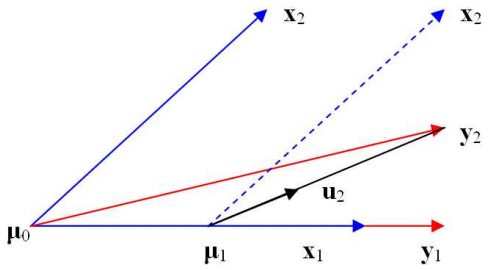

Рис. 7. Алгоритм LARS для случая $n = 2$

где $\mathbf{w}_{\mathcal{A}} = A_{\mathcal{A}}\mathcal{G}_{\mathcal{A}}^{-1}\mathbf{1}_{\mathcal{A}}$. Вектор $\mathbf{u}$ образует со столбцами матрицы $X_{\mathcal{A}}$ одинаковые углы, меньшие $\frac{\pi}{2}$. Справедливы равенства $X_{\mathcal{A}}^T\mathbf{u}_{\mathcal{A}} = A_{\mathcal{A}}\mathbf{1}_{\mathcal{A}}$ и $\|\mathbf{u}_{\mathcal{A}}\|^2 = 1$.

Действительно, $X_{\mathcal{A}}^T\mathbf{u}_{\mathcal{A}} = X_{\mathcal{A}}^T X_{\mathcal{A}}\mathbf{w}_{\mathcal{A}} = X_{\mathcal{A}}^T X_{\mathcal{A}} A_{\mathcal{A}}\mathcal{G}_{\mathcal{A}}^{-1}\mathbf{1}_{\mathcal{A}} = A_{\mathcal{A}} X_{\mathcal{A}}^T X_{\mathcal{A}}\big(X_{\mathcal{A}}^T X_{\mathcal{A}}\big)^{-1}\mathbf{1}_{\mathcal{A}} = A_{\mathcal{A}}\mathbf{1}_{\mathcal{A}}$ и норма вектора $\|\mathbf{u}_{\mathcal{A}}\|^2 = \mathbf{u}_{\mathcal{A}}^T\mathbf{u}_{\mathcal{A}} = \mathbf{w}_{\mathcal{A}}^T X_{\mathcal{A}}^T\mathbf{u}_{\mathcal{A}} = A_{\mathcal{A}}^2\mathbf{1}_{\mathcal{A}}^T(\mathcal{G}_{\mathcal{A}}^{-1})\mathbf{1}_{\mathcal{A}} = 1$.

Выполнение алгоритма. Назначим начальную оценку вектора значений зависимой переменной $\boldsymbol{\mu} = \mathbf{0}$. Вычислим текущую оценку $\boldsymbol{\mu}_{\mathcal{A}}$ и вектор корреляций

$$\mathbf{c} = X^T(\mathbf{y} - \boldsymbol{\mu}_{\mathcal{A}}).$$

Найдем текущий набор индексов $\mathcal{A}$, соответствующих признакам с наибольшими абсолютными значениями корреляций $C = \max_{j=1,\ldots,n}|c_j|$ и $\mathcal{A} = \{j : |c_j| = C\}$.

Пусть $s_j = \text{sign}(c_j)$ для $j \in \mathcal{A}$. Построим матрицу $X_{\mathcal{A}}$, вычислим $A_{\mathcal{A}}$. Вычислим вектор $\mathbf{u}_{\mathcal{A}}$ и вектор скалярных произведений

$$\mathbf{a} = X^T\mathbf{u}_{\mathcal{A}}.$$

Пересчитаем значение вектора $\boldsymbol{\mu}_{\mathcal{A}}$:

$$\boldsymbol{\mu}_{\mathcal{A}_+} = \boldsymbol{\mu}_{\mathcal{A}} + \hat{\gamma}\mathbf{u}_{\mathcal{A}}, \tag{11}$$

где

$$\hat{\gamma} = \min_{j \in \mathcal{A}^c}{}^{+} \left\{ \frac{C - c_i}{A_{\mathcal{A}} - a_j}, \frac{C + c_j}{A_{\mathcal{A}} + a_j} \right\}. \tag{12}$$

Минимум берется по всем положительным значениям аргументов для каждого j.

Добавим в множество $\mathcal{A}$ индекс j, где j доставляет минимум соответствующему значению $\hat{\gamma}$. Алгоритм повторяется n раз.

Замечание. Так как столбцы матрицы X линейно независимы, то матрица G не вырождена. В случае если матрица G является плохо обусловленной, для получения псевдообратной матрицы можно использовать сингулярное разложение.

Разъясним смысл переменной $A_{\mathcal{A}} = (\mathbf{1}_{\mathcal{A}}^T \mathcal{G}_{\mathcal{A}}^{-1} \mathbf{1}_{\mathcal{A}})^{-\frac{1}{2}}$. Пусть $\mathcal{S}_{\mathcal{A}}$ — «расширенный» симплекс

$$\mathcal{S}_{\mathcal{A}} = \left\{ \mathbf{v} = \sum_{j \in \mathcal{A}} s_j \mathbf{x}_j P_j : \sum_{j \in \mathcal{A}} P_j = 1 \right\}, \tag{13}$$

где P_j может быть отрицательным.

Лемма. Точка из $\mathcal{S}_{\mathcal{A}}$, ближайшая к началу координат, задается вектором

$$\mathbf{v}_{\mathcal{A}} = A_{\mathcal{A}} \mathbf{u}_{\mathcal{A}} = A_{\mathcal{A}} X_{\mathcal{A}} \mathbf{w}_{\mathcal{A}},$$

где $\mathbf{w}_{\mathcal{A}} = A_{\mathcal{A}} G_{\mathcal{A}}^{-1} \mathbf{1}_{\mathcal{A}}$, и $\|\mathbf{v}_{\mathcal{A}}\| = A_{\mathcal{A}}$.

Действительно, для любой точки из $\mathcal{S}_{\mathcal{A}}$ квадрат нормы вектора равен $\|X_{\mathcal{A}} P\|^2 = P^T G_{\mathcal{A}} P$. Выпишем лагранжиан с ограничением на сумму P_j:

$$P^T G_{\mathcal{A}} P - \lambda(\mathbf{1}_{\mathcal{A}}^T P - 1).$$

Минимизируем по $P_{\mathcal{A}}$, получаем $P_{\mathcal{A}} = \lambda G_{\mathcal{A}}^{-1} \mathbf{1}_{\mathcal{A}}$, суммируя, $\mathcal{G}_{\mathcal{A}}^{-1} \mathbf{1}_{\mathcal{A}}$, т.е. $P_{\mathcal{A}} = A_{\mathcal{A}}^2 G_{\mathcal{A}}^{-1} \mathbf{1}_{\mathcal{A}} = A_{\mathcal{A}} \mathbf{w}_{\mathcal{A}}$ и $\mathbf{v}_{\mathcal{A}} = X_{\mathcal{A}} P_{\mathcal{A}} \in \mathcal{S}_{\mathcal{A}}$. Норма вектора $\|\mathbf{v}_{\mathcal{A}}\|^2 = P_{\mathcal{A}}^T G_{\mathcal{A}}^{-1} P_{\mathcal{A}} = A_{\mathcal{A}}^4 \mathcal{G}_{\mathcal{A}}^{-1} \mathbf{1}_{\mathcal{A}} = A_{\mathcal{A}}^2$. Лемма доказана.

Величина γ в (12) интерпретируется следующим образом. Запишем оценку вектора зависимой переменных как функцию от γ

$$\mu(\gamma) = \mu_{\mathcal{A}} + \gamma \mathbf{u}_{\mathcal{A}}$$

при условии $\gamma > 0$. Корреляция вектора регрессионных остатков с добавляемым j-м признаком равна

$$c_j(\gamma) = \mathbf{x}_j^T(\mathbf{y} - \boldsymbol{\mu}(\gamma)) = c_j - \gamma a_j.$$

Для $j \in \mathcal{A}$ получаем

$$|c_j(\gamma)| = C - \gamma A_{\mathcal{A}}.$$

Это означает, что все рассматриваемые на данном шаге максимальные абсолютные корреляции уменьшаются на одну и ту же величину. Из предыдущих двух соотношений видно, что при $j \in \mathcal{A}^c$ корреляция $c_j(\gamma)$ принимает наибольшее значение при

$$\gamma = \frac{C - c_j}{A_{\mathcal{A}} - a_j}.$$

Аналогично корреляция $-c_j(\gamma)$ принимает наибольшее значение при

$$\gamma = \frac{C + c_j}{A_{\mathcal{A}} + a_j}.$$

Таким образом, γ в выражении (12) — минимальная положительная величина, при которой новый индекс j может быть добавлен в набор $\mathcal{A}$.

С помощью модификаций LARS можно получить решения Лассо и ступенчатой регрессии. Модификация LARS-Лассо. Пусть после шага LARS получены текущее множество индексов признаков $\mathcal{A}$ и оценка значений зависимой переменной $\boldsymbol{\mu}_{\mathcal{A}}$ и соответствующий вектор параметров $\boldsymbol{\beta}$. Пусть $\mathbf{d}$ — n-вектор, d_j равно $s_j w_{\mathcal{A}_j}$ для $j \in \mathcal{A}$, остальные компоненты равны нулю. Будем двигаться в направлении полученного решения LARS.

$$\boldsymbol{\mu}(\gamma) = X\boldsymbol{\beta}(\gamma), \text{ и } \beta_j(\gamma) = \hat{\beta}_j + \gamma \hat{d}_j,$$

где $j \in \mathcal{A}$.

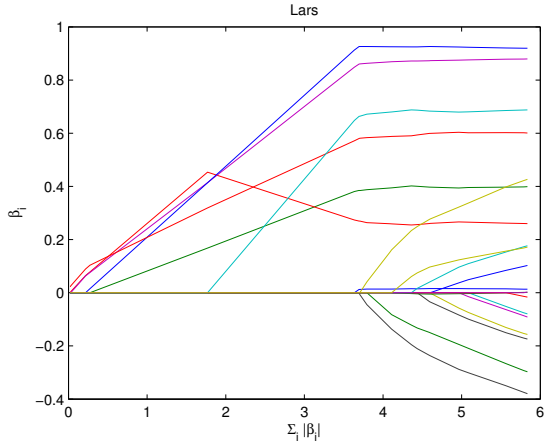

Рис. 8. Оценки коэффициентов регрессии, полученные с помощью алгоритма LARS в зависимости от L_1-нормы вектора коэффициентов

Тогда $\beta_j(\gamma)$ сменит знак при

$$\gamma_j = -\frac{\beta_j}{d_j}.$$

Первая смена знака произойдет при

$$\tilde{\gamma} = \min_{\gamma_j > 0} \gamma_j.$$

Если $\gamma_j > 0$ не выполнено ни для одного j, то по определению $\tilde{\gamma}$ полагается равным бесконечности. Если $\tilde{\gamma} < \hat{\gamma}$, то $\beta_j(\gamma)$ не является решением Лассо при $\gamma > \tilde{\gamma}$ из-за ограничений на знак.

Модификация Лассо состоит в остановке шага LARS при $\gamma = \tilde{\gamma}$, если $\tilde{\gamma} < \hat{\gamma}$ и удалении $\tilde{j}$ из вычислений следующего направления, т.е. $\boldsymbol{\mu}_{\mathcal{A}_+} = \boldsymbol{\mu}_{\mathcal{A}} + \tilde{\gamma}\mathbf{u}_{\mathcal{A}}$ и $\mathcal{A}_+ = \mathcal{A} - \tilde{j}$, вместо (44). Основным достоинством LARS является то, что он выполняется за число шагов, равное числу свободных переменных.

3.10. Оптимальное прореживание в линейной регрессии

Оптимальное прореживание — метод упрощения структуры регрессионной модели. Основная идея прореживания: элементы модели, которые оказывают малое влияние на ошибку аппроксимации (3), можно исключить из модели без значительного ухудшения качества аппроксимации.

Делается предположение, что параметры $\hat{\boldsymbol{\beta}}$ доставляют минимальное значение функции ошибки $S(\boldsymbol{\beta})$. Локальная аппроксимация функции S в окрестности точки $\hat{\boldsymbol{\beta}}$ с помощью разложения в ряд Тейлора записывается в виде

$$S(\boldsymbol{\beta}+\Delta\boldsymbol{\beta}) = S(\boldsymbol{\beta}) + \mathbf{g}^T\Delta\boldsymbol{\beta} + \tfrac{1}{2}\Delta\boldsymbol{\beta}^T H\Delta\boldsymbol{\beta} + o(\|\boldsymbol{\beta}\|^3),$$

где $\Delta\boldsymbol{\beta}$ — возмущение вектора параметров $\boldsymbol{\beta}$, $\mathbf{g} = \frac{\partial S}{\partial \boldsymbol{\beta}}$ — градиент, $H = \frac{\partial^2 S}{\partial \boldsymbol{\beta}^2}$ — матрица Гессе.

Так как квадратичная функция $S(\boldsymbol{\beta})$ достигает своего минимума при $\boldsymbol{\beta} = \hat{\boldsymbol{\beta}}$, то предыдущее выражение можно представить в виде $\Delta S = S(\boldsymbol{\beta}+\Delta\boldsymbol{\beta}) - S(\boldsymbol{\beta}) = \frac{1}{2}\Delta\boldsymbol{\beta}^T H\Delta\boldsymbol{\beta}$.

Исключение элемента модели есть обнуление одного параметра модели β_i. Исключение элемента эквивалентно выражению $\Delta\beta_i + \beta_i = 0$, иначе $\mathbf{e}_i{}^T\Delta\boldsymbol{\beta} + \beta_i = 0$, где $\mathbf{e}_i$ — вектор, i-й элемент которого равен единице, все остальные элементы равны нулю.

Требуется минимизировать квадратичную форму $\Delta\boldsymbol{\beta}^T H\Delta\boldsymbol{\beta}$ относительно $\Delta\boldsymbol{\beta}$ при ограничениях $\mathbf{e}_i{}^T\Delta\boldsymbol{\beta} + \beta_i = 0$ для всех значений i. Задача условной минимизации решается с помощью введения Лагранжиана $L = \Delta\boldsymbol{\beta}^T H\Delta\boldsymbol{\beta} - \lambda(\mathbf{e}_i{}^T + \beta_i)$, в котором λ — множитель Лагранжа. Дифференцируя Лагранжиан по приращению параметров и приравнивая его к нулю, получаем (для каждого индекса i параметра β_i)

$$\Delta\boldsymbol{\beta} = -\frac{\beta_i}{[H^{-1}]_{ii}} H^{-1}\mathbf{e}_i.$$

Этому значению вектора приращений параметров соответствует минимальное значение Лагранжиана

$$L_i = \frac{\beta_i^2}{2[H^{-1}]_{ii}}.$$

Полученное выражение называется мерой выпуклости функции ошибки S при изменении параметра β_i.

Функция L_i зависит от квадрата параметра β_i. Это говорит о том, что параметр с малым значением будет удален из модели. Однако если величина $[H^{-1}]_{ii}$ достаточно мала, это означает, что данный параметр оказывает существенное влияние на качество аппроксимации модели.

В случае линейной регрессии задача нахождения Гессиана упрощается. Так как

$$\frac{\partial S}{\partial \boldsymbol{\beta}} = -2X^T\mathbf{y} + 2X^TX\boldsymbol{\beta}$$

то Гессиан имеет вид

$$H = \frac{\partial^2 S}{\partial \boldsymbol{\beta}^2} = 2X^TX.$$

Рассмотрим один шаг алгоритма оптимального прореживания сетей для случая линейной регрессии.

1. Оцениваем параметры $\hat{\boldsymbol{\beta}}$.
2. Вычисляем матрицу, обратную Гессиану H^{-1}.
3. Находим i, соответствующее минимальному L_i.
4. Добавляем к вектору параметров $\hat{\boldsymbol{\beta}}$ вектор приращений $\Delta\beta$, соответствующий обнуляемому параметру.

Процедура повторяется до тех пор, пока значение ошибки (3) не превзойдет заранее заданное.

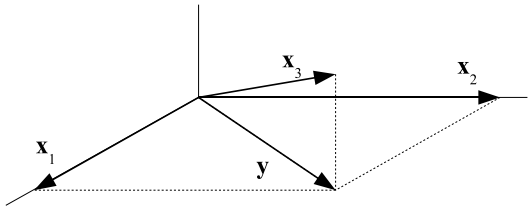

Рис. 9. Пример, иллюстрирующий последовательность выбора признаков

3.11. Модели наибольшего правдоподобия

Для иллюстрации основного недостатка алгоритма LARS рассмотрим следующий пример. Пусть матрица X состоит из столбцов значений трех признаков. Первый признак $\mathbf{x}_1$ хорошо коррелирует с вектором ответов $\mathbf{y}$, который является линейной комбинацией остальных двух признаков $\mathbf{x}_2$ и $\mathbf{x}_3$. Например, как показано на рис. 9,

$$X = [\mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3] = \begin{bmatrix} 1 & 1 & 0 \\ 1 & 0 & 1 \\ 1 & 0 & 0 \end{bmatrix}, \quad \mathbf{y} = \begin{bmatrix} 1 \\ 1 \\ 0 \end{bmatrix}.$$

LARS на первом шаге выберет первый признак, так как он сильнее коррелирует с вектором ответов, и затем присоединит остальные признаки. Ошибка модели, полученной с помощью LARS, будет отлична от нуля в то время, когда существует модель, доставляющая нулевую ошибку, и векторы-признаки, входящие в модель, ортогональны. Для разрешения этого недостатка предложен алгоритм, позволяющий удалять мультиколинеарные признаки и добавлять признаки, уменьшающие ошибку.

Предлагаемый алгоритм использует счетное число порождаемых признаков при отыскании линейной регрессионной модели. Введем процедуру пошагового нахождения модели. На каждом шаге выполняются операции добавления признаков и прорежи-

вания признаков. Под сложностью понимается число элементов линейной комбинации.

Используем два набора признаков: порожденный набор $\mathcal{Z}$ и текущий набор $\mathcal{C}$. В начале работы алгоритма $\mathcal{C} = \emptyset$.

Рассмотрим k-й шаг алгоритма.

1. Последовательно, методом Add, добавляются признаки из объединенного набора $\mathcal{C} \cup \mathcal{Z}$ в активный набор признаков $\mathcal{A}_k$. Итерации повторяются до тех пор, пока при увеличении сложности модели правдоподобие модели не будет меньше заданного порогового $\mathcal{E}_{min}$.

2. Выполняется прореживание модели: последовательно удаляются те элементы линейной комбинации, заданной набором $\mathcal{A}_k$, для которых критерий мультиколлинеарности Белсли принимает максимальное значение. Прореживание модели продолжается до тех пор пока, при уменьшении сложности модели, правдоподобие не будет меньше порогового $\mathcal{E}_{min}$. Коэффициенты полученной модели пересчитываются.

Итерации повторяются согласно критерию правдоподобия моделей. В результате получаем активный набор признаков $\mathcal{A}_k$, который на следующей итерации используется в качестве текущего набора $\mathcal{C}$.

Пороговое правдоподобие вычисляется следующим образом. Обозначим $p(\boldsymbol{\beta}) \stackrel{\text{def}}{=} p(\boldsymbol{\beta}|f)$ — априорное распределение параметров модели. Рассмотрим функцию правдоподобия $p(D|\boldsymbol{\beta}, f) \stackrel{\text{def}}{=} p(y|\{x_j\}_{j=1}^n, \boldsymbol{\beta}, f)$ — условную плотность распределения случайной величины при заданном векторе параметров.

При отыскании вектора параметров вместо максимизации функции правдоподобия $p(D|\boldsymbol{\beta}, f)$ будем максимизировать апостериорное распределение параметров

$$p(\boldsymbol{\beta}|D, f) = \frac{p(D|\boldsymbol{\beta}, f)p(\boldsymbol{\beta}|f)}{p(D|f)}. \tag{14}$$

Знаменатель $p(D|f)$ есть интеграл числителя формулы Байеса по всему пространству параметров:

$$p(D|f) = \int p(D|\boldsymbol{\beta}, f)p(\boldsymbol{\beta}|f)d\boldsymbol{\beta}. \tag{15}$$

Пусть зависимая переменная распределена нормально. Тогда функция правдоподобия принимает вид

$$p(D|\boldsymbol{\beta}, f) = \prod_{i=1}^{m} \mathcal{N}(y^i|f(\mathbf{x}^i, \boldsymbol{\beta}), \sigma_\nu^{-2}), \tag{16}$$

где σ_ν^2 — дисперсия случайной величины ν в выражении (2).

Пусть многомерная случайная величина — вектор параметров модели также имеет нормальное распределение с нулевым матожиданием и ковариационной матрицей

$$\alpha I = \frac{1}{\sigma_{\boldsymbol{\beta}}^2} I.$$

Тогда распределение вектора параметров модели

$$p(\boldsymbol{\beta}|\alpha, f) = \left(\frac{\alpha}{2\pi}\right)^n \exp\left(-\frac{\alpha}{2}\boldsymbol{\beta}^T\boldsymbol{\beta}\right). \tag{17}$$

Полученный знаменатель формулы (14) называется правдоподобием модели и служит для сравнения моделей.

Сравнение моделей выполняется с помощью связанного Байесовского вывода. Обозначим распределение моделей при фиксированных данных $p(f_i|D)$ и рассмотрим числитель формулы Байеса

$$p(f_i|D) = \frac{p(D|f_i)p(f_i)}{p(D)}, \tag{18}$$

в котором правдоподобие моделей $p(D|f_i)$ определяется выражением (15). Будем считать априорную вероятность равной для всех моделей, $p(f_i) = p(f_j)$. Так как знаменатель выражения (18) не

зависит от выбора модели, то сравнение моделей происходит через вычисление правдоподобие моделей с помощью формул (16) и (17). Порог $\mathcal{E}_{min}$ вычисляется как $\min_{i=1,\ldots,M} p(D|f_i)$ для набора из M моделей, имеющих максимальное правдоподобие. Параметр M задан.

Результатом работы алгоритма является модель удовлетворительной точности; мультикоррелирующие признаки исключены.

4. Байесовский вывод при выборе моделей

4.1. Сравнение моделей

Воспользуемся двухуровневым Байесовским выводом для оценки степени предпочтения порождаемых регрессионной моделей. Рассмотрим конечное множество моделей $f_1, \ldots, f_M$, приближающих данные D, обозначим априорную вероятность i-й модели $P(f_i)$. При появлении данных апостериорная вероятность модели $P(f_i|D)$ равна

$$P(f_i|D) = \frac{p(D|f_i)P(f_i)}{\sum_{j=1}^{M} p(D|f_j)P(f_j)}, \tag{19}$$

где $p(D|f_i)$ — функция правдоподобия моделей, определяющая, насколько хорошо модель f_i описывает данные D. Знаменатель дроби обеспечивает выполнение условия $\sum_{i=1}^{M} P(f_i|D) = 1$.

Сравним две модели с помощью апостериорных вероятностей

$$\frac{P(f_i|D)}{P(f_j|D)} = \frac{p(D|f_i)P(f_i)}{p(D|f_j)P(f_j)}. \tag{20}$$

Левая часть выражения называется отношением правдоподобия моделей. Отношение $P(f_i)/P(f_j)$ называется отношением апостериорных предпочтений моделей. Полагая априорные вероятности моделей одинаковыми, используем функции правдоподобия для выбора моделей.

Так как рассматриваемые модели f зависят от настраиваемых параметров, представим правдоподобие моделей в виде интеграла по пространству параметров

$$p(D|f) = \int p(D|\mathbf{w}, f)p(\mathbf{w}|f)d\mathbf{w}. \tag{21}$$

Априорная плотность распределения параметров $\mathbf{w}$ модели f на выборке D равна

$$p(\mathbf{w}|D, f) = \frac{p(D|\mathbf{w}, f)p(\mathbf{w}|f)}{p(D|f)}, \tag{22}$$

где $p(\mathbf{w}|f)$ — априорно заданная плотность вероятности параметров и $p(D|\mathbf{w}, f)$ — функция правдоподобия параметров. Выражения (19) и (22) называются формулами Байесовского вывода первого и второго уровня.

Рассмотрим следующую гипотезу порождения данных при восстановлении регрессии

$$y = f(\mathbf{w}, \mathbf{x}) + \nu.$$

Пусть случайная величина ν имеет нормальное распределение $\mathcal{N}(0, \sigma^2)$ с нулевым матожиданием и дисперсией σ^2, которая не зависит от свободной переменной.

В дальнейшем будет использоваться обозначение $D = (X, \mathbf{y})$, где $\mathbf{y} = [y_1, \ldots, y_N]^T$ — вектор значений зависимой переменной и X — матрица

$$X = \begin{bmatrix} \mathbf{x}_1 \\ \vdots \\ \mathbf{x}_N \end{bmatrix}.$$

Полагая $\mathbf{y}$ многомерной случайной величиной, имеем $\mathbf{y} \sim \mathcal{N}(\mathbf{f}, \beta I_N)$, где I_N единичная матрица и $\beta = \sigma^{-2}$.

Для фиксированной модели f плотность вероятности появления данных

$$p(y|\mathbf{x}, \mathbf{w}, \beta, f) \equiv p(D|\mathbf{w}, \beta, f) = \frac{\exp(-\beta E_D)}{Z_D(\beta)}, \tag{23}$$

где $\beta = \sigma^{-2}$, а коэффициент Z_D задан выражением, нормирующим функцию плотности в соответствии с гауссовым распределением

$$Z_D(\beta) = \left(\frac{2\pi}{\beta}\right)^{\frac{N}{2}}. \tag{24}$$

Функция регрессионных невязок согласно гипотезе порождения данных равна

$$E_D = \frac{1}{2}\sum_{n=1}^{N}\left(f(x_n) - y_n\right)^2. \tag{25}$$

Рассмотрим вектор параметров модели как многомерную случайную величину $\mathbf{w}$. Пусть плотность распределения параметров имеет вид многомерного нормального распределения $\mathcal{N}(\mathbf{0}, A)$ с матрицей ковариации A,

$$p(\mathbf{w}|A, f) = \frac{\exp(-E_{\mathbf{w}})}{Z_{\mathbf{w}}(A)}, \tag{26}$$

где A — ковариационная матрица случайной величины $\mathbf{w}$. Нормирующая константа $Z_{\mathbf{w}}(A)$ равна

$$Z_{\mathbf{w}}(A) = (2\pi)^{\frac{W}{2}}|A|^{\frac{1}{2}}, \tag{27}$$

где W — число параметров модели f. Функция-штраф за большое значение параметров модели при нормальном распределении равна

$$E_{\mathbf{w}} = \frac{1}{2}\mathbf{w}^T A^{-1}\mathbf{w}. \tag{28}$$

При заданной модели f и заданных значениях A и β выражение (22) принимает вид

$$p(\mathbf{w}|D, A, \beta, f) = \frac{p(D|\mathbf{w}, \beta, f)p(\mathbf{w}|A, f)}{p(D|A, \beta, f)}. \tag{29}$$

При фиксации в данном выражении гиперпараметров A и β, в знаменателе остается плотность вероятности появления данных, не зависящая от весов модели $\mathbf{w}$.

Записывая функцию ошибки

$$S(\mathbf{w}) = \frac{1}{2}\mathbf{w}^T A\mathbf{w} + \beta E_D, \tag{30}$$

получаем вместо (29) выражение

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

где Z_S — нормирующий множитель. Символ f далее будет опущен для удобства обозначений.

4.2. Сравнение элементов моделей

Предлагается итеративно найти параметры и гиперпараметры модели по отдельности. На каждой итерации сначала при фиксированных гиперпараметрах отыскиваются параметры путем оптимизации функционала (30). Используется алгоритм Левенберга-Марквардта. Затем по формулам, предложенным ниже, вычисляются гиперпараметры.

Предположим, что после очередного шага итерации нам известен локальный максимум (30) и он находится в точке $\mathbf{w}_0$. Для нахождения гиперпараметров приблизим (29) методом Лапласа. Для этого построим ряд Тейлора второго порядка логарифма числителя (29) в окрестности $\mathbf{w}_0$:

$$\ln\exp(-S(\mathbf{w})) = \ln\exp(S(\mathbf{w}_0) + \frac{1}{2}\Delta\mathbf{w}^T H\Delta\mathbf{w} + o(||\mathbf{w}||^3)), \tag{31}$$

где $\Delta\mathbf{w} = \mathbf{w} - \mathbf{w}_0$. При упрощении и отбрасывании пренебрежимо малой величины получается

$$-S(\mathbf{w}) \approx -S(\mathbf{w}_0) - \frac{1}{2}\Delta\mathbf{w}^T H\Delta\mathbf{w}. \tag{32}$$

В выражении (32) нет слагаемого первого порядка, так как предполагается, что $\mathbf{w}_0$ доставляет локальный минимум функции ошибки

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

Матрица H — матрица Гессе функции ошибок

$$H = -\nabla\nabla S(\mathbf{w})|_{\mathbf{w}=\mathbf{w}_0}. \tag{33}$$

Применяя экспоненту к обеим частям выражения (32) получаем требуемое приближение числителя (29)

$$\exp(-S(\mathbf{w})) \approx \exp(-S(\mathbf{w}_0)) \exp(-\frac{1}{2}\Delta\mathbf{w}^T H \Delta\mathbf{w}). \tag{34}$$

Таким образом, плотность распределения весов при известных гиперпараметрах принимает вид

$$p(\mathbf{w}|D, A, \beta) = \frac{\exp(-S(\mathbf{w}_0)) \exp(-\frac{1}{2}\Delta\mathbf{w}^T H \Delta\mathbf{w})}{Z_S(A, \beta)}. \tag{35}$$

Учитывая то, что интеграл выражения (29) должен равняться единице, получаем нормирующий множитель

$$Z_S = \frac{\exp(-S(\mathbf{w}_0))(2\pi)^{\frac{W}{2}}}{|H|^{\frac{1}{2}}}. \tag{36}$$

Знаменатель (29) является числитетем (19) и определяет выбор наиболее правдоподобной модели. Для нахождения гиперпараметров максимизируем функцию $p(D|A, \beta)$ относительно A и β. Запишем ее в виде

$$p(D|A, \beta) = \int p(D|\mathbf{w}, A, \beta) p(\mathbf{w}|A) d\mathbf{w}. \tag{37}$$

Используя выражения (23) и (26) перепишем (37) в виде

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

При выражении через Z_S

$$p(D|\beta, A) = \frac{Z_S}{Z_{\mathbf{w}}(A) Z_D(\beta)}.$$

Из (24), (27)и (36), логарифмируя (37), получим

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

$$\begin{aligned} \ln p(D|\beta, A) = & \underbrace{-\frac{W}{2} \ln 2\pi - \frac{1}{2} \ln |A|}_{Z_{\mathbf{w}}^{-1}(A)} \underbrace{-\frac{N}{2} \ln 2\pi + \frac{N}{2} \ln \beta}_{Z_D^{-1}(\beta)} - \\ & \underbrace{-S(\mathbf{w}_0) + \frac{W}{2} \ln 2\pi - \frac{1}{2} \ln |H|}_{Z_S}. \end{aligned} \tag{38}$$

При упрощении данного выражения получаем

$$\begin{aligned} \ln p(D|\beta, A) = & -\frac{1}{2} \ln |A| - \frac{N}{2} \ln 2\pi + \frac{N}{2} \ln \beta - \\ & \underbrace{-\beta E_D - E_{\mathbf{w}}}_{-S(\mathbf{w}_0)} - \frac{1}{2} \ln |H|. \end{aligned} \tag{39}$$

Найдем максимум выражения (38) относительно гиперпараметров, приравняв его производную поочередно по A и β к нулю. Для упрощения вычислений представим $A = I_W \| \frac{1}{\alpha} \|$. Разложим производную $\frac{\partial \ln p(D|\beta, A)}{\partial A}$ покомпонентно:

$$\frac{\partial \ln p(D|\beta, A)}{\partial \alpha_i} = \frac{1}{2\alpha_i} - \frac{w_i^2}{2} - \frac{\partial}{\partial \alpha} \ln |H|. \tag{40}$$

Производная последнего слагаемого равна

$$\begin{aligned}\frac{\partial}{\partial\alpha_i}\ln|H| &= \frac{\partial}{\partial\alpha_i}\ln(\prod_{j=1}^{W}\lambda_j+\alpha_j) = \\ &= \frac{\partial}{\partial\alpha_i}\sum_{j=1}^{W}\ln(\lambda_j+\alpha_j) = \frac{1}{\lambda_i+\alpha_i},\end{aligned} \tag{41}$$

где λ_j — собственные числа матрицы H_D — части Гессиана, не зависящей от A.

Приравнивая (40) к нулю и преобразовывая, получаем выражение для α

$$\frac{1}{\alpha_i} - w_i^2 - \frac{1}{\alpha_i+\lambda_i} = 0 \tag{42}$$

У этого уравнения два корня, однако один из них не имеет смысла, так как компоненты матрицы A не могут быть отрицательными:

$$\alpha_i = \frac{-\lambda_i + \sqrt{\lambda_i^2 + 4\frac{\lambda_i}{w_i^2}}}{2}. \tag{43}$$

Так как функция ошибки на данных не является квадратичной функцией параметров, как при линейной или RBF-регрессии, то непосредственно оптимизировать величину α невозможно, гессиан не является константой, а зависит от параметров $\mathbf{w}$. Так как мы принимаем $H = H_D + H_W$ для вектора $\mathbf{w}$, который зависит от выбора α, то собственные значения H оказываются зависящими от α.

Далее необходимо найти оптимальное значение гиперпараметра β. Обозначим через μ_j собственные значение матрицы ΔE_D. Так как $H_D = \beta\Delta E_D$, то $\lambda_j = \beta\mu_j$ и, следовательно,

$$\frac{d\lambda_j}{d\beta} = \mu_j = \frac{\lambda_j}{\beta}. \tag{44}$$

Отсюда

$$\frac{d}{d\beta}\ln|A| = \frac{d}{d\beta}\sum_{j=1}^{W}\ln(\lambda_j+\alpha_j) = \frac{1}{\beta}\sum_{j=1}^{W}\frac{\lambda_j}{\lambda_j+\alpha_j}.$$

Дифференцируя, как и в случае нахождения α, получаем, что

$$2\beta E'_D = N - \sum_{j=1}^{W}\frac{\lambda_j}{\lambda_j+\alpha_j} = N-\gamma, \tag{45}$$

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

Гиперпараметры $\boldsymbol{\alpha}$ и β_i вычисляется итеративно следующим образом — находятся оптимальные $\boldsymbol{\alpha}$, через них находятся оптимальные β. Далее новые β определяют новые значения $\boldsymbol{\lambda}$. Цикл повторяется до тех пор, пока изменение значений на соседних шагах не станет менее заранее заданной границы.

Значения функциналов ошибок $E'_{\mathbf{w}}$ и E'_D оптимизируются после каждого вычисления новых значений гиперпараметров.

При выборе моделей выполняется следующая процедура. Экспертно задается модель-претендент. Каждому элементу модели ставится в соответствие свой гиперпараметр α. Параметры и гиперпараметры модели последовательно настраиваются. Элемент модели, имеющий наименьшее значение гиперпараметра, исключается. Модель пополняется новым элементом из множества G согласно заданному правилу. Так как на каждом шаге такой модификации модели функционал качества не ухудшается, процедура выполняется до сходимости функционала качества (32).

5. Прикладные задачи

5.1. Моделирование биржевых опционов

Сравнительный анализ алгоритмов выполнен на исторических данных торгов опционом Brent Crude Oil. Срок действия оп-

Таблица 2. Результаты работы алгоритмов выбора признаков

Алгоритм	$S_{\mathcal{L}}$	$S_{\mathcal{C}}$	AIC	BIC	C_p	$\lg \kappa$	k
Генет.	0,073	0,107	-1152	-1072	337	13	26
МГУА	0,146	0,194	-1076	-1045	745	6	10
Шаг. рег.	0,128	0,154	-1092	-1055	644	7	12
Гребн.	0,111	0,146	-819	-330	832	33	160
Лассо	0,121	0,147	-1089	-1034	611	5	18
Ступ.	0,071	0,096	-1157	-1077	324	9	26
FOS	0,106	0,135	-1105	-1044	527	7	20
LARS	0,098	0,095	-1102	-1017	492	7	28
Предл.	0,097	0,123	-1118	-1054	469	5	21

циона — полгода. Тип опциона — право на продажу базового инструмента, символ CLG01. Базовый инструмент — нефть, символ NYM. Использовались ежедневные цены закрытия опциона и базового инструмента. Сетка цен исполнения опциона $\mathcal{K} = \{1400, 1425, \ldots, 1575, 1600\}$.

Регрессионная выборка

$$\big\{(\mathbf{x}^i, y^i)\big\}_{i=1}^m = \big\{(\langle K_i, t_i\rangle, \sigma_i)\big\}_{i=1}^m$$

была построена по исходным данным — историческим ценам опциона $C_{K,t}$ и базового инструмента P_t, где $K \in \mathcal{K},\ t \in T$, следующим образом. Для каждого значения K_i и t_i, $i = 1, \ldots, m$ вычислялось значение предполагаемой волатильности σ_i:

$$\sigma_i = \arg\min_{\sigma\in[0,1.5]} \big(C_{K_i,t_i} - C(\sigma, P_{t_i}, B, K_i, t_i)\big),$$

где справедливая цена опциона C вычислена по формуле Блэка-Шоулза [51]. Длина истории составляет 314 отсчетов времени.

Задано множество порождающих функций $G = \big\{1/x, \sqrt{x}, \ln(x), \tanh(x)\big\}$. Максимальная степень полинома Колмогорова-Габора равна трем. Регрессионная выборка была

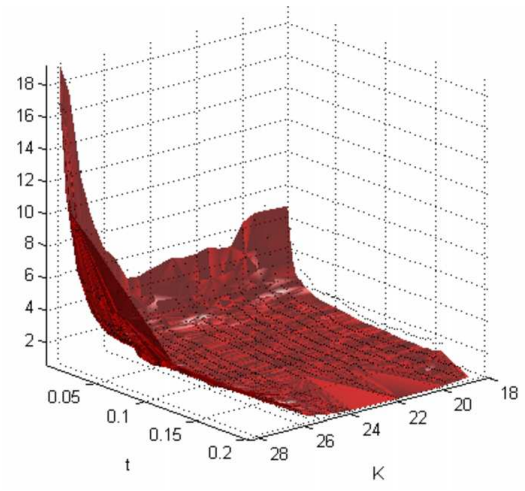

Рис. 10. Полученная модель зависимости исторической предполагаемой волатильности от цены и времени до исполнения

случайным образом разбита на контрольную и обучающую, равные по мощности. Значения ошибок на обучении и контроле были усреднены по 10 запускам алгоритмов на различных разбиениях.

Результаты экспериментов показаны в табл. 3. Для каждого алгоритма вычислялись ошибки $\mathrm{S}_{\mathcal{L}}$ и $\mathrm{S}_{\mathcal{C}}$ на обучении и контроле (3), значение информационных критериев Акаике

$$\mathrm{AIC} = m(\ln \frac{S}{m}) + 2k,$$

Байеса

$$\mathrm{BIC} = m(\ln \frac{S}{m}) + k \ln m,$$

Маллоуза, десятичный логарифм числа обусловленности κ матрицы значений отобранных признаков и сложность модели k.

На рис. 11 показана одна из полученных моделей. По оси K отложена цена исполнения опциона, по оси t отложено время до

исполнения. Точками показаны исходные данные. Полученная модель является адекватной и удовлетворительно приближает исторические данные.

5.2. Моделирование давления в двигателе

Работа алгоритма проиллюстрирована на примере данных измерения давления в двигателе внутреннего сгорания. Выборка состоит из набора временных рядов — измерений давления в камере внутреннего сгорания дизельного двигателя. Каждый временной ряд соответствует одному полному циклу работы двигателя, который состоит из рабочего и холостого тактов. Отчеты временного ряда равномерны и соответствуют углу вращения коленчатого вала. Нулевому углу соответствует верхняя мертвая точка вращения. Начало временного ряда соответствует углу в -360 градусов, конец — углу в $+359.9$ градусов. Всего один полный цикл насчитывает 7200 отсчетов. Лабораторный эксперимент включал измерения давления 122 полных циклов.

Регрессионная выборка

$$\left\{(\boldsymbol{\xi}^i, y^i)\right\}_{i=1}^m = \left\{(\langle n_i, \delta_i\rangle, P_i)\right\}_{i=1}^m$$

состоит из значений независимых переменных n_i — номера измерения и δ_i — угла вращения коленчатого вала. Каждой паре значений n_i и δ_i, $i = 1, \ldots, m$, соответствует значение давления P_i в камере внутреннего сгорания.

Задано множество порождающих функций $G = \left\{1/x, \sqrt{(|x|)}, \exp(-\frac{(x-m_i)^2}{2s_i^2}), \sin(c_i x)\right\}$. Регрессионная выборка была случайным образом разбита на контрольную и обучающую, равные по мощности. Вычислялось значение оценки скользящего контроля CV для фиксированных 10 разбиений выборки.

Результаты экспериментов показаны в табл. 3. Кроме оценки скользящего контроля вычислялись значение информационного

Таблица 3. Результаты работы алгоритмов выбора признаков

Алгоритм	CV	AIC	$\lg \kappa$	k
LARS	0,434	-1284	18	28
Предл.	0,008	-1670	4	11

критерия Акаике

$$\mathrm{AIC} = m(\ln \frac{S}{m}) + 2k,$$

десятичный логарифм числа обусловленности κ матрицы значений отобранных признаков и сложность модели k. Все величины были усреднены на 10 запусках алгоритмов.

На рис. 11 показана одна из полученных моделей. По осям отложены номера циклов, угол поворота коленчатого вала и величина давления P. Точками показаны исходные данные. Полученная модель является адекватной и удовлетворительно приближает исторические данные. Одна из полученных моделей:

$$\begin{aligned} f(\mathbf{w}, \boldsymbol{\xi}) = \quad & w_1 + w_2 \exp\left(\tfrac{-(\xi_1 - m_1)^2}{2s_1^2}\right) + \\ & + w_3 \exp\left(\tfrac{-(\xi_2 - m_2)^2}{2s_1^2}\right) + w_4 \xi_1 \exp\left(\tfrac{-(\xi_1 - m_1)^2}{2s_1^2}\right) + \\ & + w_5 \xi_1^2 + w_6 \tfrac{\xi_1}{\xi_2} + w_7 \xi_1 \exp\left(\tfrac{-(\xi_2 - m_2)^2}{2s_2^2}\right) + \\ & + w_8 \sin(c\xi_1) \exp\left(\tfrac{-(\xi_2 - m_2)^2}{2s_2^2}\right). \end{aligned}$$

5.3. Контроль состояния трубопроводов

Еще одна постановка задачи из области контроля качества состояния трубопроводов. Заданы координаты окружности (сечения трубы) — множество точек $\{(x, y)\}$, измеренных с некоторой погрешностью. Требуется найти центр (c_1, c_2) и радиус r окружности.

Запишем регерссионную модель — координаты окружности относительно центра и радиуса и выделим линейно входящие ком-

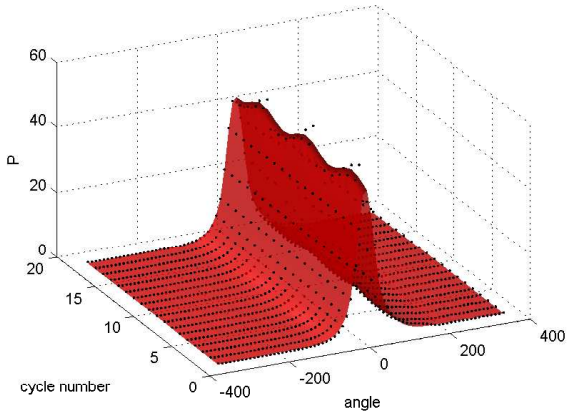

Рис. 11. Полученная модель зависимости исторической предполагаемой волатильности от цены и времени до исполнения

поненты:

$$
\begin{aligned}
&(x - c_1)^2 + (y - c_2)^2 = r^2, \\
&2xc_1 + 2yc_2 + (r^2 - c_1^2 - c_2^2) = x^2 + y^2, \\
&c_3 = (r^2 - c_1^2 - c_2^2).
\end{aligned}
$$

Тогда матрица плана линейной модели будет иметь вид

$$
\begin{pmatrix}
2x_1 & 2y_1 & 1 \\
2x_2 & 2y_2 & 1 \\
\vdots & \vdots & \vdots \\
2x_m & 2y_m & 1
\end{pmatrix}
\begin{pmatrix}
c_1 \\ c_2 \\ c_3
\end{pmatrix}
=
\begin{pmatrix}
x_1^2 + y_1^2 \\
x_2^2 + y_2^2 \\
\vdots \\
x_m^2 + y_m^2
\end{pmatrix}.
$$

Аналогичным путем получаются решения задач нахождения параметров эллипсоида, параллелограмма и других геометрических фигур.

5.4. Прогнозирование и авторегрессия

Рассмотрим постановку задачи прогнозирования временного ряда с выраженной периодической составляющей. Для прогноза используются скрытые регрессионные переменные.

Дан временной ряд $\mathbf{x} = [x_1, \ldots, x_T]^T$, $x \in \mathbb{R}^1$.

- Предполагается, что отсчеты времени сделаны через равные промежутки. Следовательно, отсчеты времени t можно без потери общности заменить на индексы — элементы натурального ряда.
- Предполагается, что ряд имеет периодическую составляющую и длина периода k известна.
- Предполагается, что ряд, возможно, имеет пропущенные значения.
- Предполагается, что длина ряда кратна периоду. Это условие удовлетворяется присоединением к началу ряда необходимого числа пропущенных значений.

Требуется построить алгоритм, выполняющий прогноз с горизонтом прогноза, равным длине одного периода.

Построение авторегрессионной матрицы. Составляется $(m \times k)$-матрица значений временного ряда:

$$X = \begin{pmatrix} x_T & x_{T-1} & \ldots & x_{T-k+1} \\ x_{(m-1)k} & x_{(m-1)k-1} & \ldots & x_{(m-2)k+1} \\ \ldots & \ldots & \ldots & \ldots \\ x_{nk} & x_{nk-1} & \ldots & x_{n(k-1)+1} \\ \ldots & \ldots & \ldots & \ldots \\ x_k & x_{k-1} & \ldots & x_1 \end{pmatrix},$$

в которой длина ряда $T = mk$. Каждый столбец матрицы содержит элементы ряда с индексами, кратными периоду.

Обозначим столбцы матрицы $\mathbf{x}_k, \ldots, \mathbf{x}_1$. Для каждого столбца i матрицы X построим набор моделей-предикторов. Для этого зафиксируем столбец $\mathbf{x}_i$, считая, что прогнозируем значение ряда в момент времени $i + k$.

Нахождение параметров авторегрессионной модели. Для каждого из прочих столбцов $\mathbf{x}_j, j = 1, \ldots, k$, и $j \neq i$ решим задачу линейной регрессии $\|\mathbf{x}_i - G_j\mathbf{w}\|^2 \to \min$, где матрица

$$G_j = \begin{pmatrix} g_1(x_{mj}) & g_2(x_{mj}) & \ldots & g_r(x_{mj}) \\ g_1(x_{(m-1)j}) & g_2(x_{(m-1)j}) & \ldots & g_r(x_{(m-1)j}) \\ \ldots & \ldots & \ldots & \ldots \\ g_1(x_j) & g_2(x_j) & \ldots & g_r(x_j) \end{pmatrix}.$$

Функции $g_1, \ldots, g_r$ заданы или определены исходя из дополнительных условий.

Выбирается заданное число p векторов $G_j\mathbf{w}$, доставляющих наибольшее значение функционалу $\rho(\mathbf{x}_i, G_j\mathbf{w})$. Обозначим P — множество выбранных индексов $\{j\}$. Строится корректор над множеством моделей-предикторов — линейная регрессия $\|\mathbf{x}_i - H_P\mathbf{b}\|^2 \to \min$ с ограничением на неотрицательность векторов $\mathbf{b}$. Матрица H_P— присоединенные векторы $G_j\mathbf{w}$, индексы $j \in P$. Прогнозируемое значение ряда $\mathbf{x}$ в момент времени $mk + i$ равно значению первого элемента вектора $H_P\mathbf{b}$.

6. Заключение

Рассмотренные алгоритмы можно разделить на группы: алгоритмы последовательного добавления или удаления признаков (FOS, LARS, Add, ступенчатая регрессия), алгоритмы с введением ограничения на функционал качества (гребневая регрессия и Лассо), алгоритмы перебора моделей (МГУА, генетический алгоритм, полный перебор), алгоритмы добавления и удаления признаков (шаговая регрессия, предложенные методы).

У алгоритмов последовательного добавления есть общий недостаток быстрая сходимость к локальному оптимуму: в примере,

показанном на рис. 9, они добавят первый признак и затем остальные, пропустив оптимальное решение — второй и третий признаки. Поскольку в алгоритмах добавления и удаления признаков вводится возможность проверки уже добавленных признаков, то становится возможным избежать неоптимального отбора признаков в этом случае. В частности, предложенный алгоритм находит оптимальное решение.

Вычислительный эксперимент показал, что увеличение числа признаков позволяет добиться улучшения качества модели. Однако при этом требуется введение дополнительных условий, позволяющих избежать появление мультиколлинеарных моделей. Предлагаемый алгоритм включает процедуру анализа мультиколлинеарности и позволяет получать хорошо обусловленные наборы порожденных признаков.

Описанные выше алгоритмы выполнены и протестированы в системе «Матлаб» и могут быть использованы для решения задач регрессионного анализа. Библиотека алгоритмов находится по адресу `http://strijov.com/examples/mdlselection.zip` и свободно распространяется на условиях лицензии GNU GPL.

Литература

1. *Efroymson M. A.* Multiple regression analysis. New York: Ralston, Wiley, 1960.
2. *Hocking R. R.* A biometrics invited paper. the analysis and selection of variables in linear regression // *Biometric.* 1976. Vol. 32, no. 1. Pp. 1–49.
3. *Akaike H.* A bayesian analysis of the minimum aic procedure // *Ann. Inst. Statist. Math.* 1978. Vol. 2, no. 30. Pp. 9–15.
4. *Mallows C. L.* Some comments on C_p // *Technometrics.* 1973. Vol. 15. Pp. 661–675.
5. *Ильин В. А.* О работах А. Н. Тихонова по методам решения некорректно поставленных задач // *Математическая*

жизнь в СССР и за рубежом. 1966. Т. 1. С. 168–175.
6. *Тихонов А. Н.* О решении некорректно поставленных задач и методе регуляризации // *Доклады академии наук СССР*. 1963. Т. 151. С. 501–504.
7. *Тихонов А. Н., Арсенин В.* . Методы решения некорректных задач. М.: Наука, 1986. 284 с.
8. *Hoerl A. E., Kennard R. W.* Ridge regression: Biased estimation for nonorthogonal problems // *Technometrics*. 1970. Vol. 3, no. 12. Pp. 55–67.
9. *Bjorkstrom A.* Ridge regression and inverse problems: Tech. rep.: Stockholm University, Sweden, 2001.
10. *Draper N. R., Smith H.* Appied Regression Analysis. John Wihley and Sons, 1998.
11. *Hastie T., Taylor J., Tibshirani R., Walther G.* Forward stagewise regression and the monotone lasso // *Electronic Journal of Statistics*. 2007. Vol. 1, no. 1. Pp. 1–29.
12. *Стрижов В. В.* Методы индуктивного порождения регрессионных моделей. М.: Вычислительный центр им. А. А. Дородницна РАН, 2008. С. 54.
13. *Malada H. R., Ivakhnenko A. G.* Inductive Learning Algorithms for Complex Systems Modeling. CRC Press, 1994. 368 pp.
14. *Ивахненко А. Г., Степашко В. С.* Помехоустойчивость моделирования. Киев: Наукова думка, 1985. 216 с.
15. *Ивахненко А. Г.* Индуктивный метод самоорганизации моделей сложных систем. Киев: Наукова думка, 1981. 296 с.
16. *Holland J. H.* Adaptation in natural and artificial systems. University of Michigan Press, 1975.
17. *Chen Y. W., Billings C. A., Luo W.* Orthogonal least squares methods and their application to non-linear system identification // *International Journal of Control*. 1989. Vol. 2, no. 50. Pp. 873–896.
18. *Chen S., Cowan C. F. N., Grant P. M.* Orthogonal least squares learning algorithm for radial basis function network // *Transaction on neural netwark*. 1991. Vol. 2, no. 2. Pp. 302–309.

19. *Berghen F.* LARS Library: Least Angle Regression Stagewise Library. No. 1. Addision-Wesley, 2005. Pp. 93–102.
20. *Guyon I., Gunn S.* Feature extraction: foundation and applications. Springer, 2006.
21. *Стрижов В. В.* Поиск параметрической регрессионной модели в индуктивно заданном множестве // *Журнал вычислительных технологий.* 2007. Т. 1. С. 93–102.
22. *Хайкин С.* Нейронные сети, полный курс. М: Вильямс, 2008. 1103 с.
23. *LeCun Y., Denker J., Solla S., Howard R. E., Jackel L. D.* Optimal brain damage // Advances in Neural Information Processing Systems II / Ed. by D. S. Touretzky. San Mateo, CA: Morgan Kauffman, 1990.
24. *Tibshirani R.* Regression shrinkage and selection via the lasso // *Journal of the Royal Statistical Society.* 1996. Vol. 32, no. 1. Pp. 267–288.
25. *Efron B., Hastie T., Johnstone I., Tibshirani R.* Least angle regression // *The Annals of Statistics.* 2004. Vol. 32, no. 3. Pp. 407–499.
26. *Lawson L., Hanson R. J.* Solving Least Squares Problems. Englewood Cliffs: Prentice Hall, 1974.
27. *Goldberg D. E.* Genetic algirothms in search, optimization, and machine learning. Addison-Wesley, 1989.
28. *Шитиков В. К., Розенберг Г. С., Зинченко Т. Д.* Количественная гидроэкология: методы системной идентификации. Тольятти: ИЭВБ РАН, 2003. 463 с.
29. *Rawlings J. O., Pantula S. G., Dickey D. A.* Applied Regression Analysis: A Research Tool. New York: Springer-Verlag, 1998.
30. *Frisch R.* Statistical Confluence Analysis by means of complete regression systems. Universitetets Okonomiske Institutt, 1934. 192 pp.
31. *Marquardt D. W.* Generalized inverses, ridge regression, biased linear estimation, and nonlinear estimation // *Technometrics.* 1996. Vol. 12, no. 3. Pp. 605–607.

32. *Belsley D. A.* Conditioning Diagnostics: Collinearity and Weak Data in Reggression. New York: John Wiley and Sons, 1991.
33. *Орлов А. И.* Эконометрика. М.: Экзамен, 2002.
34. *Kutner J., Nachtsheim C., Neter. J.* Applied Linear Regression Models. McGrow-Holl Irwin, 2004.
35. *Голуб Д., Ван-Лоан Ч.* Матричные вычисления. М.: Мир, 1999.
36. *Hogben L.* Handbook of linear algebra. CRC Press, 2007.
37. *Jolliffe I. T.* Principal component analysis. Springer, 2002. 457 pp.
38. *Isenmann A. J.* Modern multivariate statistical techniques. Springer, 2008. 734 pp.
39. *Рао С. Р.* Линейные статистические методы и их применения. М.: Наука, 1968. 547 с.
40. *Afifi A. A., Clark V., May S.* Computer-aided multivariate analysis. CRC Press, 2004.
41. *Burnham K., Anderson D. R.* Model Selection and Multimodel Inference. Springer, 2002.
42. *Ulusoy I., Bishop C. M.* Generative versus discriminative methods for object recognition // CVPR. 2005. Pp. II: 258–265.
43. *Vladislavleva E. J., Smits G. F., den Hertog D.* Order of nonlinearity as a complexity measure for models generated by symbolic regression via pareto genetic programming // *IEEE Transactions on Evolutionary Computation*. 2009. Vol. 13, no. 2. Pp. 333–349.
44. *Bishop C. M.* A new framework for machine learning // Computational Intelligence: Research Frontiers, IEEE World Congress on Computational Intelligence, WCCI 2008, Hong Kong. Springer, 2008. Pp. 1–24.
45. *Bishop C. M., Lasserre J.* Generative or discriminative? Getting the best of both worlds // In Bayesian Statistics 8 / Ed. by J. M. e. a. Bernardo. Oxford University Press, 2007. Pp. 3–23.
46. *Zelinka I., Oplatkova Z., Nolle L.* Analytic programming – symbolic regression by means of arbitrary evolutionary algorithm // *I. J. of Simulation*. 2008. Vol. 6, no. 9. Pp. 44–56.

47. *Ширяев А. Н.* Основы стохастической финансовой математики. М: ФАЗИС, 2004. Т. 1.
48. *Hastie T., Tibshirani R., Friedman J.* The Elements of Statistical Learning. Springer, 2001.
49. *Шурыгин А. М.* Прикладная стохастика: робастность, оценивание, прогноз. М: Финансы и статистика, 2000.
50. *Weisberg S.* Applied linear regression. New York: Wiley, 1980.
51. *Hull J. C.* Options, Futures and Other Derivatives. Prentice Hall, 2000.

Содержание