Последовательный выбор признаков при восстановлении регрессии*

Л. Н. Леонтьева
liubov.sanduleanu@gmail.com
Московский физико-технический институт, ФУПМ, каф. «Интеллектуальные системы»

Исследуется проблема оптимальной сложности модели в связи с ее точностью и устойчивостью. Задача состоит в нахождении наиболее информативного набора признаков в условиях их высокой мультиколлинеарности. Для выбора оптимальной модели используется модифицированный алгоритм шаговой регрессии, являющийся одним из алгоритмов добавления и удаления признаков. Для описания работы пошагового алгоритма предложена модель n-мерного куба. Проанализированы величины матожидания и дисперсии функции ошибки.

Ключевые слова: *отбор признаков, мультиколлинеарность, шаговая регрессия, метод Белсли, прогнозирование временных рядов.*

Feature selection in autoregression forecasting*

L. N. Leonteva
Moscow Institute of Physics and Technology

The authors investigate the optimal model selection problem with application to the autoregression forecasting. To solve the problem one has to select a maximum well-defined feature subset, subject to some given value of the error function. To select the feature set the modified add-del feature selection algorithm is used. This paper suggests a method of time series forecasting model selection. The computational experiment compares the electricity hourly prices forecasts.

Keywords: *feature selection, multicollinearity, stepwise regression, Belsley method, time series forecasting.*

Введение

Решается задача восстановления линейной регрессии при наличии большого числа мультиколлинеарных признаков. Термин «мультиколлинеарность» введен Р. Фишером при рассмотрении линейных зависимостей между признаками [1]. Проблема состоит в том, что количество признаков значительно превосходит число зависимых переменных, то есть мы имеем дело с переопределенной матрицей. Для решения этот задачи необходимо исключить наиболее малоинформативные признаки. Для отбора признаков предлагается использовать модифицированный метод шаговой регрессии.

Ранее для решения подобных задач использовались следующие методы: метод наименьших углов LARS [2], Лассо [3], ступенчатая регрессия [4], последовательное добавление признаков с ортогонализацией FOS [5, 6], шаговая регрессия [4, 7, 8] и другие [14]. Шаговыми методами называются методы, заключающиеся в последовательном удалении или добавлении признаков согласно определенному критерию. Существует несколько недостатков этих методов, например, важный признак может быть никогда не включен в модель, а второстепенные признаки будут включены.

Научный руководитель В. В. Стрижов

В работе предложен модифицированный метод шаговой регрессии. Существует три основых разновидности шаговых методов: метод последовательного добавления признаков, метод последовательного удаления признаков и метод последовательного добавления и удаления признаков. В работе используется последний. Метод включает два основных шага: шаг Add (последовательное добавление признаков) и шаг Del (последовательное удаление признаков). Добавление признаков производится с помощью FOS [5, 6]. Данный метод последовательно добавляет признаки, которые максимально коррелируют с вектором регрессионных остатков. Удаление признаков в нашей работе осуществляется методом Белсли [9]. Он позволяет выявить мультиколлинеарность признаков, используя сингулярное разложение матрицы признаков. Для нахождения алгоритма, который доставляет одновременно точную и устойчивую, в смысле минимизации числа мультиколлиниарных признаков, модель предложен новый критерий останова этапов Add и Del, а так же останова всего алгоритма.

Предложенный метод выбора модели проиллюстрирован задачей прогнозирования состояния здоровья людей больных диабетом. Ранее подобные задачи решались с помощью гребневой регрессии [10], метода наименьших углов, построения локальных регрессионных моделей [12, 13] и других.

Задача прогнозирования с помощью линейной регрессии

Даны временной ряд $\mathbf{s}^0 = \{x_i\}_{i=1}^{T-1}$, будем называть его целевым рядом, и временные ряды $\mathbf{s}^1, \mathbf{s}^2 \dots \mathbf{s}^p$. Необходимо спрогнозировать следующее значение s_T^0 ряда $\mathbf{s}_0$.

Предполагается, что искомая величина s_T^0 зависит от последних b значений рядов $\mathbf{s}^0, \mathbf{s}^1 \dots \mathbf{s}^p$. Параметр b называется глубиной логирования. Таким образом мы имеем дело с $b(p+1)$ признаком. Нахождение оптимальной модели, состоит в отыскании такого набора признаков, который минимизирует ошибку S при прогнозировании методом линейной регрессии.

Построим матрицу плана $\mathbf{X}^*$

$$\mathbf{X}^* = \left[\begin{array}{cccccccccc|c} s_0^1 & \dots & s_0^b & s_1^1 & \dots & s_1^b & \dots & s_p^1 & \dots & s_p^b & x_{b+1} \\ \dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots \\ s_0^{T-b-2} & \dots & s_0^{T-2} & s_1^{T-b-2} & \dots & s_1^{T-2} & \dots & s_p^{T-b-2} & \dots & s_p^{T-2} & x_{T-1} \\ \hline s_0^{T-b-1} & \dots & s_0^{T-1} & s_1^{T-b-1} & \dots & s_1^{T-1} & \dots & s_p^{T-b-1} & \dots & s_p^{T-1} & x_T \end{array}\right],$$

где s_i^j — j-ое значение ряда $\mathbf{s}_i$. То есть, строка с номером i матрицы плана X^* есть векторизованная подматрица, состоящая из значений временных рядов

$$\left[\begin{array}{ccc} s_0^i & \dots & s_p^i \\ \dots & \ddots & \dots \\ s_0^{b-i} & \dots & s_p^{b-i} \\ s_0^{b-i+1} & \dots & s_p^{b-i+1} \end{array}\right].$$

Введем обозначения:

$$\mathbf{X}^* = \left[\begin{array}{c|c} \mathbf{X} & \mathbf{y} \\ \hline \mathbf{x}_m & x_T \end{array}\right].$$

Необходимо построить линейную регрессию:

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

где $\mathbf{w}$ — вектор параметров. Тогда получим

$$x_T = \langle \mathbf{x}_m, \mathbf{w} \rangle .$$

Требуется решить задачу минимизации евклидовой нормы вектора регрессионных остатков

$$||\mathbf{X}\mathbf{w} - \mathbf{y}||^2 \to \mathbf{min}.$$

Вектор параметров $\mathbf{w}$ отыскивается с помощью метода наименьших квадратов

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

Задача выбора оптимальной модели

Опишем, в чем состоит задача выбора оптимальной модели. Задана выборка $D = (\{\mathbf{x}_i, y_i\}), i \in \mathcal{I}$, где множество свободных переменных — вектор $\mathbf{x} = [x_1, \ldots, x_j, \ldots, x_n]$, проиндексированно $j \in \mathcal{J} = \{1, \ldots, n\}$. Задано разбиение множества индексов элементов выборки $\mathcal{I} = \mathcal{L} \sqcup \mathcal{C}$. Также задан класс линейных параметрических регрессионных моделей $f(\mathbf{w}, \mathbf{x}) = \langle \mathbf{w}, \mathbf{x} \rangle$ — параметрических функций, линейных относительно параметров. Функция ошибки задана следующим образом

$$S = \sum_{i \in \mathcal{X}} (y_i - f(\mathbf{w}, \mathbf{x}_i))^2, \tag{2}$$

где $\mathcal{X} \subseteq \mathcal{I}$ —некоторое множество индексов. Требуется найти такое подмножество индексов $\mathcal{A} \subseteq \mathcal{J}$, которое бы доставляло минимум функции

$$\mathcal{A}^* = \arg\min_{\mathcal{A} \subseteq \mathcal{J}} S(f_{\mathcal{A}} | \mathbf{w}^*, \mathcal{D}_{\mathcal{C}}) \tag{3}$$

на множестве индексов $\mathcal{C}$. При этом параметры $\mathbf{w}^*$ модели должны доставлять минимум функции

$$\mathbf{w}^* = \arg\min_{\mathbf{w} \in \mathcal{W}} S(\mathbf{w} | \mathcal{D}_{\mathcal{L}}, f_{\mathcal{A}}) \tag{4}$$

на множестве индексов $\mathcal{L}$. Здесь $f_{\mathcal{A}}$ обозначает модель f, включающую только столбцы матрицы X с индексами из множества $\mathcal{A}$, а обозначение вида $S(\mathbf{w}|\mathcal{D})$ означает, что переменная $\mathcal{D}$ фиксирована, а переменная $\mathbf{w}$ изменяется.

Выбор признаков при прогнозировании

Процедура выбора оптимального набора признаков. Опишем два этапа алгоритма: Add и Del. На первом этапе последовательно добавляются признаки, согласно (4), доставляющие минимум S на обучающей выборке, заданной множеством индексов $\mathcal{L}$. На втором этапе происходит последовательное удаление признаков, согласно методу Белсли. Пусть на k-ом шаге алгоритма имеется активный набор признаков $\mathcal{A}_k \in \mathcal{J}$. На нулевом шаге $\mathcal{A}_0$ пуст.

Этап Add. Находим признак доставляющий минимум S на обучающей выборке

$$j^* = \arg\min_{j \in \mathcal{J} \setminus \mathcal{A}_{k-1}} S(\mathbf{w} | \mathcal{D}_{\mathcal{L}}, f_{\mathcal{A}_{k-1} \cup \{j\}}).$$

Затем добавляем новый признак j^* к текущему активному набору

$$\mathcal{A}_k = \mathcal{A}_{k-1} \cup \{j^*\}$$

и повторяем эту процедуру до тех пор, пока $S(f_{\mathcal{A}_k}|\mathbf{w}^*, \mathcal{D})$ превосходит свое минимальное значение на данном этапе не более, чем на некоторое заданное значение ΔS_1.

Этап Del. Находим индексы обусловленности и долевые коэффициенты для текущего набора признаков $\mathcal{A}_{k-1}$ согласно методу Белсли, описание которого приведено ниже. Далее находим количество достаточно больших индексов обусловленности. Достаточно большими будем считать индексы квадрат которых превосходит максимальный индекс обусловленности η_t, где $t = |A_{k-1}|$ количество признаков в текущем наборе $\mathcal{A}_{k-1}$.

$$i^* = \sum_{g=1}^{t} \left[\eta_g^2 > \eta_t\right]. \tag{5}$$

Затем ищем в матрице долевых коэффициентов $\mathbf{var}(\mathbf{w})$ столбец j^* с максимальной суммой по последним i^* долевым коэффициентам

$$j^* = \arg\max_{j \in \mathcal{A}_{k-1}} \sum_{g=t-i^*+1}^{t} q_g^j. \tag{6}$$

Удаляем j^*-ый признак из текущего набора

$$\mathcal{A}_k = \mathcal{A}_{k-1} \backslash j^*$$

и повторяем эту процедуру до тех пор, пока $S(f_{\mathcal{A}_k}|\mathbf{w}^*, \mathcal{D})$ превосходит свое минимальное значение на данном этапе не более, чем на некоторое заданное значение ΔS_2.

Повторение этапов Add и Del осуществляется до тех пор, пока значение $S(f_{\mathcal{A}_k}|\mathbf{w}^*, \mathcal{D})$ не стабилизируется.

Метод Белсли для удаления признаков. Рассмотрим матрицу признаков $\mathbf{X}$. Она имеет размерность $m \times n$. Выполним ее сингулярное разложение:

$$\mathbf{X} = \mathbf{U}\mathbf{\Lambda}\mathbf{V}^T,$$

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

$$\lambda_1 > \lambda_2 > \ldots > \lambda_r,$$

где r — ранг матрицы $\mathbf{X}$. Заметим, что в нашем случае $r = n$. Это связано с тем, что в алгоритме шагового выбора на каждом шаге мы имеем мультиколлиниарный, но невырожденный набор признаков. Столбцы матрицы $\mathbf{V}$ являются собственными векторами, а квадраты сингулярных чисел — собственными значениями матрицы $\mathbf{X}^T\mathbf{X}$.

$$\mathbf{X}^T\mathbf{X} = \mathbf{V}\mathbf{\Lambda}^T\mathbf{U}^T\mathbf{U}\mathbf{\Lambda}\mathbf{V}^T = \mathbf{V}\mathbf{\Lambda}^2\mathbf{V}^T,$$
$$\mathbf{X}^T\mathbf{X}\mathbf{V} = \mathbf{V}\mathbf{\Lambda}^2.$$

Отношение максимального сингулярного числа к j-му сингулярному числу назовем индексом обусловленности с номером j

$$\eta_j = \frac{\lambda_{\max}}{\lambda_j}.$$

Если матрица X неполноранговая, то значительная часть индексов обусловленности неопределено. Однако, в нашем случае, как упомяналось выше, матрица признаков X является матрицей полного ранга.

Так как модель линейна, то $\mathbf{w} = \mathbf{B}\mathbf{y}$, где $\mathbf{w}$ — вектор параметров модели. То есть $w_i = \mathbf{b}_i^T\mathbf{y}$, где

$$\mathbf{B} = \begin{pmatrix} \mathbf{b}_1^T \\ \dots \\ \mathbf{b}_n^T \end{pmatrix}.$$

Мы ищем несмещенную оценку параметров

$$\mathrm{E}(\mathbf{w}) = \mathbf{w} = \mathbf{B}\mathbf{X}\mathbf{w},$$

то есть $\mathbf{B}\mathbf{X} = \mathbf{I}$, где $\mathbf{I}$ — единичная матрица.

Тогда ковариация параметров w_i и w_j равна

$$cov(w_i, w_j) = \mathrm{E}(\mathbf{b}_i^T\mathbf{y} - \mathbf{b}_i^T\mathbf{X}\mathbf{w})(\mathbf{b}_j^T\mathbf{y} - \mathbf{b}_j^T\mathbf{X}\mathbf{w}) = \mathbf{b}_i^T\mathrm{E}((\mathbf{y} - \mathbf{X}\mathbf{w})(\mathbf{y} - \mathbf{X}\mathbf{w})^T))\mathbf{b}_j =$$

$$= \mathrm{E}(\xi_i\xi_j^T)\mathbf{b}_i^T\mathbf{b}_j = \sigma^2\mathbf{b}_i^T\mathbf{b}_j,$$

где ξ_i — i-ый регрессионный остаток, а σ^2 — дисперсия регрессионных остатков.

Мы хотим найти несмещенную оценку параметров, минимизирующую дисперсию параметров по каждой компоненте

$$\begin{cases} \sigma^2\mathbf{b}_i^T\mathbf{b}_i \longrightarrow \min\limits_{\mathbf{B}} \\ \mathbf{b}_i^T\mathbf{X} = \mathbf{e}_i^T \end{cases},$$

где $\mathbf{e}_i^T$ — i-ая строка единичной матрицы. Составим фунцию Лагранжа

$$Ł = \mathbf{b}_i^T\mathbf{b}_i + \mathbf{\Lambda}_i^T(\mathbf{X}^T\mathbf{b}_i - \mathbf{e}_i),$$

где $\mathbf{\Lambda} = (\mathbf{\Lambda}_1 \dots \mathbf{\Lambda}_n)$. Продифференцировав по $\mathbf{b}_i$, получим

$$\begin{cases} 2\mathbf{b}_i + \mathbf{X}\mathbf{\Lambda}_i \\ \mathbf{X}^T\mathbf{b}_i - \mathbf{e}_i = 0 \end{cases}$$

Из первого уравнения $\mathbf{b}_i = -\frac{1}{2}\mathbf{X}\mathbf{\Lambda}_i$, тогда $-\frac{1}{2}\mathbf{X}^T\mathbf{X}\mathbf{\Lambda}_i = \mathbf{e}_i$. То есть $\mathbf{\Lambda} = -2(\mathbf{X}^T\mathbf{X})^{-1}$, и, окончательно, для $\mathbf{B}$ получим

$$\mathbf{B} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T.$$

Для ковариационной матрицы $\mathbf{A}$ получим

$$\mathbf{A} = \sigma^2\mathbf{B}\mathbf{B}^T = \sigma^2(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{X}((\mathbf{X}^T\mathbf{X})^{-1})^T = \sigma^2\mathbf{X}^{-1}(\mathbf{X}^T)^{-1}\mathbf{X}^T\mathbf{X}((\mathbf{X}^T\mathbf{X})^{-1})^T =$$

$$= \sigma^2((\mathbf{X}^T\mathbf{X})^{-1})^T = \sigma^2(\mathbf{X}^T\mathbf{X})^{-1}.$$

В общем случае, выражение $\sigma^2(\mathbf{X}^T\mathbf{X})^{-1}$ является несмещенной оценкой ковариационной матрицы признаков, а в случае линейной модели оно в точности совпадает с ковариационной матрицей, то есть $\mathbf{A}^{-1} = \sigma^{-2}\mathbf{X}^T\mathbf{X}$.

Используя сингулярное разложение, дисперсия параметров, найденных методом наименьших квадратов $\mathbf{w} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}$, может быть записана как

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

Таким образом, дисперсия j-го регрессионного коэффициента — это j-й диагональный элемент матрицы $\mathbf{var}(\mathbf{w})$.

Для обнаружения мультиколлинеарности признаков построим таблицу, в которой каждому индексу обусловленности η_j соответствуют значения q_{ij} — долевые коэффициенты. Сумма долевых коэффициентов по индексу j равна единице.

$$\sigma^{-2}\mathbf{var}(w_i) = \sum_{j=1}^{n} \frac{\upsilon_{ij}^2}{\lambda_j^2} = (q_{i1} + q_{i2} + \ldots + q_{in}) \sum_{j=1}^{n} \frac{\upsilon_{ij}^2}{\lambda_j^2},$$

где q_{ij} — отношение соответствующего слагаемого в разложении вектора $\sigma^{-2}\mathbf{var}(w_i)$ ко всей сумме, а $\mathbf{V} = (v_{ij})$.

Таблица 1. Разложение $\mathbf{var}(\mathbf{w})$

Индекс обусловленности	$\mathbf{var}(w_1)$	$\mathbf{var}(w_2)$	$\ldots$	$\mathbf{var}(w_n)$
η_1	q_{11}	q_{21}	$\ldots$	q_{n1}
η_2	q_{12}	q_{22}	$\ldots$	q_{n2}
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$
η_n	q_{1n}	q_{2n}	$\ldots$	q_{nn}

Чем больше значение долевого коэффициента q_{ij} тем больший вклад вносит j-ый признак в дисперсию i-го регрессионного коэффициента.

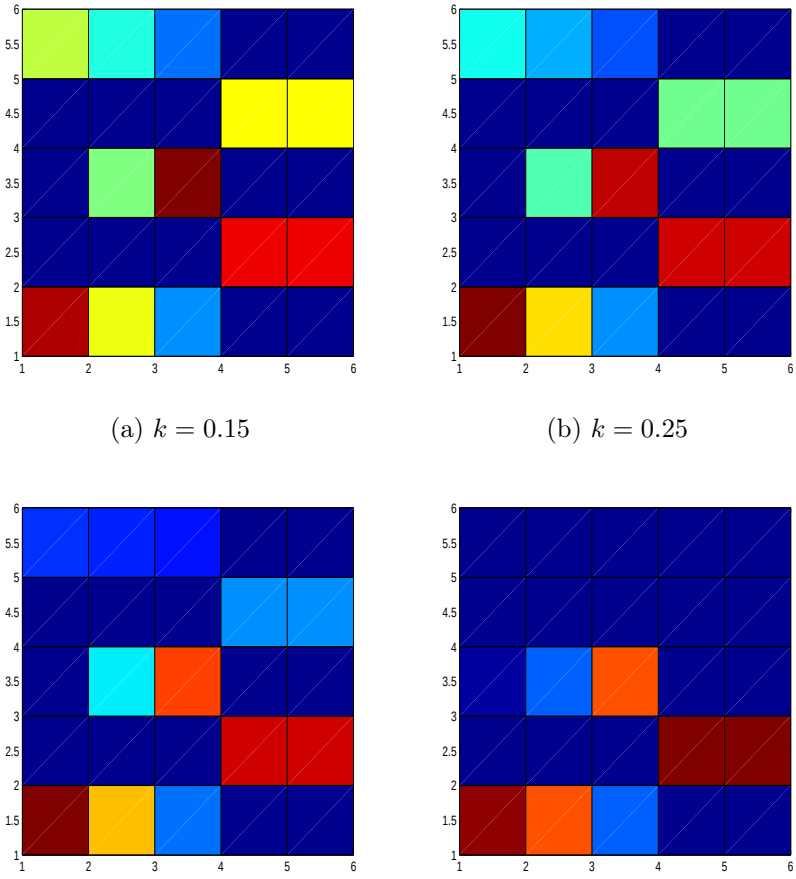
Из таблицы (1) определяется мультиколлинеарность: большие величины η_j означают, что, возможно, есть зависимость между признаками. Если присутствует только один достаточно большой индекс обусловленности, тогда возможно определение участвующих в зависимости признаков из долевых коэффициентов: признак считается вовлеченным в зависимость, если его долевой коэффициент связанный с этим индексом превышает выбранный порог (обычно 0.25). Если же присутствует несколько больших индексов обусловленности, то вовлеченность признака в зависимость определяется по сумме его дисперсионных долей, отвечающих большим значениям индекса обусловленности: если сумма превышает выбранный порог, то признак участвует как минимум в одной линейной зависимости. Для нахождения мультиколлиниарных признаков решаются задачи (5) и (6).

Проиллюстрируем метод Белсли на примере. Используется неизменные признаки x_1, x_5 и зависящие от параметра k признаки x_2, x_3, x_4. При $k = 0$ все признаки ортогональны, при увеличении k признаки x_2, x_3 приближаются к x_1, а x_4 — к x_5 вплоть до полной коллинеарности при $k = 1$. На рис. 1 приведены матрицы долевых коэффициентов в зависимости от k.

В таблице (2) приведены значения индексов обусловленности в зависимости от k.

Таблица 2. Индексы обусловленности

k	**0.15**	**0.25**	**0.4**	**0.9**
	1.0	1.0	1.0	1.0
	1.0	1.0	1.1	1.2
	1.1	1.2	1.5	21.5
	1.2	1.4	2.0	22.1
	1.2	1.5	2.1	24.0



(a) $k = 0.15$ (b) $k = 0.25$

(c) $k = 0.4$ (d) $k = 0.9$

Рис. 1. Матрицы долевых коэффициентов

Наблюдается две основных зависимости — первая между признаками x_1, x_2, x_3 и вторая между признаками x_4, x_5.

Подсчет матожижания и дисперсии функции ошибки

Функцию ошибки S можно при фиксированном наборе признаков $\mathcal{A} \in \mathcal{J}$ считать случайной величиной. Мы хотим минимизировать ее математическое ожидание и дисперсию при фиксированной сложности модели.

Сначала проведем эмпирический анализ наших реальных данных, а затем сравним полученные результаты с статистическими.

Эмпирический подход. Для данного набора признаков $\mathcal{A} \in \mathcal{J}$ будем многократно разбивать выборку на обучение $\mathcal{L}$ и контроль $\mathcal{C}$. Полученные значения функции ошибки S можно считать реализациями случайной величины. Тогда математическое ожидание и дисперсия оцениваются следующим образом

$$ES = \frac{1}{m}\sum_{i=1}^{m} S_i,$$

$$DS = \frac{1}{m}\sum_{i=1}^{m} (S_i - ES)^2,$$

где m — число разбиений выборки, а S_i — значение функции ошибки при i-ом разбиении.

Ниже представлены графики полученные по данным прогрессирования заболевания у больных диабетом. На нем отмечены все 2^n точек, где $n = 10$ — число признаков. По вертикали отложена дисперсия в логарифмическом маштабе, а по горизонтали количество признаков в наборе. При каждом значении числа признаков (сложности модели) найден набор с минимальным математическим ожиданием финкции ошибки — эти точки отмечены красным.

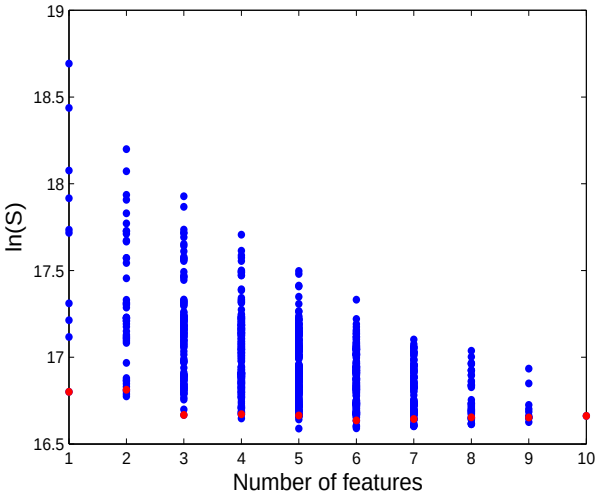


Рис. 2. Зависимость логарифма дисперсии функции ошибки от числа признаков в наборе при leave-one-out

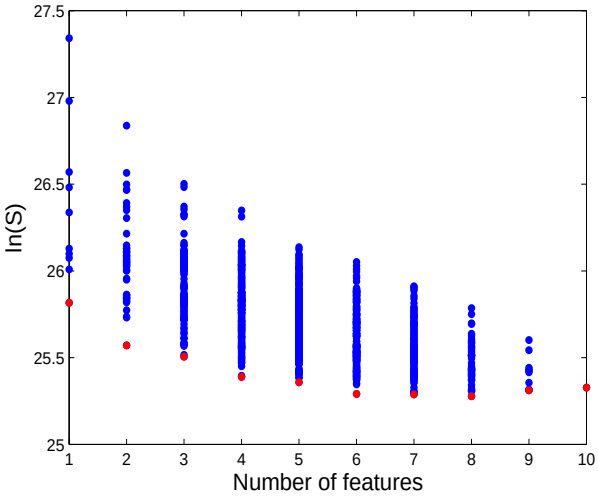


Рис. 3. Зависимость логарифма дисперсии функции ошибки от числа признаков в наборе при случайном разбиении выборки

По графикам видно, что у наборов с малым математическим ожиданием функции ошибки дисперсия тоже мала.

Статистический подход. Мы предполагаем, что данные нормальные, то есть

$$\mathbf{y} \sim \mathcal{N}(\mathbf{X}\mathbf{w}, \sigma^2\mathbf{I}),$$

где σ^2 — дисперсия регресиионных остатков, а $\mathbf{I}$ — единичная матрица.

При таких предположениях в методе наименьших квадратов функция ошибки $S(\mathbf{w}) =$ $=$ SSE имеет известное распределение

$$\frac{S}{\sigma^2} \sim \chi^2(m-n),$$

где m — число объектов (строк в матрице $\mathbf{X}$), а n — число признаков [15]. Из свойств распределения χ^2 получим

$$E\frac{S}{\sigma^2} = m - n,$$

$$D\frac{S}{\sigma^2} = 2(m-n).$$

То есть

$$ES = (m-n)\sigma^2,$$

$$DS = 2(m-n)\sigma^4,$$

причем σ^2 своя для каждого набора признаков.

Таким образом получаем, что математическое ожидание функции ошибки S достигает минимума при заданной сложности модели на том же наборе, на котором дисперсия достигает минимума. Этот результат экспериментально подтвердился на наших данных.

Путь в n-мерном кубе

В нашей задаче мы имеем дело с n признаками, то есть существует 2^n возможных наборов признаков, из которых мы пытаемся найти оптимальный. Все эти 2^n наборов

можно представить как вершины n-мерного куба. В данной работе используется шаговый алгоритм поиска оптимального набора, то есть пошагового движения по вершинам этого куба.

Приведем пример движения по вершинам куба при работе предложенного алгоритма. Всего использовалось 6 признаков $x_1, \ldots, x_6$, они изображены на рис.4. Также на нем показан вектор ответов y.

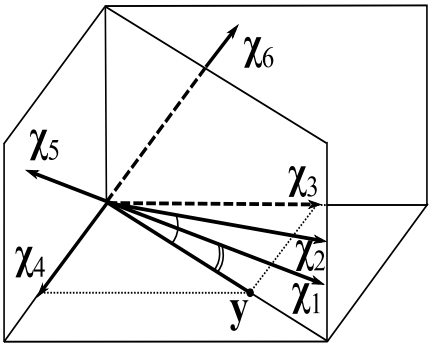


Рис. 4. Данные

На рис.5 показан путь по вершинам куба для описанных данных. По вертикали отложен номер признака, по горизонтали — номер итерации. Красная клетка означает, что признак на данной итерации вошел в набор, синяя — не вошел. Например признак номер 6 присутствовал в наборе с 3 по 8 итерацию, но в конечный набор не вошел.

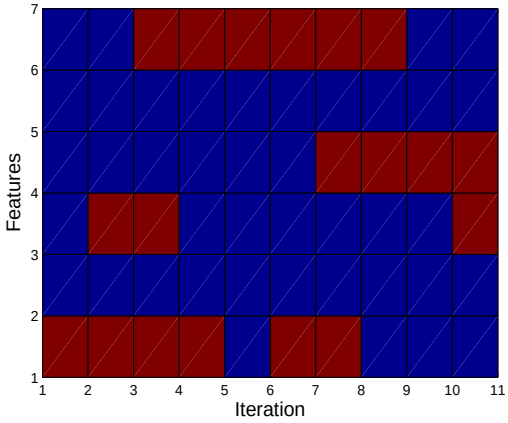


Рис. 5. Путь в кубе

Сформулируем и докажем некоторые теоретические утверждения, связанные с движением в кубе. Пошаговый алгоритм (1) выбора набора признаков.

Этап Add. Последовательно добавляются признаки, доставляющие минимум S

$$j^* = \arg \min_{j \in \mathcal{J} \setminus \mathcal{A}_{k-1}} S(\mathbf{w}|\mathcal{D}_{\mathcal{L}}, f_{\mathcal{A}_{k-1} \cup \{j\}}).$$

$$\mathcal{A}_k = \mathcal{A}_{k-1} \cup \{j^*\}$$

и повторяем эту процедуру до тех пор, пока $S(f_{\mathcal{A}_k}|\mathbf{w}^*, \mathcal{D})$ превосходит свое минимальное значение на данном этапе не более, чем на некоторое заданное значение ΔS_1.

Этап Del. Последовательно удаляем признаки, согласно методу Белсли

$$i^* = \sum_{g=1}^{t} \left[\eta_g^2 > \eta_t\right]$$

$$j^* = \arg\max_{j \in \mathcal{A}_{k-1}} \sum_{g=t-i^*+1}^{t} q_g^j$$

$$\mathcal{A}_k = \mathcal{A}_{k-1} \backslash j^*$$

и повторяем эту процедуру до тех пор, пока $S(f_{\mathcal{A}_k}|\mathbf{w}^*, \mathcal{D})$ превосходит свое минимальное значение на данном этапе не более, чем на некоторое заданное значение ΔS_2.

Повторение этапов Add и Del осуществляется до тех пор, пока значение $S(f_{\mathcal{A}_k}|\mathbf{w}^*, \mathcal{D})$ не стабилизируется.

Алгоритм (1) дает нам на выходе решение, которым является одна из вершин куба, то есть некоторый набор признаков.

Определение 1. Смежными будем называть вершины, одна из которых получается из другой добавлением одного признака.

Утверждение 1. Значение функции ошибки S в вершине-решении меньше, чем ее значение в любой смежной с ней вершине большей мощности.

Доказательство Предположим противное, и в одной из смежных вершин значение функции ошибки S меньше, чем в вершине-решении, тогда алгоритм (1) сделал бы еще один шаг добавления, но алгоритм остановился. Получили противоречие.

Определение 2. Путем в кубе будем называть последовательность вершин, соответствующую последовательности наборов признаков в пошаговом алгоритме (1). Первым членом последовательности всегда является пустой набор, а последним — решение полученное алгоритмом (1).

Определение 3. Сегментом пути назовем отрезок последовательности вершин, определенной данным путем в кубе.

Определение 4. Сегментом типа Add назовем сегмент, в котором все вершины получены добавлением нового признака.

Определение 5. Сегментом типа Del назовем сегмент, в котором все вершины получены удалением одного признака из набора.

Утверждение 2. Если путь в кубе имеет два одинаковых члена последовательности, принадлежащих некоторым сегментам одного типа, то решение полученное алгоритмом (1) равно одному из членов конечной последовательности, образованной уже пройденной частью пути.

Доказательство. Если мы попадаем на p-ом шаге в вершину в которой уже были на шаге t, причем на сегменте пути того же типа, то путь по кубу начиная с t-ого шага, совпадает с уже пройденным с p-го по t-ый шаг участком пути. Таким образом, решение, полученное алгоритмом (1), равно одному из членов конечной последовательности, образованной уже пройденной частью пути.

Утверждение 3. Значение функции ошибки S в вершине, являющейся решением, построенным с помощью алгоритма (1), меньше, чем ее значение в любой из вершин смежных с некоторой вершиной пути **a** и имеющей большее число признаков в наборе, чем вершина **a**.

Доказательство. Предположим противное, то есть значение функции ошибки S в вершине **b** смежной с вершиной пути **a**, принадлежащей сегменту типа Add меньше, чем

в вершине-решении. Но тогда вершина $\mathbf{b}$ принадлежит пути, и алгоритм (1) выдаст ее в качестве решения. Получили противоречие.

Утверждение 4. Два различных пути, проходящие через одну вершину на сегментах пути одного типа, при дальнешем движении по кубу совпадают.

Доказательство. Данное утверждение объясняется тем, что алгоритм (1) однозначно строит путь, выходящий из данной вершины и принадлежащий сегменту определенного типа.

Заключение

В работе предложен метод поиска оптимальной модели, основанный на комбинации двух стратегий: отбор признаков и выбор модели. Особенно полезен предложенный метод в случае, когда данные содержат большое число мультиколлинеарных признаков. Предложенный алгоритм позволяет получать хорошо обусловленные наборы порожденных признаков. В работе теоретически обоснованно, что математическое ожидание функции ошибки S достигает минимума при заданной сложности модели на том же наборе, на котором дисперсия достигает минимума. Этот результат так же подтвержден экспериментально на реальных данных.

Литература

[1] Frisch R. *Statistical Confluence Analysis by means of complete regression systems*, Universitetets Okonomiske Institute, 1934.

[2] Efron B., Hastie T., Johnstone I., Tibshirani R. *Least angle regression*, The Annals of Statistics, 2004, Vol. 32, no. 3., Pp. 407-499.

[3] Tibshirani R. *Regression shrinkage and selection via the lasso*, Journal of the Royal Statistical Society, 1996, Vol. 32, no. 1, Pp. 267-288.

[4] Draper N. R., Smith H. *Appied Regression Analysis*,John Wihley and Sons, 1998.

[5] Chen Y. W., Billings C. A., Luo W. *Orthogonal least squares methods and their application to non-linear system identication*, International Journal of Control, 1989, Vol. 2, no. 50, Pp. 873-896.

[6] 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.

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

[8] Rawlings J. O., Pantula S. G., Dickey D. A. *Appied Regression Analysis: A Research Tool*, New York: Springer-Verlag, 1998.

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

[10] Tarantola A. *Inverse Problem Theory*, SIAM, 2005.

[11] Johnstone I., Tibshirani R., Efron B., Hastie T. *Least Angle Regression*, 2004.

[12] McNames J. *Innovations in Local Modeling for Time Series Prediction*, 1999.

[13] Федорова В. П. *Локальные Методы Прогнозирования Временных Рядов*, 2009.

[14] Е. А. Крымова, В. В. Стрижов *Выбор моделей в линейном регрессионном анализе*, Информационные технологии, 2011.

[15] Г. И. Ивченко, Ю. И. Медведев *Введение в математическую статитсику*, ЛКИ, 2009.