

## Сглаживающие алгоритмы прогнозирования\*

*М. П. Кузнецов, А. А. Мафусалов, Н. К. Животовский, Е. Ю. Зайцев,  
Д. С. Сунгуров*

mikhail.kuznecov@phystech.edu, mafusalov@gmail.com,  
nikita.zhivotovskiy@phystech.edu, zaytsev.yevgen@gmail.com, elekt@bk.ru.  
Московский физико-технический институт, ФУПМ, каф. «Интеллектуальные системы»

Работа посвящена исследованию свойств сглаживающих алгоритмов прогнозирования и включает ряд кратких описаний и вычислительных экспериментов. В работу включены следующие алгоритмы: ядерное сглаживание, экспоненциальное сглаживание и непараметрическое сглаживание.

**Ключевые слова:** *непараметрическое сглаживание, прогнозирование, временной ряд.*

### Ядерное сглаживание

Целью проекта является прогноз временного ряда на несколько отсчетов методом ядерного сглаживания. Для достижения наилучшего качества прогноза используется выбор параметров модели. Решается задача восстановления регрессии. Задано пространство объектов  $X$  и множество возможных ответов  $Y = \mathbb{R}$ . Существует неизвестная целевая зависимость  $y^* : X \rightarrow Y$ , значения которой известны только на объектах обучающей выборки  $X^l = (x_i, y_i)_{i=1}^l$ . Требуется построить алгоритм  $a : X \rightarrow Y$ , аппроксимирующий целевую зависимость  $y^*$ .

**Пути решения задачи.**

**Определение 1.** *Гладкая, невозрастающая, ограниченная функция  $K : [0, +\infty) \rightarrow [0, +\infty)$  называется ядром.*

С помощью функции ядра зададим веса объектов:

$$w_i(x) = K\left(\frac{\rho(x, x_i)}{h}\right). \quad (1)$$

**Определение 2.** *Параметр  $h$  в формуле называется шириной окна сглаживания.*

Определим функционал качества:

$$Q(\alpha, X^l) = \sum_{i=1}^l w_i(x)(\alpha - y_i)^2 \rightarrow \min_{\alpha \in \mathbb{R}},$$

где  $w_i$  — веса обучающих объектов.

Приравняв к нулю  $\frac{\partial Q}{\partial \alpha} = 0$ , получим формулу Надарая-Ватсона:

$$a_h(x, X^l) = \frac{\sum_{i=1}^l y_i K\left(\frac{\rho(x, x_i)}{h}\right)}{K\left(\frac{\rho(x, x_i)}{h}\right)}.$$

Существует теорема, которая утверждает, что для широкого класса ядер оценка Надарая-Ватсона сходится к ожидаемому значению восстанавливаемой зависимости при неограниченном увеличении длины выборки  $l$  и одновременном уменьшении ширины окна  $h$ .

Научный руководитель В. В. Стрижов

**Выбор ширины окна и ядра.** Для выбора ширины окна производится минимизация функционала Leave-One-Out:

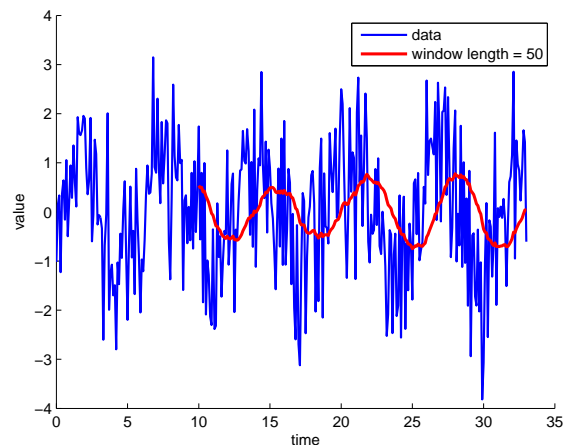
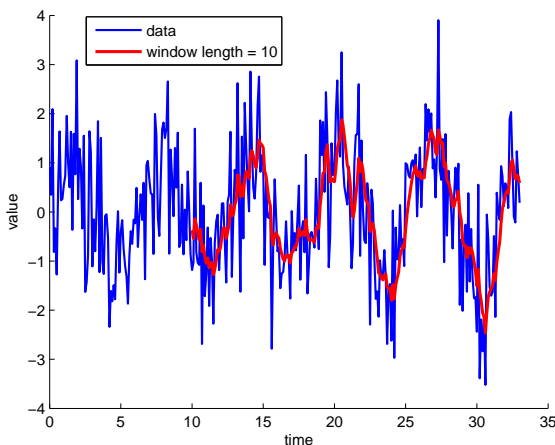
$$LOO(h, X^l) = \sum_{i=1}^l (a_h(x_i, X^l \setminus \{x_i\}) - y_i)^2 \rightarrow \min_h.$$

На практике используется несколько видов ядер:

|               |   |
|---------------|---|
| Квартическое  | $K(u) = \frac{15}{16}(1 - u^2)^2 I( u  \leq 1)$ |
| Епанечникова  | $K(u) = 0.75(1 - u^2) I( u  \leq 1)$            |
| Гаусса        | $K(u) = (2\pi)^{-1/2} \exp(-u^2/2)$             |
| Треугольное   | $K(u) = (1 -  u ) I( u  \leq 1)$                |
| Прямоугольное | $K(u) = 1/2 I( u  \leq 1)$                      |

Из 5 доступных ядер выбирается то, которое дает минимум SSE при оптимальной ширине окна.

**Результат работы алгоритма (одномерный случай).** Исходная зависимость — синус с гауссовским шумом. На левой картинке приведен результат для ширины окна, равной 10, на правой — для ширины окна, равной 50. Используемое ядро — квартическое.



## Прогнозирование временного ряда с помощью приближения производными рядами

Целью проекта является прогнозирование временного ряда на основании прогнозов для производных от него рядов.

**Постановка задачи.** Пусть дан временной ряд  $S = (s_1, \dots, s_T)$ , при значениях времени  $T = (t_1, \dots, t_T)$ . Пусть также даны производные от него ряды

$$\{S_i = (s_1^i, \dots, s_T^i)\}_{i=1}^n,$$

при тех же значениях времени, для которых существует прогноз

$$\{\bar{S}_i = (s_{T+1}^i, \dots, s_{T+\tau}^i)\}_{i=1}^n,$$

при значениях времени  $(t_{T+1}, \dots, t_{T+\tau})$ . Например, производными являются ряды, значения которых в каждой точке равны средним значениям исходного ряда в некоторой временной окрестности соответствующей точки.

Требуется построить прогноз  $\{\bar{S} = (s_{T+1}, \dots, s_{T+\tau})\}_{i=1}^n$  на интервале значений времени  $(t_{T+1}, \dots, t_{T+\tau})$ .

Будем строить прогноз как функцию прогнозов производных рядов:

$$\bar{s}_t = F(s_t^1, \dots, s_t^n), \quad F: \mathbb{R}^n \rightarrow \mathbb{R}.$$

Требуется найти  $\{\hat{F}^j(S_1, \dots, S_n)\}_{j=1}^{|P|} = \{F^{k_j}(\mathbf{w}, S_1, \dots, S_n)|_{\mathbf{w}=\mathbf{w}_j^*}\}_{k_j \in P}$ , где  $\hat{F} \in \mathfrak{F}$ ,  $\mathfrak{F}$  — множество  $n$ -местных непрерывных функций действительной переменной,  $F^k \in \mathfrak{F}_1$ ,  $\mathfrak{F}_1$  — множество функций из  $\mathfrak{F}$ , зависящих дополнительно от вектора параметров,  $k \in \mathfrak{K}$  — множество индексов функций множества  $\mathfrak{F}_1$ ,  $\mathbf{w}$  — настраиваемый вектор параметров,  $\mathbf{w} \in \mathfrak{W}(k)$ ,  $\mathfrak{W}(k)$  — множество допустимых векторов параметров функции  $F^k$ .

Множество  $\{F^{k_j}\}_{k_j \in P}$  — парето-оптимальное множество по совокупности критериев качества:

$$P = \{k \in \mathfrak{K} | POF(F^k) = 1\},$$

где  $POF(F^k)$  — номер парето-слоя, в котором лежит модель с индексом  $k$  и вектором параметров, настроенным по минимизации суммы квадратов регрессионных остатков на обучающей подвыборке:

$$\mathbf{w}^* = \arg \min_{\mathbf{w} \in \mathfrak{W}(k)} R_{J_l}(F^k).$$

Для каждой  $F^{k_j}$  вектор параметров находится как

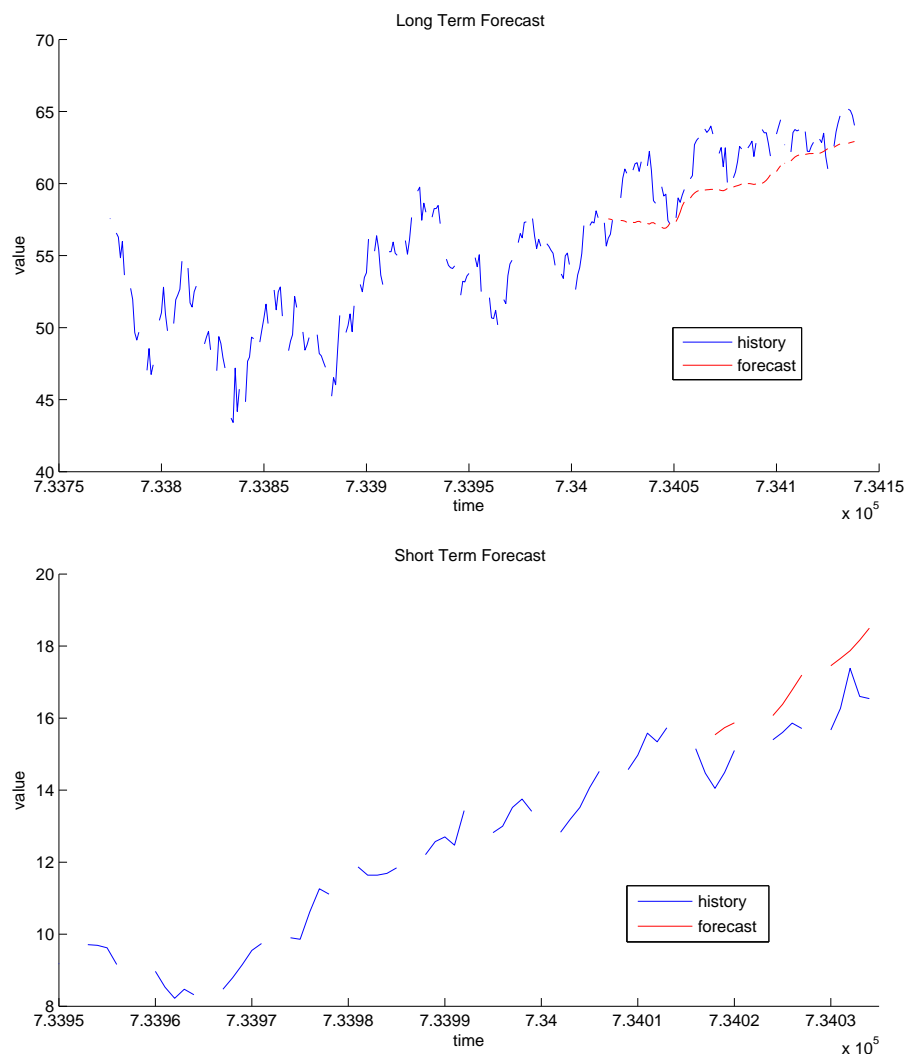
$$\mathbf{w}_j^* = \arg \min_{\mathbf{w} \in \mathfrak{W}(k_j)} R_{J_l}(F^{k_j}).$$

1. Найдём множество производных рядов  $\{S_i = (s_1^i, \dots, s_T^i)\}_{i=1}^n$  как значения функций от производного ряда:  $S_i = H_i(S)$ ,  $H_i: \mathbb{R}^n \rightarrow \mathbb{R}^n$ .
2. Продлим каждый из производных рядов каким-либо методом прогнозирования до момента времени  $T + \tau$ , получив прогнозы:  $\{\bar{S}_i = (s_{T+1}^i, \dots, s_{T+\tau}^i)\}_{i=1}^n$ .

**Метод решения задачи.** Опишем метод решения задачи, опирающийся на метод, решающий подзадачу.

1. Зададим алфавит функций:  $\{f_i(\mathbf{w}, \mathbf{x})\}_{i=1}^k$ , где  $\mathbf{x}$  — вектор-переменная,  $\mathbf{w}$  — вектор параметров, — для порождения суперпозиций.
2. Выберем подмножества производных рядов  $\{S_{i_1(\alpha)}, \dots, S_{i_{n(\alpha)}(\alpha)}\}_{\alpha \in A}$  и для каждого  $\alpha \in A$  построим суперпозиции  $F_1^\alpha, \dots, F_m^\alpha$ , приближающие исходный ряд и оптимальные в смысле постановки подзадачи.
3. Для каждого из производных рядов вычислим количество вхождений в суперпозиции по всему множеству полученных функций.
4. Выберем несколько производных рядов  $S_{i_1}, \dots, S_{i_d}$ , которые наибольшее число раз были использованы в оптимальных суперпозициях.
5. Решим подзадачу нахождения суперпозиций, приближающих ряд  $S$  производными рядами  $S_{i_1}, \dots, S_{i_d}$ , получим функции  $F_1, \dots, F_m$ .
6. На основании спрогнозированных значений производных рядов  $\{\bar{S}_i = (s_{T+1}^i, \dots, s_{T+\tau}^i)\}_{i=1}^n$  по функциям зависимости  $F_1, \dots, F_m$  получим прогнозы  $\{\hat{S}^j = (\hat{s}_{T+1}, \dots, \hat{s}_{T+\tau})\}_{j=1}^m$ .

**Результат работы алгоритма.** Ниже представлены результаты двух прогнозов на разных временных рядах, выполненных алгоритмом.



## Экспоненциальное сглаживание

Целью проекта является прогноз временного ряда на несколько отсчетов методом экспоненциального сглаживания. Для достижения наилучшего качества прогноза используется выбор параметра модели.

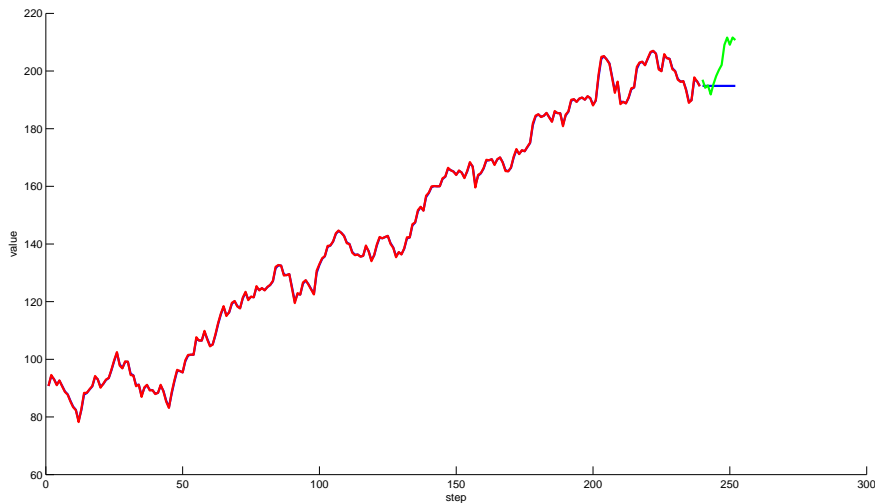
Решается задача прогнозирования временного ряда. Пусть задано пространство объектов  $X$  и задан временной ряд  $\{x_t\}_{t=1}^L$ ,  $x_t \in X$ , причем известными считаются первые  $N$  отсчетов. Требуется построить алгоритм  $a : \{1, \dots, N + D\} \rightarrow X$ , аппроксимирующий зависимость  $x_i$ ,  $i = \{1, \dots, N + D\}$ . Параметр  $D$  называется горизонтом прогнозирования.

Для решения этой задачи используется техника экспоненциального сглаживания.

**Определение 3.** Рядом экспоненциальных средних, построенным по временному ряду  $\{x_t\}_{t=0}^T$ , называется ряд  $\{S_t\}$ , построенный индуктивно:

$$S_t = \alpha x_t + (1 - \alpha)S_{t-1}, \quad t > 1.$$

В определении ряда  $S_0$  — выбранное начальное значение. Величина  $\alpha \in (0, 1)$  называется параметром сглаживания.



Значение  $S_{t-1}$  будем использовать как прогноз для  $x_t$ . Замкнутая форма для  $S_t$  получается преобразованием суммы:

$$S_t = \alpha \sum_{i=0}^N (1 - \alpha)^i x_{t-i} + (1 - \alpha)^N S_0,$$

где  $N$  — число членов ряда. Таким образом,  $S_t$  — взвешенная сумма членов ряда. Заметим, что чем «старше» член ряда, тем меньше его вес.

В задачах краткосрочного прогнозирования для повышения значимости «свежих» наблюдений параметр сглаживания следует увеличивать. Однако для сглаживания выбросов  $\alpha$  нужно уменьшать. Таким образом, задача сводится к поиску компромиссного значения параметра.

**Выбор значения  $S_0$ .** Значение  $S_0$  должно задаваться с учетом априорной информации. Часто используется среднее значение ряда или среднее значение наблюдений в предыстории, если она доступна. Также в качестве  $S_0$  может быть положено значение  $x_1$ .

В данной работе в качестве  $S_0$  используется среднее значение наблюдаемого ряда. Рассчитаем, сколько последних членов ряда вносят существенный вклад в прогноз. Пусть  $x_T$  — текущее наблюдение. Назовём возрастом наблюдения  $x_t$  относительно наблюдения  $x_T$  число  $T - t$ . В частности, возраст текущего наблюдения равен 0, предыдущего — 1.

**Определение 4.** Средним возрастом членов ряда называется взвешенная сумма возрастов с коэффициентами  $\{\alpha(1 - \alpha)^i\}_{i=0}^{\infty}$ .

Средний возраст сходится при длине ряда, стремящейся к бесконечности, к величине

$$0 \cdot \alpha + 1 \cdot \alpha(1 - \alpha) + 2 \cdot \alpha(1 - \alpha)^2 \dots = \frac{1 - \alpha}{\alpha}.$$

Значит, малые значения  $\alpha$  увеличивают средний возраст информации. Значение  $\alpha$  минимизирует Leave One Out Cross Validation:  $\alpha = \arg \min_{\alpha} \sum_{i=0}^{N-1} (S_i - x_{i+1})^2$ .

## Прогнозирование продаж групп товаров

Заданы временные ряды продаж товаров  $x_{ij}(t) \in \mathbb{R}$  — продажи  $i$ -го товара в  $j$ -том магазине за время  $t$  (где  $i \in I$ ,  $I$  — множество товаров;  $j \in J$ ,  $J$  — множество магазинов;

$t \in \mathbb{N}$ ). Значения продаж известны при  $t_0 \leq t \leq t_1$ . Также задан товарный классификатор, который разбивает товары на группы, образующие иерархическую структуру. Требуется спрогнозировать продажу заданного товара в заданном магазине на следующий день после  $t_1$ .

1. Найти все группы нижнего уровня — разбиение множества  $I$  на непересекающиеся подмножества  $I_k \subset I$ .

Для всех групп нижнего уровня  $I_k$  повторять следующие шаги.

2. Найти суммарные продажи товаров из  $I_k$  во всех магазинах

$$s_{ij}(w) = \sum_{t=t_1-w+1}^{t_1} x_{ij}(t),$$

где  $w$  - размер обучающей выборки.

Обозначим

$$S_w = \sum_{i \in I_k} \sum_{j \in J} s_{ij}(w).$$

3. Определить доли продаж отдельных товаров из группы

$$D_w(i) = \frac{1}{S_w} \sum_{j \in J} s_{ij}(w).$$

Повторять для всех  $j \in J$  и всех групп нижнего уровня  $I_k$  следующие шаги.

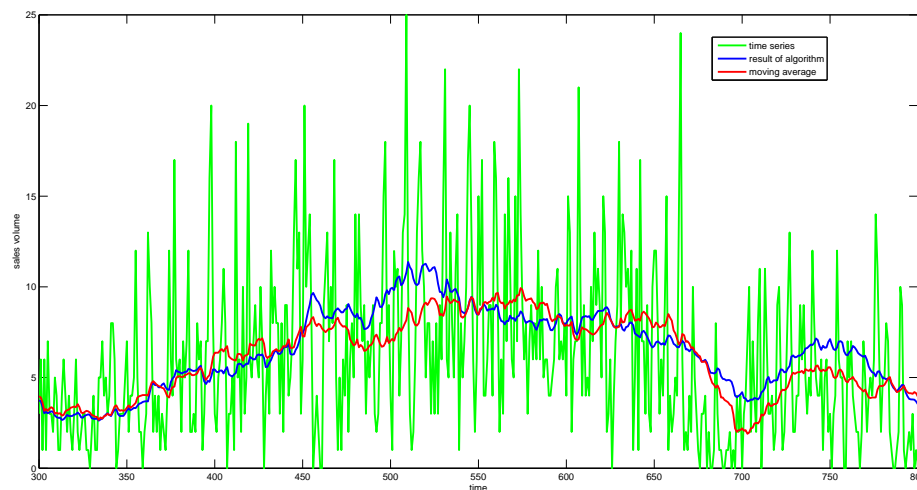
4. Вычислить по методу скользящего среднего прогноз продаж товаров из  $I_k$  в магазине  $j$  на следующий день:

$$S_{vj}(I_k) = \frac{1}{v} \sum_{i \in I_k} s_{ij}(v).$$

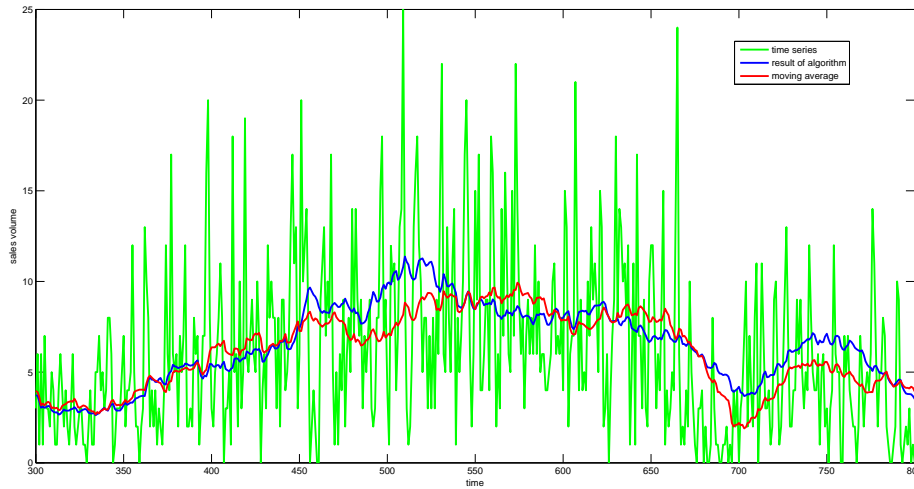
5. Распределить предсказанное число товаров согласно величинам  $D_w(i)$ :

$$\overline{y}_{ij} = S_{vj}(I_k) D_w(i).$$

Результат работы алгоритма приведен на графике.



Также применим данный алгоритм не к товарам, а к группам товаров нижнего уровня



На обоих графиках синяя линия прогноза описанного выше алгоритма и красная линия прогноза методом скользящего среднего опережают график продаж (зеленая кривая) на один шаг.

## Выбор моделей в задачах прогнозирования

**Постановка задачи и алгоритм решения.** Решается задача восстановления регрессии. Задана выборка  $D = \{(\xi^i, y^i)\}_{i=1}^m$  — множество пар вектора значений  $n$  свободных переменных  $\xi^i = (\xi_u^i)_{u=1}^U \in \mathbb{R}^U$  и значения одной зависимой переменной  $y^i \in \mathbb{R}^1$ . Задано множество порождающих функций  $G = \{g_v\}_{v=1}^V$ .

Обозначим  $a_i = g_v(\xi_u)$ , где  $i = (v-1)U + u$ . Рассмотрим декартово произведение  $\Xi \times G$  ( $\Xi$  — множество свободных переменных). Элементу  $(g_v, \xi_u)$  поставим в соответствие суперпозицию  $g_v(\xi_u)$ , однозначно определяемую индексами  $u, v$ .

В качестве модели, описывающей отношение между зависимой переменной  $y$  и свободными переменными  $a_i$ , используется полином Колмогорова-Габора:

$$y = \beta_0 + \sum_{l=1}^{UV} \beta_l a_l + \sum_{l=1}^{UV} \sum_{\zeta=1}^{UV} \beta_{l\zeta} a_l a_\zeta + \dots, \quad (2)$$

где  $\beta = (\beta_0, \beta_l, \beta_{l\zeta}, \dots)_{l,\zeta=1,2,\dots}$  — вектор коэффициентов.

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

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

Переменные  $x_i$  поставлены в однозначное соответствие мономам полинома. В матричной форме соотношение имеет вид:

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

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

Предполагается, что векторы  $x_j = (x_j^1, \dots, x_j^m)$  линейно независимы. Линейно зависимые вектора исключаются из дальнейшего рассмотрения. Для обнаружения мультиколлинеарности используется метод Белсли [4].

Пусть матрица признаков  $X$  имеет размер  $m \times n$ , центрирована и нормирована. Проводится сингулярное разложение:

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

где  $U, V$  — ортогональные матрицы,  $\Lambda$  — диагональная матрица с сингулярными числами на диагонали, упорядоченными по убыванию.

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

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

Ковариационная матрица метода наименьших квадратов может быть записана следующим образом:

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

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

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

где  $q_{ij}$  — дисперсионные доли, т.е. отношения соответствующего слагаемого в разложении  $\sigma^{-2} V(\beta_i)$  ко всей сумме.

Большие значения индексов обусловленности означает наличие мультиколлинеарности. Большие значения  $q_{ij}$ , соответствующие большим индексам обусловленности относятся к признакам, между которыми эта зависимость существует. Алгоритм отбрасывает признаки, находящиеся в зависимости с индексами обусловленности превышающими заданное значение.

Отбор признаков осуществляется одним из следующих методов

1. Полный перебор моделей;
2. Генетический алгоритм;
3. Метод группового учета аргументов;
4. Алгоритм Лассо;
5. LARS;
6. Последовательное добавление признаков с ортогонализацией;
7. Случайный поиск с адаптацией;
8. Гребневая регрессия;
9. Поиск самой длинной проекции вектора ответов на биссекторы векторов признаков.

По отобранным признакам выполняется восстановление линейной регрессии.



## Литература

- [1] В. Хардле *Прикладная непараметрическая регрессия*, 1989.
- [2] К. В. Воронцов *Лекции по алгоритмам восстановления регрессии*, 2009.
- [3] Магнус Я.Р., *Эконометрика. Начальный курс: Учеб.* - 6 изд. - М.: Дело, 2004. - 576 с.
- [4] Belsley, David A. *A Guide to using the collinearity diagnostics*, Computational Economics, 1991, Vol.4, pp.33-50.