Исследование устойчивости оценок ковариационной матрицы признаков*

А. А. Зайцев

alexey.zaytsev@datadvance.net

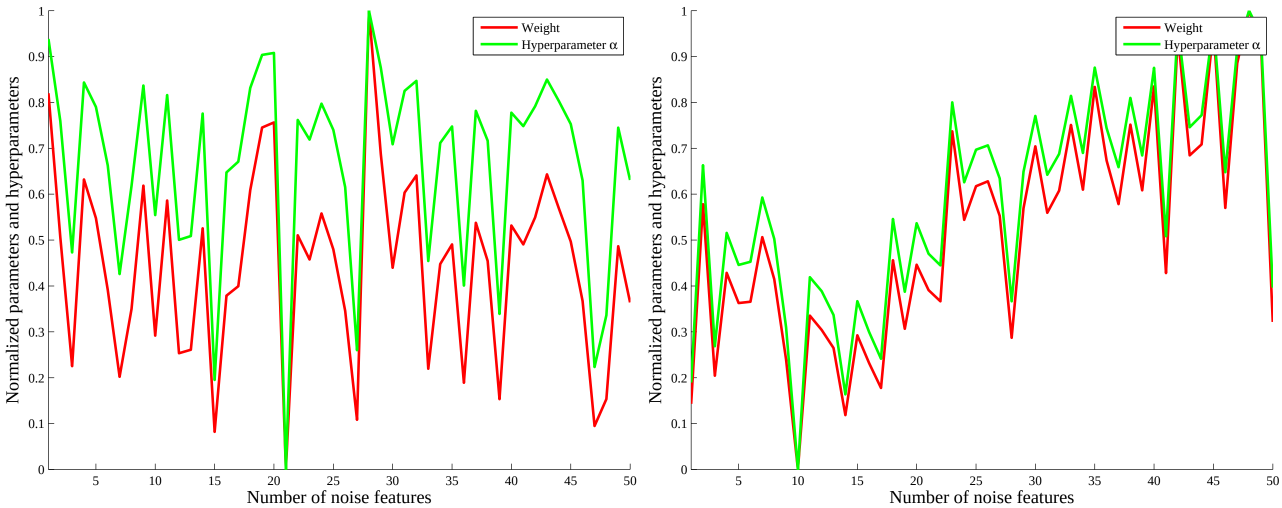
Московский физико-технический институт, ФУПМ, каф. «Интеллектуальные системы»

В данной работе исследуется устойчивость оценок ковариационной матрицы параметров модели. Рассматриваются модели линейной и существенно нелинейной регрессии. Тогда вектор параметров модели соответствует набору признаков модели. Ковариационная матрица параметров строится в предположении о вероятностном распределении вектора параметров. Исследуется, зависит ли оценка ковариационной матрицы признаков от того, являются ли признаки мультикоррелирующими и шумовыми. Для такой матрицы плана получаем расширенный вектор параметров модели и оценку матрицы ковариации параметров модели. Сравнивается ковариационная матрица для нерасширенного и расширенного вектора параметров модели. Исследуется пространство параметров для информативных признаков. Эксперименты проводятся на реальных и модельных данных.

Ключевые слова: регрессионный анализ, линейная регрессия, символьная регрессия, оценка гиперпараметров.

Введение

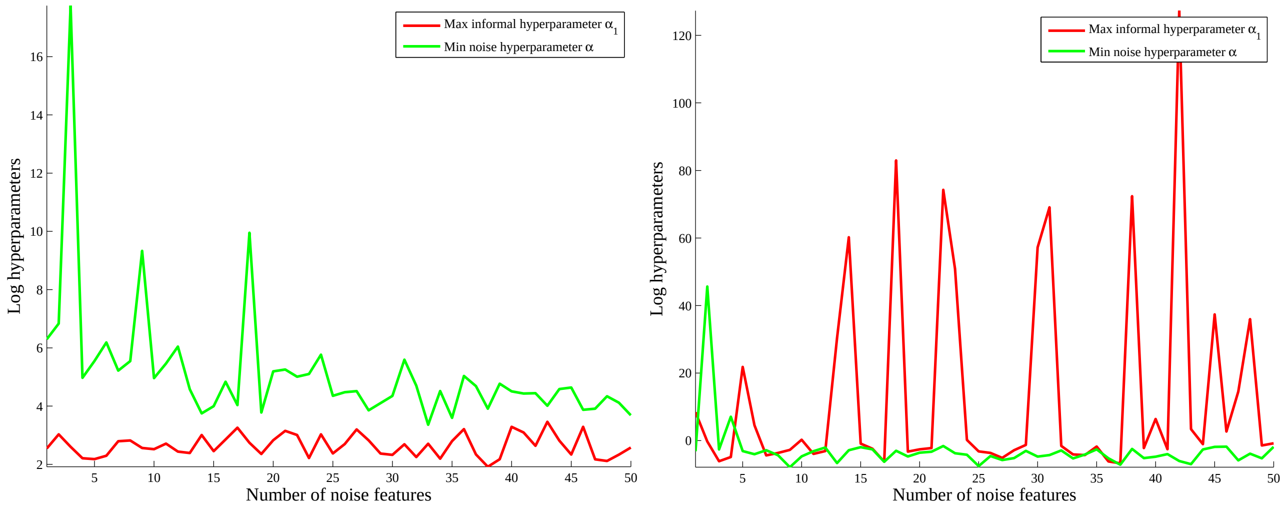
В данной работе рассматривается алгоритм выбора модели и настройки параметров модели линейной и существенно нелинейной регрессии, описанный в работе [1] для линейной и в работе [3] для существенно нелинейной регрессии.



(a) Линейная функция (b) Экспонента

Рис. 1. Зависимость параметров и гиперпараметров от числа шумовых признаков

Научный руководитель В. В. Стрижов



(a) Линейная функция (b) Экспонента

Рис. 2. Гиперпараметры для шумовых и информативного признака

Постановка задачи

Задана выборка $D = (X, \mathbf{y}) = \{(\mathbf{x}_i, y_i)\}_{i=1}^m$. Вектор свободных переменных $\mathbf{x} \in \mathbb{R}^n$, зависимая переменная $y \in \mathbb{R}$. Предполагается, что

$$y = f(\mathbf{x}, \mathbf{w}) + \varepsilon, \tag{1}$$

где $f(\mathbf{x}, \mathbf{w})$ — некоторая параметрическая функция, $\mathbf{w} \in W$ — вектор ее параметров, ε — ошибка, распределенная нормально с нулевым математическим ожиданием и дисперсией β, $\varepsilon \sim \mathcal{N}(0, \beta)$. Предполагается, что вектор параметров $\mathbf{w}$ — распределенный нормально случайный вектор с нулевым математическим ожиданием и матрицей ковариаций A.

Рассматривается класс линейных функций $f(\mathbf{x}, \mathbf{w})$. Наиболее вероятные параметры $\mathbf{w}_{MP}$ имеют вид:

$$\mathbf{w}_{MP} = \arg\max_{\mathbf{w}} p(\mathbf{w}|D, A, \beta, f). \tag{2}$$

Для такого набора параметров исследуется матрица ковариации A, который мы тоже оцениваем, используя принцип максимального правдоподобия.

Описание алгоритма оценки матрицы ковариации

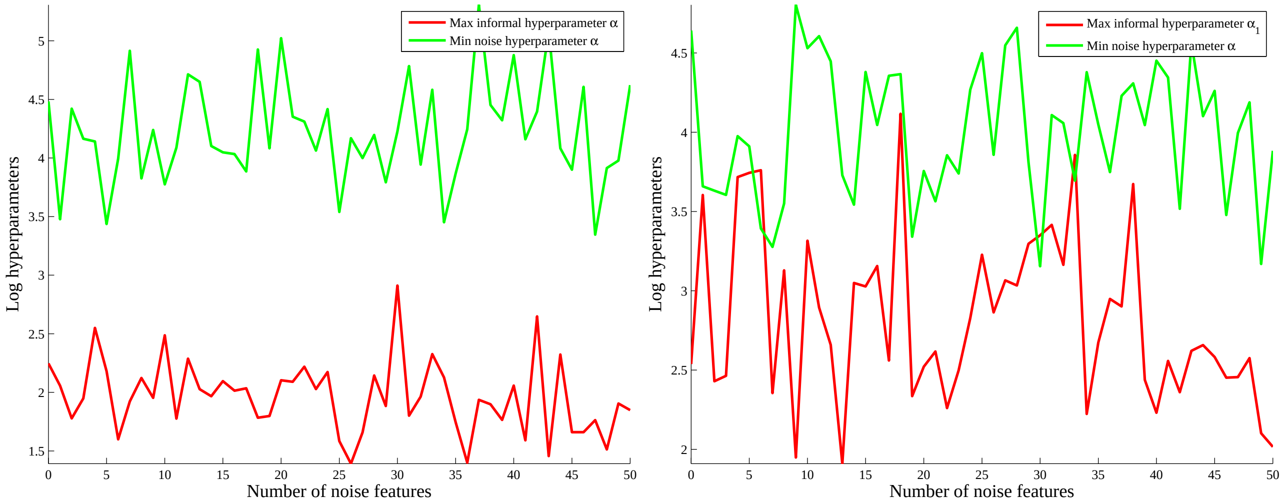
Для фиксированных гиперпараметров A, β вектор наиболее вероятных параметров минимизирует функционал

$$S(\mathbf{w}) = \mathbf{w}^T A \mathbf{w} + \beta \sum_{i=1}^{n} (y_i - \mathbf{x}_i^T \mathbf{w})^2 = E_{\mathbf{w}} + \beta E_D. \tag{3}$$

Набор наиболее вероятных гиперпараметров будем искать, максимизируя оценку правдоподобия по A, β

$$\ln p(D|A, \beta, f) = -\frac{1}{2} \ln |A| - \frac{m}{2} \ln 2\pi + \frac{m}{2} \ln \beta \underbrace{- E_{\mathbf{w}} - \beta E_D}_{S(\mathbf{w}_0)} - \frac{1}{2} \ln |H|, \tag{4}$$

здесь H — гессиан функционала (3).



(a) Линейная функция (b) Синусоида

Рис. 3. Гиперпараметры для шумовых и информативных признаков

В предположении о диагональности матрицы $A = diag(\boldsymbol{\alpha})$ и гессиана $H = diag(\mathbf{h})$, $\boldsymbol{\alpha} = \{\alpha_i\}_{i=1}^m$, $\mathbf{h} = \{h_i\}_{i=1}^m$, приравняв производные по гиперпараметрам к нулю, получаем оценку для α_i

$$\alpha_i = \frac{1}{2}\lambda_i \left(\sqrt{1 + \frac{4}{w_i^2\lambda_i}} - 1\right), \tag{5}$$

здесь $\lambda_i = \beta h_i$.

Так же получаем оценку β

$$\beta = \frac{n-\gamma}{2E_D}, \tag{6}$$

здесь

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

Используя оценки вектора параметров при фиксированных гиперпараметрах и гиперпараметров при фиксированных параметрах, выпишем итерационный алгоритм поиска наиболее вероятных параметров и гиперпараметров. Он состоит из шагов:

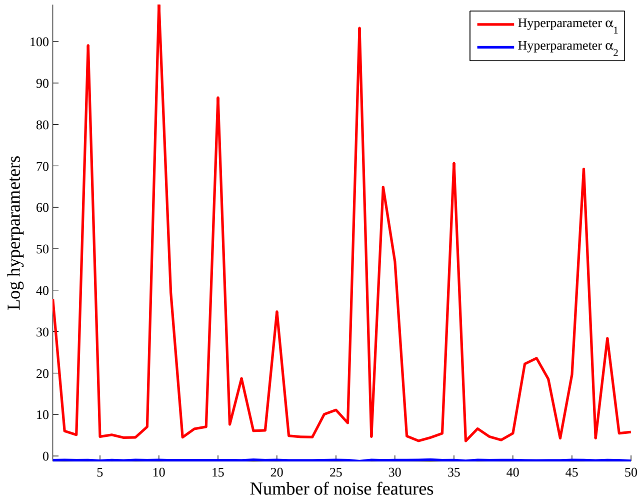
— поиск вектора параметров, максимизирующих (3),
— поиск гиперпараметров, максимизирующих правдоподобие (4),
— проверка критерия остановки.

Критерий остановки — малое изменение функционала (3) для двух последовательных итераций алгоритма.

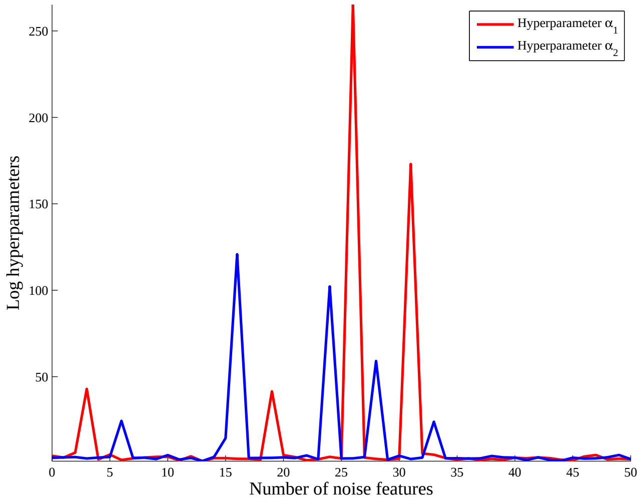
Вычислительный эксперимент: шумовые признаки

В вычислительном эксперименте исследовалась устойчивость оценок гиперпараметров при добавлении шумовых и мультиколлиенарных признаков для линейной и существенно нелинейной регрессии.

Шумовые признаки:один признак. В выборках один информативный признак и n' шумовых. Вектор свободных переменных для каждого объекта генерируется из нормального распределения с нулевым математическим ожиданием и единичной дисперсией. Рас-



(a) Линейная функция



(b) Синус

Рис. 4. Гиперпараметры для шумовых и информативных признаков

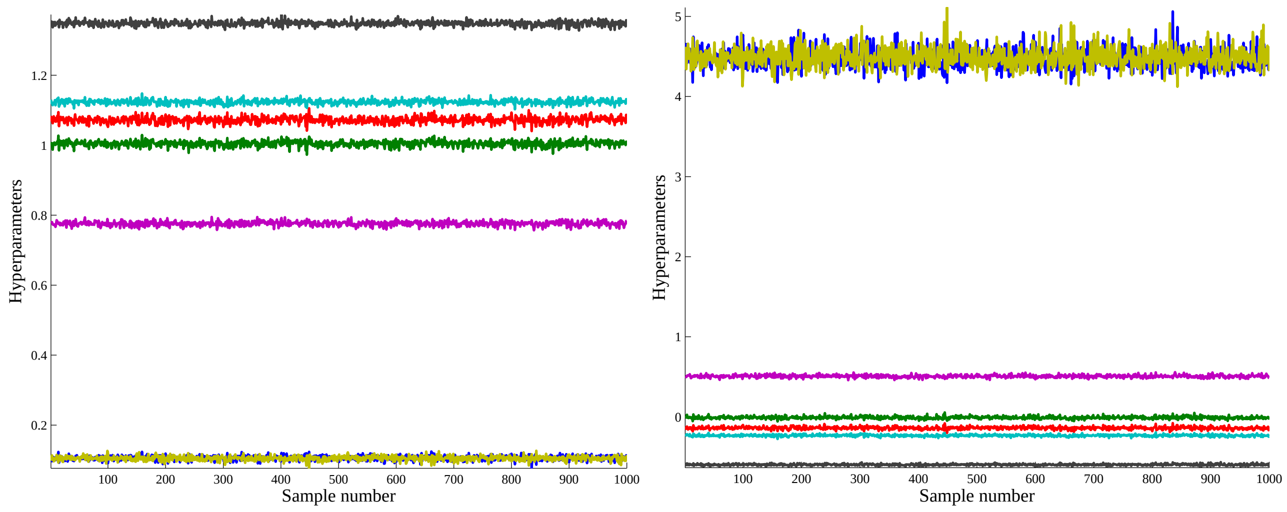
сматриваются выборки размером 100 и 1000. Зависимая переменная — зашумленная линейная или обобщенно-линейная функция входа. Рассматривались обобщенные-линейные функции $y = \exp(-\mathbf{w}^T\mathbf{x})$ и $y = \sin(\mathbf{w}^T\mathbf{x})$. Шум состоял из независимых нормальнораспределенных величин с дисперсией $\frac{1}{4}$.

Зависимость параметра от гиперпараметров. На рисунках приведена зависимость параметра w и гиперпараметра α, которые соответствуют нешумовому признаку. Мы видим, что параметр сильно коррелирует с гиперпараметром, при этом, нет зависимости от числа шумовых признаков.

Сравнение гиперпараметров для разных признаков. Гиперпараметры α_i могут [2] служить мерой информативности признаков. Сравнивались логарифм гиперпараметра значимого признака и минимальный из логарифмов гиперпарамтеров для незначимых признаков. Бралось усреднение логарифма по пяти различным выборкам. Результаты приведены на рисунках 2. На рисунке 2 видно, что в большинстве случаев значение гиперпараметра для значимого признака меньше, чем минимальное значение гиперпараметров для шумового, однако, в некоторых случаях наблюдаются выбросы.

Проводился аналогичный эксперимент для двух информативных признаков, причем сравнивался максимальное значение гиперпараметра для информативных признаков с минимальным значением признака для шумовых признаков. На рисунках 3 видно, что информативные признаки имели меньшие значения гиперпараметра α, чем информативные. Таким образом, удается выделить информативные и шумовые признаки. На рисунке 4показано сравнение информативности первого и второго информативных признаков, видно, что из-за большего веса один признак информативнее другого для линейной модели. Так же отметим, что для обобщенно-линейной функции не удается выделить наиболее информативный признак, в некоторых случаях гиперпараметры для одного из признаков стремятся к бесконечности.

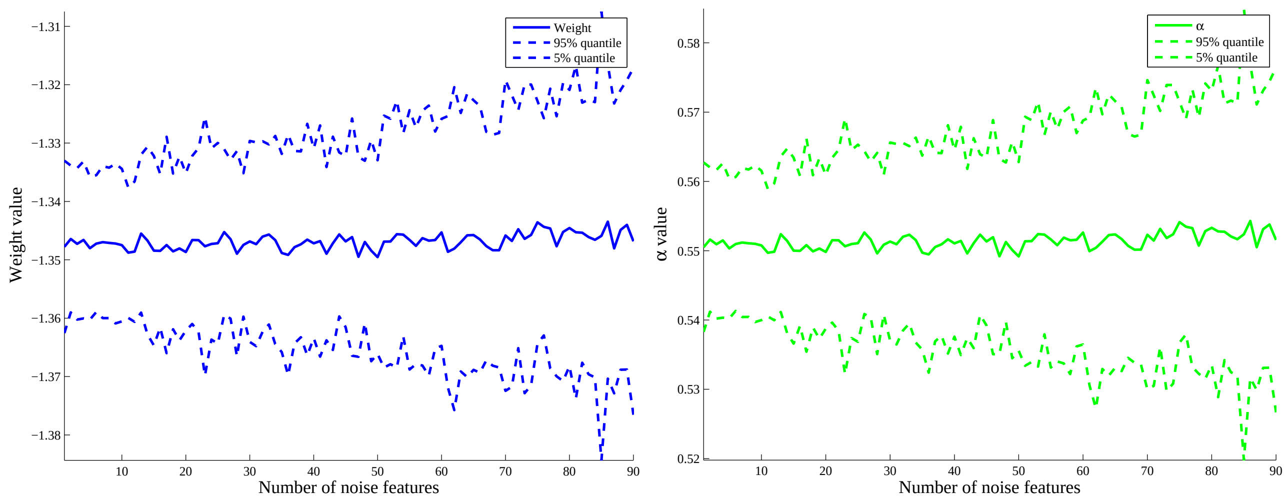
Реальные данные. Использовались реальные данные по определения характеристик цемента по его составу [4]. Данные были нормализованы так, что как у свободных, так и у зависимой переменной были нулевые математические ожидания и единичные дисперсии. Для данных без шумовых признаков алгоритм был запущен сто раз на разных подвы-



(a) Веса w_i (b) Гиперпараметры α_i

Рис. 5. Веса и гиперпараметры для выборки без шумовых признаков

борках размера 90 (размер полной выборки — 103). Результаты приведены на рисунке 5. Видно, что признаки разделяются по информативности и что информативность почти всегда эквивалента модулю веса.



(a) Вес w_7 (b) Гиперпараметр α_7

Рис. 6. Зависимость квантили оценки параметров и гиперпараметров при добавлении шумовых признаков

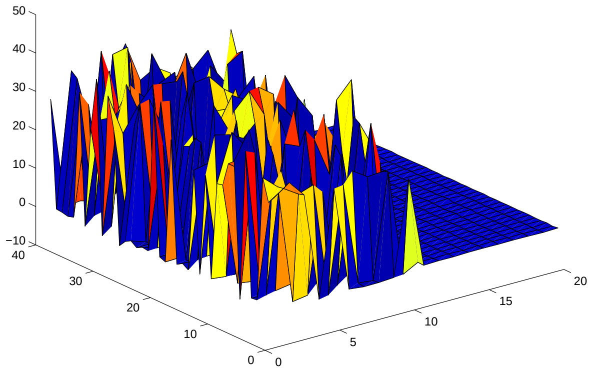
Так же был проведен следующий эксперимент. К начальному набору свободных переменных был добавлен ряд шумовых признаков, затем на ста запусках была оценена 95-процентная квантиль рассматриваемой величины. На рисунке 6 видно, что увеличение числа шумовых признаков увеличивает, хоть и не сильно, квантиль как оценки параметра, так и оценки гиперпараметра для разных признаков. Отметим, что, тем не менее, это не влияет на разделимость признаков по информативности.

Вычислительный эксперимент: мультиколлинеарные признаки

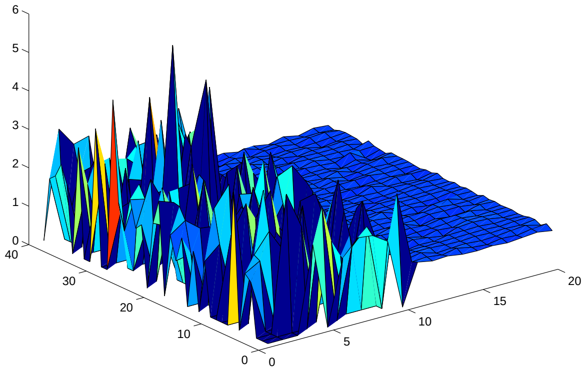
Модельные данные . Рассматривался следующий набор данных. Была сгенерирована выборка из нормального распределения размером 100 точек, количество признаков — двадцать. Ковариационная матрица первых десяти признаков имела вид:

$$\begin{pmatrix} 1.1 & 1 & 1 & \dots & 1 \\ 1 & 1.1 & 1 & \dots & 1 \\ \dots & \dots & \dots & \dots & \dots \\ 1 & 1 & 1 & \dots & 1.1 \end{pmatrix}$$

Ковариационная матрица для последних десяти признаков была единичной. Первая и вторая десятки признаков были порождены независимо.



(a) Гиперпараметры α_i



(b) Параметры w_i

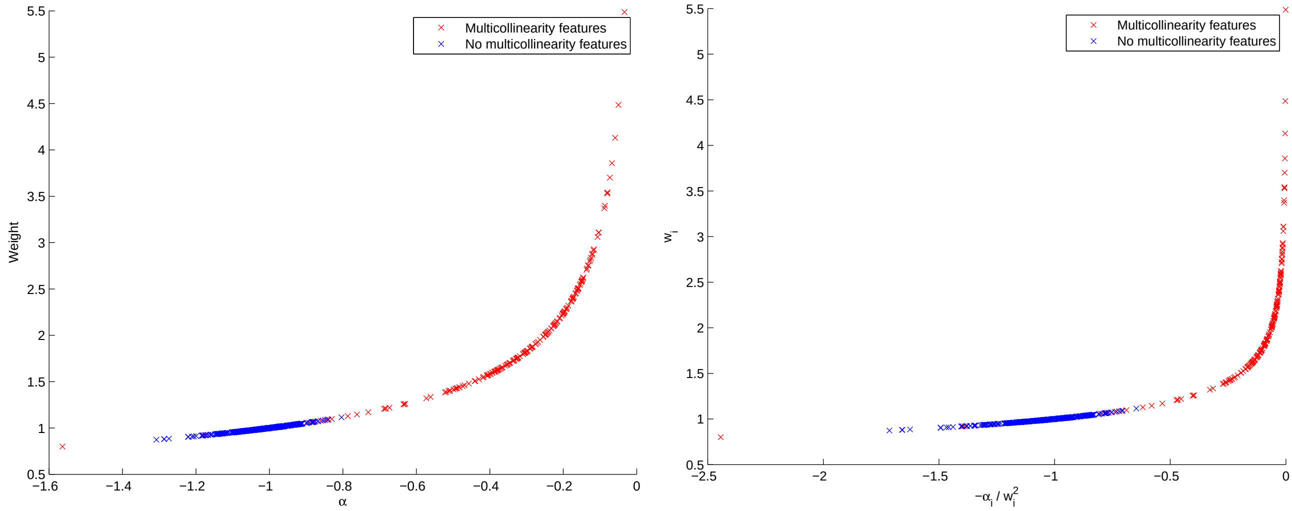
Рис. 7. Полученные значения параметров и гиперпараметров

Вектор откликов имел вид:

$$y = \sum_{i=1}^{n} x_i.$$

Было сделано 50 запусков эксперимента. Полученные значения логарифмов гиперпараметров α_i и параметров w_i изображены на рисунке 7. Ближе к читателю расположены оценки, полученные для коррелирующих признаков, дальше — для не коррелирующих признаков. Видно, что для признаков, не являющихся мультиколлинеарными, оценки значений гиперпараметров и параметров мало зависят от обучающей выборки. В то же время, для мультиколлинеарных признаков значения гиперпараметров и параметров сильно менялись от запуска к запуску.

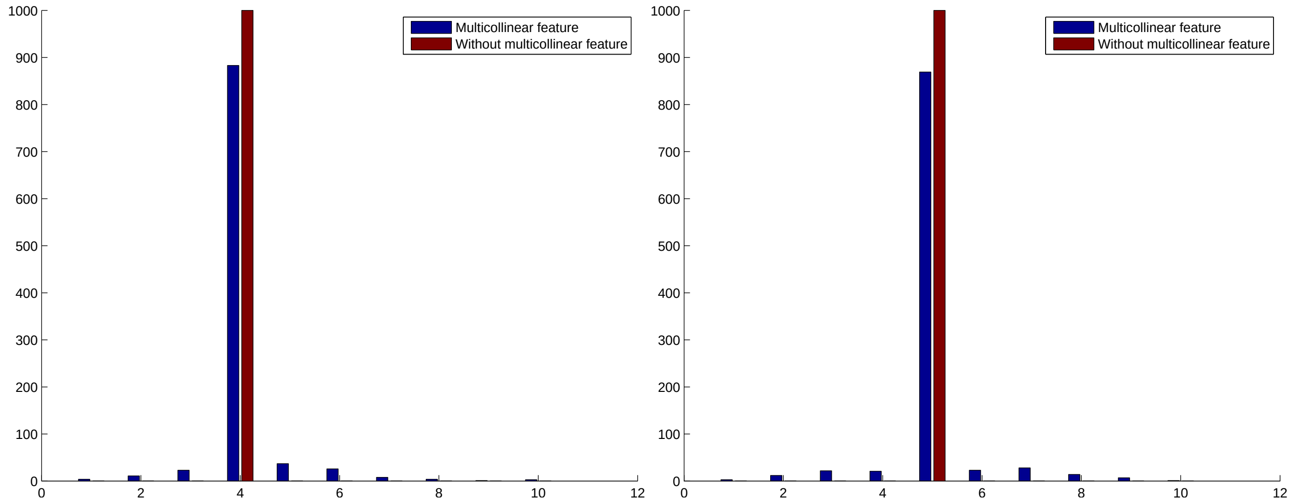
Для признаков с ненулевыми весами была построена кривая зависимости значений параметров от гиперпараметров (отметим, что истинное значение всех параметров равно единице). Полученная кривая приведена на рисунке 8 для тех признаков, параметры которых больше нуля. Мы видим, что при нормализации гиперпараметра α_i на w_i^2 признаки разделяются на две группы, в которых примерно одинаковые информативности. Таким образом, алгоритм верно классифицировал, что информативность признака, связанного с другими признаками посредством корреляции выше, чем информативность независимых



(a) Зависимость w_i от α_i (b) Зависимость w_i от $\frac{\alpha_i}{w_i^2}$

Рис. 8. Зависимость значения параметра w_i от гиперпараметра α_i

признаков. Отметим так же, что для некоторых признаков вес получался равным нулю. Все такие признаки принадлежали группе мультиколлинеарных.



(a) Первый признак (b) Второй признак

Рис. 9. Сравнительные гистограммы значений гиперпараметров

Реальные данные. Использовались реальные данные [4]. К данным добавлялся признак, сильно коррелирующий с одним из предложенных. Такой признак равнялся зашумленному стандартным нормальным шумом признаку. Сравнительные гистограммы значений гиперпараметров приведены на рисунке 9. Видно, что добавление мультикоррелирующего признака влияет на значение информативности исходного признака.

Существенно нелинейная регрессия. В этом эксперименте порождались модели существенно нелинейной регрессии [2, 3], затем рассматривалось распределение параметров и гиперпараметров для полученных моделей. Размер обучающей выборки — 10000

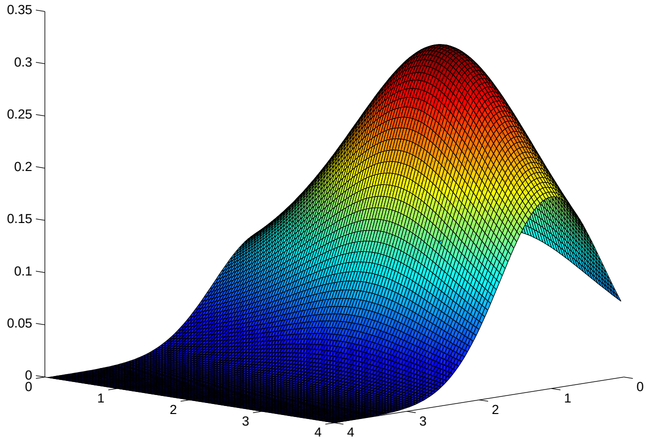


Рис. 10. Функция Котанчека

точек, делалась попытка аппроксимации функции, предложенной Котанчеком

$$f(x_1, x_2) = \frac{e^{-(x_1-1)^2}}{(x_2-2.5)^2+3.2}.$$

Вид функции показан на рисунке 10. Полученное распределение значений параметров и гиперпарамтеров — на рисунке 11. Видно, что значения параметров для разных моделей получают похожие значения, не зависимые от значения гиперпараметра. Это связано с линейным членом нелинейной модели, который появляется достаточно часто в функциях, точно приближающих искомую зависимость.

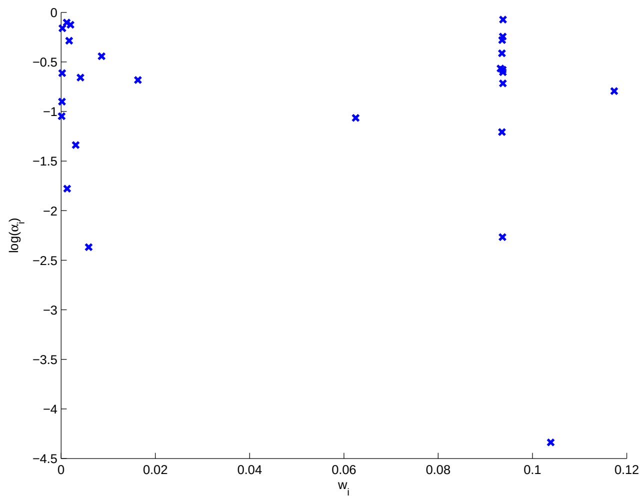


Рис. 11. Зависимость значения параметра w_i от гиперпараметра α_i для существенно нелинейных моделей

Выводы

Полученные результаты говорят о том, что предложенный подход является устойчивым к добавлению шумовых и мультиколлинеарных признаков.

Литература

[1] В. В. Стрижов, Р. А. Сологуб, *Индуктивное порождение регрессионных моделей предполагаемой волатильности для опционных торгов*. Вычислительные технологии, 14, 2009.

[2] В. В. Стрижов, Р. А. Сологуб, *Алгоритм выбора нелинейных регрессионных моделей с анализом гиперпараметров*. ММРО-14, 2009.

[3] А. А Зайцев, *Выбор моделей нелинейной регрессии с анализом гиперпараметров*. Конференция МФТИ, 2010.

[4] Yeh, I. and others, *Modeling slump flow of concrete using second-order regressions and artificial neural networks*. Cement and Concrete Composites, 29, 2007.