Министерство образования и науки Российской Федерации
Московский физико-технический институт (государственный университет)
Факультет управления и прикладной математики
Вычислительный центр им. А. А. Дородницына РАН
Кафедра «Интеллектуальные системы»

Токмакова Александра Алексеевна

Выбор устойчивых прогностических моделей в задачах нелинейного регрессионного анализа

010656 — Математические и информационные технологии

Выпускная квалификационная работа магистра

Научный руководитель:
к. ф.-м. н. Стрижов Вадим Викторович

Москва

2014

Содержание

Аннотация

В работе рассматривается проблема построения оптимальных устойчивых моделей в задаче нелинейного регрессионного анализа. В условиях мультиколлинеарности выборки выбор устойчивых моделей из значительного числа затруднен в связи с необходимостью оценки большого числа параметров. Оценка глобально оптимального значения параметров невозможна, так как функция ошибок нелинейных регрессионных моделей имеет большое количество локальных экстремумов как относительно параметров, так и относительно состава признаков. Предлагается стратегия порождения устойчивых моделей с помощью последовательного добавления и удаления элементов модели и оценки её структурных параметров.

Ключевые слова: *байесовский вывод, метод белсли, пошаговый отбор признаков, нелинейные модели, структурная сложность.*

1 Введение

Основной задачей регрессионной анализа является восстановление параметров регрессионной модели, при которых выбранная модель наилучшим образом описывает данные. Для оценки качества приближения данных вводится функция ошибки, вид которой определяется статистическими предположениями о характере распределения зависимой переменной и вектора параметров регрессионной модели. Гипотеза порождения данных играет центральную роль в выборе функции ошибки и, как следствие, в методе оценки параметров модели. Для подтверждения или опровержения этой гипотезы выполняются статистические тесты, называемые анализом регрессионных остатков [6, 3, 4, 5]. При этом считается, что независимая переменная не является случайной величиной, не содержит ошибок и не нуждается в дополнительных статистических гипотезах.

При оценке параметров модели необходимо учитывать сложность модели, так как из-за наличия в регрессионной выборке шумовых или мультикоррелирующих признаков можно получить неустойчивую оценку. Для получения устойчивых оценок параметров ранее предлагались регуляризующие методы и методы отбора признаков, например LASSO [15], LARS [17], RidgeRegression [14, 21], ElasticNet [20], также при прореживании некоторых нелинейных моделей [18, 19].

В работе предлагается использовать функцию ошибки, содержащую в качестве регуляризующих множителей ковариационные матрицы зависимой переменной и параметров модели. Принята гипотеза о нормальном распределении данных и параметров. Подобный вид функции ошибки получен с помощью байесовского вывода и рассмотрен ранее в работах [17, 3, 4, 5].

Функция ошибки используется для максимизации правдоподобия данных, максимизации вероятности параметров модели, максимизации правдоподобия самой регрессионной модели [7, 8, 9] и несмещенности оценки параметров. В таком случае ковариационная матрица вектора параметров регрессионной модели $\mathbf{A}^{-1}$ может быть использована для суждения об устойчивости найденного решения и отбора признаков. Задача оценки ковариационных матриц рассматривалась с различных точек зрения. Так в книгах [1, 2] приведена оценка ковариационной матрицы параметров, полученная методом наименьших квадратов. В работе описана процедура несмещен-

ной оценки ковариационной матрицы параметров модели, путем разбиения регрессионной выборки на непересекающиеся подмножества.

В работе рассматриваются нелинейные модели, принадлежащие классу двухслойных нейронных сетей. Обобщающая способность и прогностические возможности нейронных сетей позволяют использовать их для решения реальных задач регрессии [22]. Однако такой подход подразумевает использование сетей с большим количеством слоев и нейронов, а соответственно и связей между ними. Большое количество связей требует использования больших обучающих выборок, увеличивает время обучения и работы сети. Ввиду этого возникает задача минимизации размера сети без потери точности классификации.

Существуют два базовых подхода к решению вышеописанной проблемы: *наращивание структуры сети* (network growing) [23] и *прореживание структуры сети* (network pruning) [24, 19, 18].

Согласно первому подходу в качестве начальной модели выбирается простая сеть с небольшим числом нейронов, возможностей которой недостаточно для решения поставленной задачи, после чего в сеть добавляются новые нейроны и связи между ними. В алгоритмах метода прореживания модифицируется многослойная сеть с избыточным числом нейронов и связей между ними. Классическими алгоритмами прореживания нейронных сетей являются «optimal brain damage» [19] и «optimal brain surgery» [18], основанные на вычислении вторых производных функции ошибки [16] и методе «back propagation» [25] обучения нейроной сети. Также получили развитие *гибридные алгоритмы*, в которых объединяются оба упомянутых выше подхода. Они позволяют улучшить качество результата, однако обладают своими границами применимости [26, 27].

Предлагается набор критериев прореживания и наращивания нейронной сети. Основными идеями являются принцип локально-оптимального выбора и использование информации о ковариационной матрице параметров сети. В качестве иллюстрирующего примера работа процедуры пошаговой модификации структуры модели выбрана задача регрессии на данных winequality, UCI [28].

2 Постановка задачи

Рассмотрим выборку $\mathfrak{D} = \{\mathbf{x}_i, y_i\}$, $i \in \mathcal{I} = \{1, \ldots, m\}$, где $\mathbf{x}_i^\mathsf{T} \in \mathbb{R}^n$, $\mathbf{y} \in \mathbb{R}^m$. Рассмотрим нелинейную модель, которая по крайней мере один раз дифференцируема относительно параметров $\mathbf{w}$:

$$\mathbf{f} : (\mathbf{w}, \mathbf{X}) \mapsto \mathbf{y},$$

где $\mathbf{X} \in \mathbb{R}^{m \times n}$ — матрица плана, $\mathbf{w} = [w_1, \ldots, w_j, \ldots, w_t]^\mathsf{T}$, $j \in \mathcal{J} = \{1, \ldots, t\}$.

Требуется найти такое множество индексов $\mathcal{A}^* \subseteq \mathcal{J}$, которое является решением задачи многокритериальной опитимизации:

$$\begin{aligned} \mathcal{A}^* &= \operatorname*{argmin}_{\mathcal{A} \subseteq \mathcal{J}} \mathfrak{S}(\mathbf{w}^*_{\mathcal{A}} | \mathfrak{D}, \mathbf{f}_{\mathcal{A}}), \\ \mathbf{w}^*_{\mathcal{A}} &= \operatorname*{argmin}_{\mathbf{w} \in \mathbb{R}^t} \big\{ S(\mathbf{w} | \mathfrak{D}_{\mathcal{L}}, \mathbf{f}) \big\}, \end{aligned} \tag{2.1}$$

где $\mathfrak{S}$ — вектор, состоящий из функций S, U и F (функции оценки ошибки, устойчивости и сложности модели соответственно); $\mathbf{f}_{\mathcal{A}}$ — модель, оптимизированная на параметрах с индексами из множества $\mathcal{A}$; $\mathbf{w}_{\mathcal{A}}$ — вектор параметров с индексами из множества $\mathcal{A}$. В данной работе рассмотрена однокритериальная оптимизация функции ошибки $S(\mathbf{w})$.

Пусть нелинейная модель $\mathbf{f}$ принадлежит классу двуслойных нейронных сетей с $M + 1$ нейронами в скрытом слое. Функцией активации нейронов скрытого слоя является гиперболический тангенс:

$$a_k = \mathbf{tanh}\big(\Sigma_{\ell=0}^n W_{k\ell}^{(1)} x_{p\ell}\big),$$

где $k = [1, \ldots, M]$, $p = [1, \ldots, m]$, $\mathbf{W}^{(1)} \in \mathbb{R}^{M \times (n+1)}$, а $\mathbf{x}_p$ — строка матрицы плана $\mathbf{X}$.

Функцией активации нейронов выходного слоя является линейная функция:

$$y_p = \Sigma_{k=0}^M W_k^{(2)} a_k,$$

где $\mathbf{W}^{(2)} \in \mathbb{R}^{M+1}$.

Вектор параметров $\mathbf{w}$ модели $\mathbf{f}$ представим в виде конкатенации векторизованных матриц $\mathbf{w} = \mathrm{vec}(\mathbf{W}^{(1)} | \mathbf{W}^{(2)})$. Тогда модель $\mathbf{f}$ можно записать в следующем виде:

$$f(\mathbf{x}, \mathbf{w}) = \mathbf{W}^{(2)} \mathbf{tanh}(\mathbf{W}^{(1)} \mathbf{x}),$$

где $\mathbf{x}$ — строка матрицы плана $\mathbf{X}$.

3 Оценка правдоподобия модели

Пусть многомерная случайная величина $\mathbf{y} \sim \mathcal{N}(\mathbf{f}, \mathbf{B})$ имеет нормальное распределение

$$p(\mathbf{y}) = \frac{1}{(2\pi)^{\frac{m}{2}} \det^{\frac{1}{2}}(\mathbf{B}^{-1})} \exp\left(-\frac{1}{2}(\mathbf{y}-\mathbf{f})^{\intercal}\mathbf{B}(\mathbf{y}-\mathbf{f})\right). \tag{3.1}$$

В данной работе считается, что матрица $\mathbf{B} = \beta\mathbf{I}$. Так как правая часть выражения (3.1) зависит от вида регрессионной модели $\mathbf{f}$, вектора параметров $\mathbf{w}$, независимой переменной $\mathbf{X}$ и от ковариационной матрицы $\mathbf{B}$, перепишем его в виде

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

где Z_D — нормирующий коэффициент для плотности нормального распределения.

$$Z_D = (2\pi)^{\frac{m}{2}} \det^{\frac{1}{2}}(\mathbf{B}^{-1}).$$

Функция ошибки, соответствующая математическому ожиданию регрессионной модели при данной гипотезе, определена как

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

Пусть вектор параметров $\mathbf{w}$ модели $\mathbf{f}$ — многомерная случайная величина с математическим ожиданием $\mathbf{w}_0$, ковариационной матрицей $\mathbf{A}^{-1}$ с распределением

$$p(\mathbf{w}|\mathbf{A}, \mathbf{f}) = \frac{1}{(2\pi)^{\frac{t}{2}} \det^{\frac{1}{2}}(\mathbf{A}^{-1})} \exp\left(-\frac{1}{2}(\mathbf{w}-\mathbf{w}_0)^{\intercal}\mathbf{A}(\mathbf{w}-\mathbf{w}_0)\right) = \frac{\exp(-E_{\mathbf{w}})}{Z_{\mathbf{w}}(\mathbf{A})}. \tag{3.2}$$

Нормирующий коэффициент $Z_{\mathbf{w}}(\mathbf{A})$ равен

$$Z_{\mathbf{w}}(\mathbf{A}) = (2\pi)^{\frac{t}{2}} \det^{\frac{1}{2}}(\mathbf{A}^{-1}),$$

где t — число параметров модели $\mathbf{f}$. Функция-штраф за большое значение параметров модели для принятого распределения определена как

$$E_{\mathbf{w}} = \frac{1}{2}(\mathbf{w}-\mathbf{w}_0)^{\intercal}\mathbf{A}(\mathbf{w}-\mathbf{w}_0).$$

Для нахождения наиболее вероятных параметров модели $\mathbf{f}(\mathbf{w}, \mathbf{x})$ используем Байесовский вывод [11, 12, 13]. При заданной модели $\mathbf{f}$ и заданных значениях $\mathbf{A}$ и $\mathbf{B}$ апостериорное распределение параметров модели имеет вид:

$$p(\mathbf{w}|\mathfrak{D}, \mathbf{A}, \mathbf{B}) = \frac{p(\mathfrak{D}|\mathbf{w}, \mathbf{B})p(\mathbf{w}|\mathbf{A})}{p(\mathfrak{D}|\mathbf{A}, \mathbf{B})}, \text{ где} \tag{3.3}$$

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

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

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

$p(\mathfrak{D}|\mathbf{A}, \mathbf{B}) = \int_{\mathbb{R}^t} p(\mathfrak{D}|\mathbf{w}, \mathbf{B}) p(\mathbf{w}|\mathbf{A}) d\mathbf{w}$ — функция правдоподобия модели.

Определим функцию ошибки $S(\mathbf{w})$ как

$$S(\mathbf{w}) = E_{\mathbf{w}} + E_D = \frac{1}{2}(\mathbf{w} - \mathbf{w}_0)^{\mathsf{T}} \mathbf{A} (\mathbf{w} - \mathbf{w}_0) + \frac{1}{2}(\mathbf{y} - \mathbf{f})^{\mathsf{T}} \mathbf{B} (\mathbf{y} - \mathbf{f}). \tag{3.4}$$

В данной работе правдоподобие модели будет оцениваться с помощью метода Монте-Карло. В общем случае вычисление интеграла $p(\mathfrak{D}|\mathbf{A}, \mathbf{B})$ невозможно или сильно затруднено, так как пространство параметров обладает высокой размерностью [29]. Метод Монте-Карло позволяет приблизить значение интеграла усреднённым значением подынтегральной функции.

Рассмотрим м.с.в $\mathbf{w} \sim \mathcal{N}(\mathbf{w}_0, \mathbf{A}^{-1})$ и сформируем выборку размера b из этого распределения.

Запишем математическое ожидание функции правдоподобия модели $p(\mathfrak{D}|\mathbf{A}, \mathbf{B}, \mathbf{f})$

$$\mathbb{E}\big[p(\mathfrak{D}|\mathbf{A}, \mathbf{B}, \mathbf{f})\big] = \int p(\mathfrak{D}|\mathbf{w}, \mathbf{B}) p(\mathbf{w}|\mathbf{A}) d\mathbf{w}.$$

Из условия нормировки следует, что $\int p(\mathbf{w}|\mathbf{A}) = 1$. Воспользовавшись законом больших чисел получим:

$$\hat{\mathbb{E}}\big[p\big] = \int p(\mathfrak{D}|\mathbf{w}, \mathbf{B}) p(\mathbf{w}|\mathbf{A}) d\mathbf{w} \approx \frac{1}{b} \sum_{k=1}^{b} p(\mathfrak{D}|\mathbf{w}_k, \mathbf{B}),$$

где b — размер сэмплирующей выборки.

4 Метод оценки ковариационной матрицы

Рассмотрим случайную величину $\mathbf{w} \in \mathbb{R}^t$. По определению ковариационная матрица величины $\mathbf{w}$ вычисляется как:

$$\mathbf{A}^{-1} = \mathbb{E}[(\mathbf{w} - \mathbb{E}\mathbf{w})(\mathbf{w} - \mathbb{E}\mathbf{w})^{\mathsf{T}}] = \mathbb{E}[\mathbf{w}\mathbf{w}^{\mathsf{T}}] - \mathbb{E}[\mathbf{w}] \cdot \mathbb{E}[\mathbf{w}^{\mathsf{T}}].$$

Разобьем множество индексов выборки $\mathcal{I}$ на b непересекающихся подмножеств, на каждой подвыборке оценим вектор параметров $\mathbf{w}$ модели $\mathbf{f}$ и получим реализацию случайной величины вектора параметров модели $\tilde{\mathbf{w}}$. Сформируем матрицу $\mathbf{W}$:

$$\mathbf{W} = [\tilde{\mathbf{w}}_1, \ldots, \tilde{\mathbf{w}}_b] \in \mathbb{R}^{t \times b}.$$

Отцентрировав каждую строку $\mathbf{W}_j - \mathbb{E}\mathbf{W}_j \mapsto \mathbf{W}_j$, оценим ковариационную матрицу параметров модели $\mathbf{f}$:

$$\mathbf{A}^{-1} = \frac{1}{b}\mathbf{W}\mathbf{W}^{\intercal}.$$

Для устойчивого обращения матрицы $\mathbf{A}^{-1}$ предлагается сначала удалить из неё столбцы и строки, соответствующие параметрам с дисперсией меньше заданного порога, а затем добавить к диагонали небольшую величину [14, 21].

4.1 Подбор величины b

Для использования способа оценки ковариационной матрицы $\mathbf{A}^{-1}$, описанного в разделе выше, необходимо найти оптимальное число разбиений выборки b. Для каждого фиксированного b будем минимизировать функцию ошибки (3.4), используя случайные приближения параметров $\mathbf{A}$ и $\mathbf{w}_0$ априорного распределения вектора парамеров $\mathbf{w}$ (3.2) модели $\mathbf{f}$. Рассмотрим итерационный процесс:

1. Зададим число разбиений выборки b;

2. Зададим случайные приближения для $\mathbf{w}_0$, $\mathbf{A}$ и $\mathbf{B} = \beta\mathbf{I}$;

3. Оптимизируем значение функции ошибки $S(\mathbf{w})$ на каждом из b разбиений $(\mathbf{X}_v, \mathbf{y}_v)$:

$$\tilde{\mathbf{w}}_v = \operatorname*{argmin}_{\mathbf{w} \in \mathbb{R}^t} S(\mathbf{w}|\mathbf{X}_v, \mathbf{y}_v);$$

4. Сформируем матрицу реализаций параметров:

$$\mathbf{W} = [\tilde{\mathbf{w}}_1, \ldots, \tilde{\mathbf{w}}_b] \in \mathbb{R}^{t \times b};$$

5. Оценим параметры априорных распределений:

- $\mathbf{w}_0$ как среднее по реализациям каждой компоненты вектора параметров:

$$\hat{w}_{0j} = \mathbb{E}\mathbf{W}_j,$$

где $\mathbf{W}_j$ — строка матрицы $\mathbf{W}$ с индексом j;

- $\mathbf{A}$ в соответствии с разделом 4;

- скаляр β:

$$\hat{\beta} = \frac{1}{b}\sum_{v=1}^{b} \frac{1}{\mathbf{var}\big(\mathbf{y}_b - \mathbf{f}(\tilde{\mathbf{w}}_v, \mathbf{X}_v)\big)},$$

где **var** — обозначения для дисперсии.

5 Подбор начального приближения

Введем понятие структуры двухслойной неронной сети. В данной работе структурой модели будет называться ациклический направленный граф с тремя множествами вершин фиксированной максимальной валентности:

(0) — множество вершин свободных переменных: содержит $n+1$ вершину, каждая из которых принимает значение X_{ij} (x_0 принимает начение 1).

(1) — множество вершин функции активации первого слоя: содержит $M+1$ вершину, каждая из которых принимает значение a_k (a_0 принимает значение 1).

(2) — множество вершин функции активации выходного слоя: содержит одну вершин, которая принимает значение y_i.

Максимальная валентность вершин из множества свободных переменных равна M, вершин функции активации первого слоя — $n+2$, вершин функции активации выходного слоя — $M+1$.

Ребра графа могут идти только из множества вершин с меньшим индексом во множество вершин с большим индексом, то есть (0) $\to$ (1) или (1) $\to$ (2).

Пусть вектор $\mathbf{z} \in \{0,1\}^t$

$$z_j = \begin{cases} 0, & \text{если } w_j + \Delta w_j = 0, \text{ т.е. ребра нет в графе;} \\ 1, & \text{иначе.} \end{cases} \tag{5.1}$$

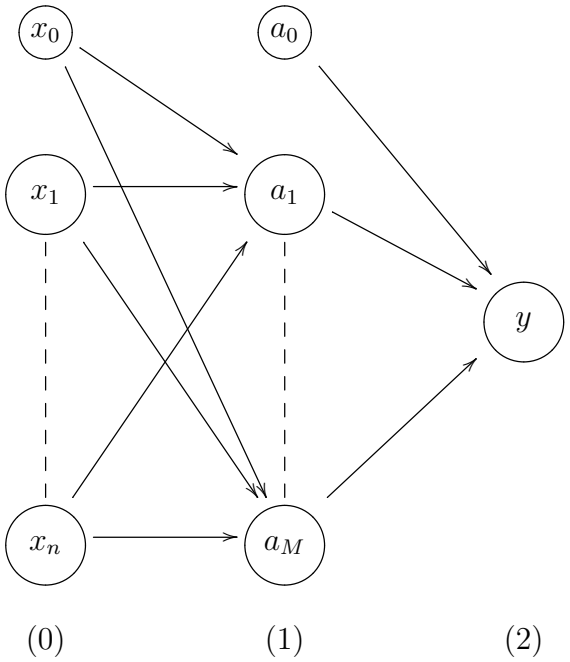

Рис. 1: Структура двухслойной нейронной сети

Тогда вектор $\mathbf{z}$ и величины n и M однозначно задают вид графа **??**, а значит и структуру модели.

Полный граф, описывающий структуру модели двуслойной нейронной сети, имеет $(n+1)\cdot(M+1)^2$ ребер, то есть для нахождения оптимальной в смысле минимума функции ошибки модели необходимо перебрать $2^{(n+1)\cdot(M+1)^2} - const$ моделей.

Предлагается найти максимально правдоподобную модель последовательным изменением структуры модели $\mathbf{f}$, причем на каждом шаге изменять структуру модели на единицу.

Представим процедуру последовательной модификации структуры модели в виде пути внутри t-мерного гиперкуба $\mathcal{Z}$, каждая вершина которого является бинарным вектором (5.1). Так как соседние вершины гиперкуба $\mathcal{Z}$ отличаются единственной компонентой вектора $\mathbf{z}$, на каждом шаге алгоритма изменения модели происходит выбор из $(n+1)\cdot(M+1)^2$ моделей-претендентов.

Для оценки параметров и гиперпараметров $\mathbf{A}$, $\mathbf{w}_0$ и β модели $\mathbf{f}$ применим последовательный Байесовский вывод: в качестве априорного распределения параметров модели используется модификация апостериорных вероятностей параметров модели, полученных на предыдущем шаге.

Рассмотрим последовательно порожденные модели $\mathbf{f}_{\mathcal{A}_q}$ и $\mathbf{f}_{\mathcal{A}_{q+1}}$. Апостериорное распределение параметров модели $\mathbf{f}_{\mathcal{A}_q}$ в соответствии с (3.3) имеет вид:

$$p(\mathbf{w}_{\mathcal{A}_q}|\mathfrak{D}, \mathbf{A}_q, \mathbf{B}_q) = \frac{p(\mathfrak{D}|\mathbf{w}_{\mathcal{A}_q}, \mathbf{B}_q)p(\mathbf{w}_{\mathcal{A}_q}|\mathbf{A}_q)}{p(\mathfrak{D}|\mathbf{A}_q, \mathbf{B}_q)}.$$

Начальные приближения гиперпараметров для модели $\mathbf{f}_{\mathcal{A}_{q+1}}$ будут выглядеть следующим образом:

$$\tilde{\mathbf{w}}_{\mathcal{A}_q} = [\mathbf{w}_{\mathcal{A}_q}; 0]^\intercal;\ \tilde{\mathbf{A}}_q = \begin{pmatrix} \mathbf{A}_q & 0 \\ 0 & 1 \end{pmatrix}.$$

Тогда функция правдоподобия данных (3.1) и априорное распределение параметров (3.2) модели $\mathbf{f}_{\mathcal{A}_{q+1}}$ будут выглядеть следующим образом:

$$p(\mathfrak{D}|\mathbf{w}_{\mathcal{A}_{q+1}}, \mathbf{B}_q) = \frac{1}{(2\pi)^{\frac{m}{2}} \det^{\frac{1}{2}}(\mathbf{B}_q^{-1})} \exp\left(-\frac{1}{2}(\mathbf{y} - \mathbf{f}_{\mathcal{A}_{q+1}})^\intercal \mathbf{B}_q (\mathbf{y} - \mathbf{f}_{\mathcal{A}_{q+1}})\right);$$

$$p(\mathbf{w}_{\mathcal{A}_{q+1}}|\tilde{\mathbf{A}}_q) = \frac{1}{(2\pi)^{\frac{t}{2}} \det^{\frac{1}{2}}(\tilde{\mathbf{A}}_q^{-1})} \exp\left(-\frac{1}{2}(\mathbf{w}_{\mathcal{A}_{q+1}} - \tilde{\mathbf{w}}_{\mathcal{A}_q})^\intercal \tilde{\mathbf{A}}_q (\mathbf{w}_{\mathcal{A}_{q+1}} - \tilde{\mathbf{w}}_{\mathcal{A}_q})\right).$$

5.1 Вычислительный эксперимент

Рассмотрим процедуру построения модели, с помощью последовательного добавления признаков, в которой в качестве априорного распределения параметров модели используется модификация апостериорных вероятностей параметров моделей, полученных на предыдущем шаге.

Пусть на шаге q известна правдоподобная нелинейная модель субоптимальной сложности, то есть известна оценка её параметров $\mathbf{w}^*_{\mathcal{A}_q}$ и гиперпараметров $\mathbf{A}^*_q$ и $\mathbf{w}^*_{0q}$.

Рассмотрим шаг алгоритма с номером q.

1. Найдем такой элемент множества $\mathcal{J} \setminus \mathcal{A}_q$, при добавлении которого в модель функция ошибки $S(\mathbf{w}_{\mathcal{A}_q})$ минимальна:

$$j^* = \underset{j \in \mathcal{J} \setminus \mathcal{A}_q}{\operatorname{argmin}} S(w_j|\mathfrak{D}, \mathbf{f}_{\mathcal{A}_q \cup \{j\}}, \mathbf{w}_{\mathcal{A}_q}).$$

   Причем при вычислении функции ошибки оптимизируется только компонента вектора параметров $\mathbf{w}_{\mathcal{A}_q}$ и элементы матрицы $\mathbf{A}_q$, соответствующие параметру с индексом j.

2. Обновим оценки гиперпараметров $\mathbf{w}^*_{0(q+1)}$, $\mathbf{A}^*_{q+1}$;

$$\mathbf{w}^*_{0(q+1)} = [\mathbf{w}^*_{0q}; w_{j*}]^\mathsf{T};\ \mathbf{A}^*_{q+1} = \begin{pmatrix} \mathbf{A}^*_q & 0 \\ 0 & \mathbf{A}_q(j^*, j^*) \end{pmatrix}.$$

3. Добавим новый элемент j^* к текущему набору

$$\mathcal{A}_{q+1} = \mathcal{A}_q \cup \{j^*\}$$

и будем повторять эту процедуру до тех пор, пока

$$|S(\mathbf{w}^*_{\mathcal{A}_{q+1}}|\mathfrak{D}, \mathbf{f}_{\mathcal{A}_{q+1}}) - S(\mathbf{w}^*_{\mathcal{A}_q}|\mathfrak{D}, \mathbf{f}_{\mathcal{A}_q})| \leqslant \Delta S_1.$$

В качестве иллюстрации описанного алгоритма рассмотрим следующий пример. Пространство признаков матрицы плана $\mathbf{X}$ изображено на рисунке 2. Здесь векторы $\boldsymbol{\chi}_4$, $\boldsymbol{\chi}_5$ и $\boldsymbol{\chi}_6$ ортогональны друг другу, а векторы $\boldsymbol{\chi}_1$, $\boldsymbol{\chi}_2$ и $\boldsymbol{\chi}_3$ сильно коррелируют. Зададим вектор зависимой переменной $\mathbf{y}$ в виде линейной комбинации:

$$\mathbf{y} = 0.3\boldsymbol{\chi}_5 + 0.3\boldsymbol{\chi}_6 + 0.3\boldsymbol{\chi}_3 + \mathcal{U}[0, 1],$$

где $\mathcal{U}[0, 1]$ — равномерное распределение на отрезке $[0, 1]$.

На первом шаге зафиксируем модель с единственным признаком $\boldsymbol{\chi}_4$. На рисунке 3 изображены изменения вектора параметров модели.

Рассмотрим первый столбец. На первом графике изображены оптимальные значения параметров w_1—w_3, w_5 и w_6 при фиксированном параметре w_4. Парой с минимальным значением ошибки на этом шаге является $(\boldsymbol{\chi}_4, \boldsymbol{\chi}_6)$. Начиная с третьего графика, параметры добавляемых признаков равны нулю (изображено темно-синим цветом): то есть к $\boldsymbol{\chi}_4$, $\boldsymbol{\chi}_6$ и $\boldsymbol{\chi}_5$ невозможно добавить признаки таким образом, чтобы функция ошибки уменьшилась (рис. 4) (при фиксированных значениях параметров и матрицы ковариации уже добавленных признаков). Итоговая модель, полученная с помощью описанного алгоритма, выглядит следующим образом:

$$\mathbf{y} = 0.3\boldsymbol{\chi}_5 + 0.8\boldsymbol{\chi}_6 + 0.5\boldsymbol{\chi}_4.$$

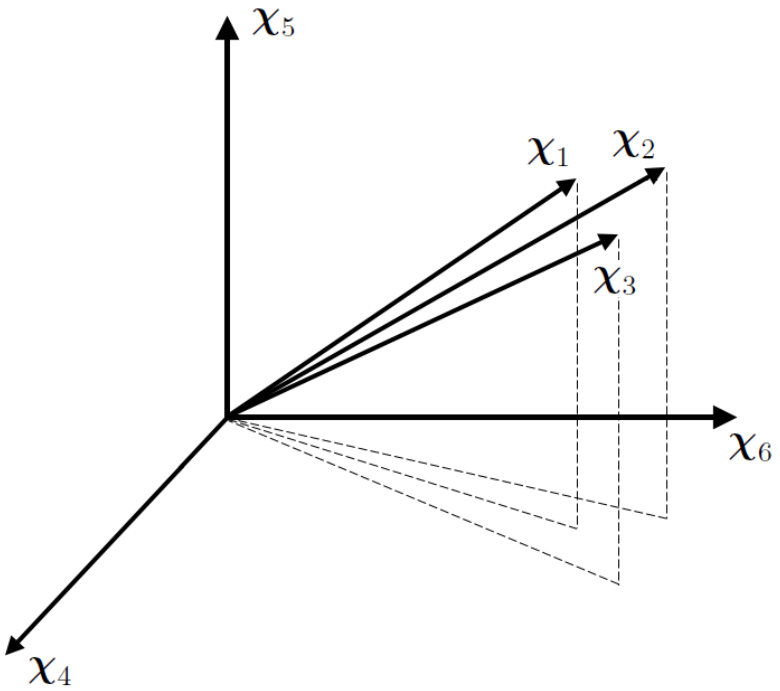

Рис. 2: Пространство признаков

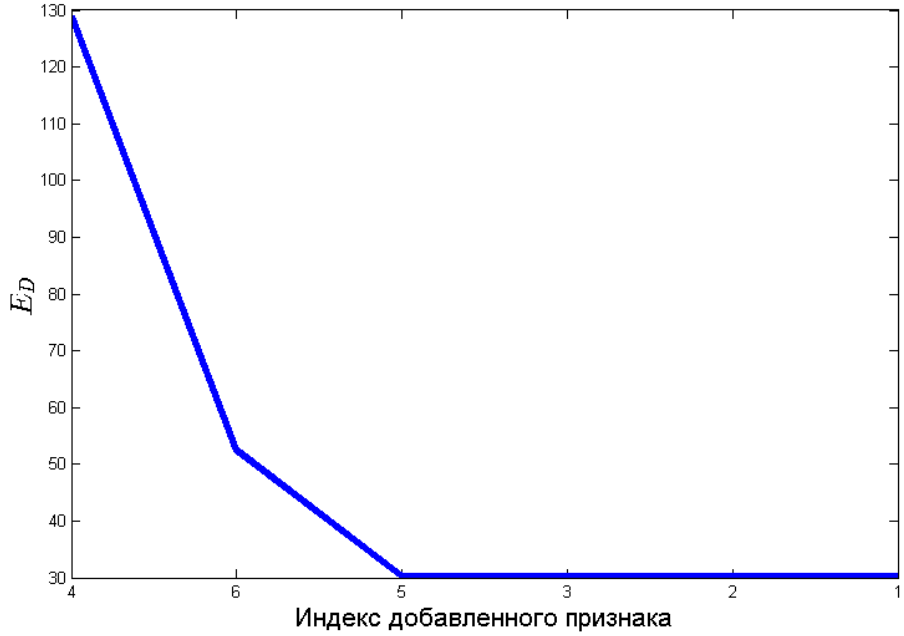

Рис. 4: График функции ошибки

Расширим предложенный алгоритм на шаге (1). Для каждого индекса $j \in \mathcal{J} \setminus \mathcal{A}_q$

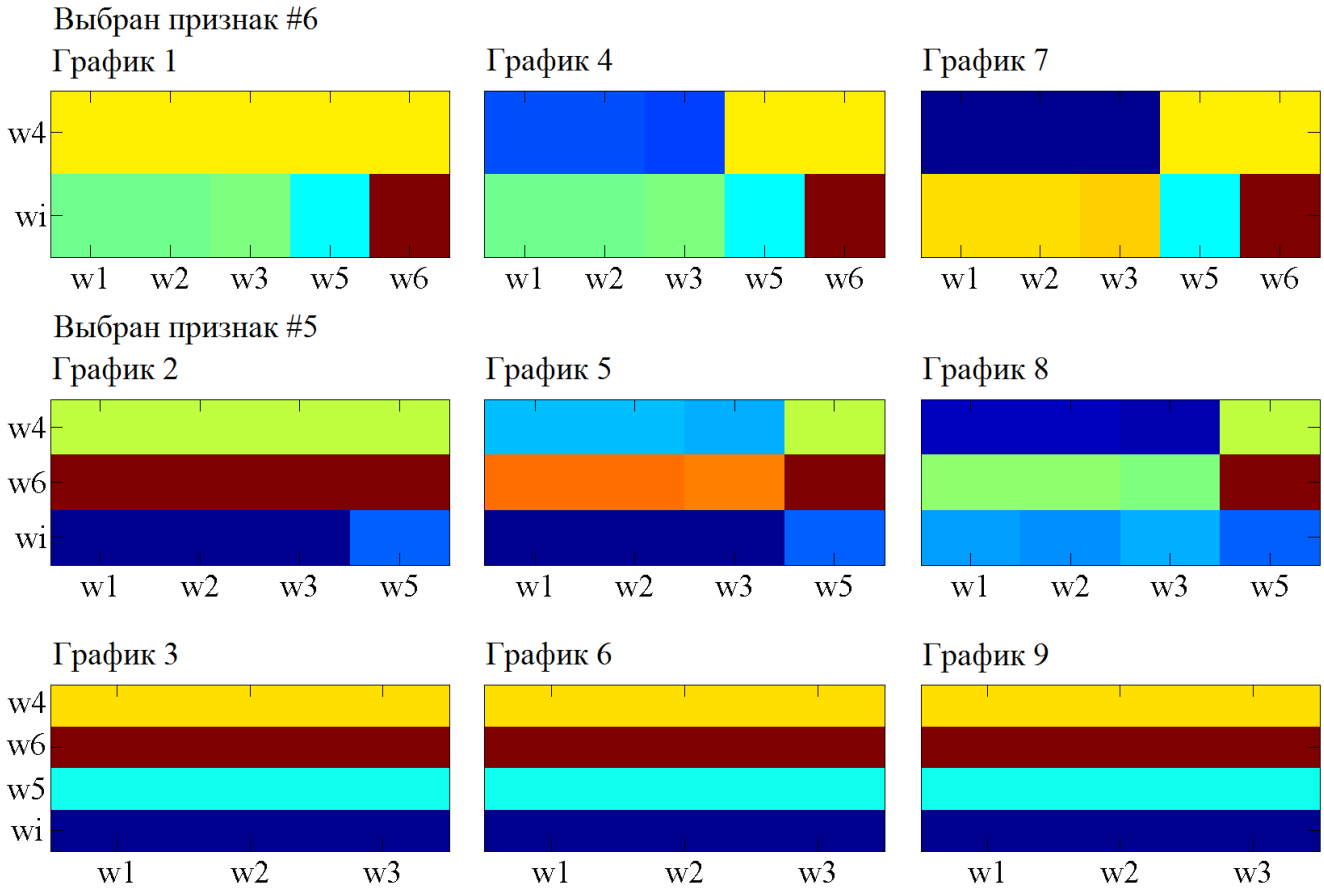

Рис. 3: Изменения вектора параметров

будем производить «обратную» оптимизацию:

$$\mathbf{w}^*_{\mathcal{A}_q \cup \{j\}} = \underset{\mathbf{w} \in \mathbb{R}^{\mathcal{A}_q}}{\operatorname{argmin}} S(\mathbf{w}_{\mathcal{A}_q} | \mathfrak{D}, \mathbf{f}_{\mathcal{A}_q}, w_j),$$

где w_j — оптимальное значение параметра j на множестве признаков $\mathcal{A}_q \cup \{j\}$. Такая модификация позволяет судить о близости к оптимальному значению функции ошибки при выбранных параметрах.

Во втором столбце рисунка 3 показаны результаты «обратной» оптимизации. Нижняя строка графиков из первого столбца в точности равна нижней строке графиков из второго столбца.

Третий столбец рисунка 3 соответствует параметрам модели настроенным без каких-либо фиксированных элементов.

При рассмотрении ортогональных векторов ($(\boldsymbol{\chi}_4, \boldsymbol{\chi}_6)$, $(\boldsymbol{\chi}_4, \boldsymbol{\chi}_5)$ в первой строке рисунка 3 и $(\boldsymbol{\chi}_4, \boldsymbol{\chi}_6, \boldsymbol{\chi}_5)$ во второй строке) видно, что найденные оптимальные значения параметров доставляют глобальный минимум функции ошибки на этих при-

знаках, для коррелирующих же комбинаций существуют значения фиксированных параметров, при которых значение функции ошибки будет уменьшаться.

6 Процесс оптимизации структуры модели

Стратегией пошаговой модификации модели называет процедура модификации модели, в которой на каждом шаге решается оптимизационная задача вида:

$$j^* = \underset{j \in \mathcal{A}}{\operatorname{argopt}} Q(\mathbf{w}^*_{\mathcal{A}}),$$

где Q — некоторый критерий качества, а $\mathbf{w}^*_{\mathcal{A}}$ определяется выражением (2.1).

Стратегия задается следующими математическими объектами:

- набором критериев оптимизации $\{Q\}$;
- набором ограничений на структуру и параметры модели $\mathcal{A} \in \mathcal{J}$, $\mathbf{w} = \mathbf{w}^*_{\mathcal{A}}$;
- критериями останова шагов добавления (6.1) и удаления (6.2) структурных единиц в модель;
- критерием останова процедуры выбора модели (6.3).

То есть действуя согласно стратегии, мы будем изменять структуру модели, удаляя из неё элементы и добавляя их до тех пор, пока значение критериев оптимизациине стабилизируется.

6.1 Добавление структурных связей

Описанная процедура позволяет найти индекс параметра, при добавлении которого в сеть функция ошибки минимальна.

Пусть на шаге q известна правдоподобная нелинейная модель субоптимальной сложности, то есть известна оценка её параметров $\mathbf{w}^*_{\mathcal{A}_q}$ и гиперпараметров $\mathbf{w}^*_{0q}$ и $\mathbf{A}^*_q$:

$$\mathbf{f}_{\mathcal{A}_q}|\mathbf{w} = \mathbf{w}^*_{\mathcal{A}_q} : \mathbf{X} \mapsto \mathbf{y}.$$

Рассмотрим шаг алгоритма с номером $q+1$:

1. найдем такой элемент множества $\mathcal{J} \setminus \mathcal{A}_q$, при котором функция ошибки $S(\mathbf{w}_{\mathcal{A}_q \cup \{j\}})$ минимальна:

$$j^* = \underset{j \in \mathcal{J} \setminus \mathcal{A}_q}{\operatorname{argmin}} S(\mathbf{w}_{q+1} | \mathfrak{D}, \mathbf{f}_{\mathcal{A}_q \cup \{j\}});$$

2. обновим оценки для $\mathbf{w}^*_{0(q+1)}$ и $\mathbf{A}^*_{q+1}$ в соответствии с разделом 5;

3. добавим новый элемент j^* к текущему набору

$$\mathcal{A}_{q+1} = \mathcal{A}_q \cup \{j^*\}.$$

Будем повторять эту процедуру до тех пор, пока изменение функции ошибки на соседних итерациях не превысит заранее заданного порога ΔS_1

$$|S(\mathbf{w}^*_{\mathcal{A}_{q+1}} | \mathfrak{D}, \mathbf{f}_{\mathcal{A}_{q+1}}) - S(\mathbf{w}^*_{\mathcal{A}_q} | \mathfrak{D}, \mathbf{f}_{\mathcal{A}_q})| \leqslant \Delta S_1. \tag{6.1}$$

6.2 Удаление структурных связей

Для удаления связей воспользуемся модфикацией метода Белсли. Рассмотрим множество индексов связей между нейронами $\mathcal{A}_q$.

1. Вычислим индексы обусловленности η_j (6.5) и матрицу долевых коэффициентов $\mathfrak{R} = [r_{gj}]$ (6.6) для $\mathbf{w}_{\mathcal{A}_q}$;

2. Найдем индекс g^* максимального индекса обусловленности $\eta_{\max}$;

3. В матрице долевых коэффициентов $\mathfrak{R}$ найдем индекс столбца j^*:

$$j^* = \underset{j \in \mathcal{A}_q}{\operatorname{argmax}} r_{g^* j}.$$

4. Удалим j^* индекс из множества $\mathcal{A}_q$:

$$\mathcal{A}_{q+1} = \mathcal{A}_q \setminus \{j^*\}$$

и будем повторять эту процедуру до тех пор, пока изменение функции ошибки на соседних итерациях не превысит заранее заданного порога ΔS_2

$$|S(\mathbf{w}^*_{\mathcal{A}_{q+1}} | \mathfrak{D}, \mathbf{f}_{\mathcal{A}_{q+1}}) - S(\mathbf{w}^*_{\mathcal{A}_q} | \mathfrak{D}, \mathbf{f}_{\mathcal{A}_q})| \leqslant \Delta S_2. \tag{6.2}$$

6.2.1 Метод Белсли для удаления признаков

Рассмотрим матрицу $\mathbf{A}_q$, полученную методом из раздела 4. Она невырождена по построению и имеет ранг $\text{rank}(\mathbf{A}_q) = |\mathcal{A}_q|$. Так как матрица $\mathbf{A}_q$ положительно определеная, воспользуемся разложением Холецкого [10] и представим её в виде:

$$\mathbf{A}_q = \mathbf{L}\mathbf{L}^\intercal, \tag{6.3}$$

где $\mathbf{L}$ — нижняя треугольная матрица.

Запишем сингулярное разложение матрицы $\mathbf{L}$:

$$\mathbf{L} = \mathbf{U}\mathbf{\Lambda}\mathbf{V}^\intercal, \tag{6.4}$$

где $\mathbf{U}$ и $\mathbf{V}$ — ортогональные матрицы, а $\mathbf{\Lambda}$ — диагональная матрица с собственными значениями λ_j на диагонали, такими что

$$\lambda_1 > \lambda_2 > \ldots > \lambda_q > 0.$$

Индексом обусловленности с индексом j будем называть отношение максимального собственного числа матрицы $\mathbf{L}$ к j-ому собственному числу:

$$\eta_j = \frac{\lambda_{\max}}{\lambda_j}. \tag{6.5}$$

Большие значения индексов обусловленностей указывают на зависимость между признаками, при чем чем больше η_j, тем сильнее зависимость.Поэтому на этапе удаления связей необходимо найти такой индекс j^*, что:

$$j^* = \underset{j \in \mathcal{A}_q}{\text{argmax}}\, \eta_j.$$

Далее воспользуемся выражениями (6.3) и (6.4) и представим ковариационную матрицу $\mathbf{A}_q^{-1}$ в виде:

$$\mathbf{A}_q^{-1} = (\mathbf{L}\mathbf{L}^\intercal)^{-1} = (\mathbf{V}\mathbf{\Lambda}^\intercal\mathbf{U}^\intercal\mathbf{U}\mathbf{\Lambda}\mathbf{V}^\intercal)^{-1} = \mathbf{V}\mathbf{\Lambda}^{-2}\mathbf{V}^{-1}.$$

Оценками дисперсии параметров будут диагональные элементы матрицы $\mathbf{A}_q^{-1}$:

$$\alpha_{ii} = \sigma(w_i) = \sum_{j=1}^{t} \frac{v_{ij}^2}{\lambda_j^2}.$$

Дисперсионной долей r_{ij} будем называть вклад j-го признака в дисперсию i-го элемента вектора параметров $\mathbf{w}_{\mathcal{A}_q}$:

$$r_{ij} = \frac{v_{ij}^2/\lambda_j^2}{\sum_{j=1}^{t} v_{ij}^2/\lambda_j^2}. \tag{6.6}$$

На рисунке 5 изображен вектор индексов обусловленности $\boldsymbol{\eta}$ и матрица долевых коэффициентов $\mathfrak{R}$ для выборки, исследуемой в работе. Здесь матрица ковариаций была оценена как $\mathbf{A}^{-1} = (\mathbf{X}\mathbf{X}^{\mathsf{T}})^{-1}$.

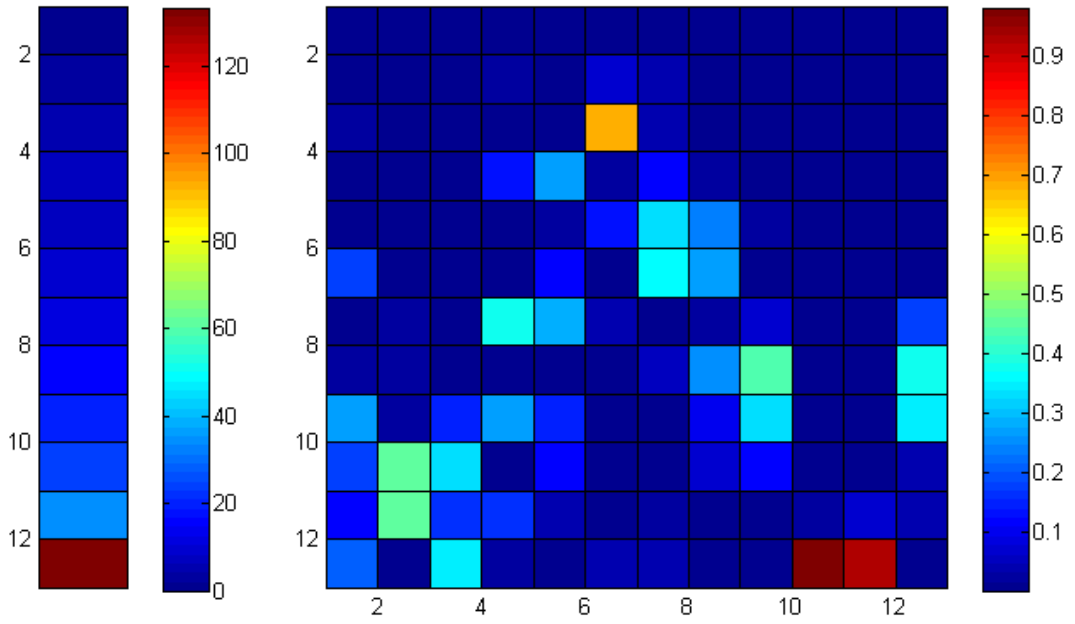

Рис. 5: Вектор индексов обусловленности $\boldsymbol{\eta}$ и матрица долевых коэффициентов $\mathfrak{R}$

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

По индексам обусловленности и матрице долевых коэффициентов определяется мультиколлинеарность: большие величины η_j означают, что, возможно, есть зависимость между признаками. 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 зависимости.

6.3 Критерий стабилизации процедуры Add-Del

В качестве критерия стабилизации структуры модели в процедуре Add-Del предлагается использовать энтропию

$$H(\mathcal{A}, \mathcal{A}') = -\rho(\mathbf{z}, \mathbf{z}') \ln\big(\rho(\mathbf{z}, \mathbf{z}')\big),$$

где $\rho(\cdot, \cdot)$ — функция расстояния Хэмминга между векторами $\mathbf{z}$ и $\mathbf{z}'$, где

$$z_j = \begin{cases} 0, & \text{если } w_j + \Delta w_j = 0, \text{ т.е. } j \notin \mathcal{A}; \\ 1, & \text{иначе.} \end{cases}$$

Процесс считается стабильным, если изменение энтропии $H(\mathcal{A}, \mathcal{A}')$ не превосходит заданного порога.

7 Модификация алгоритма Левенберга-Марквардта

Для минимизации функции ошибки воспользуемся алгоритмом Левенберга-Марквардта, который предназначен для оптимизации параметров нелинейных регрессионных моделей. Алгоритм заключается в последовательном приближении заданных начальных значений параметров к искомому локальному оптимуму и является обобщением метода сопряжённых градиентов и алгоритма Ньютона-Гаусса.

Рассмотри функцию ошибки вида:

$$S = \frac{1}{2}(\mathbf{w} + \Delta\mathbf{w})^T \mathbf{A}(\mathbf{w} + \Delta\mathbf{w}) + \frac{1}{2}\big(\mathbf{f}(\mathbf{w} + \Delta\mathbf{w}, \mathbf{X}) - \mathbf{y}\big)^T \mathbf{B}\big(\mathbf{f}(\mathbf{w} + \Delta\mathbf{w}, \mathbf{X}) - \mathbf{y}\big). \quad \text{(eq: S)}$$

На нулевой итерации алгоритма задаётся начальное приближение для $\mathbf{w}$. Для оценки приращения $\Delta\mathbf{w}$ используется линейное приближение функции $\mathbf{f}(\mathbf{w}, \mathbf{X} + \Delta\mathbf{w}) \approx \mathbf{f}(\mathbf{w}, \mathbf{X}) + \mathbf{J}\Delta\mathbf{w}$, где $\mathbf{J}$ — якобиан функции $\mathbf{f}(\mathbf{w}, \mathbf{X})$ в точке $\mathbf{w}$:

$$\mathbf{J} = \begin{pmatrix} \frac{\partial \mathbf{f}(\mathbf{w}, \mathbf{x}_1)}{\partial w_1} & \cdots & \frac{\partial \mathbf{f}(\mathbf{w}, \mathbf{x}_1)}{\partial w_n} \\ \vdots & \ddots & \vdots \\ \frac{\partial \mathbf{f}(\mathbf{w}, \mathbf{x}_m)}{\partial w_1} & \cdots & \frac{\partial \mathbf{f}(\mathbf{w}, \mathbf{x}_m)}{\partial w_n} \end{pmatrix}.$$

Приращение $\Delta\mathbf{w}$ в точке оптимума для функции ошибки (eq: S) равно нулю. Поэтому для нахождения экстремума приравняем вектор частных производных S по $\mathbf{w}$ к нулю. Для этого представим S в виде двух слагаемых:

$$S_1 = \frac{1}{2}(\mathbf{w} + \Delta\mathbf{w})^T \mathbf{A}(\mathbf{w} + \Delta\mathbf{w});$$

$$S_2 = \frac{1}{2}\big(\mathbf{f}(\mathbf{w} + \Delta\mathbf{w}, \mathbf{X}) - \mathbf{y}\big)^T \mathbf{B}\big(\mathbf{f}(\mathbf{w} + \Delta\mathbf{w}, \mathbf{X}) - \mathbf{y}\big).$$

После дифференцирования получим следующие выражения:

$\frac{\partial S_1}{\partial \mathbf{w}} = \frac{1}{2}(\mathbf{w} + \Delta\mathbf{w})^T(\mathbf{A} + \mathbf{A}^T)$,

$\frac{\partial S_2}{\partial \mathbf{w}} = \frac{1}{2}\big[\big(\mathbf{J}(\mathbf{w} + \Delta\mathbf{w}) - \mathbf{y}\big)^T \mathbf{B}^T\mathbf{J} + \big(\mathbf{J}(\mathbf{w} + \Delta\mathbf{w}) - \mathbf{y}\big)^T \mathbf{B}\mathbf{J}\big]$.

Таким образом, чтобы найти приращение $\Delta\mathbf{w}$ необходимо решить систему линейных уравнений:

$\nabla S = \frac{1}{2}(\mathbf{w}+\Delta\mathbf{w})^T(\mathbf{A}+\mathbf{A}^T) + \frac{1}{2}\big[\big(\mathbf{J}(\mathbf{w}+\Delta\mathbf{w})-\mathbf{y}\big)^T \mathbf{B}^T\mathbf{J} + \big(\mathbf{J}(\mathbf{w}+\Delta\mathbf{w})-\mathbf{y}\big)^T \mathbf{B}\mathbf{J}\big] = 0.$

Раскроем скобки и приведём подобные слагаемые:

$\mathbf{w}^T\mathbf{A} + \Delta\mathbf{w}^T\mathbf{A} + \mathbf{w}^T\mathbf{A}^T + \Delta\mathbf{w}^T\mathbf{A}^T + (\mathbf{w}^T\mathbf{J}^T\mathbf{B}^T + \Delta\mathbf{w}^T\mathbf{J}^T\mathbf{B}^T - \mathbf{y}^T\mathbf{B}^T + \mathbf{w}^T\mathbf{J}^T\mathbf{B} + \Delta\mathbf{w}^T\mathbf{J}^T\mathbf{B} - \mathbf{y}^T\mathbf{B})\mathbf{J} = 0.$

Сгруппируем и перенесём в одну сторону члены, содержащие приращение параметров $\Delta\mathbf{w}$:

$\Delta\mathbf{w}^T(\mathbf{A}+\mathbf{A}^T+\mathbf{J}^T\mathbf{B}^T\mathbf{J}+\mathbf{J}^T\mathbf{B}\mathbf{J}) = -\mathbf{w}^T\mathbf{A}-\mathbf{w}^T\mathbf{A}^T-\mathbf{f}^T\mathbf{B}^T\mathbf{J}+\mathbf{y}^T\mathbf{B}^T\mathbf{J}-\mathbf{f}^T\mathbf{B}\mathbf{J}+\mathbf{y}^T\mathbf{B}\mathbf{J}.$

Выразив приращение $\Delta\mathbf{w}$, получим следующую рекуррентную формулу:

$\Delta\mathbf{w} = \big[\big(\mathbf{A}+\mathbf{A}^T+\mathbf{J}^T(\mathbf{B}^T+\mathbf{B})\mathbf{J}\big)^{-1}\big]^T\big(-\mathbf{w}^T(\mathbf{A}+\mathbf{A}^T)+(\mathbf{y}-\mathbf{f}(\mathbf{w},\mathbf{X}))^T(\mathbf{B}^T+\mathbf{B})\mathbf{J}\big)^T.$

Алгоритм останавливается, в том случае, если приращение $\Delta\mathbf{w}$ в последующей итерации меньше заданного значения, либо если параметры $\mathbf{w}$ доставляют ошибку S меньшую заданной величины. Значение вектора $\mathbf{w}$ на последней итерации считается искомым.

8 Вычислительный эксперимент

Описанная процедура Add-Del применялась к данным winequality, UCI [28]. Данные содержат 12 числовых признаков, вещественный откликб 6497 объектов. В качестве нелинейной модели рассматривалась двухслойная нейронная сеть с пятью нейронами в скрытом слое. Общее число структурных связей в такой сети: (12+1)*(5+1) + (5+1) = 71. На рисунках 6, 7 и 8 изображены изменения функции ошибки, числа обусловленностей матрицы $\mathbf{A}^{-1}$ и вектора $\mathbf{z}$, описывающего структуру модели.

Рис. 6: Значение функции ошибки во время процедуры Add-Del

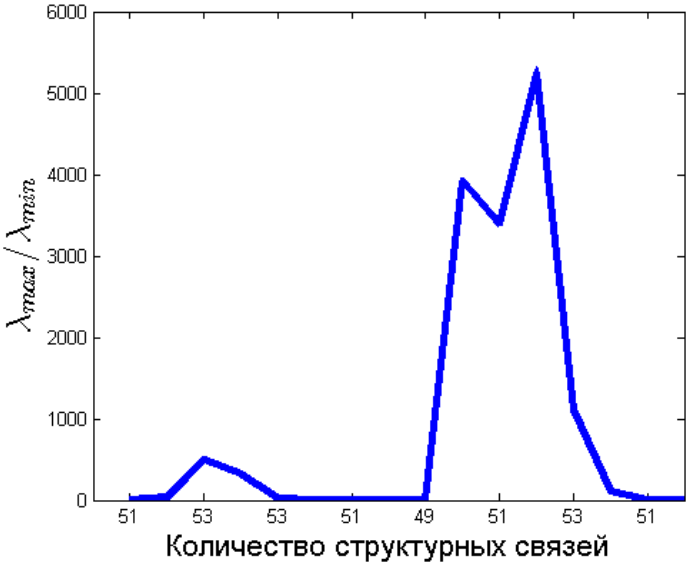

Рис. 7: Изменение числа обусловленностей матрицы $\mathbf{A}^{-1}$ во время процедуры Add-Del

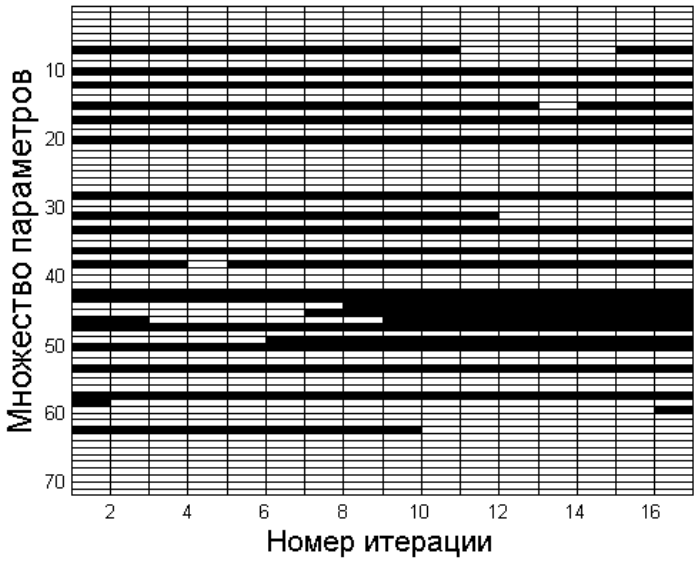

Рис. 8: Изменение множества параметров во время процедуры Add-Del

9 Заключение

В работе предложена процедура пошаговой оптимизации структуры нейронной сети для получения точных и устойчивых моделей, описан алгоритм получения несмещенной оценки ковариационной матрицы параметров модели. Проведен вычислительный эксперимент на реальных данных, численно подтверждающий полученные результаты.

Список литературы

[1] *Seber G. A. F., Wild C.* Nonlinear Regression. — Wiley-IEEE, 2003. — 768 pp.

[2] *Айвазян С. А.* Основы эконометрики. — М.: ЮНИТИ-ДАНА, 1998. — 51–67 с.

[3] *Lehmann E. L., Romano J. P.* Testing Statistical Hypothesis. — Springer, 2005. — 784 pp.

[4] *Pagan A. R., Hall A. D.* Diagnostic tests as residual analysis. — Australian National University, 1983. — 55 pp.

[5] *Montgomery D. C.* Design and analysis of experiments. — John Wiley and Sons, 2008. — 696 pp.

[6] *Anscombe F. J., Tukey J. W.* The examination and analysis of residuals // *Technometrics.* — 1963. — Vol. 5. — Pp. 141–160.

[7] *Le Cam L., Lo Yang G.* Asymptotics in statistics: some basic concepts. — Springer, 2000. — 285 pp.

[8] *Le Cam L.* Maximum likelihood — an introduction // *ISI Review.* — 1990. — Vol. 58 (2). — Pp. 153–171.

[9] *Hald A.* On the history of maximum likelihood in relation to inverse probability and least squares // *Statistical Science.* — 1999. — Vol. 14 (2). — Pp. 214–222.

[10] *Садовничий В. А.* Теория операторов. — Дрофа, 2001. — 352 с.

[11] *Bickel P. J., Doksum K. A.* Mathematical Statistics, Volume 1: Basic and Selected Topics. — Pearson Prentice–Hall, 2007. — 556 pp.

[12] *Jaynes E.* Probability Theory: The Logic of Science. — CUP, 2003. — 758 pp.

[13] *Howson C., Urbach P.* Scientific Reasoning: the Bayesian Approach. — Open Court Publishing Company, 2005. — 327 pp.

[14] *Hoerl A.E., Kennard R.W.* Ridge regression: Biased estimation for nonortthogonal problems // *Technometrics.* — 1970. — Vol. 3, no. 12. — Pp. 55–67.

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

[16] *Bishop C.* Exact calculation of the Hessian matrix for the multilayer perceptron // *Neural Computation.* — 1992. — Vol. 4. — Pp. 494–501.

[17] *Efron B., Hastie T., Johnstone I., Tibshirani R.* Least angle regression // *The Annals of Statistics.* — 2004. — Vol. 32, no. 2. — P. 407–499.

[18] *Hassibi B., Stork D. G.* Second order derivatives for network pruning: Optimal brain surgeon // Advances in Neural Information Processing Systems / edited byS. J. Hanson, J. D. Cowan, C. L. Giles. — Morgan Kaufmann, San Mateo, CA, 1993. — Vol. 5. — Pp. 164–171.

[19] *LeCun Y., Denker J., Solla S., Howard R. E., Jackel L. D.* Optimal brain damage // Advances in Neural Information Processing Systems II / edited byD. S. Touretzky. — San Mateo, CA: Morgan Kauffman, 1990. — Pp. 598–605.

[20] *Zou H., Hastie T.* Regularization and variable selection via the elastic net // Journal of the Royal Statistical Society: Series B (Statistical Methodology). — 2005. — Vol. 5. — Pp. 301–320.

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

[22] *Adya M., Collopy F.* How Effective are Neural Networks at Forecasting and Prediction? A Review and Evaluation // Journal of Forecasting. — 1998. — Vol. 17. — Pp. 481–495.

[23] *MacLeod C., Maxwell G.* Incremental Evolution in ANNs: Neural Nets which Grow // Artif. Intell. Rev. — 2001. — Vol. 16, no. 3. — Pp. 201–224.

[24] *Karnin E. D.* A simple procedure for pruning back–propagation trained neural networks // IEEE Transactions on Neural Networks. — 1990. — Vol. 1, no. 2. — Pp. 239–242.

[25] *Haykin S.* Нейронные сети: полный курс. — Издательский дом «Вильямс», 2006.

[26] *Yang S.-H., Chen Y.-P.* An evolutionary constructive and pruning algorithm for artificial neural networks and its prediction applications // Neurocomputing. — 2012. — Vol. 86. — Pp. 140–149.

[27] *Pu X., Sun P.* A New Hybrid Pruning Neural Network Algorithm Based on Sensitivity Analysis for Stock Market Forcast // Journal of Information and Computational Science. — 2013. — Vol. 3. — Pp. 883-892.

[28] *Cortez P., Cerdeira A., Almeida F., Matos T. and Reis J.* Modeling wine preferences by data mining from physicochemical properties // Decision Support Systems, Elsevier. — 2009. — Vol. 47, no. 4. — Pp. 547-553.

[29] *Nabney I.* Netlab. Algorithms for pattern recognition. — Springer, 2002. — 420 pp.