Методы инициализации параметров нелинейной регрессионной модели

Беляев М. Г.[1,2,3], ***Бурнаев Е. В.***[1,2,3], ***Ерофеев П. Д.***[1,3], ***Приходько П. В.***[1,2,3]
belyaevmichael@gmail.com, burnaev@iitp.ru, erofeev.paul@gmail.com, prikhodkop@gmail.com
Москва, [1]Институт Проблем Передачи Информации РАН им. Харкевича, [2]DATADVANCE,
[3]Московский Физико-технический Институт (Государственный Университет)

При построении нелинейной регресионной модели необходимо правильно нициализировать её параметры. В данной работе проводится сравнение некоторых широко распространенных методов и нескольких новых подходов к инициализации аппроксимационной модели, представляющей из себя разложение по словарю параметрических функций специального вида. Результаты численных экспериментов позволяют утверждать, что один из предложенных подходов к инициализации дает улучшение качества конечной аппроксимации на специфическом классе функций (двумерные гладкие и разрывные функции с множеством особенностей в области определения). Однако, в общем случае ни один из методов инициализации, в том числе и общепризнанных, не показал сколько-нибудь значимого улучшения качества аппроксимации или времени обучения.

При построении нелинейных регрессионных моделей возникает несколько типичных задач [1]: 1) первичная обработка исходных данных; 2) выбор и инициализация параметров используемой регрессионной модели; 3) обучение (подстройка параметров) модели и 4) оценка точности полученной аппроксимации. Обучение чаще всего является итеративным процессом, так как в общем (нелинейном) случае не существует явных формул, позволяющих точно оценить параметры регрессионной модели по данным. Известно, что начальный выбор архитектуры модели и значений параметров влияет не только на общее время обучения, но и на качество конечной аппроксимации [2, 3]. В данной работе исследуется влияние инициализации параметров на примере нелинейной регрессионной модели, представляющей из себя разложение по словарю параметрических функций.

Введение

Введем некоторые обозначения. Пусть задана выборка данных: $S = \{(\mathbf{X}_i, y_i); i = 1, \ldots, N\}$, $\mathbf{X}_i \in \mathbb{R}^n$, $y_i \in \mathbb{R}^1$, порожденная неизвестной функцией $y = f(\mathbf{X})$. Необходимо построить функцию $\hat{f}(\mathbf{X})$, которая будет близка к исходной функции $f(\mathbf{X})$ в смысле некоторой нормы (обычно это среднеквадратичная ошибка). Мы будем рассматривать аппроксимирующие функции вида:

$$\hat{f}(\mathbf{X}) = \sum_{j=1}^{p} V_j \sigma(\mathbf{X} \times \mathbf{W}_j^T + d_j) + V_0,\ \mathbf{W}_j \in \mathbb{R}^n \quad (1)$$

- разложение по словарю параметрических функций [4]. Здесь в качестве функции $\sigma(\cdot)$ выступают функции специального вида - сигмоиды (гиперболический тангенс). Настраиваемые параметры модели: p, V_0, V_j, $\mathbf{W}_j$ и d_j, $(j = 1, \ldots, p)$. Веса V_j могут быть однозначно определены по остальным параметрам модели с помощью решения линейной регрессионной задачи. Для обучения модели используется алгортим RProp (Resilient Propagation).

Таким образом, в контексте рассматриваемой проблемы инициализации модели необходимо решать сразу две задачи: 1) подбор количества p функций для словаря; 2) инициализация параметров этих функций. При этом порядок решения этих задач может быть разным.

В данной работе в рамках решения поставленных задач рассмотрены два принципиально разных подхода: **рандомизированные методы инициализации**, широко используемые в подобных задачах [2, 5, 9], предполагают случайные значения параметров модели; **детерминированные методы инициализации**, учитывающие характерные особенности аппроксимируемой выборки, являются более предпочтительными в смысле повторяемости результатов. Рандомизированные алгоритмы имеют следующие преимущества: простота реализации и незначительные вычислительные затраты. Это позволило рандомизированному подходу получить широкое распространение [2]. Однако, если качество полученной модели и время, затраченное на обучение, оказываются приемлемым, то в случае рандомизированной инициализации, как показали эксперименты, не представляется возможным добиться приемлимой повторяемости результатов даже на одних и тех же данных.

Статья организована следующим образом. Во втором и третьем разделах подробно описаны рассматриваемые алгоритмы в свете предложенной классификации. Четвертый раздел посвящен ре-

зультатам численных экспериментов. В последнем, пятом разделе, подведены итоги работы.

Рандомизированная инициализация

Широкое распространение для инициализации моделей типа (1) получили алгоритмы рандомизированной инициализации. В этом разделе будут рассмотрены некоторые наиболее известные из них.

Инициализация Нгуена-Видроу. Самым распространенным способом инициализации нелинейных моделей типа (1) является рандомизированный алгоритм NW, предложенный Нгуеном и Видроу [9]. Весам $\mathbf{W}_j$ и b_j присваиваются начальные значения так, чтобы активные области соответствующих сигмоидов были распределены примерно равномерно в пространстве регрессоров. Таким образом? каждый элемент матрицы весов $\mathbf{W}$ инициализируется числом из равномерного распределения:[1]

$$\mathbf{W}_j \sim U[-I,\, I]^n, \tag{2}$$

где $I = p^{\frac{1}{N}}$. Компоненты вектора $\mathbf{b}$ также выбираются из равномерного распределения $U[-I, I]$.

Инициализация SCAWI. Подход к инициализации весов, используемый Драго и Риделла[2] [5], напоминает алгоритм NW, но с другой границей значений:

$$\mathbf{W}_j \sim U[-I,\, I]^n, \tag{3}$$

где $I = 1.3/\sqrt{1 + N\nu^2}$, $\nu^2 = \frac{1}{n\,N} \sum_{i=1}^{N} \sum_{j=1}^{n} x_{ij}^2$. Такая инициализация позволяет гарантировать, что значения аргументов сигмоидов будут находится в области ненасыщения сигмоида, при этом оказываясь значительно отличными от нуля. Компоненты вектора $\mathbf{b}$ выбираются аналогично предыдущему методу из равномерного распределения.

Сферическая инициализация. Для многомерных пространств покомпонентная случайная генерация векторов приводит к их кластеризации, порождая кластеризацию направлений сигмоидов. Существуют теоретические результаты [6], согласно которым наилучшая аппроксимация получается в случае равномерного распредления направлений по сфере. Представим веса модели (1) в виде $\mathbf{W}_j = R_j \mathbf{S}_j$, где $\mathbf{S}_j$ - случайный вектор, расположенный на единичной сфере, а R_j - некоторый радиус. Предлагается использовать следующую схему сферической инициализации весов (SWI). На первом этапе получаем значения углов φ_k $(k = 1, \ldots, n-1)$ из случайного распределения $U[-\pi, \pi]$, а затем переходим в декартовы координаты:

$$\begin{gathered} w_1 = R\cos(\varphi_1), \\ w_2 = R\sin(\varphi_1)\cos(\varphi_2), \\ \ldots \\ w_{n-1} = R\sin(\varphi_1)\ldots\sin(\varphi_{n-2})\cos(\varphi_{n-1}), \\ w_n = R\sin(\varphi_1)\ldots\sin(\varphi_{n-2})\sin(\varphi_{n-1}). \end{gathered}$$

Радиус по аналогии со SCAWI предлагается выбирать равным $R = \frac{1.3}{\sqrt{1+N\nu^2}}$. Компоненты вектора $\mathbf{b}$ выбираются аналогично методу Нгуена-Видроу из равномерного распределения.

Подбор числа сигмоидов. Приведенные алгоритмы не позволяют ответить на вопрос, сколько сигмоидов необходимо для построения аппроксимации. Предлагается два варианта решения этой проблемы: 1) подбор числа сигмоидов по сетке (по минимальной ошибке на валидационном множестве); 2) использование жадного набора сигмоидов [7] по критерию минимальной ошибки (до тех пор пока ошибка аппроксимации на валидационном множестве не начнет возрастать) или наибольшей корреляции. Последний подход предполагает решение задачи: $\min_{\mathbf{V}} \|\mathbf{V}\|_0$ при условии $\Xi\mathbf{V} = \mathbf{Y}$, где Ξ - матрица, состоящая из значений построенных сигмоидов в точках выборки: $\Xi_{ij} = \sigma(\mathbf{X}\mathbf{W}_j^T + d_j)$, $i = = 1, \ldots, N$, $j = 1, \ldots, p$; $\mathbf{Y} = [y_1, \ldots, y_N]^T$ - вектор-столбец целевых переменных.

Выбор начального положения центров сигмоидов.

Во всех описанных ранее алгоритмах центры сигмоидов выбираются случайно (вследствие случайности выбора вектора $\mathbf{b}$). Однако, в связи с тем, что аппроксимируемая функция задана в ограниченном числе точек, разумно предположить, что мы сможем построить эффективное приближение функции только вблизи от этих точек. Поэтому правильно было бы располагать активные области сигмоидов рядом с точками выборки. Это легко сделать, если переписать функцию активации в следующем виде: $\sigma(\mathbf{X}\mathbf{W}_j^T + b_j) = \sigma((\mathbf{X} - \mathbf{X}_0)\mathbf{W}_j^T)$, где $b_j = \mathbf{X_0}\mathbf{W}_j^T$, $\mathbf{X}_0$ – некоторая точка выборки.

Детерминированная инициализация

Рандомизированные алгоритмы инициализации имеют существенный недостаток: качество и время обучения могут существенно отличаться для двух разных запусков на одних и тех же данных. Также

[1] Здесь и далее будем считать, что в ходе предварительной обработки данных пространство регрессоров ограничено гиперкубом $[-1, 1]^n$.

[2] Алгоритм также известен под названием SCAWI (Statistically Controlled Activation Weight Initialization).

остается открытым вопрос подбора оптимального числа p функций в конечном словаре. В этом разделе приведено описание разработанных детерминированных алгоритмов инициализации, позволяющих решить эти проблемы.

Инициализация на основе локальных особенностей исходных данных.

В основе этого алгоритма лежит идея о том, что исходно сигмоиды необходимо располагать в тех областях, где аппроксимируемая функция имеет локальные особенности. Алгоритм может быть описан следующим образом.

Входные параметры алгоритма инициализации: выборка для обучения S и (опционально) количество требуемых сигмоидов: p[3].

Выходные параметры алгоритма: количество сигмоидов p, матрица весов $\mathbf{W} = [\mathbf{W}_j]_{j=1}^{p}$ и вектор весов $\mathbf{d} = [d_1, ..., d_p]$.

Локальная аппроксимация. На этом шаге для каждой точки $\mathbf{X}_i \in S$ исходной обучающей выборки строится локальная аппроксимация с помощью одного сигмоида. Для этого формируется веса $p_j = \frac{\exp\left(-\sum_{m=1}^{d} \frac{(x_{im}-x_{jm})^2}{h_m^2}\right)}{\sum_{l=1}^{N} \exp\left(-\sum_{m=1}^{d} \frac{(x_{im}-x_{lm})^2}{h_m^2}\right)}$, где h_m – ширина ядра, которую можно оценить, например, по классической формуле Боумана-Аззалини [8]: $h_m = s_m \{4/(n+2)N\}^{\frac{1}{n+4}}$, где s_m - оценка стандартного отклонения по m-ой компоненте выходных векторов обучающей выборки, $j = 1, \ldots, N$. Затем решается задача линейной аппроксимации: $\min_{\mathbf{W}_i} \sum_{j=1}^{N} p_j^2 \|\sigma^{-1}\left(\frac{y_j}{V_i}\right) - (\mathbf{X}_j - \mathbf{X}_i)\mathbf{W}_i^T\|_2^2$, где V_i подбирается по равномерной сетке. Таким образом, в каждую точку обучающей выборки ставится свой сигмоид, описывающий локальные особенности аппроксимируемой функции вблизи этой точки.

Отбор сигмоидов. После того как построены все сигмоиды из них необходимо выбрать наиболее коррелированные с заданными целевыми переменными. Если задан параметр p, то путем жадного набора, предложенного ранее формируется словарь из p сигмоидов. Если этот параметр не задан, то подбирается оптимальное (по ошибке на валидационном множестве) число сигмоидов для начальной аппроксимации.

Как показали опыты, качество конечной аппроксимаци существенно зависит от ширины ядра h_m. И даже небольшие изменения ширины ядра h_m могут привести к значительному изменению качества аппроксимации.

Численные эксперименты

В этом разделе приведены описание и результаты численных экспериментов для разных вариантов инициализации: Нгуена-Видроу (NW), Драго-Риделла (SCAWI(1,2)), сферической инициализации (SWI) и инициализации на основе локальных особенностей исходных данных (DWI). Инициализация SCAWI(1) отличается от SCAWI(2) тем, что в первой сигмоиды располагаются случайно (оригинальный алгоритм), а во второй - в точках выборки. Для сравнения выбранных алгоритмов были использованы данные, полученные с помощью типичных в задачах нелинейной регрессии двумерных функций ($x_i \in [-1, 1], i = 1, 2$):

$$f_1(x_1, x_2) = \frac{\sin^2(\sum_{i=1}^{2}(x_i + 0.6)^2 - 0.3)}{\tanh[\sum_{i=1}^{2}((x_i + 0.6)^2 - 0.3)^2 + 0.4]};$$

$$f_2(x_1, x_2) = \frac{x_1 + x_2}{1 + 4(x_1^2 + x_2^2)};$$

$$f_3(x_1, x_2) = \sum_{i=1}^{2} x_i + 1 \cdot (\sum_{i=1}^{2} x_i^2 < 0.25) - 2 \cdot (\sum_{i=1}^{2}(x_i - 0.7)^2 < 1);$$

$$f_4(x_1, x_2) = ((6x_1)^2 + 6x_2 - 11)^2 + (6x_1 + (6x_2)^2 - 7)^2;$$

Результаты численных экспериментов приведены в таблицах 1, 2 и 3. Эксперименты проводились на 20 случайных выборках мощностью 300 точек по 10 запусков на каждой - для рандомизированных алгоритмов и по одному запуску - для детерминированных. Все значения, приведенные в таблицах, являются десятичными логарифмами отношения соответствующих абсолютных значений к значениям, полученным при эталонной инициализации Нгуена-Видроу (NW)[4]. В таблице 1 приведены значения логарифмов отношения медиан среднеквадратичных ошибок конечной аппроксимации к соответствующим значениям при эталонной инициализации по всем запускам для каждой функции и

[3] Этот параметр задается, если подбор числа сигмоидов осуществляется по сетке.

[4] Таким образом, если некоторое значение близко к 0, то данная инициализация не отличается от эталонной, по данной характеристике; если значение имеет порядок 1, то данная инициализация имеет конечную ошибку аппроксимации (или время обучения), которая на порядок больше, чем при эталонной инициализации; если значение близко к -1, то инициализация имеет конечную ошибку аппроксимации (или время обучения), которая на порядок меньше, чем при эталонной инициализации.

Таблица 1. Качество конечной аппроксимации (ошибка среднеквадратичная) по отношению к эталонной.

Данные	SCAWI(1)	SCAWI(2)	SWI	DWI
f_1	0.08	0.07	0.20	0.57
f_2	-0.04	-0.08	-0.07	0.57
f_3	0.01	-0.00	-0.00	-0.03
f_4	0.14	-0.08	0.19	0.28

Таблица 2. Качество конечной аппроксимации (95 % квантиль абсолютной ошибки) по отношению к эталонной.

Данные	SCAWI(1)	SCAWI(2)	SWI	DWI
f_1	0.15	0.12	0.16	0.53
f_2	-0.03	-0.05	-0.10	0.55
f_3	0.00	-0.01	0.00	-0.01
f_4	0.06	-0.18	0.20	0.21

Таблица 3. Время обучения модели по отношению к эталонному.

Данные	SCAWI(1)	SCAWI(2)	SWI	DWI
f_1	-0.08	-0.14	-0.19	-0.61
f_2	0.00	-0.05	-0.03	-0.77
f_3	0.12	0.15	0.07	-0.34
f_4	-0.06	-0.04	-0.14	-0.07

для каждого метода инициализации. Аналогичные значения для 95 % квантилей приведены в таблице 2. Очевидно, что ни один из алгоритмов инициализации не дает существенного выигрыша в качестве конечной модели. Тем не менее расстановка центров сигмоидов в точках выборки дает небольшое стабильное улучшение. Если проводить сравнение по времени обучения, то несомненно лидирует детерминированный алгоритм инициализации, в то время как остальные алгоритмы по этой характеристике значимо не отличаются. Однако детерминированный алгоритм проигрывает по качеству аппроксимации конечной модели. На рассматриваемых двумерных функциях также не было выявлено существенного отличия сферической инициализации от других рандомизированных алгоритмов.

Заключение

В данной статье были предложены новые методы инициализации нелинейной регрессионной модели, представляющей из себя разложение по словарю параметрических функций специального вида. Также проведено сравнение предложенных алгоритмов с наиболее распространенными алгоритмами инициализации моделей подобного рода. Предложенный подход расстановки центров сигмоидов в точках выборки оказался эффективным на классе рассматриваемых функций. Однако, несмотря на разумный подход к проведению детерминированной инициализации, позволивший существенно сократить время последующего обучения, качество аппроксимации конечной модели при такой инициализации оказывалось хуже.

Следует отметить, что обучение модели проводилось с помощью алгоритма RProp. Использование других алгоритмов обучения может изменить характер зависимости конечной апроксимации от начальной инициализации.

Литература

[1] *Bates D. M., Watts D. G.* Nonlinear Regression Analysis and Its Applications // Wiley Series in Probability and Statistics. — New York: Wiley, 1988. — Vol. 32. — P. 365.

[2] *Fernandez-Redondo M., Hernandez-Espinosa C.* Weight initialization methods for multilayer feedforward // Proc. of the 9th European Symposiumon Artificial Neural Networks ESANN. — 2001. — Pp. 25–27.

[3] *Thimm G., Fiesler E.* Optimal Setting of Weights, Learning Rate, and Gain // DIAP Research Rep. — 1997. — Pp. 97–04.

[4] *Burnaev E., Belyaev M., Prikhodko P.* About hybrid algorithm for tuning of parameters in approximation based on linear expansion in parametric functions // Intellectualization of information processing conference. — Vol. 1. — 2010.

[5] *Drago G., Ridella S.* Possibility and Necessity Pattern Classification using an Interval Arithmetic Perceptron // Neural Computing & Applications. — 1999. — Vol. 8, — Pp. 40–52.

[6] *Maiorov V., Oskolkov K., Temlyakov V.* Gridge approximation and Radon compass // Approxim. Theory, Ed. B. Bojanov, DARBA, — 2002. — Pp. 284–309.

[7] *Bruckstein A. M., Donoho D. L., Elad M.* From Sparse Solutions of Systems of Equations to Sparse Modeling of Signals and Images // SIAM Review. — 2009. — Vol. 51, — P. 34.

[8] *Bowman A. W., Azzalini A.* Applied Smoothing Techniques for Data Analysis: The Kernel approach with S-Plus Illustrations // Oxford University Press, USA, — 1997.

[9] *Nguyen D., Widrow B.* Improving the learning speed of 2-layer neural networks by choosing initial values of the adaptive weights // IJCNN International Joint Conference on NN, 1990. — Pp. 21–26.