ДОКЛАДЫ РОССИЙСКОЙ АКАДЕМИИ НАУК. МАТЕМАТИКА, ИНФОРМАТИКА, ПРОЦЕССЫ УПРАВЛЕНИЯ, 2020, том 490, с. 74–77

ИНФОРМАТИКА

УДК 519.651

ОТСУТСТВИЕ УЗКИХ ГОРЛОВИН В АРХИТЕКТУРЕ НЕЙРОННОЙ СЕТИ ОПРЕДЕЛЯЕТ ЕЕ СВОЙСТВА КАК ФУНКЦИИ ОБЩЕГО ПОЛОЖЕНИЯ

© 2020 г. С. В. Курочкин*

Представлено академиком РАН К.В. Рудаковым 06.11.2019 г.
Поступило 08.11.2019 г.
После доработки 08.11.2019 г.
Принято к публикации 08.11.2019 г.

Доказано, что искусственная нейронная сеть с гладкими функциями активации является функцией Морса для почти всех, в смысле меры Лебега, наборов весов в случае, если в сети нет слоев с количеством нейронов меньшим, чем в предшествующих и последующих слоях.

Ключевые слова: нейронная сеть, функция Морса

DOI: 10.31857/S2686954320010166

ВВЕДЕНИЕ

Искусственные нейронные сети являются эффективным инструментом для решения различных задач анализа данных. Возможность с их помощью распознавать/аппроксимировать сложные нелинейные зависимости в данных убедительно подтверждена практикой. В качестве теоретической основы здесь выступают результаты об универсальных аппроксимативных свойствах нейронных сетей [1, 2]. В алгебраическом подходе аналогом является свойство полноты семейства алгоритмов, см. [3].

В настоящей работе дано теоретическое обоснование другого важного и реально используемого свойства функций, получаемых в результате аппроксимации точечных или дискретных данных посредством нейросетей – возможности получать информацию об исследуемом объекте, анализируя структуру линий уровня и/или индексы критических точек аппроксимирующей функции. Примером реализации такого подхода в задачах анализа изображений является [4].

Свойство дифференцируемой функции нескольких переменных (или функции на многообразии) быть функцией Морса обеспечивает регулярное строение ее линий уровня, их перестройка при прохождении критического уровня подчинена определенным правилам, а в количестве и индексах критических точек содержится важная информация, см., например, [5]. В смысле своих качественных свойств, т.е. с точностью до диффеоморфизмов области определения и интервала значений, функция Морса является дискретным объектом и поддается анализу алгебраическими методами, см. [3]. Так, на сфере S^2 имеется 17746 топологически различных функций Морса с 4 седлами [6]. Для применения в прикладных задачах, где пространство всевозможных функций описывается конечным (возможно, как в случае глубоких нейронных сетей, очень большим) числов параметров, важно иметь уверенность в том, что в данном пространстве дополнение к функциям Морса имеет нулевую меру. Практически это будет означать, что при решении реальной задачи функции, получаемые на всех шагах так называемого обучения нейронной сети, будут функциями Морса.

В данной работе получено условие на архитектуру нейронной сети, при котором для почти всех наборов параметров реализуемое сетью отображение будет функцией Морса. Смысл условия в том, что в сети не должно быть узких горловин (bottleneck) – когда в каком-то слое количество нейронов строго меньше, чем в слоях до и после. Сети с горловиной используются в специальных целях – как автокодировщики, см. [7, гл. 14]. Сети без горловин являются обычной практикой, для них выполняются теоремы об универсальной аппроксимации.

Близким к доказанному результату является утверждение, что почти все (в смысле меры Лебега в пространстве коэффициентов) многочлены нескольких переменных являются функциями Морса [8, 9].

Национальный исследовательский университет "Высшая школа экономики", Москва, Россия
**E-mail: skurochkin@hse.ru*

1. ПРЕДВАРИТЕЛЬНЫЕ СВЕДЕНИЯ, ТЕРМИНОЛОГИЯ

1.1. Нейронные сети

Нейронная сеть – это вещественнозначная функция нескольких вещественных переменных, являющаяся композицией нескольких отображений вида: аффинное преобразование, затем покоординатное применение фиксированной нелинейной функции (так называемой функции активации). Коэффициенты аффинных преобразований (так называемые веса) являются настраиваемыми параметрами. Каждое отдельное взятие аффинной формы с последующим применением функции активации называется нейроном. В качестве функций активации могут выбираться сигмоидная $\varphi(x) = \dfrac{1}{[1+\exp(-x)]}$ и различные другие варианты. Количество слоев и нейронов в каждом из них и вид функций активации являются гиперпараметрами.

Задача нахождения наилучшего набора весов ставится как задача минимизации ошибки аппроксимации на обучающем наборе данных, который содержит входные векторы x_k, $k = 1, 2, ..., N$, и соответствующие им целевые значения y_k. Эта задача глобальной безусловной невыпуклой оптимизации решается методами типа градиентного спуска, с регуляризацией, препятствующей чрезмерной подгонке к данным (overfitting), см., например, [7].

1.2. Функции Морса

Пусть $U \subset \mathbb{R}^n$ – область, $f: U \to \mathbb{R}$ – дифференцируемая функция. Точка $x \in U$ называется регулярной, если градиент f в этой точке не равен нулю, и критической в противном случае. Число $y \in \mathbb{R}$ является критическим значением для f, если $y = f(x)$ для некоторой критической точки x. Если в $f^{-1}(y)$ нет критических точек (в частности, если прообраз пуст), то такое значение y называется регулярным для f. Все остальные значения являются критическими. Согласно теореме Сарда, множество критических значений имеет меру ноль. Критическая точка называется невырожденной, если в этой точке гессиан f является невырожденной матрицей. Невырожденные критические точки изолированны. Функцией Морса называется дифференцируемая функция, имеющая только невырожденные критические точки. Подробно см., например, [5].

Вопрос формулируется так: дана архитектура нейронной сети; можно ли, и при каких условиях, утверждать, что для почти всех наборов весов (в смысле меры Лебега в пространстве весов) соответствующая сеть является функцией Морса. Далее будет получен такой критерий, а также рассмотрены контрпримеры.

2. НЕЙРОННАЯ СЕТЬ БЕЗ УЗКИХ ГОРЛОВИН ПОЧТИ НАВЕРНОЕ ЯВЛЯЕТСЯ ФУНКЦИЕЙ МОРСА

Т е о р е м а. *Пусть $x \in U \subset \mathbb{R}^n$, $w \in W \subset \mathbb{R}^p$, U, W – области соответственно в пространстве признаков и пространстве весов, $f: U \times W \to \mathbb{R}$ – нейронная сеть с произвольным количеством слоев и нейронов в слоях, функциями активации φ типа сигмоидной и условием, что в ней нет промежуточного слоя, количество нейронов в котором строго меньше, чем в некоторых слоях выше и ниже его по ходу распространения сигнала. Входной вектор в данном случае считается слоем с количеством нейронов, равным размерности пространства признаков. Тогда для почти всех наборов весов w сеть $f(x, w)$ является функцией Морса.*

Д о к а з а т е л ь с т в о. Пусть w_{ki}^j – вес j-го нейрона k-го слоя, соответствующий i-му нейрону предшествующего слоя, при $i = 0$ это порог. По правилу дифференцирования сложной функции, производная $\dfrac{\partial f(x,w)}{\partial x}$ представляется в виде произведения нескольких (по числу слоев) матричных сомножителей вида $\Psi(s_{(t)})W_{(t)}$, где t – номер слоя, $W_{(t)}$ – матрица весов этого слоя (без порогов), $\Psi(s_{(t)}) = \operatorname{diag}(\varphi'(s_{(t)}^j))$ – диагональная матрица с положительными элементами, зависящими через свои непосредственные аргументы $s_{(t)}^j$ от аргумента x и весов текущего и предшествующих, но не последующих слоев сети:

$$\frac{\partial f(x,w)}{\partial x} = [\Psi(s_{(L)})W_{(L)}] \ldots [\Psi(s_{(1)})W_{(1)}].$$

Крайний левый сомножитель, соответствующий выходному нейрону сети, является строкой.

Множество $\mathbb{W}$ таких наборов весов w, что все матрицы $W_{(k)}$, $k = 1, 2, \ldots, L$, имеют полный ранг, является в пространстве весов дополнением к объединению конечного числа многообразий меньшей размерности ([10, теорема 17.3]), т.е. множеству меры ноль. При этом $\mathbb{W}$, как и сами матрицы $W_{(k)}$, не зависит от x. Далее рассмотрение производится поочередно для каждой из компонент $\mathbb{W}$.

Пусть $A_{(k)}(x, w) = \Psi(s_{(k)})W_{(k)}$. Рассмотрим малую вариацию весов k-го слоя. Для порогов

$$\frac{\partial^2 f(x,w)}{\partial w_{k0}^j \partial x} = \left[\frac{\partial}{\partial s_k^j} u(x,w)\right] W_{(k)} \ldots,$$

где

$$u(x,w) = A_{(L)}(x,w)\ \ldots\ A_{(k+1)}(x,w)\Psi(s_{(k)})$$

и многоточием обозначены предшествующие (по ходу сигнала в сети) члены, которые не зависят от весов текущего слоя. Аналогично для производных по весам из матрицы $W_{(k)}$

$$\frac{\partial^2 f(x,w)}{\partial w_{ki}^j \partial x} = \left\{\left[\frac{\partial}{\partial s_k^j}u(x,w)\right]W_{(k)} + u(x,w)E_i^j\right\}\ldots,$$

где E_i^j – матрица, в которой элемент (i,j) равен единице, а остальные нулю. Отсюда следует, что используя различные вариации весов k-го слоя, можно получать возмущения градиента сети вида

$$A_{(L)}(x,w)\ \ldots\ A_{(k+1)}(x,w)ZA_{(k-1)}(x,w)\ \ldots\ A_{(1)}(x,w),$$

где Z – произвольная матрица соответствующего размера (здесь учтено, что $\Psi(s_{(k)})$ всегда обратима).

Л е м м а 1. *Матричное уравнение*

$$ADX + YDB = C,$$

где D – положительная диагональная матрица и все размеры матриц считаются согласованными, разрешимо относительно X, Y для тех и только тех C, которые, рассматриваемые как линейные отображения, отображают ядро B в образ A.

Утверждение следует из [11], см. также [12, теорема 8.6.1.1].

Далее полагая последовательно $k = L, L-1, ..., 1$, рассмотрим произведения $A_{(L)}(x,w)\ ...\ A_{(k)}(x,w)$ на предмет эпиморфности производной по w. При $k = L$ она имеет место. При домножении на каждую очередную матрицу $A_{(k-1)}$ могут представиться следующие случаи:

1. $\prod_{j=k,\ldots L} A_{(j)} \neq 0$ и горизонтальный размер матрицы $A_{(k-1)}$ не меньше вертикального. Тогда из элементарных соотношений для рангов $\prod_{j=k-1,\ldots L} A_{(j)} \neq 0$ и из леммы 1 возмущениями весов слоев с $k-1$ по L можно получить произвольное возмущение результата, т.е. эпиморфность сохраняется.

2. $\prod_{j=k,\ldots L} A_{(j)} \neq 0$ и горизонтальный размер матрицы $A_{(k-1)}$ меньше вертикального. Тогда из леммы 1 эпиморфность сохраняется, но может оказаться, что $\prod_{j=k-1,\ldots L} A_{(j)} = 0$.

3. $\prod_{j=k,\ldots L} A_{(j)} = 0$ и горизонтальный размер матрицы $A_{(k-1)}$ не больше вертикального. Тогда эпиморфность сохраняется и $\prod_{j=k-1,\ldots L} A_{(j)} = 0$.

4. $\prod_{j=k,\ldots L} A_{(j)} = 0$ и горизонтальный размер матрицы $A_{(k-1)}$ больше вертикального. Тогда $\prod_{j=k-1,\ldots L} A_{(j)} = 0$ и образ производной совпадает с линейной оболочкой строк матрицы $A_{(k-1)}$, т.е. эпиморфность нарушается.

Следовательно, для того чтобы эпиморфность производной по весам не имела места, необходимо и достаточно, чтобы сначала встретился случай типа 2, и затем типа 4. Так происходит в точности в тех случаях, когда в сети имеется узкая горловина.

Л е м м а 2. *Пусть U, W – области соответственно в $\mathbb{R}^n$ и $\mathbb{R}^p$, $p \geq n$, $f(x,w)$ – дифференцируемая функция, $f: U \times W \to \mathbb{R}$. Обозначим $f_w(x) = f(x,w)$, $f_w: U \to \mathbb{R}$, $df_w: U \to \mathbb{R}^n$ – ее производная по x,*

$$df_w(x) = \left(\frac{\partial f_w(x)}{\partial x_1}, \ldots, \frac{\partial f_w(x)}{\partial x_n}\right),$$

и $F(x,w) = df_w(x)$, $F: U \times W \to \mathbb{R}^n$. Пусть известно, что в любой точке (x,w) производная от F по w является сюръекцией. Тогда для почти всех w (в смысле обычной меры Лебега в $\mathbb{R}^p$) f_w является функцией Морса.

Доказательство леммы может быть выведено из [13, теорема 1.2.4].

Доказательство теоремы заканчивается применением леммы 2 с учетом того, что для нейронной сети всегда $p > n$.

Следующий пример показывает, как наличие горловины в сети связано с ее дифференциальными свойствами как отображения.

К о н т р п р и м е р. Рассмотрим сеть архитектуры 2-1-2-1:

$$f(x,w) = \varphi(\hat{w}_0 + \hat{w}_1\varphi(\tilde{w}_0^1 + \tilde{w}_1^1\varphi(w_0 + w_1x^1 + w_2x^2)) + \\ + \hat{w}_2\varphi(\tilde{w}_0^2 + \tilde{w}_1^2\varphi(w_0 + w_1x^1 + w_2x^2))).$$

Поскольку при преобразовании в первом слое размерность входного вектора понижается, все критические точки f, если они есть, будут вырожденными. Пусть функция активации φ сигмоидная, и рассмотрим набор весов: $w_0 = 0$, $w_1 = 1$, $w_2 = 0$, $\tilde{w}_0^1 = -10$, $\tilde{w}_1^1 = 1$, $\tilde{w}_0^2 = 10$, $\tilde{w}_1^2 = -1$, $\hat{w}_0 = 0$, $\hat{w}_1 = 1$, $\hat{w}_2 = 1$. Точка $x = (0,0)$ является локальным минимумом, который устойчив, т.е. существует и мало меняется при произвольных малых возмущениях всех весов. Следовательно, в пространстве весов существует множество положительной меры такое, что все соответствующие сети имеют вырожденную критическую точку.

Далее, если убрать один слой со стороны входа, т.е. взять 1-2-1-сеть, то она почти для всех весов будет функцией Морса, и при этом для множества весов положительной меры будет иметь критические точки.

Если убрать еще один слой, то полученная 2-1-сеть для всех ненулевых наборов весов будет иметь только регулярные значения.

ЗАКЛЮЧЕНИЕ

Топологический анализ данных существенно использует алгебраические методы для исследования многомерных данных. Полученный в данной работе результат показывает, что наряду с подходом, основанным на вычислении устойчивых (persistent) гомологий клеточных комплексов [14], для этих целей применимы и могут быть полезны также методы дифференциальной топологии. Искусственные нейронные сети наиболее распространенной архитектуры почти наверное (т.е. при решении практических задач всегда) удовлетворяют условиям применимости таких методов, что открывает некоторые новые возможности для применения алгебраического подхода в задачах анализа данных.

ИСТОЧНИКИ ФИНАНСИРОВАНИЯ

Результаты получены в рамках НИР, реализуемой в Центре хранения и анализа больших данных МГУ им. М.В. Ломоносова.

СПИСОК ЛИТЕРАТУРЫ

1. *Cybenko G.V.* Approximation by Superpositions of a Sigmoidal function // Mathematics of Control Signals and Systems. 1989. V. 2. № 4. P. 303–314. https://doi.org/10.1007/bf02551274
2. *Pinkus A.* Approximation Theory of the MLP Model in Neural Networks // Acta Numerica. 1999. V. 8. P. 143–195. https://doi.org/10.1017/S0962492900002919
3. *Журавлев Ю.И., Рудаков К.В.* Об алгебраической коррекции процедур обработки (преобразования) информации // Проблемы прикладной математики и информатики. М.: Наука, 1987. С. 187–198.
4. *Курочкин С.В.* Распознавание гомотопического типа объекта с помощью дифференциально-топологических инвариантов аппроксимирующего отображения // Компьютерная оптика. 2019. Т. 43. № 4. С. 611–617. https://doi.org/10.18287/2412-6179-2019-43-4-611-617
5. *Постников М.М.* Введение в теорию Морса. М.: Наука, 1971.
6. *Arnold V.* Smooth functions statistics // Funct. Anal. Other Math. 2006. № 1. P. 111–118. https://doi.org/10.1007/s11853-007-0008-6
7. *Гудфеллоу Я., Бенджио И., Курвилль А.* Глубокое обучение. М.: ДМК Пресс, 2017.
8. *Le C.* A note on Optimization with Morse Polynomials // Commun. Korean Math. Soc. 2018. V. 33. № 2. P. 671–676. https://doi.org/10.4134/CKMS.c170221
9. *Banyaga A., Hurtubise D.* Lectures on Morse Homology. Kluwer Texts Math. Sci. Dordrecht: Kluwer Acad. Publ., 2004. V. 29.
10. *Прасолов В.В.* Элементы комбинаторной и дифференциальной топологии. М.: МЦНМО, 2014.
11. *Baksalary J.K., Kala R.* The Matrix Equation AX – – YB = C // Linear Algebra and Its Applications. 1979. V. 30. P. 41–43. https://doi.org/10.1016/0024-3795(79)90004-1
12. *Прасолов В.В.* Задачи и теоремы линейной алгебры. М.: МЦНМО, 2015.
13. *Nicolaescu L.* An Invitation to Morse Theory. Springer, 2011. ISBN 978-1-4614-1105-5.
14. *Carlsson G.* Topology and Data // Bulletin of the American Mathematical Society. 2009. V. 46. № 2. P. 255–308. https://doi.org/10.1090/S0273-0979-09-01249-X

ABSENCE OF BOTTLENECKS IN A NEURAL NETWORK DETERMINES ITS GENERIC FUNCTIONAL PROPERTIES

S. V. Kurochkin

National Research University Higher School of Economics, Moscow, Russian Federation

Presented by Academician of the RAS K.V. Rudakov November 6, 2019

Received November 8, 2019

An artificial neural network with smooth activation functions and without bottlenecks is a Morse function for almost all, with respect to Lebesgue measure, sets of weights.

Keywords: neural network, Morse function