УДК 004.032.26
DOI: 10.36979/1694-500X-2022-22-8-93-100

ИСКУССТВЕННЫЕ НЕЙРОННЫЕ СЕТИ: АРХИТЕКТУРА, СОЗДАНИЕ, ПРОВЕРКА

Е.В. Куцев

Аннотация. Актуальность применения искусственных нейронных сетей во всех сферах жизнедеятельности человека объясняется их востребованностью и практичностью в использовании. Исследованы принципы проектирования, построения и проверки качества работы искусственных нейронных сетей в общем виде. Проведен анализ принципов построения искусственных нейронных сетей, а также обобщение теоретического материала. В ходе проведенного анализа и обобщения информации представлены основные сведения, принципы работы и проверки качества получаемых результатов, необходимые для проектирования искусственных нейронных сетей. Результаты исследования могут быть применены на этапе проектирования и построения искусственных нейронных сетей начинающими разработчиками, что положительно повлияет на их дальнейшее развитие.

Ключевые слова: нейрон; градиентный спуск; регрессия; классификация.

ЖАСАЛМА НЕЙРОН ТАРМАКТАРЫ: АРХИТЕКТУРАСЫ, ТҮЗҮҮ, ТЕКШЕРҮҮ

Е.В. Куцев

Аннотация. Адамдын жашоосунун бардык чөйрөлөрүндө жасалма нейрон тармактарын колдонуунун актуалдуулугу аларга суроо-талаптын болушу жана колдонуудагы практикалуулугу менен түшүндүрүлөт. Макалада жалпысынан жасалма нейрон тармактарын долбоорлоонун, куруунун жана иштөө сапатын сыноонун принциптери изилденген. Жасалма нейрон тармактарын куруу принциптерин талдоо, ошондой эле теориялык материалды жалпылоо жүргүзүлдү. Маалыматтарга талдоо жүргүзүүнүн жана жалпылоонун жүрүшүндө жасалма нейрон тармактарын долбоорлоо үчүн зарыл болгон негизги маалыматтар, иштөө принциптери жана алынган натыйжалардын сапатын текшерүү көрсөтүлдү. Изилдөөнүн натыйжалары жасалма нейрон тармактарын долбоорлоо жана куруу баскычында баштапкы иштеп чыгуучулар тарабынан колдонулушу мүмкүн, бул алардын андан аркы өнүгүшүнө оң таасирин тийгизет.

Түйүндүү сөздөр: нейрон; градиенттик түшүү; регрессия; классификация.

ARTIFICIAL NEURAL NETWORKS: ARCHITECTURE, CREATION, VERIFICATION

E.V. Kutsev

Abstract. Currently, there is an increase in the widespread use of artificial neural networks in all spheres of human life. The relevance of their application is explained by demand and practicality in use. The purpose of this article is to study the principles of designing, building and testing the quality of artificial neural networks work in a general way. To systematize the information obtained in the study of neural networks, an analysis of construction's principles, as well as a generalization of theoretical material, was carried out. In the course of the analysis and generalization necessary basic information, principles of operation and quality control of the obtained results are presented for the design of artificial neural networks. The results of this article can be applied at the stage of designing and building artificial neural networks by novice developers, which will positively affect their further development.

Keywords: neuron; gradient descent; regression; classification.

Введение. В основе создания искусственных нейронных сетей лежит человеческий мозг, в котором в процессе сложного взаимодействия между нейронами, соединенными между собой синаптической связью, обеспечивается выполнение огромного количества различных функций организма. Роль нейронов в искусственных нейронных сетях выполняют простейшие процессоры, моделирующие некоторые функции биологических нейронов, которые объединены в огромную сеть и поэтому способны решать сложные задачи [1].

Одним из перспективных направлений в развитии информационных технологий всех областей является применение искусственных нейронных сетей, их востребованность и практичность использования [2]. К задачам, которые решаются с применением нейронных сетей, относятся вопросы классификации, регрессии.

Материалы и методы исследования. Нейронные сети состоят из нейронов, связанных между собой, поэтому нейрон является основой, фундаментом нейронной сети. Для полного понимания работы искусственных нейронных сетей необходимо начинать с изучения отдельно взятого искусственного нейрона (рисунок 1).

Значение искусственного нейрона вычисляется следующим образом:

$$S = \sum_{i=1}^{n} x_i \times w_i + w_0, \quad (1)$$

где x_i – входные значения нейрона, поступившие извне или с выхода другого нейрона; w_i – вес входа нейрона; w_0 – вес смещения (необходим для достижения порога возбуждения нейрона).

Затем к результату применяется функция активации (ФА) $Y = F$.

Если значение нейрона превышает заданную величину, то применяется функция активации.

У искусственного нейрона бывает несколько входов. Каждый вход умножается на свой вес, затем полученные значения суммируются, и прибавляется вес смещения (рисунок 2).

В качестве функции активации все большую популярность приобретает так называемая функция ReLU – Rectified linear unit. Главными достоинствами данной функции является простота и плавная числовая «шкала возбуждения» нейрона (рисунок 3).

Нейронная сеть состоит из множества нейронов, соединенных таким образом, что выход одного нейрона является входом другого (рисунок 4).

Предсказываемый признак Y может быть:

- количественным – задача регрессии;
- меткой класса – задача классификации.

В зависимости от требуемой задачи будет зависеть архитектура нейронной сети. Для задачи регрессии выходной слой состоит из одного нейрона (рисунок 5, а), а для задачи классификации выходной слой нейронной сети имеет такое количество, сколько имеется классов (рисунок 5, б).

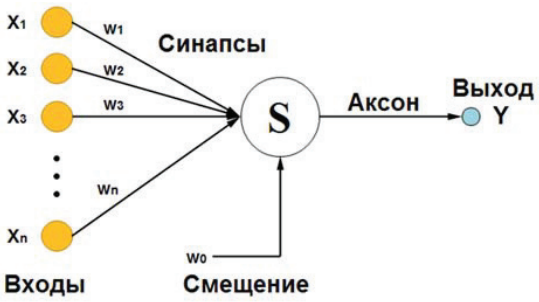

Рисунок 1 – Искусственный нейрон

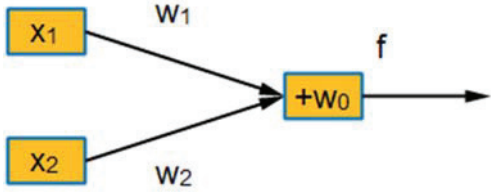

Рисунок 2 – Несколько входов искусственного нейрона

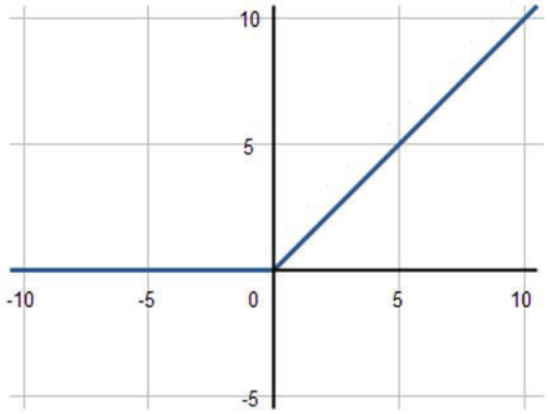

Рисунок 3 – Функция ReLU

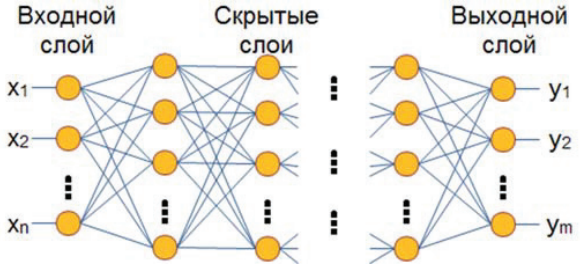

Рисунок 4 – Нейронная сеть

Рисунок 5 – Нейронные сети для задач: а – регрессии; б – классификации

$F_{NN}(x)$ – функция, которая преобразует вход нейронной сети в выход (рисунок 2).

$$F_{NN}(x) = x_1 \times w_1 + x_2 \times w_2 + w_0 . \quad (2)$$

Операция «softmax» – операция перехода от чисел к классам (вероятностным принадлежностям).

Пусть a, b – значения, которые выдала нейронная сеть в выходном слое (рассмотрим бинарную классификацию). Тогда числа можно рассматривать как вероятности принадлежности классам. Эта операция и называется операцией «softmax».

$$\frac{e^a}{(e^a + e^b)}, \frac{e^b}{(e^a + e^b)} . \quad (3)$$

Для многоклассовой операции «softmax» (см рисунок 5, б) используется следующая формула:

$$\Pr(Y = 0) = \frac{e^{a_0}}{e^{a_0} + \ldots e^{a_{k-1}}}, \; Pr(Y = k - 1) = \frac{e^{(a_{(k-1)})}}{(e^{(a_0)} + e^{(a_{(k-1)})})}, \quad (4)$$

где $a_0, \ldots, a_{k-1}$ – выходы нейронов последнего слоя.

Теперь, когда рассмотрены основные типы нейронных сетей, необходимо определить планы решения задач регрессии и классификации.

План решения задач регрессии и классификации с помощью нейронной сети:

1. Взять тренировочную выборку (ТВ) – набор объектов с известными значениями целевого признака Y .

2. Задать основные параметры нейронной сети: количество слоев, количество нейронов на каждом слое, тип связи между нейронами и т. д.

3. Составить функцию $F_{NN}(x)$. Пусть y_i точное значение целевого признака Y для i -го объекта из ТВ; $F_{NN}(x_i)$ – значение функции нейронной сети для i -го объекта из ТВ – для задачи регрессии.

Пусть ТВ состоит из объектов $x_{1,\ldots,}x_m$, для которых известны их точные метки классов $y_{1,\ldots,}y_m$ и нейронная сеть дает вероятности $p_{1,\ldots,}p_m$ принадлежности к истинному классу. Выписать выражения для вероятностей $\Pr(Y=0)$ и $\Pr(Y=k-1)$ – для задачи классификации.

4. Составить функцию потерь (суммарная ошибка на ТВ) для задачи регрессии:

$$L(w)=\left(F_{NN}(x_1)-y_1\right)^2+\left(F_{NN}(x_2)-y_2\right)^2+\ldots+\left(F_{NN}(x_m)-y_m\right)^2. \quad (5)$$

Для задачи классификации:

$$L(w)=-\ln p_1-\ln p_2-\ldots-\ln p_m. \quad (6)$$

5. Относительно переменных w_i – значения весов, необходимо найти точку минимума функции $L(w)$. Точка минимума будет определять оптимальные веса нейронной сети.

6. Присвоить весам нейронной сети найденные оптимальные значения.

7. Практически применять нейронную сеть на объектах, которые не принадлежат ТВ.

Общий план тренировки нейронных сетей выглядит следующим образом:

1. Задание архитектуры нейронной сети.

2. Составление функции потерь $L(w)$ по заданной ТВ.

3. Нахождение оптимальных значений весов нейронной сети путем нахождения минимума функции потерь.

4. Предсказание нейронной сетью значений для новых объектов.

Для нахождения минимума функции потерь $L(w)$ используется так называемый «метод градиентного спуска» (ГС). Его можно сравнить с обычным шариком, который под действием силы тяжести скатывается в самую глубокую область поверхности (рисунок 6).

Теперь необходимо смоделировать движение шарика. Для этого зададим длину шага h. Сделав шаг длины h, шарик будет останавливаться и думать, в какую сторону ему дальше катиться.

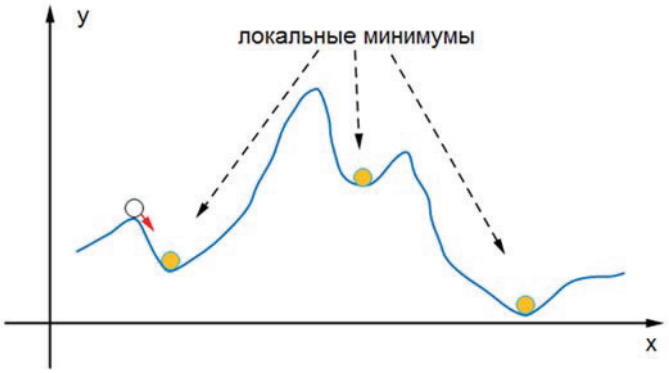

Рисунок 6 – Локальные минимумы

Пусть шарик находится в точке с координатой a_0. Тогда новая точка a_1 равна:

$$a_1 := a_0 - f'(a_0) \times h. \quad (7)$$

После этого процесс повторяется до тех пор, пока некоторая a_n не будет близка к 0 или если задано количество итераций – пока они не закончатся.

В ГС необходимо выбрать начальную точку для спуска (начальные значения весов нейронной сети). Обычно эта точка выбирается случайно, но от нее зависит успех ГС. Если начальные значения весов нейронной сети подобрать неудачно, то можно оказаться на плато, а, следовательно, ГС остановится и никуда не пойдет, также можно попасть в зону бесконечного спуска.

Для борьбы с такими случаями используются следующие алгоритмы:

- инициализация весов и нормализация данных – попадание на плато;
- регуляризация – бесконечный спуск.

Суть инициализации весов состоит в следующем – начальные значения весов связи между нейронами полносвязного слоя необходимо задавать случайными числами из диапазона $[-b; b]$, где $b \geq 1 / n$, n – количество нейронов в каждом из двух полносвязных слоев.

Если число нейронов в двух соседних полносвязных слоях разное, то n находится как среднее арифметическое числа нейронов в соседних слоях: $n = (n_1 + n_2) / 2$. Рассмотренный вид инициализации весов – инициализация Ксавье.

Нормализация данных – это приведение данных к одному масштабу. После нормализации все данные становятся безразмерными и принадлежат одному диапазону. Данные ТВ нужно нормализовать перед тренировкой нейронной сети [3].

Рассмотрим два способа нормализации:

1. Пусть $X = \{x_1, \ldots, x_m\}$ все данные из столбца X, $a = \min\{x_1, \ldots, x_m\}$, $b = \max\{x_1, \ldots, x_m\}$ – минимальное и максимальное значение в столбце.

Тогда нормализованный столбец состоит из значений:

$$x_i' := \frac{x_i - a}{b - a}. \quad (8)$$

После нормализации значения в столбце X принадлежат интервалу $[0; 1]$.

2. Пусть $X = \{x_1, \ldots, x_m\}$ все данные из столбца X. Тогда величина и называется средним значением признака X.

$$\bar{x} := \frac{x_1 + \ldots + x_m}{m}. \quad (9)$$

Величина называется отклонением признака X.

$$s_x = \sqrt{\frac{1}{m-1}\left((x_1 - \bar{x})^2 + \ldots + (x_m - \bar{x})^2\right)}. \quad (10)$$

Тогда нормализованный столбец состоит из значений:

$$x_i' := \frac{x_i - \bar{x}}{s_x}. \quad (11)$$

С вероятностью 95–99 % значения нормализованного столбца принадлежат интервалу $[-3; 3]$.

Данные ТВ необходимо нормализовать перед началом тренировки нейронной сети. Когда сеть будет натренирована, то при получении нового объекта для предсказания его также необходимо нормализовать, используя числовые характеристики ТВ.

После обучения нейронной сети веса у нее могут оказаться по абсолютной величине очень большими. В результате небольшое изменение в нецелевых признаках приведет к большому изменению предсказанного признака [4].

Запрет нейронной сети на поиск слишком больших весов необходимо вписать в функцию потерь (5) или (6) следующим образом:

$$L_{reg} = L(w) + C\left(w_1^2 + \ldots + w_k^2\right), \tag{12}$$

где C – некоторая неотрицательная константа (ее необходимо выбрать); $w_1^2 + \ldots + w_k^2$ – квадраты весов связи и смещения.

Константа C – это значение весов нейронной сети. Она отражает компромисс между двумя противоположными задачами:

- поиск минимума функции потерь $L(w)$;
- уменьшение весов сети.

Чем больше константа C, тем выше приоритет второй задачи, и наоборот – чем ближе константа C к нулю, тем выше приоритет первой задачи.

Одним из способов повышения точности предсказания нейронной сети является создание множества нейронных сетей (ансамбля) на основе созданной. Данный способ достигается путем применения дропаута.

Дропаут – это преднамеренная деактивация части нейронов на шаге обучения (рисунок 7).

В качестве ответа нейронной сети необходимо взять усредненный ответ каждого члена ансамбля. Точность всего ансамбля выше точности каждого из его членов.

Перед тем, как произвести один шаг ГС, случайным образом необходимо временно исключить из нейронной сети часть нейронов. При этом возникнет новая сеть с новой функцией нейронной сети $F_{NN}(x)$ и новой функцией потерь $L(w)$. По данной функции потерь $L(w)$ осуществляется один шаг ГС и пересчитываются веса сети. И так необходимо рассчитать для каждого члена ансамбля.

Если после исключения части нейронов нейронная сеть перестает быть связной, то есть отсутствуют пути от входного слоя к выходному, то формируется новое множество нейронов для исключения до тех пор, пока сеть не станет связной.

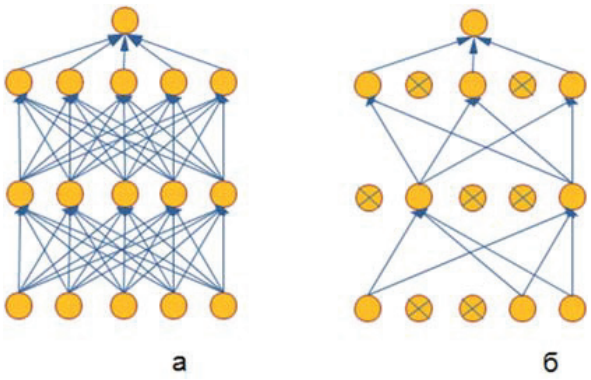

Рисунок 7 – Обычная (а) и после применения дропаута (б) нейронные сети

Алгоритм исключения нейронов:

1. Фиксируется вероятность исключения нейрона p (обычно в интервале $[0.2; 0.5]$).
2. На каждом шаге ГС для каждого нейрона нейронной сети проводится случайное испытание – исключать его или нет.
3. Формируется нейронная сеть из оставшихся нейронов, для нее выписывается своя функция потерь. После этого осуществляется один шаг ГС и дропаут повторяется.

Когда нейронная сеть натренирована с помощью дропаута для проверки предсказания сети на новом объекте, необходимо выход каждого нейрона умножить на вероятность исключения p этого нейрона.

Натренированную нейронную сеть необходимо проверить на точность и адекватность. Для этого необходимо множество объектов с известными ответами (размеченную выборку – РВ) разбить на две части: тренировочную (ТВ) и валидационную (ВВ) выборки (обычно 80 и 20 % соответственно). При этом тренировать нейронную сеть необходимо на ТВ, а проверять точность и адекватность на ВВ.

Формула для определения точности нейронной сети существенно зависит от типа задачи – задача регрессии или задача классификации.

Точность задачи регрессии определяется по следующей формуле:

$$MAE = \frac{1}{n}\left(\left|y_1 - z_1\right| + \ldots + \left|y_t - z_t\right|\right), \tag{13}$$

где MAE – mean absolute error – средняя абсолютная ошибка; n – количество выходных нейронов; y_t – истинное значение (признак Y); z_t – значение, предсказанное нейронной сетью.

Для определения точности задачи классификации необходимо информацию занести в матрицу ошибок (таблица 1).

Таблица 1 – Матрица ошибок

Матрица ошибок 0		Истинный класс 1	
Предсказанный класс	0	TN	FN
	1	FP	TP

TP – true positive – классификатор верно отнёс объект к рассматриваемому классу.

TN – true negative – классификатор верно утверждает, что объект не принадлежит к рассматриваемому классу.

FP – false positive – классификатор неверно отнёс объект к рассматриваемому классу.

FN – false negative – классификатор неверно утверждает, что объект не принадлежит к рассматриваемому классу.

Точность определяется по формуле:

$$precision = \frac{TP}{TP + FP}. \tag{14}$$

Полнота определяется по формуле:

$$recall = \frac{TP}{TP + FN}. \tag{15}$$

Стремятся, чтобы точность и полнота были равны 1 (100 %).

Если ошибка нейронной сети даже на ТВ высокая, то это означает, что имеет место недообучение сети (underfitting). В этом случае необходимо взять более мощную нейронную сеть (с большим числом нейронов).

Если ошибка нейронной сети на ТВ низкая, а ошибка на ВВ высокая, то это означает, что произошло переобучение (overfitting). В этом случае необходимы:

- регуляризация;
- дропаут;
- использование больших данных.

Заключение. Результаты теоретического исследования позволяют сделать следующие выводы:

1. В зависимости от требуемой задачи (регрессии или классификации) будет зависеть архитектура нейронной сети.
2. Для решения различных задач с помощью искусственных нейронных сетей необходимо их правильно и достаточно натренировать, используя при этом методы проверки точности предсказания.
3. Для борьбы с попаданием начальной точки ГС на плато и зону бесконечного спуска при минимизации функции потерь <<Eqn0437.wmf>>, необходимо использовать алгоритмы инициализации весов, нормализации данных, а также регуляризации.
4. Повышение точности предсказания искусственной нейронной сети можно добиться путем применения дропаута (преднамеренной деактивации части нейронов на шаге обучения).
5. «С помощью нейронных сетей возможно реализовать невероятные технологии. Главное преимущество нейронных сетей – их самообучаемость» [5].
6. Искусственные нейронные сети в перспективе по многим направлениям будут использоваться все шире во многих областях, где не нужны слишком интеллектуальные решения, выполняемые живым человеком. Насколько бы не была умна искусственная нейронная сеть, она всегда будет зависеть от своего создателя – человека.

Поступила: 02.06.22; рецензирована: 16.06.22; принята: 20.06.22.

Литература

1. *Лысцов Н.А.* Нейронные сети: применение и перспективы / Н.А. Лысцов, А.И. Мартышкин // Научное обозрение. Педагогические науки. 2019. № 3 (часть 2). С. 35–38.
2. *Лисовский А.Л.* Применение нейросетевых технологий для разработки систем управления. Стратегические решения и риск-менеджмент / А.Л. Лисовский. 2020; 11(4). С. 378–389. URL: https://doi.org/10.17747/2618-947X-923
3. *Николенко С.* Глубокое обучение. Погружение в мир нейронных сетей / С. Николенко, А. Кадурин, Е. Архангельская. СПб.: Питер, 2018. 480 с.
4. *Маслов А.С.* Нейронные сети / С. Николенко, А. Кадурин, Е. Архангельская // Международный студенческий научный вестник. 2018. № 3-1.
5. *Иванько А.Ф.* Информационные нейронные сети / А.Ф. Иванько, М.А. Иванько, О.Д. Колесникова // Научное обозрение. Технические науки. 2019. № 4. С. 11–16.