Глава 7

Нейронные сети прямого распространения

В данной главе будет продолжено изучение моделей искусственных нейронов применительно к задачам обучения. Рассмотрение таких моделей было начато в главе 3. В качестве примеров были описаны модели персептрона и логистического нейрона.

Успешное использование индивидуального нейрона в качестве предиктора базируется на предположении о существовании некоторой линейной зависимости между признаками рассматриваемых объектов. Очевидно, что такое предположение не всегда является обоснованным. Поэтому соответствующие алгоритмы обучения по своей природе являются слабыми учителями.

При описании алгоритма адаптивного бустинга было показано, что с использованием слабого учителя, может быть построен составной предиктор. Такой составной предиктор состоит из набора промежуточных предикторов, построенных с помощью слабого учителя, и усиливает их распознающие свойства. Если в качестве слабого учителя взять алгоритм обучения индивидуального нейрона, то получившийся составной предиктор будет представлять собой пример, так называемой, нейронной сети.

Изучению нейронных сетей [48] и будет посвящена настоящая глава.

7.1 Искусственный нейрон и функции активации

Описание архитектуры нейронной сети должно в себя включать описание её элементов, являющихся индивидуальными нейронами, а также задание её топологии, определяющей связи между элементами. Начнём с напоминания и обобщения определений, которые были даны в разделе 3.4.

Определение 7.1. *Искусственным нейроном* (*нейроном*) будем называть функцию вида

$$\mathbf{u} \longmapsto \sigma(\langle \mathbf{w}, \mathbf{u} \rangle + b) = \sigma\left(\sum_{j=1}^{m} w_j u_j + b \right) \qquad \big(\mathbf{w}, \mathbf{u} \in \mathbb{R}^m; m \in \mathbb{N}; b \in \mathbb{R}\big). \tag{7.1}$$

Функция $\sigma : \mathbb{R} \longrightarrow \mathbb{R}$ называется *функцией активации*, а величина b называется *смещением*. При этом будем говорить, что *входу* нейрона с номером j приписан *вес* w_j, и на этот вход поступает число u_j. Значение, вычисленное нейроном, поступает на его *выход*. Смещение b будем интерпретировать как вес нулевого входа, на который всегда поступает число 1. Поэтому для смещения будет также использоваться обозначение w_0.

Как правило, мы будем отождествлять нейрон с парой $(\mathbf{w}, b)$, состоящей из вектора весов его входов и смещения.

Нейрон, использующий функцию активации σ будем называть *σ-нейроном*, а в контексте рассматриваемой нейронной сети будем называть также *σ-узлом*.

Используемые на практике функции активации могут обладать совершенно разными свойствами. Они могут быть ограниченными и неограниченными, гладкими и разрывными. Большинство из них являются монотонно неубывающими функциями. Хотя и это свойство не является обязательным.

Приведём примеры функций активаций. Графики некоторых из них представлены на рис. 7.1.

В определении персептрона фигурирует функция активации sign. Другим примером ограниченной и кусочно-постоянной функции активации является $\sigma_{01} \coloneqq \mathbf{1}_{u \geqslant 0}$.

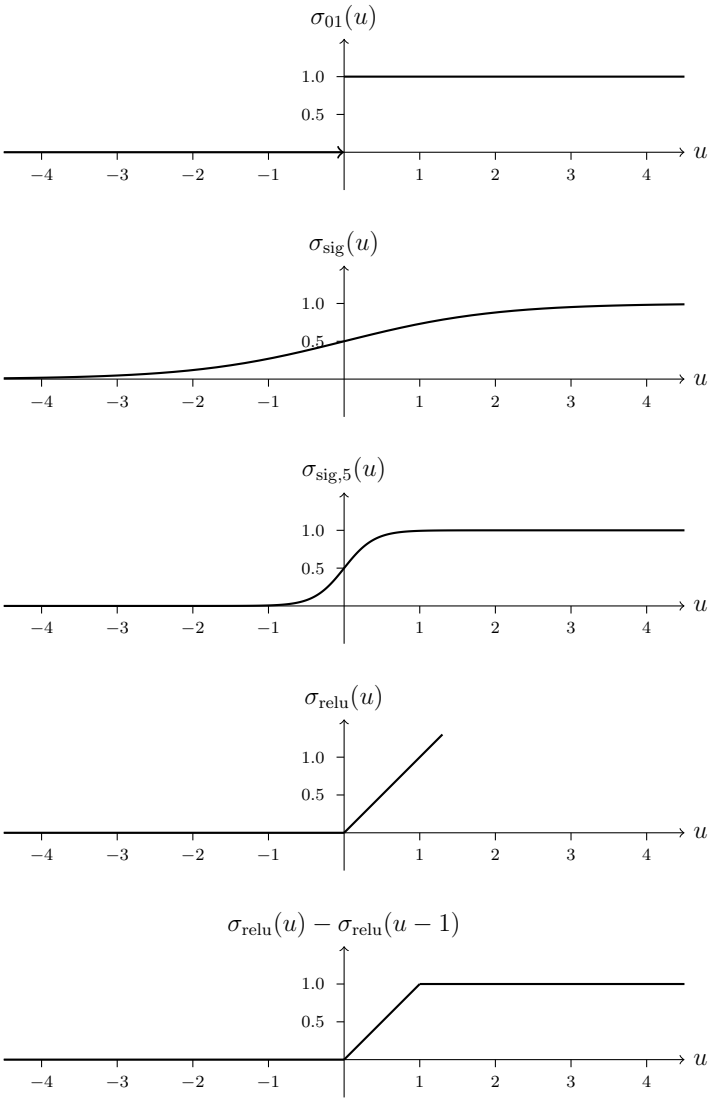


Рис. 7.1: Графики функций активации $\sigma_{01}(u)$, $\sigma_{\mathrm{sig}}(u)$, $\sigma_{\mathrm{sig},5}(u)$, $\sigma_{\mathrm{relu}}(u)$ и $\sigma_{\mathrm{relu}}(u) - \sigma_{\mathrm{relu}}(u-1)$.

Примером гладкой и строго возрастающей функции активации является сигмоида σ_{sig}. С помощью сигмоиды может быть определён целый класс гладких функций активации вида

$$\sigma_{\mathrm{sig},c}(u) \coloneqq \sigma_{\mathrm{sig}}(cu) = \frac{e^{cu}}{1 + e^{cu}} \quad (u \in \mathbb{R}; c > 0).$$

С их помощью, сколь угодно точно может, может быть приближена кусочно-постоянная функция σ_{01}.

Дальнейшее обобщение приводит к понятию сигмоидной функции активации.

Определение 7.2. Непрерывная функция $\sigma : \mathbb{R} \longrightarrow \mathbb{R}$ называется *сигмоидной*, если

$$\lim_{u \to -\infty} \sigma(u) = 0, \qquad \lim_{u \to +\infty} \sigma(u) = 1.$$

Гладкость функции активации играет важную роль при использовании методов численной оптимизации в процессе обучения. В частности наличие у неё производной позволяет использовать метод обратного распространения ошибки.

Производная функции активации $\sigma_{\mathrm{sig},c}$ по правилу

$$\sigma_{\mathrm{sig},c}{}'(u) = c\sigma_{\mathrm{sig},c}(u)(1 - \sigma_{\mathrm{sig},c}(u)) \quad (u \in \mathbb{R}; c > 0).$$

Другим примером гладкой функции активации служит гиперболический тангенс

$$\sigma_{\mathrm{t}}(u) \coloneqq \tanh(u) = \frac{e^u - e^{-u}}{e^u + e^{-u}} \quad (u \in \mathbb{R}).$$

Его производная вычисляется по правилу

$$\sigma'_{\mathrm{t}}(u) = 1 - \sigma^2_{\mathrm{t}}(u) \quad (u \in \mathbb{R}).$$

Гиперболический тангенс и сигмоида связаны соотношением

$$\sigma_{\mathrm{t}}(u) = 2\sigma_{\mathrm{sig},2}(u) - 1 \quad (u \in \mathbb{R}).$$

Примером неограниченной функции активации может служить *функция положительной срезки* (*rectified linear unit*)

$$\sigma_{\mathrm{relu}}(u) \coloneqq \max\{u, 0\} \quad (u \in \mathbb{R}).$$

Использование неограниченных функций активации безусловно оправданно в случае решения задачи регрессии. Например, использование нейрона с тождественной функцией активации в точности соответствует задаче линейной регрессии. В дальнейшем, тождественную функцию активации будем обозначать через σ_+.

В то же время использование неограниченной функции активации для решения задачи классификации, на первый взгляд, может показаться необычным. Забегая вперёд, отметим, что существует такое явление, как *насыщенность* (*saturation*) нейронной сети. Оно проявляется у многослойных (глубоких) нейронных сетей и связанно с особенностями их обучения. В случае использования сигмоиды в качестве функции активации нейронов в скрытых слоях их выходы всегда будут принимать значения очень близкие либо к 0, либо к 1. Это в свою очередь будет приводить к невозможности или существенному замедлинию процесса обучения. Как правило, итоговая нейронная сеть будет переобученной. Использование функции положительной срезки в качестве функции активации нейронов, находящихся в скрытых слоях сети, позволяет избежать этой проблемы.

Понятие функции активации может быть обобщено за счёт использования отображений вида $\mathbb{R}^m \longrightarrow \mathbb{R}^m$ $(m \in \mathbb{N})$. В качестве примера можно привести *многопеременную логистическую функцию*

$$\sigma_{\mathrm{softmax},c}(\mathbf{u}) \coloneqq \frac{1}{\sum_{i=1}^m e^{cu_i}} (e^{cu_1}, \ldots, e^{cu_m})^t \qquad (\mathbf{u} \in \mathbb{R}^m; c > 0).$$

В дальнейшем, мы будем часто доопределять обычную функцию активации вида $\sigma : \mathbb{R} \longrightarrow \mathbb{R}$, полагая

$$\sigma : \mathbf{u} \longmapsto \big(\sigma(u_1), \ldots, \sigma(u_m)\big)^t \qquad (\mathbf{u} \in \mathbb{R}^m).$$

7.2 Дискриминационные функции

Установим важное свойство сигмоидных функций, которое будет использоваться при изучении аппроксимационных свойств нейронных сетей.

Определение 7.3. Конечная мера (со знаком) μ, заданная на некоторой борелевской σ-алгебре $\mathcal{B}(\Omega)$, называется *регулярной* [49], если для любого измеримого множества $A \in \mathcal{B}(\Omega)$ и $\varepsilon > 0$ существуют измеримые множества $F, G \in \mathcal{B}(\Omega)$ такие, что одновременно выполняются следующие условия:

- $\overline{F} \subseteq A$;
- $A \subseteq \operatorname{int}(G)$;
- для любого измеримого множества $C \subseteq G \setminus F$ выполняется условие $\mu(C) < \varepsilon$.

Обозначим

$$I_m \coloneqq [0,1]^m \qquad (m \in \mathbb{N}).$$

Определение 7.4. Непрерывная функция $\sigma : \mathbb{R} \longrightarrow \mathbb{R}$ называется *дискриминационной*, если для любой регулярной меры $\mu \in \mathcal{M}_{\pm}(I_m)$ из условия

$$\int\limits_{I_m} \sigma\big(\langle \mathbf{w}, \mathbf{u}\rangle + b\big)\, \mu(d\mathbf{u}) = 0 \qquad \big(\text{для всех } \mathbf{w} \in \mathbb{R}^m;\ b \in \mathbb{R}\big) \tag{7.2}$$

следует, что $\mu = 0$.

Теорема 7.1. *Любая сигмоидная функция является дискриминационной.*

В начале докажем вспомогательное утверждение.

Утверждение 7.1. *Обозначим*

$$\begin{aligned} H^{+}_{\mathbf{w},b} &\coloneqq \big\{\mathbf{u} \in I_m \,:\, \langle \mathbf{w}, \mathbf{u}\rangle + b > 0\big\},\\ H^{-}_{\mathbf{w},b} &\coloneqq \big\{\mathbf{u} \in I_m \,:\, \langle \mathbf{w}, \mathbf{u}\rangle + b < 0\big\},\\ P_{\mathbf{w},b} &\coloneqq \big\{\mathbf{u} \in I_m \,:\, \langle \mathbf{w}, \mathbf{u}\rangle + b = 0\big\} \qquad \big(\mathbf{w} \in \mathbb{R}^m;\ b \in \mathbb{R}; m \in \mathbb{N}\big). \end{aligned}$$

Если для регулярной меры $\mu \in \mathcal{M}_{\pm}(I_m)$ выполняется условие

$$\mu(H^{+}_{\mathbf{w},b}) = 0,\ \mu(H^{-}_{\mathbf{w},b}) = 0,\ \mu(P_{\mathbf{w},b}) = 0 \qquad \big(\textit{для всех } \mathbf{w} \in \mathbb{R}^m;\ b \in \mathbb{R}\big),$$

то $\mu = 0$.

◄ Зафиксируем произвольный вектор $\mathbf{w} \in \mathbb{R}^m$. Обозначим через K образ множества I_m при отображении $\mathbf{u} \longmapsto \langle \mathbf{w}, \mathbf{u}\rangle$. Заметим, что $K \subset \mathbb{R}$ является компактом.

Рассмотрим нормированное пространство $(\mathcal{L}^{\infty}(K), \|\cdot\|_{\mathcal{L}^{\infty}(K)})$ и определим на нём линейный функционал F по правилу

$$F(f) \coloneqq \int\limits_{I_m} f(\langle \mathbf{w}, \mathbf{u}\rangle)\, \mu(d\mathbf{u}) \qquad \big(f \in \mathcal{L}^{\infty}(K)\big).$$

Этот функционал является непрерывным, так как

$$F(f) \leqslant \int\limits_{I_m} \big|f(\langle \mathbf{w}, \mathbf{u}\rangle)\big|\, \mu(d\mathbf{u}) \leqslant \|f\|_{\mathcal{L}^{\infty}(K)} \left| \int\limits_{I_m} \mu(d\mathbf{u}) \right| = \mu(I_m)\, \|f\|_{\mathcal{L}^{\infty}(K)}.$$

Для любого $b \in \mathbb{R}$ выполняются равенства

$$F\big(\mathbf{1}_{[b,\infty)}\big) = \int\limits_{\{\mathbf{u}\in I_m\,:\,\langle \mathbf{w},\mathbf{u}\rangle \geqslant b\}} \mu(d\mathbf{u}) = \mu(H^+_{\mathbf{w},-b}) + \mu(P_{\mathbf{w},-b}) = 0$$

и

$$F\big(\mathbf{1}_{(b,\infty)}\big) = \int\limits_{\{\mathbf{u}\in I_m\,:\,\langle \mathbf{w},\mathbf{u}\rangle > b\}} \mu(d\mathbf{u}) = \mu(H^+_{\mathbf{w},-b}) = 0.$$

Для любых $c, d \in \mathbb{R}$ выполняются равенства $\mathbf{1}_{[c,d]} = \mathbf{1}_{[c,\infty)} - \mathbf{1}_{(d,\infty)}$ и $\mathbf{1}_{[c,d)} = \mathbf{1}_{[c,\infty)} - \mathbf{1}_{[d,\infty)}$, а значит

$$F\big(\mathbf{1}_{[c,d]}\big) = 0 \quad \text{и} \quad F\big(\mathbf{1}_{[c,d)}\big) = 0.$$

Обозначим через $\mathcal{H} \subset \mathcal{L}^\infty(K)$ множество всех конечных линейных комбинаций функций вида $\mathbf{1}_{[c,d]}$ и $\mathbf{1}_{[c,d)}$. Очевидно, что

$$F\big(h\big) = 0 \qquad \big(\text{для всех } h \in \mathcal{H}\big).$$

Любая непрерывная функция f, заданная на K, равномерно приближается некоторой последовательностью $\{h_n \in \mathcal{H}\}_{n\in\mathbb{N}}$. Следовательно, учитывая напрерывность линейного функционала F, получим

$$F\big(f\big) = F\big(\lim_{n\to\infty} h_n\big) = \lim_{n\to\infty} F\big(h_n\big) = 0.$$

Вычислим преобразование Фурье $\widehat{\mu}$ для меры μ. Получим

$$\begin{aligned}\widehat{\mu}(\mathbf{w}) &= \int\limits_{I_m} e^{i\langle \mathbf{w},\mathbf{u}\rangle}\,\mu(d\mathbf{u}) = \int\limits_{I_m} \cos(\langle \mathbf{w},\mathbf{u}\rangle)\,\mu(d\mathbf{u}) + i\int\limits_{I_m} \sin(\langle \mathbf{w},\mathbf{u}\rangle)\,\mu(d\mathbf{u}) \\ &= F\big(\cos\big) + iF\big(\sin\big) = 0 \qquad \big(\text{для всех } \mathbf{w} \in \mathbb{R}^m\big).\end{aligned}$$

Равенство преобразований Фурье конечных мер, заданных на $\mathcal{B}(I_m)$, означает равенство самих мер. Учитывая этот факт, заключим, что $\mu = 0$. ■

◀ (доказательство теоремы 7.1) Рассмотрим произвольные сигмоидную функцию σ и регулярную меру $\mu \in \mathcal{M}_\pm(I_m)$. Предположим, что для них выполняется условие (7.2).

Далее, зафиксируем произвольные $\mathbf{w} \in \mathbb{R}^m$, $b \in \mathbb{R}$. Определим параметризованное семейство непрерывных функций, заданных на I_m, полагая

$$\sigma_{\lambda,\theta}(\mathbf{u}) \coloneqq \sigma\big(\lambda(\langle \mathbf{w},\mathbf{u}\rangle + b) + \theta\big) \qquad \big(\lambda, \theta \in \mathbb{R}\big).$$

Заметим, что при фиксированных $\mathbf{u}, \theta$ выполняются условия

$$\lim_{\lambda\to\infty} \sigma_{\lambda,\theta}(\mathbf{u}) = \begin{cases} 1, & \text{если} \quad \langle \mathbf{w},\mathbf{u}\rangle + b > 0; \\ 0, & \text{если} \quad \langle \mathbf{w},\mathbf{u}\rangle + b < 0; \\ \sigma(\theta), & \text{иначе.} \end{cases}$$

Определим параметризованное семейство ограниченных измеримых функций, заданных на I_m, полагая

$$\gamma_\theta(\mathbf{u}) \coloneqq \begin{cases} 1, & \text{если} \quad \mathbf{u} \in H^+_{\mathbf{w},b}; \\ 0, & \text{если} \quad \mathbf{u} \in H^-_{\mathbf{w},b}; \\ \sigma(\theta), & \text{если} \quad \mathbf{u} \in P_{\mathbf{w},b} \end{cases} \qquad \big(\theta \in \mathbb{R}\big).$$

Заметим, что имеет место поточечная сходимость $\sigma_{\lambda,\theta}(\mathbf{u}) \longrightarrow \gamma_\theta(\mathbf{u})$ на I_m при $\lambda \longrightarrow \infty$. Применяя теорему Лебега о мажорированной сходимости, получим

$$\lim_{\lambda\to\infty} \int\limits_{I_m} \sigma_{\lambda,\theta}(\mathbf{u})\,\mu(d\mathbf{u}) = \int\limits_{I_m} \gamma_\theta(\mathbf{u})\,\mu(d\mathbf{u})$$
$$= \mu(H^+_{\mathbf{w},b}) + \sigma(\theta)\mu(P_{\mathbf{w},b}) \qquad \big(\theta \in \mathbb{R}\big).$$

Из предположения (7.2) следует, что

$$\int\limits_{I_m} \sigma_{\lambda,\theta}(\mathbf{u})\,\mu(d\mathbf{u}) = 0 \qquad \big(\lambda, \theta \in \mathbb{R}\big),$$

а значит

$$\mu(H^+_{\mathbf{w},b}) + \sigma(\theta)\mu(P_{\mathbf{w},b}) = 0 \qquad \big(\theta \in \mathbb{R}\big). \tag{7.3}$$

Переходя в левой части (7.3) к пределу при $\lambda \longrightarrow -\infty$, получим $\mu(H^+_{\mathbf{w},b}) = 0$. Переходя в правой части (7.3) к пределу при $\lambda \longrightarrow \infty$, получим $\mu(H^+_{\mathbf{w},b}) + \mu(P_{\mathbf{w},b}) = 0$, а значит $\mu(P_{\mathbf{w},b}) = 0$.

Учитывая произвльность выбора $\mathbf{u}, \theta$ и тождество $H^-_{\mathbf{w},b} = H^+_{-\mathbf{w},-b}$, получим $\mu(H^-_{\mathbf{w},b}) = 0$.

В заключение, применяя утв. 7.1, получим дискриминационность функции σ. ■

7.3 Определение нейронной сети

Перейдём к определению понятия нейронной сети прямого распространения (feedforward neural network), которое является основным объектом изучения этой и следующей глав. Мы не будем рассматривать другие типы нейронных сетей. Поэтому в дальнейшем нейронные сети прямого распространения будем называть просто нейронными сетями.

Определение 7.5. *Нейронной сетью прямого распространения* (*нейронной сетью*) называется набор $N = (V, E, F_\sigma, F_\omega, F_\beta)$, где (V, E) – ациклический ориентированный граф, а $F_\sigma, F_\omega, F_\beta$ – функции разметки.

Вершины графа (V, E), не имеющие входящих рёбер, называются *входами*, а вершины, не имеющие исходящих рёбер, называются *выходами*.

Функция F_σ ставит в соответствие каждой вершине $v \in V$, не являющейся входом, некоторую функцию активации σ_v. При этом вершина v называется *σ_v-узлом*.

Функция F_ω ставит в соответствие каждому ребру $(v', v) \in E$ *вес* $\omega_{v',v} \in \mathbb{R}$.

Функция F_β ставит в соответствие каждой вершине $v \in V$, не являющейся входом, *смещение* $\beta_v \in \mathbb{R}$.

Определение 7.6. Пусть $N = (V, E, F_\sigma, F_\omega, F_\beta)$ – нейронная сеть. Существует разбиение множества вершин

$$V = \bigcup_{l=0}^{L} V_l, \tag{7.4}$$

где

- $V_l \neq \varnothing$ и $V_l \cap V_k = \varnothing$ при $l \neq k$ $(l, k = 0, 1, \ldots, L)$;
- если $(v', v) \in E$, $v' \in V_l$ и $v \in V_k$, то $k > l$;
- V_0 состоит из всех входов, а V_L состоит из всех выходов.

Множество V_l $(l = 0, 1, \ldots, L)$ называется *l-м слоем*. Слой V_0 называется *входным*. Слой V_L называется *выходным*. Слои $V_1, \ldots, V_{L-1}$ называются *внутренними* или *скрытыми*. Входной слой является вспомогательным. Говорят, что нейронная сеть N является *L слойной*.

Для краткости, количество элементов в l-м слое будем обозначать через $m_l \coloneqq |V_l|$.

Будем предполагать, что внутри каждого слоя все узлы занумерованы. Через $v_i^{(l)}$ будем обозначать i-й узел внутри l-го слоя.

Приведённое формальное определение хорошо согласуется с интуитивным представлением о нейронной сети, как совокупности нейронов, чьи входы и выходы соединены друг с другом по некоторым правилам. Однако в нём ничего не говорится о возможности вычисления (задания) функциональных зависимостей с помощью нейронных сетей.

Определение 7.7. Пусть $N = (V, E, F_\sigma, F_\omega, F_\beta)$ – нейронная сеть с разбиением на слои (7.4). Поставим в соответствие каждому узлу $v \in V$ функцию вида

$$h(\cdot; v) : \mathbb{R}^{m_0} \longrightarrow \mathbb{R}.$$

Входам будут соответствовать проекции

$$h\big(\mathbf{u}; v_i^{(0)}\big) \coloneqq u_i \qquad \big(\mathbf{u} \in \mathbb{R}^{m_0}; i = 1, \ldots, m_0\big).$$

Вершине $v \in V \setminus V_0$, будет соответствовать функция

$$h\big(\mathbf{u}; v\big) \coloneqq \sigma_v\Bigg(\sum_{v':(v',v)\in E} \omega_{v',v}\, h(\mathbf{u}; v') + \beta_v\Bigg) \qquad \big(\mathbf{u} \in \mathbb{R}^{m_0}\big).$$

Определим отображение

$$h_N : \mathbb{R}^{m_0} \longrightarrow \mathbb{R}^{m_L},$$

полагая

$$h_N\big(\mathbf{u}\big) \coloneqq \Big(h\big(\mathbf{u}; v_1^{(L)}\big), \ldots, h\big(\mathbf{u}; v_{|V_L|}^{(L)}\big)\Big)^t \qquad \big(\mathbf{u} \in \mathbb{R}^{m_0}\big).$$

Будем говорить, что отображение h_N *задаётся* (*определяется*) нейронной сетью N.

Замечание. В дальнейшем, мы обычно будем отождествлять нейронную сеть N и задаваемую ею функцию h_N. Соответственно, будут также отождествляться и классы нейронных сетей с семейством задаваемых этими нейронными сетями функций.

Пример 7.1. Рассмотрим нейронную сеть N_1, изображённую на рис. 7.2 а). Входы изображены маленькими чёрными кружками. Узлы изображены белыми кружками, внутри которых помещены символы соответствующих функций активации. Если узел имеет ненулевое смещение, то соответствующее значение помещается рядом с белым кружком (вверху и справа). Веса рёбер изображаются вверху или внизу соответствующей стрелки. Нумерация узлов внутри слоя осуществляется сверху вниз.

Нейронная сеть N_1 является 3-х слойной. Первый слой содержит два σ_{relu}-узла. Второй слой содержит три σ_{relu}-узла. Третий выходной слой содержит один σ_{sig}-узел.

Узлам первого слоя соответствуют функции

$$\begin{aligned} h\big(u_1, u_2; v_1^{(1)}\big) &= \sigma_{\text{relu}}\big(u_1 + 0.5\big) \\ h\big(u_1, u_2; v_2^{(1)}\big) &= \sigma_{\text{relu}}\big(u_1 - u_2 + 0.5\big) \qquad \big(u_1, u_2 \in \mathbb{R}\big). \end{aligned}$$

Узлам второго слоя соответствуют функции

$$h\big(u_1,u_2;v_1^{(2)}\big)=\sigma_{\mathrm{relu}}\big(2h\big(u_1,u_2;v_1^{(1)}\big)\big)$$
$$h\big(u_1,u_2;v_2^{(2)}\big)=\sigma_{\mathrm{relu}}\big(h\big(u_1,u_2;v_2^{(1)}\big)\big)$$
$$h\big(u_1,u_2;v_3^{(2)}\big)=\sigma_{\mathrm{relu}}\big(2h\big(u_1,u_2;v_2^{(1)}\big)\big)\qquad\big(u_1,u_2\in\mathbb{R}\big).$$

Таким образом, нейронная сеть N_1 задаёт функцию

$$h_{N_1}\big(u_1,u_2\big)=\sigma_{\mathrm{sig}}\big(h\big(u_1,u_2;v_1^{(1)}\big)+h\big(u_1,u_2;v_1^{(2)}\big)$$
$$+2h\big(u_1,u_2;v_2^{(2)}\big)+h\big(u_1,u_2;v_3^{(2)}\big)-0.5\big)\qquad\big(u_1,u_2\in\mathbb{R}\big).$$

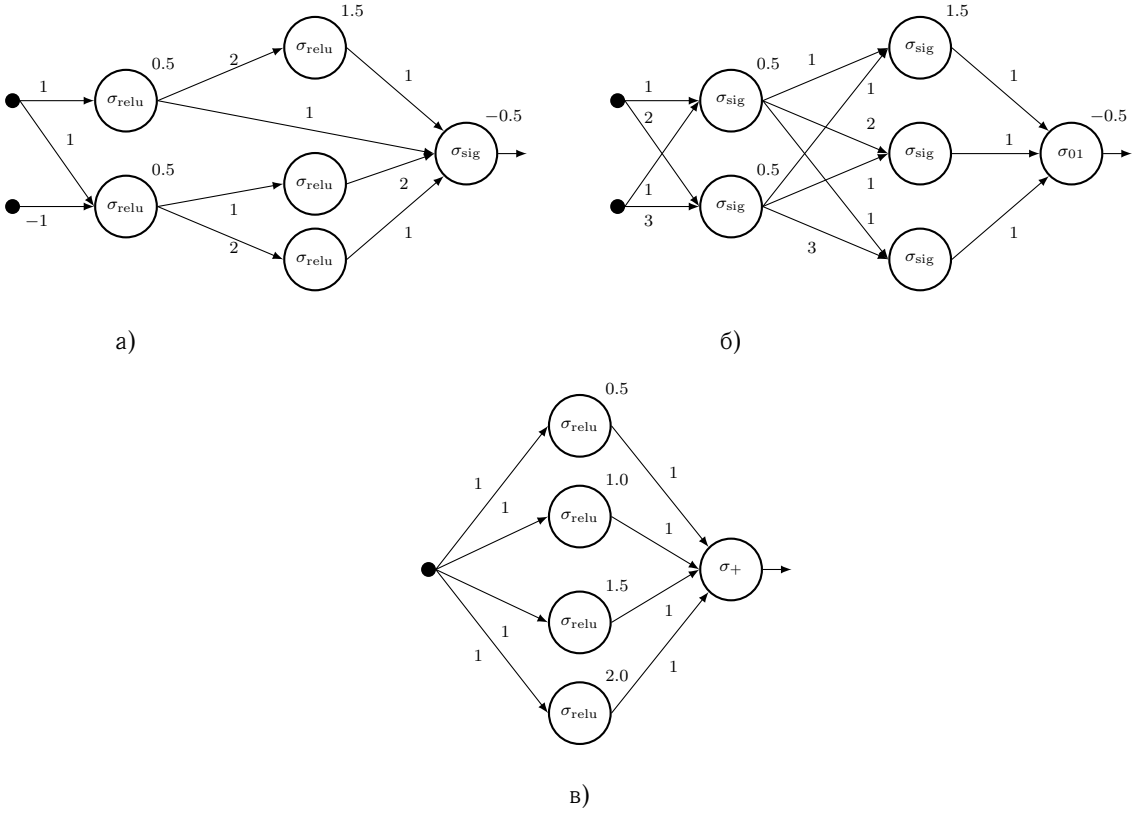


Рис. 7.2: Примеры нейронных сетей.

Разделение узлов нейронной сети на слои приводит к идее о том, что только узлы из соседних слоёв могут непосредственно взаимодействовать друг с другом могут. Очевидно, что нейронная сеть N_1 из предыдущего примера не обладает этим свойством. Узел из первого слоя напрямую соединён с узлом из третьего слоя. Поэтому данное условие требует дополнительной формализации.

Определение 7.8. Пусть $N=(V,E,F_\sigma,F_\omega,F_\beta)$ – нейронная сеть с разбиением на слои (7.4). Будем говорить, что N обладает свойством *полносвязности относительно соседних слоёв*, если одновременно выполняются следующие условия:

- если $v'\in V_{l-1}$ и $v\in V_l$ $(1\leqslant l\leqslant L)$, то $(v',v)\in E$;
- если $(v',v)\in E$, то найдётся l $(1\leqslant l\leqslant L)$ такой, что $v'\in V_{l-1}$ и $v\in V_l$.

Утверждение 7.2. *Пусть* $m_0, m_1, \ldots, m_L \in \mathbb{N}$ *и* $m \coloneqq m_0$. *Предположим, что функция* $h : \mathbb{R}^m \longrightarrow \mathbb{R}^{m_L}$ *имеет следующее представление*

$$h\big(\mathbf{u}\big) = \sigma_L\big(\mathbf{W}^{(L)}\sigma_{L-1}\big(\mathbf{W}^{(L-1)} \cdots \sigma_1\big(\mathbf{W}^{(1)}\mathbf{u} + \mathbf{b}^{(1)}\big) \cdots + \mathbf{b}^{(L-1)}\big) + \mathbf{b}^{(L)}\big), \tag{7.5}$$

где

- $\mathbf{u} \in \mathbb{R}^m$;
- $\mathbf{W}^{(l)} \in \mathbb{R}^{m_l \times m_{l-1}}$, $\mathbf{b}^{(l)} \in \mathbb{R}^{m_l}$ $(l = 1, \ldots, L)$;
- $\sigma_1, \ldots, \sigma_{L-1}, \sigma_L$ – *функции активации.*

Тогда существует полносвязная относительно соседних слоёв нейронная сеть N *такая, что* $h = h_N$.

Верно и обратное утверждение. Для функции, задаваемой полносвязной относительно соседних слоёв нейронной сетью, имеет место представление (7.5).

Опишем классы нейронных сетей, которые будут изучаться в дальнейшем.

Определение 7.9. Класс $\mathcal{F}^{m_L,\ldots,m_0}_{\sigma_L,\ldots,\sigma_1}$ состоит из всех полносвязных относительно соседних слоёв нейронных сетей, обладающих следующими свойствами. Каждая нейронная сеть

- является L слойной;
- l-й $(l = 0, 1, \ldots, L)$ слой содержит m_l узлов;
- узлам невходного l-го $(l = 1, \ldots, L)$ слоя приписана функция активации σ_L.

В дальнейшем, подобные классы будем называть *моделями нейронных сетей*.

Таким образом, модель нейронной сети состоит из всех нейронных сетей, имеющих одинаковую топологию и отличающихся только значениями весов рёбер и смещений.

Функции, задаваемые нейронными сетями из класса $\mathcal{F}^{m_L,\ldots,m_0}_{\sigma_L,\ldots,\sigma_1}$, образуют параметризованной семейство. Каждая такая функция $h(\mathbf{u};\boldsymbol{\theta})$ имеет вид (7.5) относительно набора своих параметров

$$\boldsymbol{\theta} \coloneqq \big\{\mathbf{W}^{(L)}, \ldots, \mathbf{W}^{(1)}; \mathbf{b}^{(L)}, \ldots, \mathbf{b}^{(1)}\big\}. \tag{7.6}$$

При этом, *разметностью* набора параметров $\boldsymbol{\theta}$ (*разметностью* модели нейронной сети $\mathcal{F}^{m_L,\ldots,m_0}_{\sigma_L,\ldots,\sigma_1}$) будем называть величину

$$|\boldsymbol{\theta}| \coloneqq \sum_{l=1}^{L} m_l(m_{l-1} + 1), \tag{7.7}$$

которая равна общему числу элементов матриц и векторов, содержащихся в $\boldsymbol{\theta}$.

Пример 7.2. Нейронная сеть N_2, изображённая на рис. 7.2 б) имеет модель $\mathcal{F}^{1,3,2,2}_{\sigma_{01},\sigma_{\text{sig}},\sigma_{\text{sig}}}$. Она задаёт функцию

$$h_{N_2}\big(\mathbf{u}\big) = \sigma_{01}\big(\mathbf{W}^{(3)}\sigma_{\text{sig}}\big(\mathbf{W}^{(2)}\sigma_{\text{sig}}\big(\mathbf{W}^{(1)}\mathbf{u} + \mathbf{b}^{(1)}\big) + \mathbf{b}^{(2)}\big) + \mathbf{b}^{(3)}\big) \qquad \big(\mathbf{u} \in \mathbb{R}^2\big),$$

где

$$\mathbf{W}^{(1)} = \begin{pmatrix} 1 & 1 \\ 2 & 3 \end{pmatrix}, \mathbf{W}^{(2)} = \begin{pmatrix} 1 & 1 \\ 2 & 1 \\ 1 & 3 \end{pmatrix}, \mathbf{W}^{(3)} = \begin{pmatrix} 1 & 1 & 1 \end{pmatrix},$$

$$\mathbf{b}^{(1)} = \begin{pmatrix} 0.5 \\ 0.5 \end{pmatrix}, \mathbf{b}^{(2)} = \begin{pmatrix} 1.5 \\ 0 \\ 0 \end{pmatrix}, \mathbf{b}^{(3)} = \begin{pmatrix} -0.5 \end{pmatrix}.$$

Определение 7.10. Класс $\mathcal{F}^{L}_{+,\sigma,m}$ состоит из всех полносвязных относительно соседних слоёв нейронных сетей, обладающих следующими свойствами. Каждая нейронная сеть

- является $L+1$ слойной;
- выходной слой состоит из одного σ_+-узла;
- всем узлам из внутренних слоёв приписана функция активации σ.

Для краткости класс $\mathcal{F}^{1}_{+,\sigma,m}$ будем обозначать через $\mathcal{F}_{+,\sigma,m}$. С учётом ранее сделанного замечания об отождествлении класса нейронных сетей и семейства задаваемых ими функций можно дать эквивалентное определение

$$\mathcal{F}_{+,\sigma,m} \coloneqq \left\{ \mathbf{u} \longmapsto \sum_{j=1}^{M} a_j \sigma\big(\langle \mathbf{w}_j, \mathbf{u}\rangle + b_j\big) \;:\; \mathbf{u}, \mathbf{w}_j \in \mathbb{R}^m \,; a_j, b_j \in \mathbb{R} \,; M \in \mathbb{N} \right\}.$$

Пример 7.3. Нейронная сеть N_3, изображённая на рис. 7.2 в) принадлежит классам $\mathcal{F}^{1,4,1}_{\sigma_+,\sigma_{\mathrm{relu}}}$ и $\mathcal{F}_{+,\sigma_{\mathrm{relu}},1}$. Она задаёт функцию

$$h_{N_3}(u) = \sigma_{\mathrm{relu}}(u+0.5) + \sigma_{\mathrm{relu}}(u+1.0) + \sigma_{\mathrm{relu}}(u+1.5) + \sigma_{\mathrm{relu}}(u+2.0) \qquad (u \in \mathbb{R}).$$

7.4 Обучение нейронных сетей

Напомним, что в соответствии с декомпозицией ошибки обучения процесс обучения сводится к минимизации эмпирического риска. В такой постановке задачи предполагается, что имеется некоторый набор обучающих примеров $\mathbf{z}$ длины $n \in \mathbb{N}$. Функция эмпирического риска определяется как

$$r(l; h, \mathbf{z}) \coloneqq \frac{1}{n} \sum_{i=1}^{n} l(h(x_i), y_i),$$

где l – используемая функция потерь, а $z_i = (x_i, y_i)$ – i-й пример в наборе $\mathbf{z}$.

Соответственно, минимизация эмпирического риска означает нахождение предиктора

$$h_{\mathbf{z}} \in \operatorname*{arg\,min}_{h \in \mathcal{H}} r(h, \mathbf{z}). \tag{7.8}$$

В качестве класса предикторов будем рассматривать некоторую фиксированную модель нейронной сети $\mathcal{F}^{m_L,\ldots,m_0}_{\sigma_L,\ldots,\sigma_1}$. У этого класса в качестве множества объектов X выступает $\mathbb{R}^{m_0}$, а в качестве множества меток Y выступает $\mathbb{R}^{m_L}$.

Данный класс представляет собой параметризованное семейство функций. Поэтому задача минимизации (7.8) может быть переформулирована

$$\boldsymbol{\theta}_{\mathbf{z}} \in \operatorname*{arg\,min}_{\boldsymbol{\theta} \in \mathbb{R}^p} r(\boldsymbol{\theta}), \qquad r(\boldsymbol{\theta}) \coloneqq r(h(\cdot; \boldsymbol{\theta}), \mathbf{z}), \tag{7.9}$$

где p – размерность набора параметров (7.7) модели нейронной сети.

Метод градиентного спуска

Рассмотрим *градиентный метод* (см. [27]) поиска точки минимума произвольной дифференцируемой функции $r : \mathbb{R}^p \longrightarrow \mathbb{R}$ $(p \in \mathbb{N})$. Метод предполагает построение последовательности приближений к точке минимума

$$\boldsymbol{\theta}^{(0)}, \boldsymbol{\theta}^{(1)}, \ldots, \boldsymbol{\theta}^{(n)}, \boldsymbol{\theta}^{(n+1)}, \ldots \tag{7.10}$$

Задаётся начальное приближение $\boldsymbol{\theta}^{(0)}$, а каждое следующее приближение строится на основе рекуррентной формулы

$$\boldsymbol{\theta}^{(n+1)} = \boldsymbol{\theta}^{(n)} - h_n \nabla r\big(\boldsymbol{\theta}^{(n)}\big) \qquad (n \in \mathbb{N}). \tag{7.11}$$

Скалярный множитель $h_n \geqslant 0$ перед градиентом называется *длиной шага*. Существует две основные стратегии выбора длины шага.

Невозрастающая последовательность $\{h_n\}_{n\in\mathbb{N}}$ выбирается заранее. Например, может быть выбрана постоянная длина шага $h_n = h > 0$.

В качестве длины шага может быть выбрано решение задачи одномерной оптимизации

$$h_n = \arg\min_{h\geqslant 0} r\big(\boldsymbol{\theta}^{(n)} - h\nabla r\big(\boldsymbol{\theta}^{(n)}\big)\big).$$

Подобная задача может решаться, например, методом золотого сечения или бисекций.

Применимость градиентного метода базируется на следующем факте. Для дифференцируемой функции r антиградиент $-\nabla r$ является направлением её наискорейшего локального убывания. Поэтому градиентный метод с подходящей стратегией выбора длины шага приведёт к нахождению локального минимума. Следует подчеркнуть, что в общем случае градиентный метод не гарантирует нахождение глобального минимума.

Вернёмся к рассмотрению случая, когда функция r является функцией минимизации эмпирического риска (7.9) и, соответственно, имеет вид

$$r(\boldsymbol{\theta}) = \frac{1}{n}\sum_{i=1}^{n} r_i(\boldsymbol{\theta}). \tag{7.12}$$

В этом случае, вычисление её градиента ∇r сводится к вычислению градиентов ∇r_i слагаемых в правой части равенства (7.12). Такие градиенты могут быть вычислены с помощью так называемого *метода обратного распространения ошибки* [50].

Прежде, чем перейти к его описанию, заметим, что на практике для решения рассматриваемой оптимизационной задачи (7.9) обычно используется так называемый метод *стохастического градиентного спуска* [51]. Его особенность состоит в том, что на каждом шаге вычислительного процесса (7.11) вычисляется градиент не по всем примерам из обучающей выборки, а по одному случайно выбранному примеру или по случайно выбранному набору примеров. Это позволяет существенно снизить как вычислительную погрешность каждого шага вычислительного процесса, так и общий объём вычислений. Для более подробного ознакомления с этим подходом можно обратиться, например, к обзору [52].

Вычисление градиента

Рассмотрим произвольную функцию $f : \mathbb{R}^p \longrightarrow \mathbb{R}$, которая может быть представлена в виде

$$f(\boldsymbol{\theta}) = C(h(\mathbf{u}; \boldsymbol{\theta})) \qquad (\boldsymbol{\theta} \in \mathbb{R}^p),$$

где $h(\cdot;\boldsymbol{\theta}) \in \mathcal{F}^{m_L,\ldots,m_0}_{\sigma_L,\ldots,\sigma_1}$, а $\mathbf{u} \in \mathbb{R}^{m_0}$ – фиксированный вектор. Именно такой вид имеют слагаемые в правой части равенства (7.12). Дополнительно будем предполагать гладкость функции C.

Нашей целью является вычисление градиента

$$\nabla f = \left\{ \ldots, \frac{\partial f}{\partial w_{jk}^{(l)}}, \ldots, \frac{\partial f}{\partial b_j^{(l)}}, \ldots \right\}.$$

Введём обозначения

$$\begin{array}{rcll} \mathbf{u}^{(0)} & \coloneqq & \mathbf{u}, & \\ \mathbf{u}^{(l)} & \coloneqq & \sigma_l\big(\mathbf{s}^{(l)}\big), & \\ \mathbf{s}^{(l)} & \coloneqq & \mathbf{W}^{(l)}\mathbf{u}^{(l-1)} + \mathbf{b}^{(l)} & (l = 1, \ldots, L), \end{array} \tag{7.13}$$

которые можно переписать в координатном виде

$$\begin{array}{rcll} u_j^{(l)} & \coloneqq & \sigma_l\big(s_j^{(l)}\big), & \\ s_j^{(l)} & \coloneqq & \displaystyle\sum_{k=1}^{m_{l-1}} w_{jk}^{(l)} u_k^{(l-1)} + b_j^{(l)} & (j = 1, \ldots, m_l). \end{array}$$

Величина

$$\delta_j^{(l)} \coloneqq \frac{\partial f}{\partial s_j^{(l)}}$$

интерпретируется как ошибка, возникшая в процессе обучения в j-ом нейроне l-го слоя. Вычислим эти ошибки.

Утверждение 7.3 (правило BP_1)**.** *Для выходного слоя L выполняются равенства*

$$\delta_j^{(L)} = \frac{\partial C}{\partial u_j^{(L)}} \cdot \sigma_L'\big(s_j^{(L)}\big) \qquad (j = 1, \ldots, m_L).$$

◀ Используя правило дифференцирования сложной функции, запишем

$$\delta_j^{(L)} = \frac{\partial C}{\partial s_j^{(L)}} = \sum_{k=1}^{m_L} \frac{\partial C}{\partial u_k^{(L)}} \cdot \frac{\partial u_k^{(L)}}{\partial s_j^{(L)}} = \frac{\partial C}{\partial u_j^{(L)}} \cdot \frac{\partial u_j^{(L)}}{\partial s_j^{(L)}} = \frac{\partial C}{\partial u_j^{(L)}} \cdot \sigma_L'\big(s_j^{(L)}\big).$$

Здесь был использован тот факт, что $\frac{\partial u_k^{(L)}}{\partial s_j^{(L)}} = 0$ при $k \neq j$. ■

Утверждение 7.4 (правило BP_2)**.** *Для каждого внутреннего слоя l $(l = 1, \ldots, L-1)$ выполняются равенства*

$$\delta_j^{(l)} = \sum_{k=1}^{m_{l+1}} w_{kj}^{(l+1)} \cdot \delta_k^{(l+1)} \cdot \sigma_l'\big(s_j^{(l)}\big) \qquad (j = 1, \ldots, m_l).$$

◀ Используя правило дифференцирования сложной функции, запишем

$$\delta_j^{(l)} = \frac{\partial f}{\partial s_j^{(l)}} = \sum_{k=1}^{m_{l+1}} \frac{\partial f}{\partial s_k^{(l+1)}} \cdot \frac{\partial s_k^{(l+1)}}{\partial s_j^{(l)}} = \sum_{k=1}^{m_{l+1}} \delta_k^{(l+1)} \cdot \frac{\partial s_k^{(l+1)}}{\partial s_j^{(l)}}.$$

По определению

$$s_k^{(l+1)} = \sum_{j=1}^{m_l} w_{kj}^{(l+1)} \cdot u_j^{(l)} + b_k^{(l+1)} = \sum_{j=1}^{m_l} w_{kj}^{(l+1)} \cdot \sigma_l\big(s_j^{(l)}\big) + b_k^{(l+1)},$$

а значит

$$\frac{\partial s_k^{(l+1)}}{\partial s_j^{(l)}} = w_{kj}^{(l+1)} \cdot \sigma_l'\big(s_j^{(l)}\big).$$

■

Перейдём к вычислению координат градиента ∇f.

Утверждение 7.5 (правило BP_3)**.** *Для каждого слоя l $(l = 1, \ldots, L)$ выполняются равенства*

$$\frac{\partial f}{\partial b_j^{(l)}} = \delta_j^{(l)} \qquad (j = 1, \ldots, m_l).$$

◄ Используя правило дифференцирования сложной функции, запишем

$$\frac{\partial f}{\partial b_j^{(l)}} = \sum_{k=1}^{m_l} \frac{\partial f}{\partial s_k^{(l)}} \cdot \frac{\partial s_k^{(l)}}{\partial b_j^{(l)}} = \sum_{k=1}^{m_l} \delta_k^{(l)} \cdot \frac{\partial s_k^{(l)}}{\partial b_j^{(l)}} = \delta_j^{(l)} \cdot \frac{\partial s_j^{(l)}}{\partial b_j^{(l)}} = \delta_j^{(l)}.$$

Здесь был использован тот факт, что $\frac{\partial s_k^{(l)}}{\partial b_j^{(l)}} = 0$ при $k \neq j$. ■

Утверждение 7.6 (правило BP_4)**.** *Для каждого слоя l $(l = 1, \ldots, L)$ выполняются равенства*

$$\frac{\partial f}{\partial w_{jk}^{(l)}} = u_k^{(l-1)} \cdot \delta_j^{(l)} \qquad (j = 1, \ldots, m_l; k = 1, \ldots, m_{l-1}).$$

◄ Используя правило дифференцирования сложной функции, запишем

$$\frac{\partial f}{\partial w_{jk}^{(l)}} = \sum_{p=1}^{m_l} \frac{\partial f}{\partial s_p^{(l)}} \cdot \frac{\partial s_p^{(l)}}{\partial w_{jk}^{(l)}} = \sum_{p=1}^{m_l} \delta_p^{(l)} \cdot \frac{\partial s_p^{(l)}}{\partial w_{jk}^{(l)}} = \delta_j^{(l)} \cdot \frac{\partial s_j^{(l)}}{\partial w_{jk}^{(l)}} = \delta_j^{(l)} \cdot u_k^{(l-1)}.$$

Здесь был использован тот факт, что $\frac{\partial s_p^{(l)}}{\partial w_{jk}^{(l)}} = 0$ при $p \neq j$. ■

Объединяя вместе сказанное выше, сформулируем итоговый алгоритм вычисления градиента ∇f.

Вход:

$\mathbf{u} \in \mathbb{R}^{m_0}, \quad \boldsymbol{\theta} \in \mathbb{R}^p.$

Прямое распространение:

В цикле по $l = 1, \ldots, L$, используя формулы (7.13), вычислим вектора

$$\mathbf{u}^{(l)} = \big(u_1^{(l)}, \ldots, u_{m_l}^{(l)}\big)^t, \quad \mathbf{s}^{(l)} = \big(s_1^{(l)}, \ldots, s_{m_l}^{(l)}\big)^t.$$

Вычисление выходной ошибки:

Используя правило BP_1, вычислим величины

$$\delta_1^{(L)}, \ldots, \delta_{m_L}^{(L)}.$$

Обратное распространение ошибки:

В цикле по $l = L - 1, \ldots, 1$, используя правило BP_2, вычислим величины

$$\delta_1^{(l)} \ldots, \delta_{m_l}^{(l)}.$$

Выход:

Используя правила BP_3 и BP_4, вычислим координаты градиента

$$\nabla f = \left\{ \ldots, \frac{\partial f}{\partial w_{jk}^{(l)}}, \ldots, \frac{\partial f}{\partial b_j^{(l)}}, \ldots \right\}.$$

Сходимость градиентного метода

Перейдём к рассмотрению вопросов, связанных со сходимостью градиентного метода. Интересно заметить, что в окрестности локального минимума целевой функции градиентный метод сходится со скоростью геометрической прогрессии. При этом, что удивительно, можно использовать постоянную длину шага.

Теорема 7.2 (см. [27]). *Пусть* $r : \mathbb{R}^p \longrightarrow \mathbb{R}$ $(p \in \mathbb{N})$ *– дважды дифференцируемая функция с липшицевым гессианом. Это означает, что для некоторой константы* $M > 0$ *выполняется неравенство*

$$\left\|\nabla^2 r(\boldsymbol{\theta}) - \nabla^2 r(\boldsymbol{\theta}')\right\|_2 \leqslant M\left\|\boldsymbol{\theta} - \boldsymbol{\theta}'\right\|_2 \qquad \big(\textit{для всех } \boldsymbol{\theta}, \boldsymbol{\theta}' \in \mathbb{R}^p\big).$$

Пусть $\boldsymbol{\theta}^*$ *– точка локального минимума функции* r*. Предположим, что существуют константы* $c, C > 0$ *такие, что*

$$c\mathbf{I} \prec \nabla^2 r(\boldsymbol{\theta}) \prec C\mathbf{I},$$

где $\mathbf{I}$ *– единичная матрица размерности, а запись вида* $\mathbf{A} \prec \mathbf{B}$ *означает, что матрица* $\mathbf{B} - \mathbf{A}$ *положительно определена.*

Рассмотрим последовательность приближений (7.10), построенных градиентным методом (7.11) с постоянной длиной шага

$$h_n = \frac{2}{c + C} \qquad (n \in \mathbb{N}).$$

Предположим, что для начального приближения $\boldsymbol{\theta}^{(0)}$ *выполняется неравенство*

$$d_0 \coloneqq \left\|\boldsymbol{\theta}^{(0)} - \boldsymbol{\theta}^*\right\|_2 < \bar{d} \coloneqq \frac{2c}{M}.$$

Тогда

$$\left\|\boldsymbol{\theta}^{(n)} - \boldsymbol{\theta}^*\right\|_2 \leqslant \frac{\bar{d} d_0}{\bar{d} - d_0}\left(1 - \frac{2c}{C + 3c}\right)^n \qquad (n \in \mathbb{N}).$$

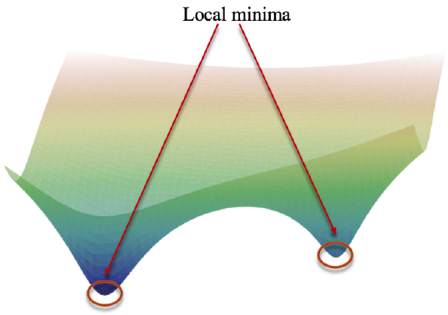
Наибольший интерес представляет нахождение глобального минимума. В этом случае определяющую роль начинают играть свойства целевой функции. Например, если функция выпуклая, то её локальный минимум является также и глобальным.

Однако в общем случае ожидать, что функция эмпирического риска, используемая в обучении нейронной сети, будет выпуклой, не приходится. Она может иметь несколько локальных минимумов. Это значит, что выбор «неправильного» начального приближения может привести к нахождению локального, а не голобального минимума.

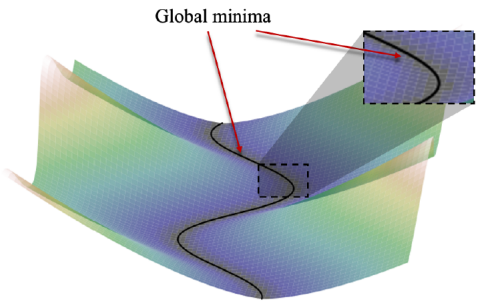
В то же время было давно замечено, что глубокие нейронные сети с большим числом параметров хорошо обучаются. Некоторый свет на эту ситуацию проливает работа [53]. В ней показывается (см. рис. 7.3), что с увеличением числа параметров модели нейронной сети все локальные минимумы целевой функции углубляются до одного уровня и образуют связную область. Таким образом, вне зависимости от выбора начального приближения градиентный метод сходится к одному из таких глобальных минимумов.

7.5 Обобщающие свойства нейронных сетей

Напомним, что в случае задачи бинарной классификации ключевую роль в исследовании равномерной сходимости эмпирического риска играет понятие размерности Вапника-Червоненкиса рассматриваемого класса гипотез. Конечность этой характеристики говорит



(a) Loss landscape of under-parameterized models



(b) Loss landscape of over-parameterized models

Рис. 7.3: Рисунок из статьи [53], на котором изображены графики двух целевых функций. Рисунок слева соответствует ситуации, когда количество параметров модели нейронной сети не превосходит размера обучающей выборки. Рисунок справа соответствует сверх параметризованной модели нейронной сети, у которой число параметров многократно превышает размер обучающей выборки.

о наличии свойства равномерной сходимости эмпирического риска, а её величина влияет на скорость сходимости.

Таким образом, размерность Вапника-Червоненкиса моделей неронных сетей напрямую характеризует их обобщающие свойства. Последовательно рассмотрим существующие оценки этой характеристики для моделей неронных сетей, использующих кусочно-постоянную, кусочно-полиномиальную и сигмоидную функцию активации в своих внутренних узлах.

Кусочно-постоянная функция активации

Определение 7.11. Пусть $p \in \mathbb{N}$ и σ – функция активации. Через $\mathcal{C}_{\sigma;p}$ обозначим семейство всех моделей нейронных сетей обладающих следующим свойством. Каждая из таких моделей содержит нейронные сети, у которых

- выходной слой содержит ровно один σ_{01}-узел;
- внутренние слои состоят только из σ-узлов;
- размер параметров равен p.

Для оценки размерности Вапника-Червоненкиса моделей нейронных сетей в зависимости от размера их параметров введём функцию

$$\mathrm{vc}_1(\sigma; p) \coloneqq \max_{\mathcal{F} \in \mathcal{C}_{\sigma;p}} \mathrm{vc}(\mathcal{F}) \qquad (p \in \mathbb{N}).$$

Теорема 7.3 (Кавер [54], Баум и Хаусслер [55]). *Справедлива верхняя оценка*

$$\mathrm{vc}_1(\sigma_{01}; p) = \mathcal{O}(p \ln p). \tag{7.14}$$

На самом деле оценка (7.14) является асимптотически оптимальной. В работе [56] на основе теории сложности булевых схем строится последовательность моделей нейронных сетей, размерность Вапника-Червоненкиса которых растёт пропорционально $p \ln p$.

Теорема 7.4 (Маасс [56]). *Справедлива оценка*

$$\mathrm{vc}_1(\sigma_{01}; p) = \Theta(p \ln p). \tag{7.15}$$

Интересно отметить, что оценка (7.15) никак не зависит от глубины нейронной сети. Кроме того, она останется справедливой для любой кусочно-постоянной функции активации, которая отлична от постоянной функции и имеет конечное число промежутков, на которых она принимает постоянные значения.

Кусочно-полиномиальная функция активации

Определение 7.12. Функция $\sigma : \mathbb{R} \longrightarrow \mathbb{R}$ называется *кусочно-полиномиальной*, если она может быть представлена в виде

$$\sigma(u) = \sum_{i=1}^{s} g_i(u)\mathbf{1}_{J_i}(u) \qquad (s \in \mathbb{N}; u \in \mathbb{R}),$$

где $J_1, \ldots, J_s$ – попарно непересекающиеся промежутки, образующие разбиение вещественной прямой $\mathbb{R}$, а $g_1, \ldots, g_s$ – полиномы с вещественными коэффициентами.

Под *степенью* функции σ будем понимать максимум степеней полиномов g_i $(i = 1, \ldots, s)$.

Кусочно-линейной называется функция степени 1, а *кусочно-постоянной* называется функция степени 0.

Например, функция активации σ_{relu} является кусочно-линейной, а функция σ_{01} является кусочно-постоянной.

Существующие верхние оценки размерности Вапника-Червоненкиса моделей нейронных сетей, использующих кусочно-полиномиальные функции активации, во многом базируются на следующем результате.

Теорема 7.5 (Голдберг и Джеррам [57])**.** *Пусть $m, p, t \in \mathbb{N}$ и класс функций $\mathcal{H} \subseteq \{0,1\}^{\mathbb{R}^m}$ может быть параметризован*

$$\mathcal{H} = \left\{\mathbf{u} \longmapsto h(\mathbf{u}; \boldsymbol{\theta}) \,:\, \boldsymbol{\theta} \in \mathbb{R}^p\right\}$$

с помощью некоторой функции $h : \mathbb{R}^m \times \mathbb{R}^p \longrightarrow \{0,1\}$.

Предположим, что существует алгоритм, который для любой входной пары $(\mathbf{u}, \boldsymbol{\theta}) \in \mathbb{R}^m \times \mathbb{R}^p$ вычисляет выходное значение $h(\mathbf{u}; \boldsymbol{\theta})$, выполняя при этом не более t операций следующих типов:

- *вещественные арифметические операции $+$, $-$, $\times$ и $/$;*
- *условные переходы относительно вещественные сравнений $>$, $\geqslant$, $<$, $\leqslant$, $=$ и $\neq$.*

Тогда

$$\mathrm{vc}(\mathcal{H}) \leqslant 4p(t+2).$$

Функция, задаваемая нейронной сетью с кусочно-постоянной функцией активации, может быть вычислена за $\mathcal{O}(p)$ операций. Из теоремы 7.5 сразу следует верхняя оценка $\mathcal{O}(p^2)$ для размерности Вапника-Червоненкиса. Она отличается в худшую сторону от оптимальной оценки (7.15) из теоремы 7.4.

В то же время приведённые рассуждения остаются справедливыми и для нейронных сетей с кусочно-полиномиальной функцией активации во внутренних узлах.

Теорема 7.6 (Голдберг и Джеррам [57])**.** *Пусть σ – кусочно-полиномиальная функция активации. Тогда*

$$\mathrm{vc}_1(\sigma; p) = \mathcal{O}(p^2). \tag{7.16}$$

Более аккуратное использование теоремы 7.4 позволяет существенно улучшить оценку (7.16). Однако для этого придётся учитывать в качестве дополнительного параметра количество слоёв нейронной сети.

Определение 7.13. Пусть $p, L \in \mathbb{N}$ и σ – функция активации. Через $\mathcal{C}_{\sigma;p,L}$ обозначим семейство всех моделей нейронных сетей обладающих следующим свойством. Каждая из таких моделей содержит нейронные сети, у которых

- L слоёв;
- выходной слой содержит ровно один σ_{01}-узел;
- внутренние слои состоят только из σ-узлов;
- размер параметров равен p.

Для оценки размерности Вапника-Червоненкиса моделей нейронных сетей в зависимости от размера их параметров и числа слоёв введём функцию

$$\mathrm{vc}_2(\sigma; p, L) \coloneqq \max_{\mathcal{F} \in \mathcal{C}_{\sigma;p,L}} \mathrm{vc}(\mathcal{F}) \qquad (p, L \in \mathbb{N}).$$

Теорема 7.7 (Бартлетт и др. [58])**.** *Пусть σ – кусочно-полиномиальная функция активации. Тогда*

$$\mathrm{vc}_2(\sigma; p, L) = \mathcal{O}\big(pL \ln p + pL^2\big).$$

В работе [59] доказываются почти оптимальные оценки для случая кусочно-линейной функции активации. Приведём их.

Теорема 7.8 (Бартлетт и др. [59])**.** *Пусть σ – кусочно-линейная функция активации. Тогда существуют константы $c, C > 0$ такие, что*

$$c \cdot pL \ln(p/L) \leqslant \mathrm{vc}_2(\sigma; p, L) \leqslant C \cdot pL \ln p \qquad (p, L \in \mathbb{N}).$$

Определение 7.14. Пусть $p, k \in \mathbb{N}$ и σ – функция активации. Через $\mathcal{C}'_{\sigma;p,k}$ обозначим семейство всех моделей нейронных сетей обладающих следующим свойством. Каждая из таких моделей содержит нейронные сети, у которых

- общее число внутренних узлов равно k;
- выходной слой содержит ровно один σ_{01}-узел;
- внутренние слои состоят только из σ-узлов;
- размер параметров равен p.

Для оценки размерности Вапника-Червоненкиса моделей нейронных сетей в зависимости от размера их параметров и числа внутренних узлов введём функцию

$$\mathrm{vc}_3(\sigma; p, k) \coloneqq \max_{\mathcal{F} \in \mathcal{C}'_{\sigma;p,k}} \mathrm{vc}(\mathcal{F}) \qquad (p, k \in \mathbb{N}).$$

Теорема 7.9 (Бартлетт и др. [59])**.** *Пусть σ – кусочно-линейная функция активации, $d \in \mathbb{N}$ – степень σ и $s \in \mathbb{N}$ – количество промежутков разбиения вещественной прямой $\mathbb{R}$ из определения σ. Тогда*

$$\mathrm{vc}_3(\sigma; p, k) = \mathcal{O}(pk \ln(d+1)s).$$

Сигмоидная функция активации

Рассмотрение сигмоидных функций активации логично начать с сигмоиды σ_{sig}. Однако непосредственное применение методов, разработанных для случая кусочно-полиномиальных функций активации, будет затруднительно. Это в первую очередь касается ключевой теоремы 7.5. В работе [60] доказывается аналог этой теоремы, применимый для случая сигмоиды.

Теорема 7.10 (Карпински и Макинтайр [60]). *Пусть $m, p, t \in \mathbb{N}$ и класс функций $\mathcal{H} \subseteq \{0,1\}^{\mathbb{R}^m}$ может быть параметризован*

$$\mathcal{H} = \left\{ \mathbf{u} \longmapsto h(\mathbf{u}; \boldsymbol{\theta}) \,:\, \boldsymbol{\theta} \in \mathbb{R}^p \right\}$$

с помощью некоторой функции $h : \mathbb{R}^m \times \mathbb{R}^p \longrightarrow \{0,1\}$.

Предположим, что существует алгоритм, который для любой входной пары $(\mathbf{u}, \boldsymbol{\theta}) \in \mathbb{R}^m \times \mathbb{R}^p$ вычисляет выходное значение $h(\mathbf{u}; \boldsymbol{\theta})$, выполняя при этом не более t операций следующих типов:

- *вещественные арифметические операции $+$, $-$, $\times$ и $/$;*
- *условные переходы относительно вещественные сравнений $>$, $\geqslant$, $<$, $\leqslant$, $=$ и $\neq$;*
- *взятие вещественной экспоненты $\alpha \mapsto e^{\alpha}$.*

Тогда

$$\mathrm{vc}(\mathcal{H}) = \mathcal{O}(t^2 p^2).$$

Непосредственным из теоремы 7.10 вытекает следующая оценка.

Теорема 7.11 (Карпински и Макинтайр [60]). *Справедлива верхняя оценка*

$$\mathrm{vc}_1(\sigma_{\mathrm{sig}}; p) = \mathcal{O}(p^4).$$

Обобщить теорему 7.11 на случай произвольных сигмоидных функций активации не представляется возможным. Существует модель 2-х слойных нейронных сетей с одним входом и двумя внутренними узлами с сигмоидной функции активации, у которой размерность Вапника-Червоненкиса бесконечна.

Теорема 7.12 (см. [61]). *Пусть*

$$\hat{\sigma}(u) \coloneqq \sigma_{\mathrm{sig}}(u) + cu^3 e^{-u^2} \sin(u) \qquad (u \in \mathbb{R})$$

для некоторого $c > 0$.

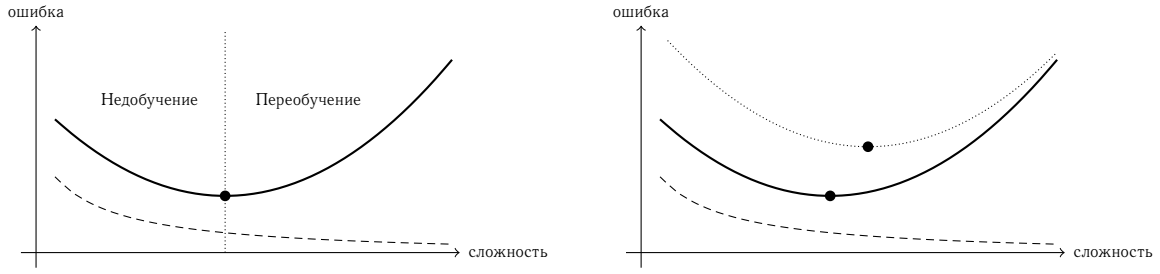
Определим класс функций

$$\mathcal{H} \coloneqq \left\{ u \longmapsto \sigma_{01}\big(w_1\, \hat{\sigma}(a_1 u) + w_2\, \hat{\sigma}(a_2 u) + b\big) \,:\, w_1, w_2, a_1, a_2, b \in \mathbb{R} \right\}.$$

Тогда

$$\mathrm{vc}(\mathcal{H}) = \infty.$$

Любопытно отметить, что при малых значениях параметра c график функции $\hat{\sigma}$ является незначительным искажением графика функции σ_{sig}. При этом, обобщающие свойства моделей нейронных сетей, построенных на основе этих функций активации, кардинальным образом отличаются.



а) б)

Рис. 7.4: Графики зависимости между ошибкой обучения и сложностью модели обучения.

7.6 Глубокие сети

Вернёмся к рассмотрению базовых понятий статистической теории обучения [1], на этот раз вооружившись математическим аппаратом, который был введён в предыдущих главах, и применительно к моделям искусственных нейронных сетей.

Для этого обратимся к графикам, изображённым на рис. 7.4 а). Сплошная линия соответствует ожидаемому риску, а пунктирная линия соответствует эмпирическому риску. Эти графики носят концептуальный характер и призваны отразить статистическую интуицию. При фиксированной обучающей выборке с ростом сложности модели эмпирический риск должен монотонно убывать, а ожидаемый риск должен сначала убывать до своей точки минимума, а затем начать расти. Промежуток до точки минимума соответствует режиму недообучения, а промежуток после точки минимума соответствует режиму переобучения.

Поэтому, с точки зрения классической статистики, очень маленький эмпирический риск, скорее всего, говорит о большом ожидаемом риске и неудовлетворительном решении задачи обучения. Таким образом, следует стремиться выбирать не очень простую, но и не очень сложную модель. В идеале сложность модели должна совпадать с точкой минимума графика функции ожидаемого риска. Точное нахождение такой точки минимума с практической точки зрения является очень сложной задачей.

Приближённое решение можно попытаться найти с помощью так называемого метода *структурной минимизации риска*. Для этого достаточно взять эмпирическую оценку, правая часть которой зависит от сложности модели. В качестве приближённого решения можно взять точку минимума графика правой части этой эмпирической оценки. На рис. 7.4 б) он изображён точечной линией.

Напомним, что в общем случае эмпирическая оценка имеет вид

$$R(h) \leqslant r(h, \mathbf{z}) + c(n, \mathcal{H}; \delta), \tag{7.17}$$

где $h \in \mathcal{H}$, n – размер обучающей выборки $\mathbf{z}$, а δ – параметр достоверности (неравенство выполняется с вероятностью $1 - \delta$). Предполагается, что $c(n, \mathcal{H}; \delta) \longrightarrow 0$ при $n \longrightarrow \infty$.

Ранее, в главе 5 фактически была получена эмпирическая оценка вида

$$R(h) \leqslant r(h, \mathbf{z}) + 62 \sqrt{\frac{\mathrm{vc}(\mathcal{H})}{n}} + \sqrt{\frac{1}{2n} \ln\left(\frac{1}{\delta}\right)}, \tag{7.18}$$

правая часть которой зависит от размерности Вапника-Червоненкиса модели.

В предыдущем разделе 7.5 размерность Вапника-Червоненкиса моделей нейронных сетей была выражена через их параметры (количество узлов и слоёв). Таким образом, неравенство (7.18) может быть преобразовано к эмпирической оценке вида (7.17), у которой слагаемое $c(n, \mathcal{H}; \delta)$ выражается через такие параметры. Возникает закономерный

вопрос о практической применимости подобных оценок как с точки зрения описанной выше задачи выбора модели, так и с точки зрения оценки неизвестного значения ожидаемого риска $R(h)$ через известный эмпирический риск $r(h, \mathbf{z})$.

Если используемая функция потерь принимает свои значения из отрезка $[0, 1]$, то ожидаемый и эмпирический риски также будут принимать свои значения из этого отрезка. Повторяя рассуждения, которые проводились в примере 5.6 из раздела 5.5, применительно к глубоким нейронным сетям с большим числом параметров можно сделать вывод. Если ориентироваться на размеры обучающих выборок, с которыми можно встретиться на практике, то слагаемое $c(n, \mathcal{H}; \delta)$ будет на порядки превосходить значения $R(h)$ и $r(h, \mathbf{z})$, а значит, о практической применимости такой эмпирической оценки не может быть и речи.

Сделанный вывод справедлив для эмпирической оценки, в основу которой изначально было положено неравенство, в котором фигурировала размерность Вапника-Червоненкиса, и которая потом оценивалась сверху через параметры модели нейронной сети. Возможно, подобная последовательность действий и является причиной низкой эффективности, и существуют другие эффективные эмпирические оценки. Однако, такое предположение, скорее всего, неверно.

В известной работе [62] описывается следующий эксперимент. Авторы этой статьи брали популярные модели глубоких нейронных сетей и популярные наборы данных. Наборы данных намеренно портились. В одном случае перемешивались метки, а объекты оставались неизменными. В другом случае, наоборот, метки оставались неизменными, а в объекты добавлялся случайный шум. Рассматривались также случаи, когда одновременно модифицировались и метки, и объекты. По исходным и модифицированным наборам данных формировались обучающие выборки и выделялись тестовые данные. По этим обучающим выборкам производилось обучение нейронных сетей. Тестовые данные, которые не участвовали в процессе обучения, использовались для приближённого вычисления ожидаемых рисков.

Во всех случаях эмпирический риск был нулевым. Все нейронные сети фактически запоминали обучающие примеры. С ожидаемым риском картина была другой. В случае нейронных сетей, обученных по исходным (неиспорченным) данным, ожидаемый риск был близок к нулю. Для других нейронных сетей ожидаемый риск сильно отличался от нуля.

Применяя этот результат к произвольной эмпирической оценке, получим неравенство вида

$$R(h_{\mathbf{z}}) \leqslant \underbrace{r(h_{\mathbf{z}}, \mathbf{z})}_{(=0)} + c(n, \mathcal{H}; \delta),$$

где правая часть не зависит от обучающей выборки, а левая часть в зависимости от обучающей выборки может быть как близкой к нулю, так и сильно отличаться от него. Можно ли сделать вывод о непригодности практического применения эмпирических оценок к глубоким нейронным сетям и противоречит ли этот результат классической теории статистического обучения?

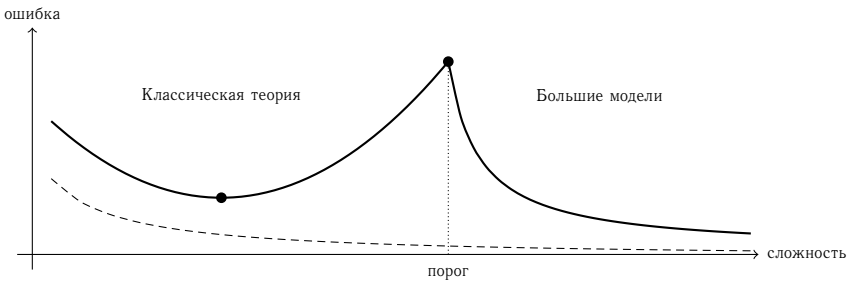


Рис. 7.5: Кривая двойного спуска.

Смысл глубокого обучения состоит в том, что необходимо использовать большие (сверх параметризованные) модели. Число параметров должно на порядки превосходить размер обучающий выборки. В то же время классическая теория начинает работать в противоположной ситуации, когда размер обучающей выборки «больше» сложности модели.

Такой вывод можно сделать из работы [63], в которой была обнаружено, что график ожидаемого риска на самом деле имеет форму кривой с двойным спуском рис. 7.5. Оказывется, что после того как сложность модели достигает определённого порогового значения, ожидаем риск перестаёт возрастать и начинает вновь монотонно убывать. Судя по всему, в этот момент перестают работать статистические закономерности классической теории. При этом обучение нейронной сети трансформируется в решение некоторой интерполяционной задачи, аппроксимирующей примеры из обучающей выборки.

Этот феномен получил название *безвредного* (*benign*) переобучения. В настоящее время он вляется объектом активного исследования научного сообщества.