Постановка задач и выбор моделей в машинном обучении

Вадим Викторович Стрижов

Московский физико-технический институт

Осенний семестр 2019

Однослойная нейронная сеть

Нейронная сеть и задача регрессии

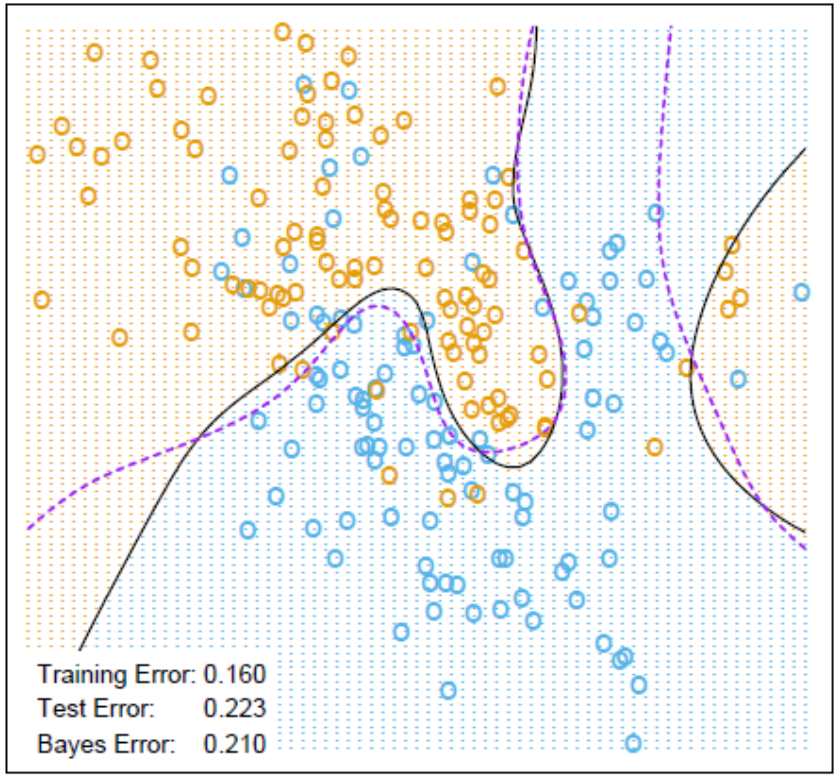
Нейронная есть — универсальная модель, решающая широкий класс прикладных задач.
Рассмотрим выборку $(\mathbf{x}_i, y_i)$, $\mathbf{x}_i \in \mathbb{R}^d$, $y_i \in \mathbb{R}^1$, $i = 1, \ldots, \ell$,
$\mathbf{x}$ — описание объекта, вектор из d элементов — признаков x_j,
y — зависимая переменная.
Требуется построить аппроксимирующую поверхность $a(\mathbf{x})$.

Нейронная сеть и задача классификации

Требуется построить разделяющую поверхность $a(\mathbf{x}) \mapsto y \in \{-1, +1\}$, значения этой функции

$$a(\mathbf{x}) > 0, \text{ если } y = +1,$$

$$a(\mathbf{x}) \leq 0, \text{ если } y = -1.$$



Нейронная сеть как универсальная модель

Теорема (А. Н. Колмогоров, 1957)

Каждая непрерывная функция $a(\mathbf{x})$, заданная на единичном кубе d-мерного пространства, представима в виде

$$a(\mathbf{x}) = \sum_{i=1}^{2d+1} \sigma_i \left(\sum_{j=1}^{d} f_{ij}(x_j) \right), \text{ где } \mathbf{x} = [x_1, \ldots, x_d]^\mathsf{T},$$

функции $\sigma_i(\cdot)$, $f_{ij}(\cdot)$ непрерывны, причем f_{ij} не зависят от выбора a.

Нейронная сеть как универсальная модель

Иначе, функцию от d аргументов можно представить в виде комбинации $d(2d + 1)$ функций одного аргумента.

Нейронная сеть как универсальная модель

Важно. Колмогоров не указал, какими именно должны быть функции σ_i, f_{ij}.

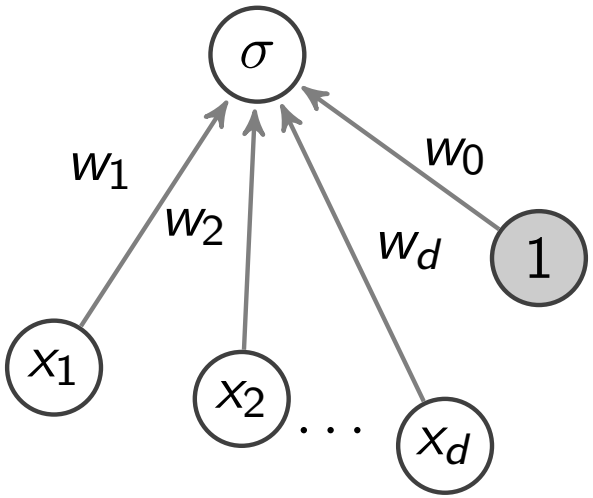
Однослойная нейронная сеть как нейрон

Однослойная нейронная сеть (или нейрон) — это комбинация

$$a(\mathbf{x}, \mathbf{w}) = \sigma(\mathbf{w}^{\mathsf{T}}\mathbf{x}) = \sigma\left(\sum_{j=1}^{d} w_j^{(1)} x_j + w_0^{(1)}\right),$$

σ — функция активации, непрерывная монотонная функция, желательно дифференцируемая,
$\mathbf{w}$ — вектор параметров (весов),
$\mathbf{x}$ — объект, вектор с присоединенным элементом 1 для веса w_0.

Однослойная нейронная сеть как нейрон

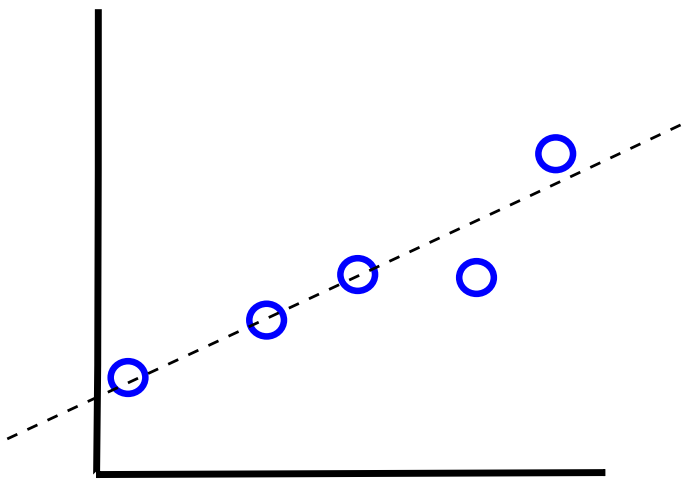


Однослойная нейронная сеть — линейная модель

Сеть с линейной функцией активации, задача восстановления регрессии $\mathbf{x}$ на y:

$$a(\mathbf{x}, \mathbf{w}) = \mathbf{w}^{\mathsf{T}}\mathbf{x}.$$

Функция активации $\sigma = \mathrm{id}$.

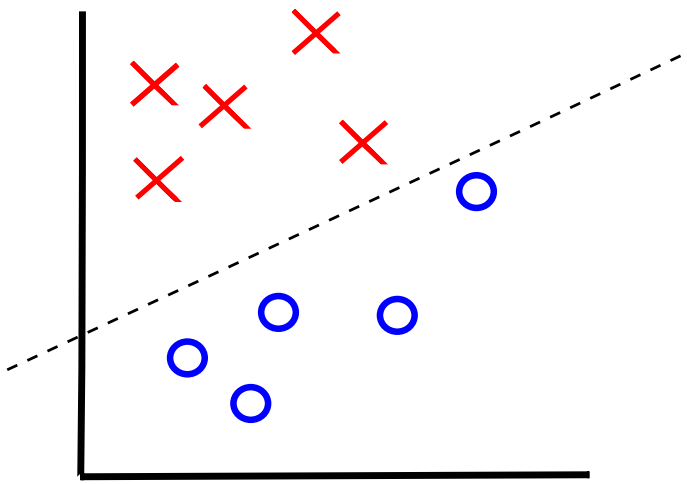


Однослойная нейронная сеть — линейная модель

Сеть с пороговой функцией активации, задача классификации:

$$a(\mathbf{x}, \mathbf{w}) = \text{sign}(\mathbf{w}^{\mathsf{T}}\mathbf{x}).$$

Функция активации $\sigma = \text{sign}(\cdot)$ — знак скалярного произведения.



Однослойная сеть — модель логистической регрессии

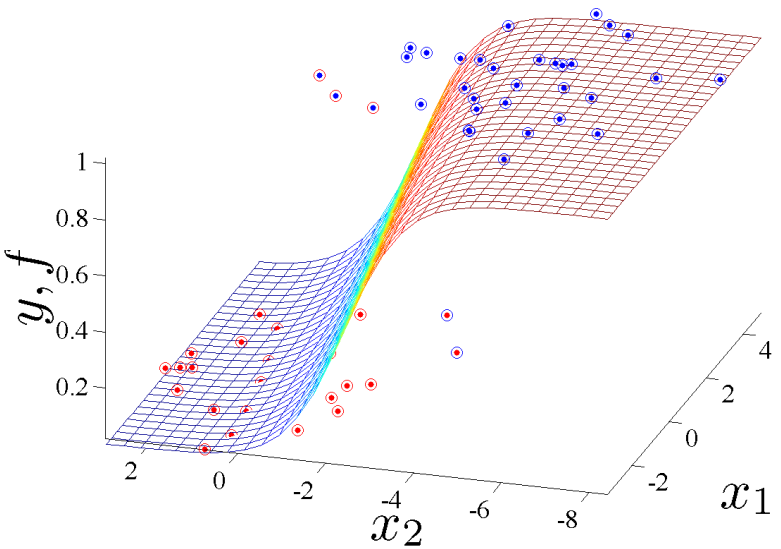
Сигмоидная функция активации

$$a(\mathbf{x}, \mathbf{w}) = \sigma(\mathbf{w}^\mathsf{T}\mathbf{x}) = \frac{1}{1 + \exp(-\mathbf{w}^\mathsf{T}\mathbf{x})}.$$

Функция определяет вероятность принадлежности заданного объекта $\mathbf{x}$ классу $y = 1$

$$\sigma(\mathbf{w}^\mathsf{T}\mathbf{x}) = P(y = 1|\mathbf{w}, \mathbf{x})$$

при заданных параметрах $\mathbf{w}$.



Многоклассовая классификация

Обобщение для многоклассового случая $\mathbf{y} = [y^1, \ldots, y^K]^\mathsf{T}$

$$\sigma = \text{softmax}(\mathbf{w}_1^\mathsf{T}\mathbf{x}, \ldots, \mathbf{w}_K^\mathsf{T}\mathbf{x}) = \frac{exp(\mathbf{w}_k^\mathsf{T}\mathbf{x})}{\sum_{i=1}^K \exp(\mathbf{w}_i^\mathsf{T}\mathbf{x})}.$$

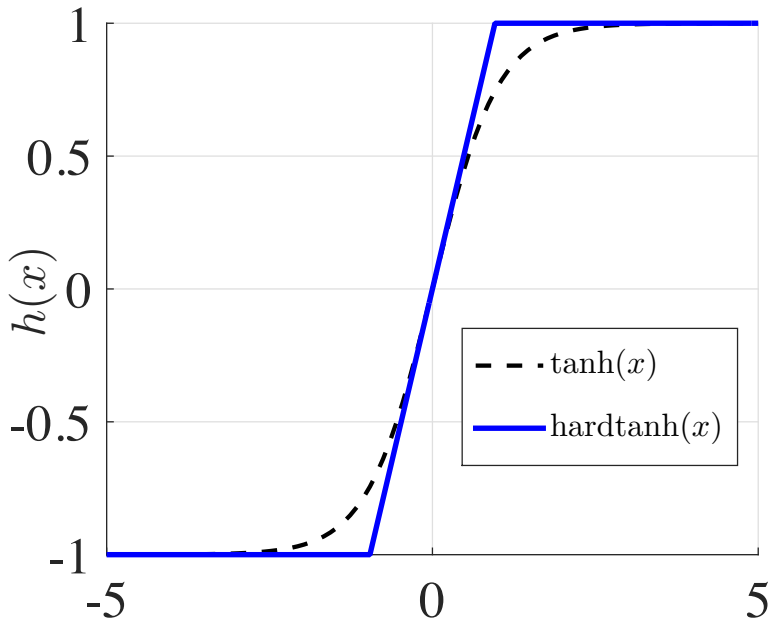
Сеть состоит из K нейронов, и вычисляет вероятности принадлежности объекта $\mathbf{x}$ к различным K классам одновременно.

Виды функций активации

Гиперболический тангенс

$$\tanh(x) = \frac{\exp(2x) - 1}{\exp(2x) + 1}, \text{ вариант } \operatorname{hardtanh}(x) = \begin{cases} -1, & \text{при } x < -1, \\ 0, & \text{при } -1 \le x \le 1, \\ 1, & \text{при } x > 1. \end{cases}$$

Эта функция активации — обобщение пороговой фунции sign, которая может быть продифференцирована.

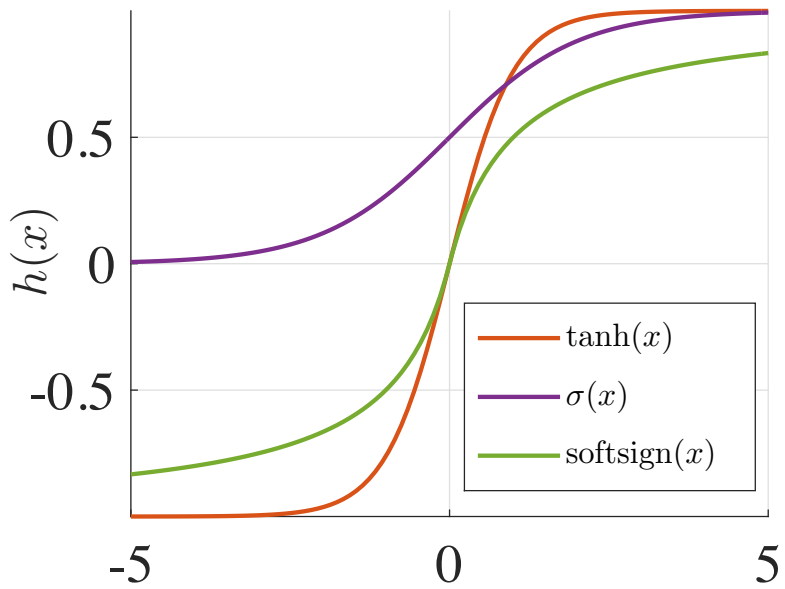


Виды функций активации

Функция активации $\sigma(x)$

$$\text{softsign}(x) = \frac{x}{1 + |x|}$$

сходится к $+1$ или -1 медленнее, чем $\tanh(x)$.



Виды функций активации

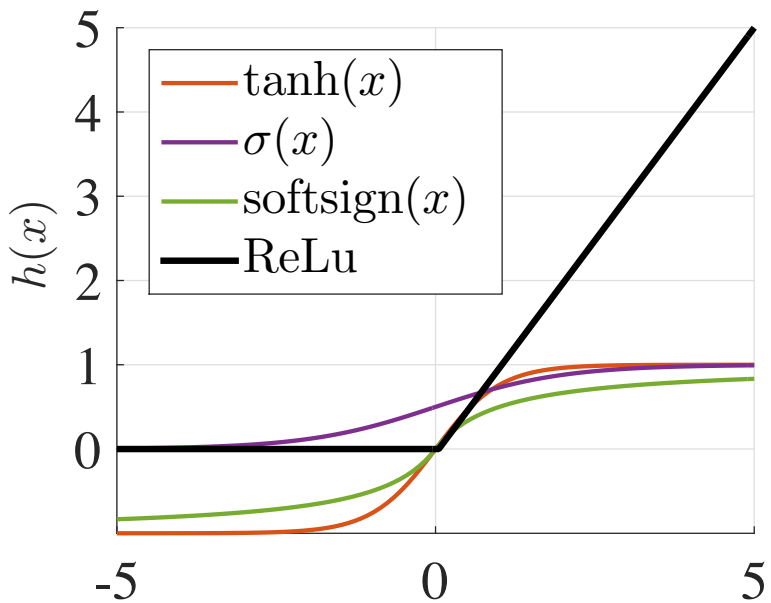
Линейный выпрямитель, Rectified linear unit (ReLu),

$$\mathrm{ReLu}(x) = \max(0, x)$$

и его приближение

$$\ln(1 + \exp(x))$$

используется в сетях глубокого обучения при обработке изображений.



Резюме

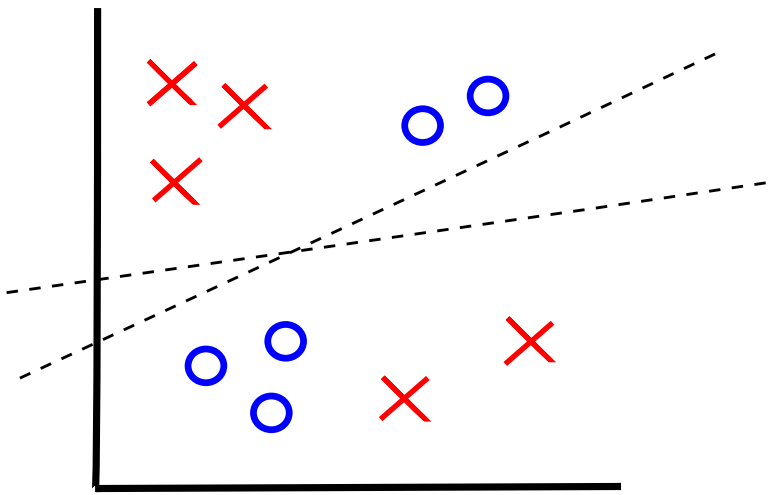
- Нейронная сеть предназначена для решения широкого класса прикладных задач регрессии и классификации (обучение с учителем).
- Она является универсальной моделью, так как может приближать функции любой сложности.
- Нейрон, или однослойная нейронная сеть — функция активации от линейной комбинации признаков объекта.
- При построении сети используются функции активации различных видов.

Далее: двуслойные и многослойные нейронные сети.

Многослойная нейронная сеть
Функция ошибки

Пределы применимости однослойных сетей

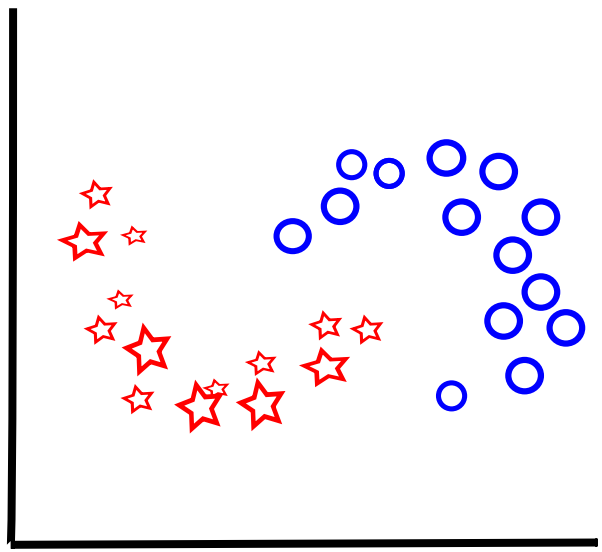
Однослойные сети применимы только для линейно разделимых выборок.



Не существует разделяющей плоскости, разделяющей данную выборку на два класса корректно (без ошибок).

Пределы применимости однослойных сетей

Для корректного разделения этой выборки подходит кривая линия.



Двухслойная нейронная сеть как комбинация однослойных

Эта сеть состоит из линейной комбинации нейронов (однослойных нейронных сетей).

$$a(\mathbf{x}, \mathbf{w}) = \sigma^{(2)} \left(\sum_{i=1}^{D} w_i^{(2)} \cdot \sigma^{(1)} \left(\sum_{j=1}^{d} w_{ij}^{(1)} x_j + w_{i0}^{(1)} \right) + w_0^{(2)} \right),$$

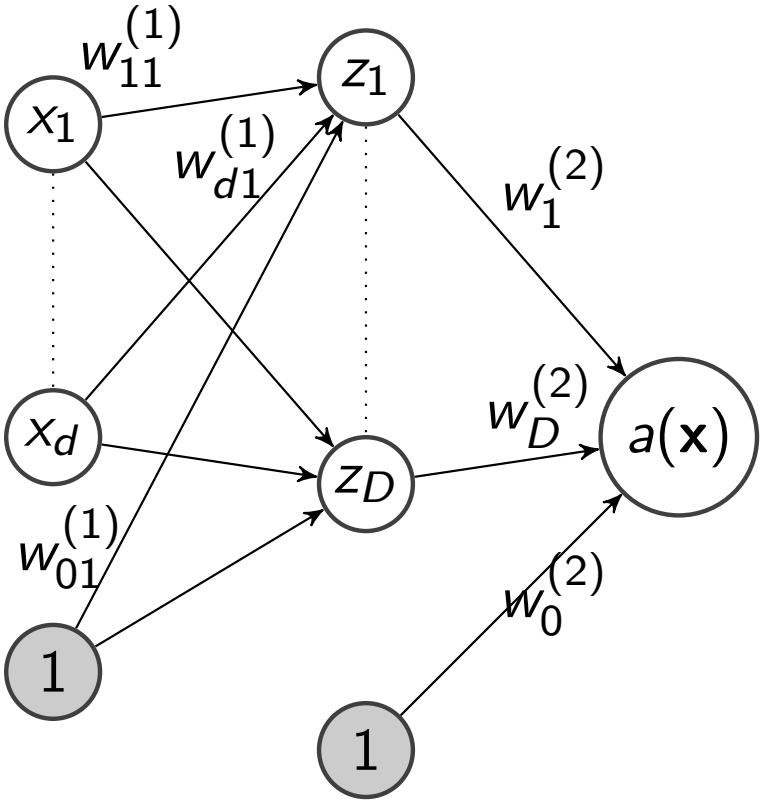
или в векторных обозначениях

$$a(\mathbf{x}, \mathbf{w}) = \sigma^{(2)} \left(\mathbf{w}^{\mathsf{T}(2)} \boldsymbol{\sigma}^{(1)} \left([\mathbf{w}_1^{\mathsf{T}(1)} \mathbf{x}, \ldots, \mathbf{w}_D^{\mathsf{T}(1)} \mathbf{x}] \right) \right).$$

Соединенный вектор параметров $\mathbf{w} = \{w_i^{(2)}, w_{ij}^{(1)}, w_{i0}^{(1)}, w_0^{(2)}\}$.

Двухслойная нейронная сеть

Двуслойная сесть представима в виде двудольных направленных графов, где исходящая вершина графа связана со всеми входящими вершинами.



Граф можно продолжать вправо и дальше для получения многослойной нейронной сети.

Разделяющая способность многослойной нейронной сети

Теорема (Универсальная теорема аппроксимации, K. Hornik, 1991)

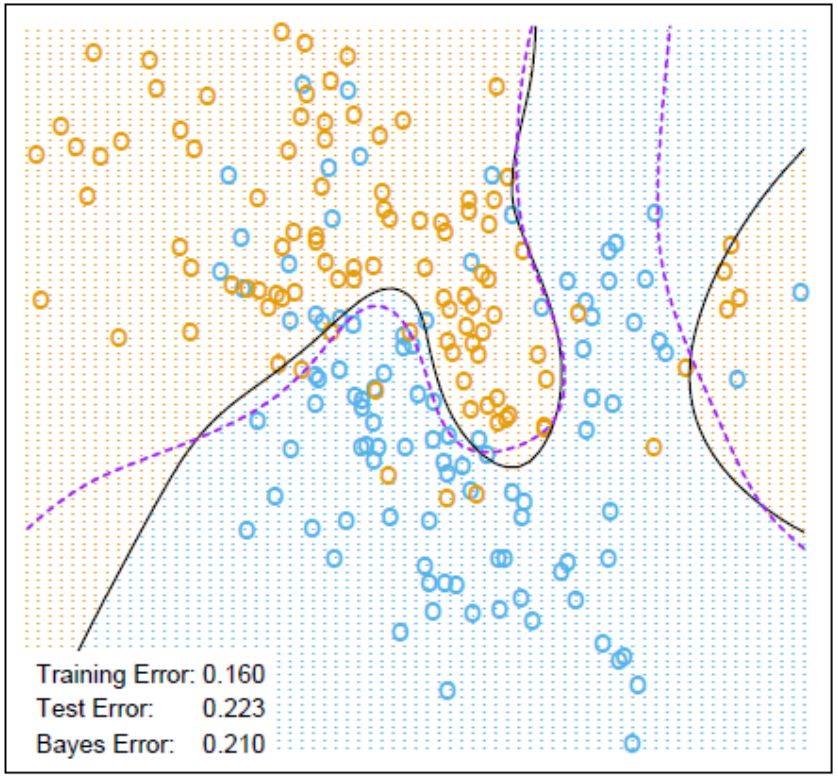
для любой непрерывной функции найдется нейронная сеть $a(\mathbf{x})$ *с линейным выходом, аппроксимирующая* $f(\mathbf{x})$ *с заданной точностью.*

Разделяющая способность многослойной нейронной сети

- Теорема выполняется для $\sigma(f) = \text{sigmoid}(f)$, $\sigma(f) = \tanh(f)$ и ряда других функций активации.
- Для получения этой заданной точности необходимо определить оптимальные параметры $\mathbf{w}^*$.

Разделяющая поверхность при оптимальных значениях параметров

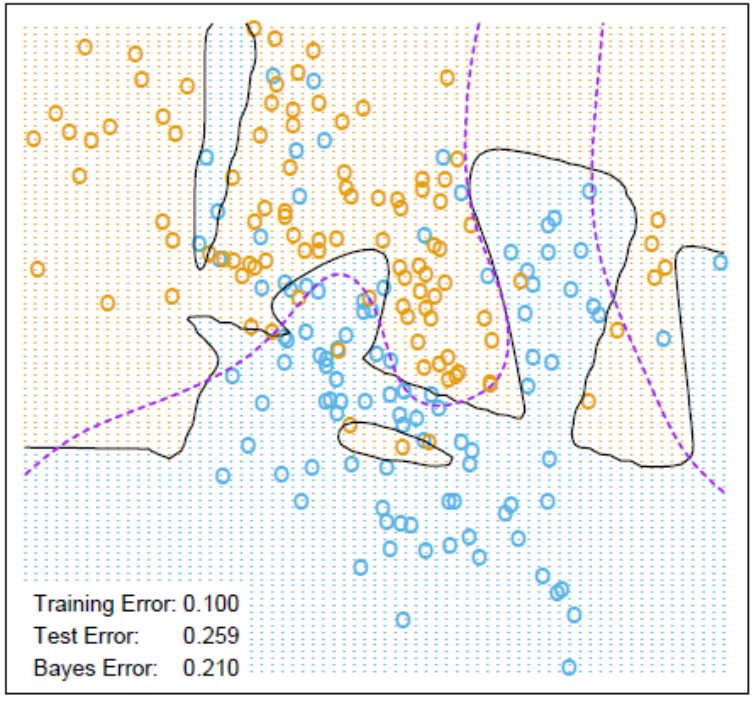
Качество аппроксимации y функцией $a(\mathbf{x}, \mathbf{w})$ зависит от оптимизации параметров $\mathbf{w}$.



Несколько объектов попали в область, принадлежащую другому классу вследствие случайной природы выборки.

Разделяющая поверхность при неоптимальных значениях параметров

Параметры **w** нейросети настроены таким образом, что разделяет текущую выборку корректно, однако при изменении состава выборки разделение будет неверным.



Переобучение нейронной сети

Сеть, корректно апроксимирующая обучающую выборку и плохо аппроксимирующая контрольную, называется переобученной.

Функции ошибки нейронной сети

Оптимизируем параметры так, чтобы минимизировать значение функции ошибки

$$\mathbf{w}^* = \arg\min_{\mathbf{w}} Q(\mathbf{w}).$$

Задача оптимизации

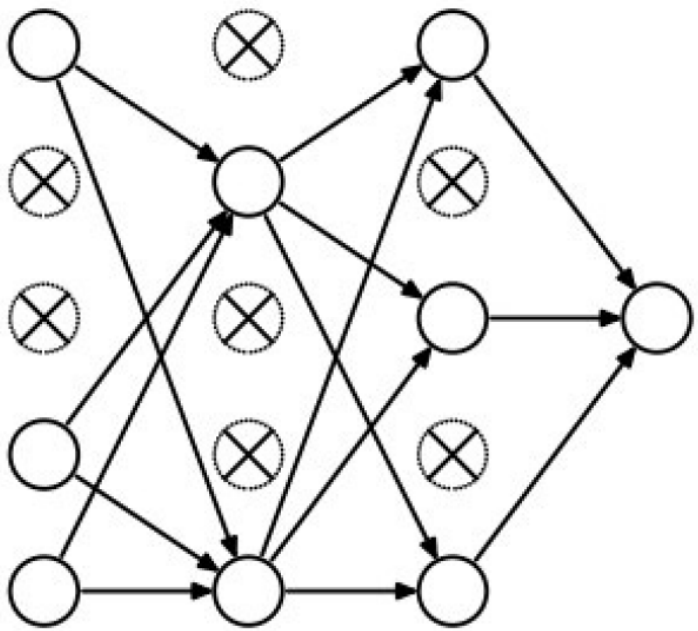
$$\mathbf{w}^* = \arg\min_{\mathbf{w}} Q(\mathbf{w})$$

Функция ошибки Q зависит от

- выборки $(\mathbf{x}_i, y_i)$, $i = 1, \ldots, \ell$,
- структуры нейросети (числа слоев, нейронов, видов функций активации),
- значения вектора параметров $\mathbf{w}$.

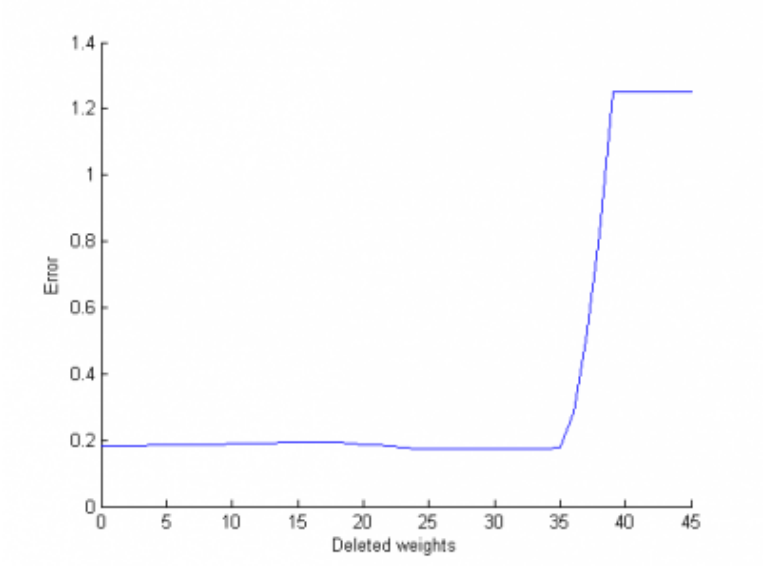
Прореживание сети

Для снижения числа параметров предлагается исключить некоторые нейроны или связи.



Прореживание сети

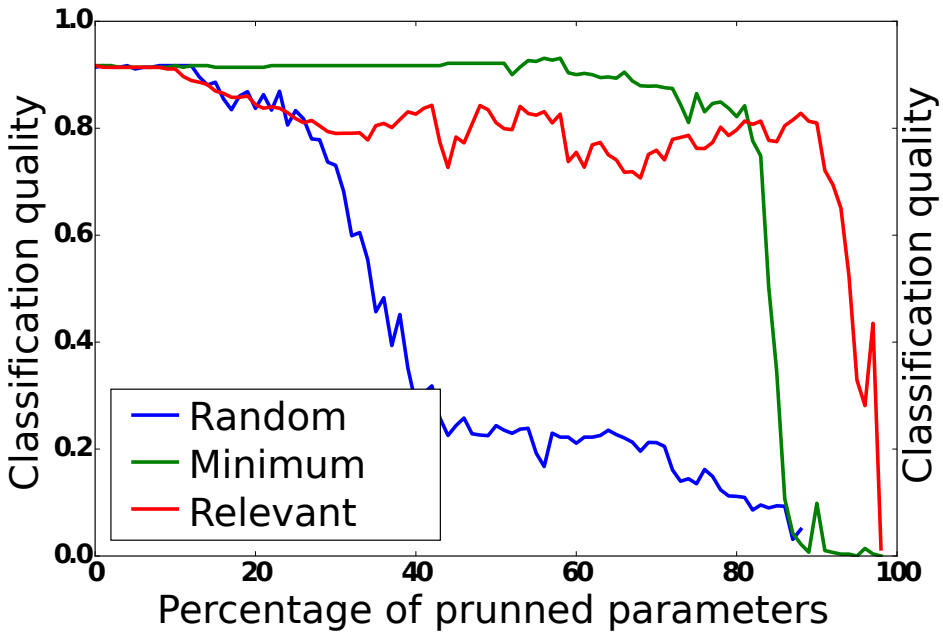
Если функция ошибки не изменяется, сеть можно упрощать и дальше.



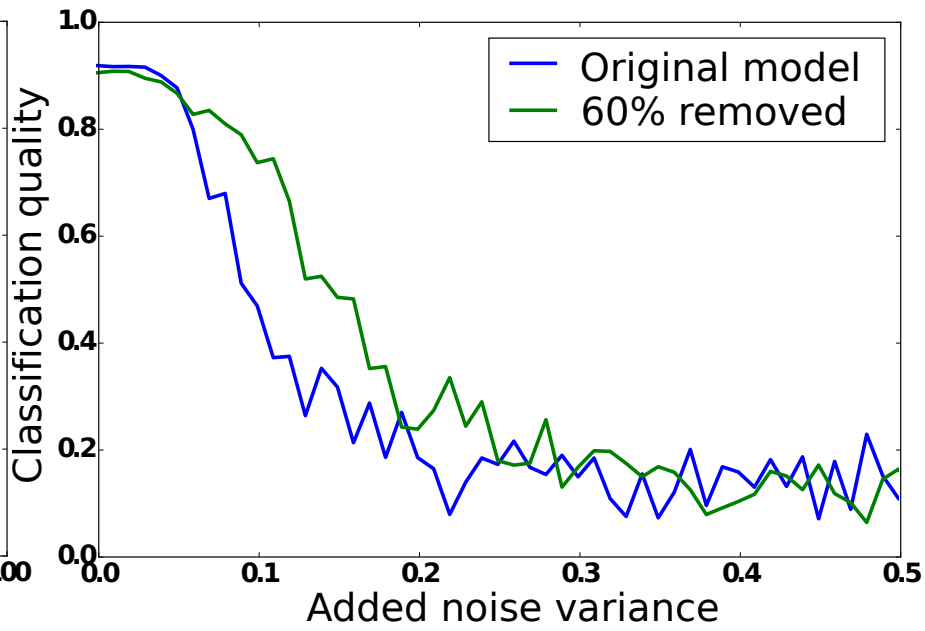
Зависимость функции ошибки $Q(\mathbf{w})$ от числа удаленных нейронов.

The problem of optimal model structure selection

The evidence of models with excessive number of parameters does not change significantly after the model structure pruning procedure.



Excessive of the model parameters



Model stability

The deep learning paradigm presumes optimization of models with excessive complexity.

Двоичное представление структуры модели

Модель f выбирается из множества моделей-претендентов $\mathfrak{F}$ путем оптимизации двоичного вектора $\mathbf{a} \in \mathbb{B}^n$,

$$\hat{y} = f(\mathbf{w}, \mathbf{x}) = a_1 w_1 x_1 + \cdots + a_n w_n x_n$$

для линейной модели $f(\mathbf{w}, \mathbf{x}) = \mathbf{x}^\mathsf{T}\mathbf{w}$
и для нейронной сети

$$\mathbf{f}(\mathbf{w}, \mathbf{x}) = \frac{\exp\big(\mathbf{h}(\mathbf{x})\big)}{\sum_k \exp(h_k(\mathbf{x}))}, \quad \mathbf{h}(\mathbf{x}) = \mathbf{W}_2^\mathsf{T}\mathbf{tanh}(\mathbf{W}_1^\mathsf{T}\mathbf{x}), \quad \mathbf{w} = \mathbf{vec}(\mathbf{W}_1 \vdots \mathbf{W}_2)$$

путем зануления соовтествующего параметра

$$w_j = 0$$

или согласно методу оптимального прореживания

$$\mathbf{e}_j^\mathsf{T}\Delta\mathbf{w} + w_j = 0$$

где j-й элемент вектора $\mathbf{e}$ равен 1, прочие равны 0.
Модель задана вершиной двоичного n-мерного куба.

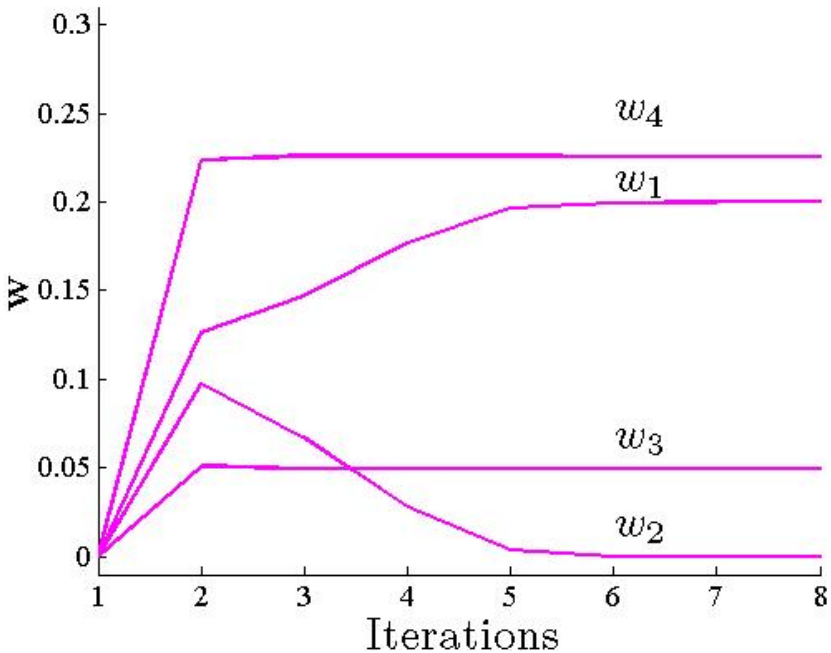
Стратегии прореживания параметров сети

Параметр можно удалить если

- он имеет значение, близкое к нулю,
- его значение сильно изменяется при изменении выборки (большая дисперсия),
- его удаление меньше всего влияет на изменение значения функции ошибки.

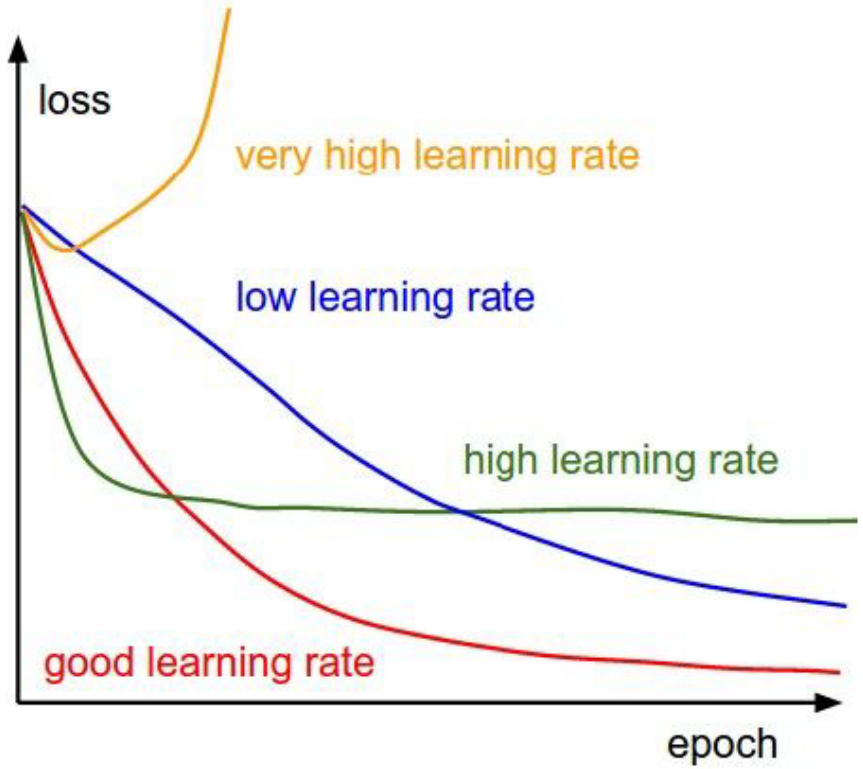
Стабилизация параметров сети

При нахождении минимума (он может оказаться локальным) параметры стабилизируются.



Скорость обучения сети

По скорости обучения можно судить о соответствии выборки и нейронной сети.



Обучение и контроль

Разница между значениями функции ошибки на обучении и контроле не должна быть существенной.