Модели машинного обучения. Суперпозиции

Плетнев Никита Вячеславович

Московский физико-технический институт

Курс: Математические методы прогнозирования
(В. В. Стрижов)/Группа 574, весна 2019

Литература

- Последовательный выбор моделей глубокого обучения оптимальной сложности, Бахтеев О. Ю., диссертация;
- Шпаргалка по всем сетям, их классификация и строгое описание, Жариков И. Н..

Постановка задачи выбора структуры модели

Проблема выбора структуры модели глубокого обучения формулируется следующим образом: решается задача классификации или регрессии на заданной или пополняемой выборке $\mathfrak{D}$. Требуется выбрать структуру нейронной сети, доставляющей минимум ошибки на этой выборке и максимум качества на некотором внешнем критерии.
Под моделью глубокого обучения понимается суперпозиция дифференцируемых по параметрам нелинейный функций. Под структурой модели понимается значения структурных параметров модели, т.е. величин, задающих вид итоговой суперпозиции.

Основные определения

Объект

пара $(\mathbf{x}, y), \mathbf{x} \in \mathrm{X} = \mathrm{R}^n, y \in \mathrm{Y}$.
В случае задачи классификации y является распределением вероятностей принадлежности объекта $\mathbf{x} \in \mathrm{X}$ множеству классов $\{1, \ldots, Z\}$: $\mathrm{Y} \subseteq [0, 1]^Z$, где Z — число классов.
В случае задачи регрессии Y является некоторым подмножеством вещественных чисел $\mathrm{Y} \subseteq \mathrm{R}$.
$\mathbf{x}$ — признаковое описание, y — метка объекта.

Модель $\mathbf{f}$

дифференцируемая по параметрам функция из множества признаковых описаний объекта во множество меток:

$$\mathbf{f} : \mathrm{X} \times \mathrm{W} \to \mathrm{Y},$$

где W — пространство параметров функции $\mathbf{f}$.

Семейство моделей

Пусть задан направленный граф (V, E). Пусть для каждого ребра $(j, k) \in E$ определен вектор базовых функций мощности $K^{j,k}$: $\mathbf{g}^{j,k} = [\mathbf{g}_0^{j,k}, \ldots, \mathbf{g}_{K^{j,k}}^{j,k}]$. Пусть также для каждой вершины v определена функция аггрегации $\mathbf{agg}_v$. Граф (V, E) в совокупности со множестом векторов базовых функций $\{\mathbf{g}^{j,k}, (j, k) \in E\}$ и множеством функций аггрегаций $\{\mathbf{agg}_v, v \in V\}$ называется *семейством моделей* $\mathfrak{F}$, если функция, задаваемая рекурсивно как

$$\mathbf{f}_k(\mathbf{x}) = \mathbf{agg}_{v_k}\left(\left\{\langle \gamma^{j,k}, \mathbf{g}^{j,k}\rangle(\mathbf{f}_j(\mathbf{x})) \mid j \in \mathrm{Adj}(v_k)\right\}\right), \quad \mathbf{f}_0(\mathbf{x}) = \mathbf{x},$$

является моделью при любых значениях векторов $\gamma^{j,k} \in [0, 1]^{K^{j,k}}$.

Примерами функций агрегации являются функции суммы и конкатенации векторов.

Основные определения

Слои, или подмодели

Функции $\mathbf{f}_1, \ldots \mathbf{f}_{|V|}$ из определения семейства моделей называются слоями модели $\mathbf{f}$.

Параметры модели $\mathbf{f}$ из семейства моделей $\mathfrak{F}$

Конкатенация векторов параметров всех базовых функций $\{\mathbf{g}^{j,k} \mid (j,k) \in E\}, \mathbf{w} \in \mathrm{W}$. Вектор параметров базовой функции $\mathbf{g}_l^{j,k}$ будем обозначать как $\mathbf{w}_l^{j,k}$.

Структура $\mathbf{\Gamma}$ модели $\mathbf{f}$ из семейства моделей $\mathfrak{F}$

Конкатенация векторов $\gamma^{j,k}$.

Множество всех возможных значений структуры $\mathbf{\Gamma}$ будем обозначать Γ. Векторы $\{\gamma^{j,k} \mid (j,k) \in E\}$ назовем *структурными параметрами модели*.

Параметризация множества моделей M

Семейство моделей $\mathfrak{F}$, такое что для каждой модели $\mathbf{f} \in M$ существует значение структуры модели $\mathbf{\Gamma}$, при котором функция $\mathbf{f}$ совпадает с функцией из определения семейства моделей.

Варианты множества структур Γ

1. Вершины булева куба: $\gamma^{j,k} \in \{0,1\}^{K^{j,k}}$;
2. Внутренность булева куба: $\gamma^{j,k} \in [0,1]^{K^{j,k}}$;
3. Вершины симплекса: $\gamma^{j,k} \in \overline{\Delta}^{K^{j,k}-1}$;
4. Внутренность симплекса: $\gamma^{j,k} \in \Delta^{K^{j,k}-1}$.

Варианты множества структур Γ

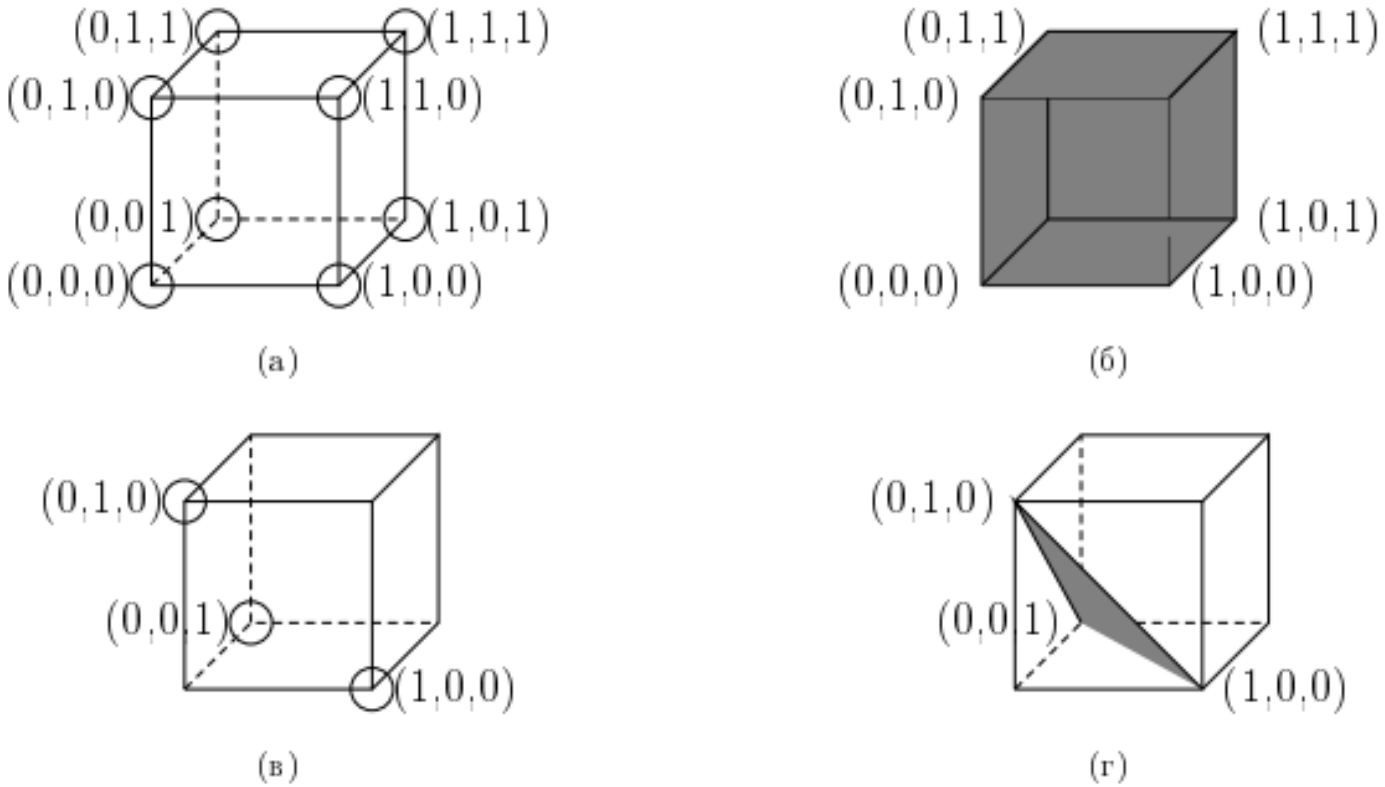
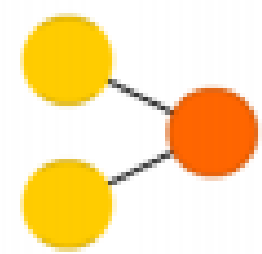


Рис. 1.2. Примеры ограничений для одного структурного параметра $\boldsymbol{\gamma}, |\boldsymbol{\gamma}| = 3$. а) структурный параметр лежит на вершинах куба, б) структурный параметр лежит внутри куба, в) структурный параметр лежит на вершинах симплекса, г) структурный параметр лежит внутри симплекса.

Графовое представление структуры модели

Одним из возможных представлений структуры моделей глубокого обучения является графовое представление, в котором в качестве ребер графа выступают нелинейные функции, а в качестве вершин графа — представление выборки под действием соответствующих нелинейных функций. Такой подход реализован в библиотеках TensorFlow, PyTorch.
В то же время, существуют и другие способы представления модели. В ряде работ, посвященных байесовской оптимизации, модель рассматривается как черный ящик, над которым производится ограниченный набор операций типа «произвести оптимизацию параметров» и «предсказать значение зависимой переменной по независимой переменной и параметрам модели». Подход, описанный в данных работах, также коррелирует с библиотеками машинного обучения, например sklearn.

Пример: перцептрон



Самая простая нейронная сеть.
Входные элементы напрямую соединены
с выходными с помощью системы весов.

$$\mathbf{f}_0(\mathbf{x}) = \mathbf{x} \xrightarrow{\mathbf{g}_0^{0,1}(\mathbf{x}) = \sigma(\langle \mathbf{w}, \mathbf{x} \rangle)} \mathbf{f}_1(\mathbf{x})$$

$\mathbf{x} \in \mathrm{R}^n$; $\mathbf{w} \in \mathrm{R}^n$.
$\mathbf{g}_0^{0,1} : \mathrm{R}^n \to \mathrm{R}$.
Модель задается формулой:

$$\mathbf{f}_1 = \mathbf{agg}_1(\gamma_0^{0,1}\mathbf{g}_0^{0,1}(\mathbf{x})).$$

Здесь функция аггрегации тождественна, а $\gamma_0^{0,1} = 1$.

Пример: многослойный перцептрон и RBF

Перцептрон, в котором присутствует дополнительный скрытый слой.
Нейроны одного слоя между собой не связаны, при этом каждый нейрон связан с каждым нейроном соседнего слоя.

$$\mathbf{f}_0(\mathbf{x}) = \mathbf{x} \xrightarrow{\mathbf{g}_0^{0,1}(\mathbf{x})} \mathbf{f}_1(\mathbf{x}) \xrightarrow{\mathbf{g}_0^{1,2}(\mathbf{u})} \mathbf{f}_2(\mathbf{x})$$

$\mathbf{x} \in \mathrm{R}^n$; $\mathbf{W} \in \mathrm{R}^{h \times n}$; $\mathbf{u} \in \mathrm{R}^h$; $\mathbf{w} \in \mathrm{R}^h$.
$\mathbf{g}_0^{0,1}(\mathbf{x}) = \sigma_h(W\mathbf{x}) : \mathrm{R}^n \to \mathrm{R}^h$.
$\mathbf{g}_0^{1,2}(\mathbf{u}) = \sigma(\langle \mathbf{w}, \mathbf{u} \rangle) : \mathrm{R}^h \to \mathrm{R}$.
Модель задается формулой:

$$\mathbf{f}_2(\mathbf{x}) = \gamma_0^{1,2}\mathbf{g}_0^{1,2}(\gamma_0^{0,1}\mathbf{g}_0^{0,1}(\mathbf{x})).$$

Пример: Deep Feed Forward

$$\mathbf{f}_0(\mathbf{x}) = \mathbf{x} \xrightarrow{\mathbf{g}_0^{0,1}(\mathbf{x})} \mathbf{f}_1(\mathbf{x}) \xrightarrow{\mathbf{g}_0^{1,2}(\mathbf{u}_1)} \mathbf{f}_2(\mathbf{x}) \xrightarrow{\mathbf{g}_0^{2,3}(\mathbf{u}_2)} \mathbf{f}_3(\mathbf{x}) = a(\mathbf{x})$$

Выходные значения сети $\mathbf{a} \in \mathbb{R}^m$ на объекте $\mathbf{x}$:

$$\mathbf{a}(\mathbf{x}) = \sigma_m\left(\langle \mathbf{w}, \mathbf{u}_2 \rangle\right) \quad \mathbf{u}_2 = \sigma_{h_2}\left(\mathbf{W}_2\mathbf{u}_1\right) \quad \mathbf{u}_1 = \sigma_{h_1}\left(\mathbf{W}_1\mathbf{x}\right),$$

где
$$\sigma_m : \mathbb{R} \to \mathbb{R}, \qquad \mathbf{w} \in \mathbb{R}^{h_2},$$
$$\sigma_{h_2} : \mathbb{R}^{h_2} \to \mathbb{R}^{h_2}, \qquad \mathbf{W}_2 \in \mathbb{R}^{h_2 \times h_1},$$
$$\sigma_{h_1} : \mathbb{R}^{h_1} \to \mathbb{R}^{h_1}, \qquad \mathbf{W}_1 \in \mathbb{R}^{h_1 \times n}.$$

Модель задается формулой:

$$\mathbf{f}_3(\mathbf{x}) = \gamma_0^{2,3}\mathbf{g}_0^{2,3}(\gamma_0^{1,2}\mathbf{g}_0^{1,2}(\gamma_0^{0,1}\mathbf{g}_0^{0,1}(\mathbf{x}))).$$

Вариация — ELM со случайными связями между нейронами, то есть в функциях $\mathbf{g}$ компоненты зависят не от всех аргументов.

Пример: сверточная нейросеть

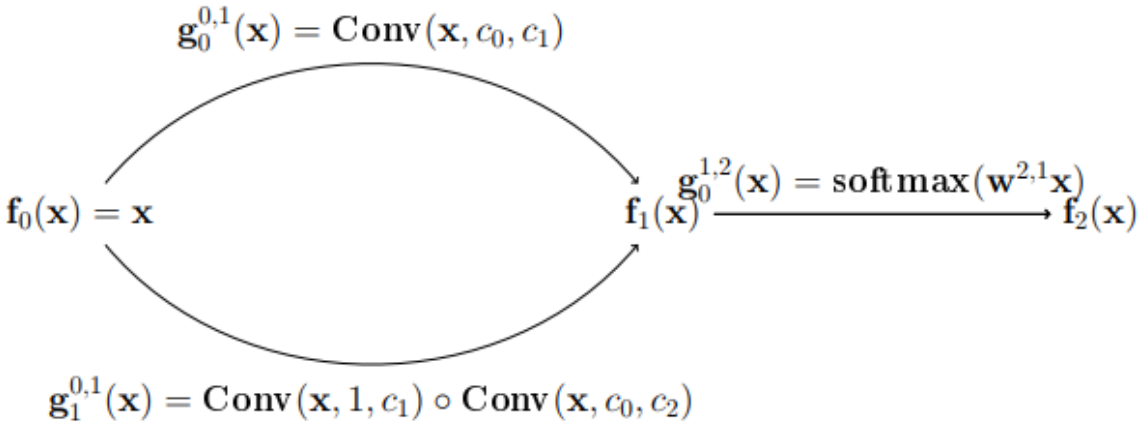


Рис. 1.1. Пример семейства моделей глубокого обучения: семейство описывает сверточную нейронную сеть.

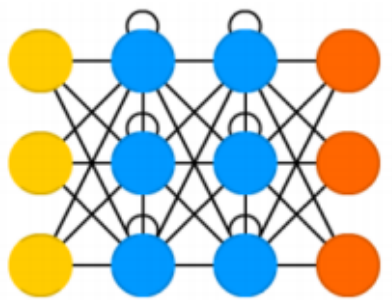
Пример семейства моделей, которое описывает сверточную нейронную сеть, представлена на Рис. 1.1. Семейство задает множество моделей с двумя операциями свертки с одинаковым размером фильтра c_0 и различным числом каналов c_1 и c_2. Единичная свертка с c_1 каналами $\mathbf{Conv}(\mathbf{x}, c_1, 1)$ требуется для выравнивания размерностей скрытых слоев. Каждая модель семейства задается формулой:

$$\mathbf{f} = \mathbf{agg}_2\left(\left\{\boldsymbol{\gamma}_0^{1,2}\mathbf{g}_0^{1,2}\left(\mathbf{agg}_1\left(\{\boldsymbol{\gamma}_0^{0,1}\mathbf{g}_0^{0,1}(\mathbf{x}), \boldsymbol{\gamma}_1^{0,1}\mathbf{g}_1^{0,1}(\mathbf{x})\}\right)\right)\right\}\right).$$

Положим, что функции аггрегации $\mathbf{agg}_1, \mathbf{agg}_2$ являются операциями суммы. Заметим, что к вершине "2" ведет только одно ребро, поэтому операцию суммы можно опустить. Итоговая формула модели задается следующим образом:

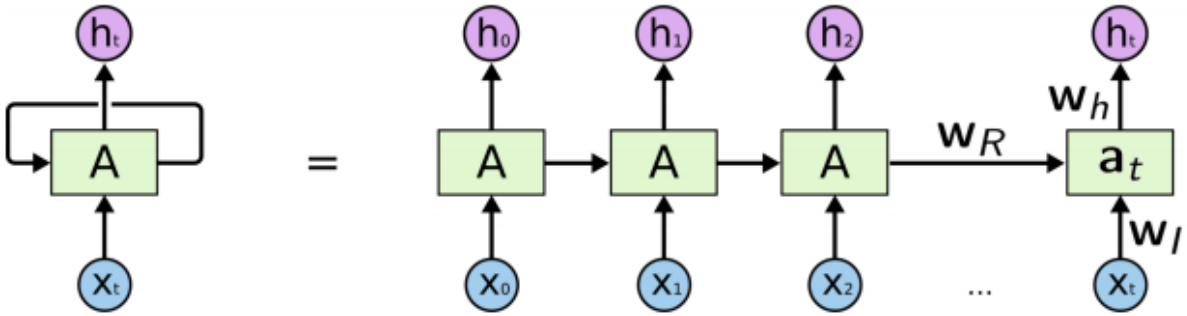
$$\mathbf{f} = \boldsymbol{\gamma}_0^{1,2}\mathbf{softmax}\left(\boldsymbol{\gamma}_0^{0,1}\mathbf{Conv}(\mathbf{x}, c_0, c_1)(\mathbf{x}) + \boldsymbol{\gamma}_1^{0,1}\mathbf{Conv}(\mathbf{x}, 1, c_1) \circ \mathbf{Conv}(\mathbf{x}, c_0, c_2)(\mathbf{x})\right).$$

Пример: рекуррентная нейросеть



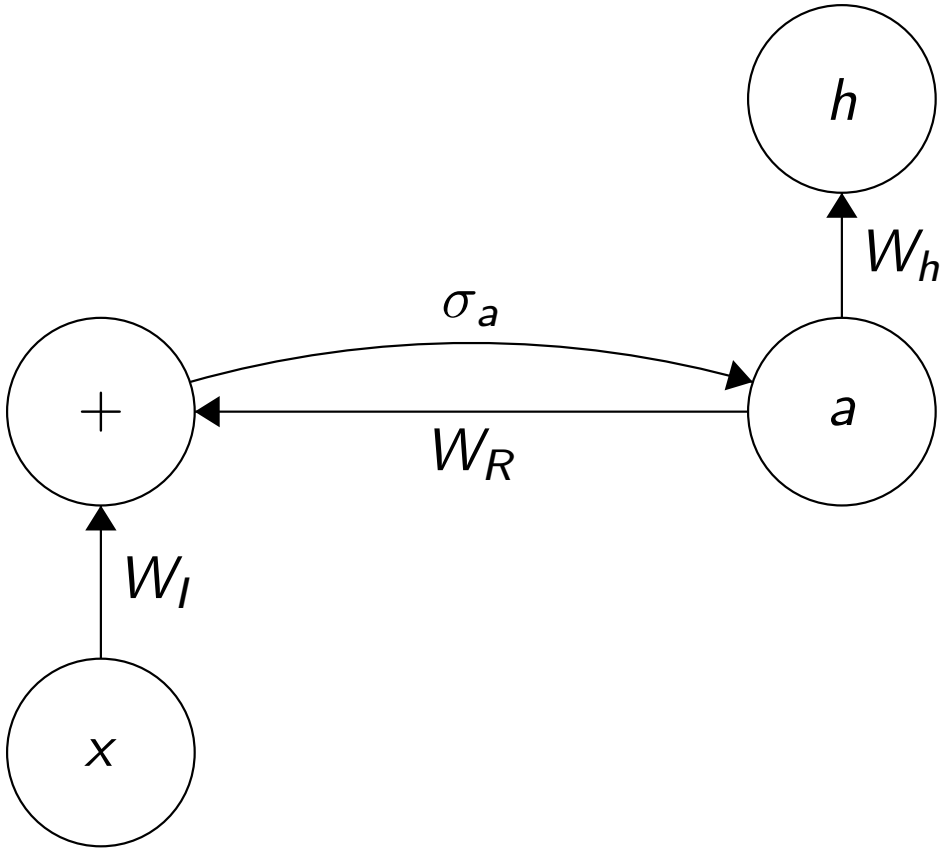
Нейроны получают информацию не только от предыдущего слоя, но и от самих себя в результате предыдущего прохода.

Развернем обратную связь одного нейрона:



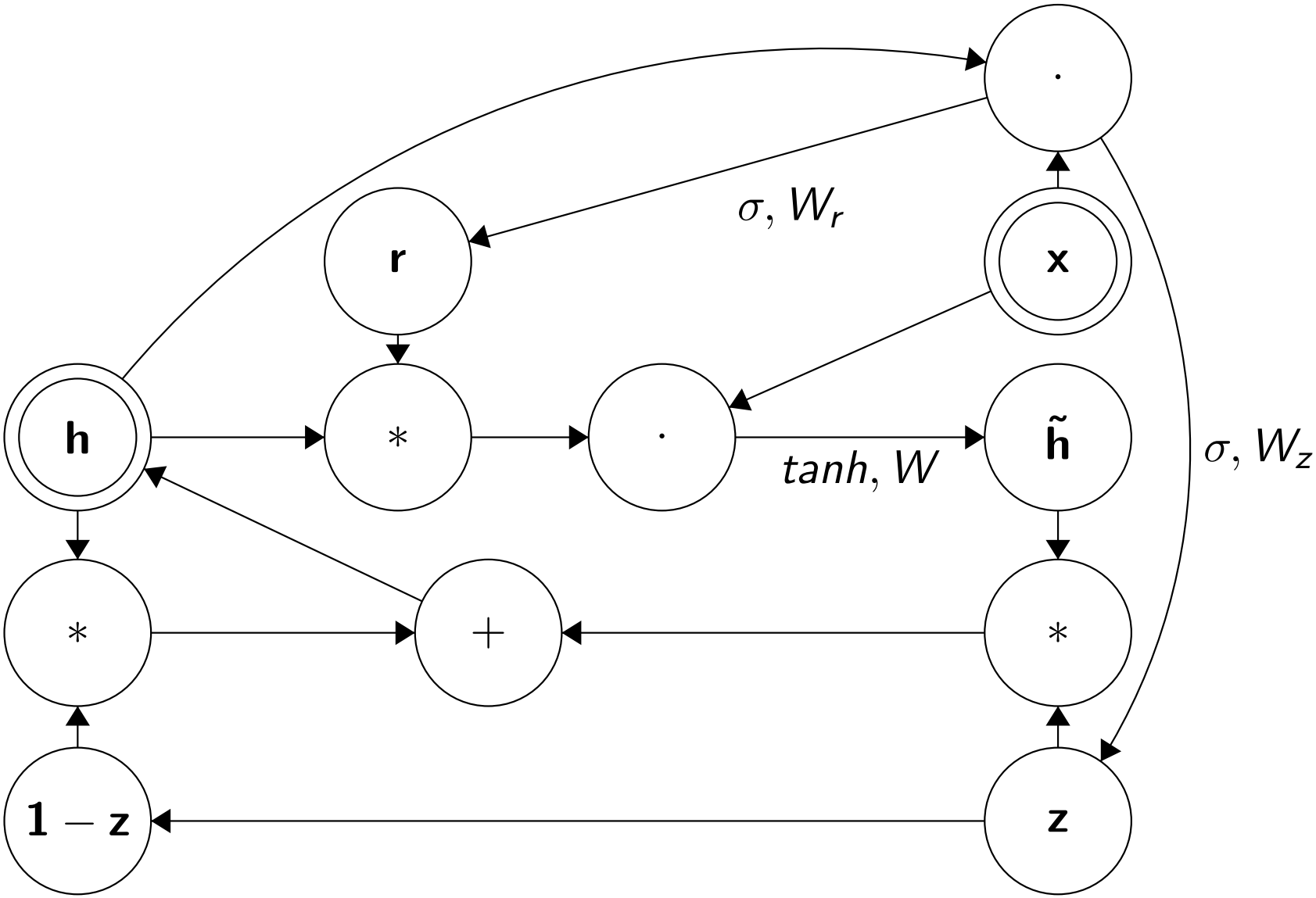
$$\mathbf{h}_t = \sigma_h(\mathbf{W}_h\mathbf{a}_t) \qquad \mathbf{a}_t = \sigma_a(\mathbf{W}_I\mathbf{x}_t + \mathbf{W}_R\mathbf{a}_{t-1})$$

Пример: рекуррентная нейросеть



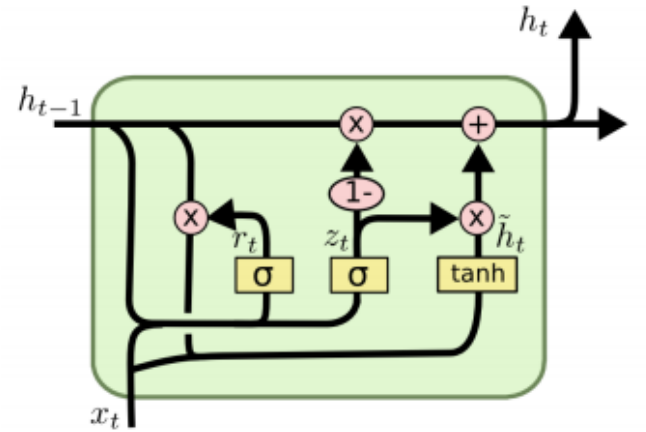
В этом случае все структурные параметры равны единице.
$\mathbf{x} \in \mathrm{R}^n$; $\mathbf{W_I} \in \mathrm{R}^{a \times n}$; $\mathbf{W_R} \in \mathrm{R}^{a \times a}$; $\mathbf{W_h} \in \mathrm{R}^{h \times o}$; $\mathbf{a} \in \mathrm{R}^a$; $\mathbf{h} \in \mathrm{R}^h$.

Пример: GRU — упрощенный вариант LSTM



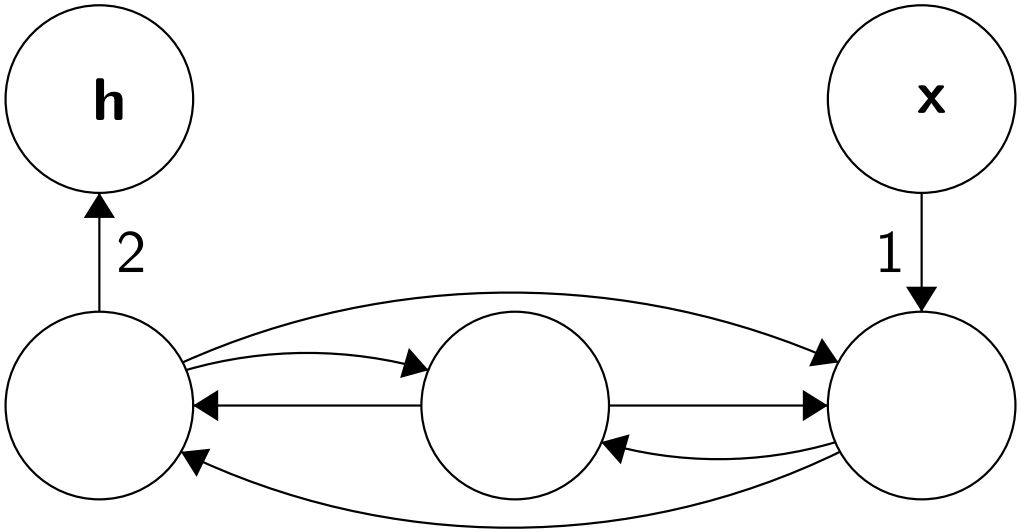
Пример: GRU — упрощенный вариант LSTM

$\mathbf{x} \in \mathrm{R}^n$; $\mathbf{h} \in \mathrm{R}^h$;
$W_z, W_r, W \in \mathrm{R}^{h \times (n+h)}$;
$\mathbf{r}, \mathbf{z}, \tilde{\mathbf{h}}, \mathbf{1} \in \mathrm{R}^h$.



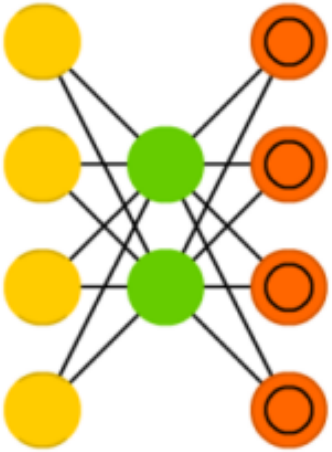
$$\mathbf{z}_t = \sigma\left(\mathbf{W}_z \left[\mathbf{h}_{t-1}, \mathbf{x}_t\right]\right),$$
$$\mathbf{r}_t = \sigma\left(\mathbf{W}_r \left[\mathbf{h}_{t-1}, \mathbf{x}_t\right]\right),$$
$$\widetilde{\mathbf{h}}_t = \tanh\left(\mathbf{W}\left[\mathbf{r}_t * \mathbf{h}_{t-1}, \mathbf{x}_t\right]\right),$$
$$\mathbf{h}_t = (1 - \mathbf{z}_t) * \mathbf{h}_{t-1} + \mathbf{z}_t * \widetilde{\mathbf{h}}_t.$$

Пример: ESN и LSM



1: $\mathrm{R}^n \to \mathrm{R}^m$, веса инициализируются;
2: $\mathrm{R}^p \to \mathrm{R}^h$, веса настраиваются;
Средний уровень — схематическое изображение динамического резервуара, связи в котором задаются один раз случайным образом.
У ESN сигмоиды, а у LSM — пороговые функции; по достижении порога нейрон посылает импульс другим нейронам.

Пример: автокодировщик



Другой способ использования FFNN.
Идея — автоматическое кодирование.

Конструируется таким образом, чтобы не иметь возможность точно скопировать вход на выходе.

$$\mathbf{f}_0(\mathbf{x}) = \mathbf{x} \xrightarrow{\mathbf{g}_0^{0,1}(\mathbf{x})} \mathbf{f}_1(\mathbf{x}) \xrightarrow{\mathbf{g}_0^{1,2}(\mathbf{u})} \mathbf{f}_2(\mathbf{x})$$

$\mathbf{x} \in \mathrm{R}^n$; $\mathbf{W_h} \in \mathrm{R}^{h \times n}$; $\mathbf{W_n} \in \mathrm{R}^{n \times h}$; $\mathbf{u} \in \mathrm{R}^h$.
$\mathbf{g}_0^{0,1}(\mathbf{x}) = \sigma_h(W_h\mathbf{x}) : \mathrm{R}^n \to \mathrm{R}^h$.
$\mathbf{g}_0^{1,2}(\mathbf{u}) = \sigma_n(W_n\mathbf{u}) : \mathrm{R}^h \to \mathrm{R}^n$.
Модель задается формулой:

$$\mathbf{f}_2(\mathbf{x}) = \gamma_0^{1,2}\mathbf{g}_0^{1,2}(\gamma_0^{0,1}\mathbf{g}_0^{0,1}(\mathbf{x})).$$

Значения структурных параметров — единицы.

Пример: DAE

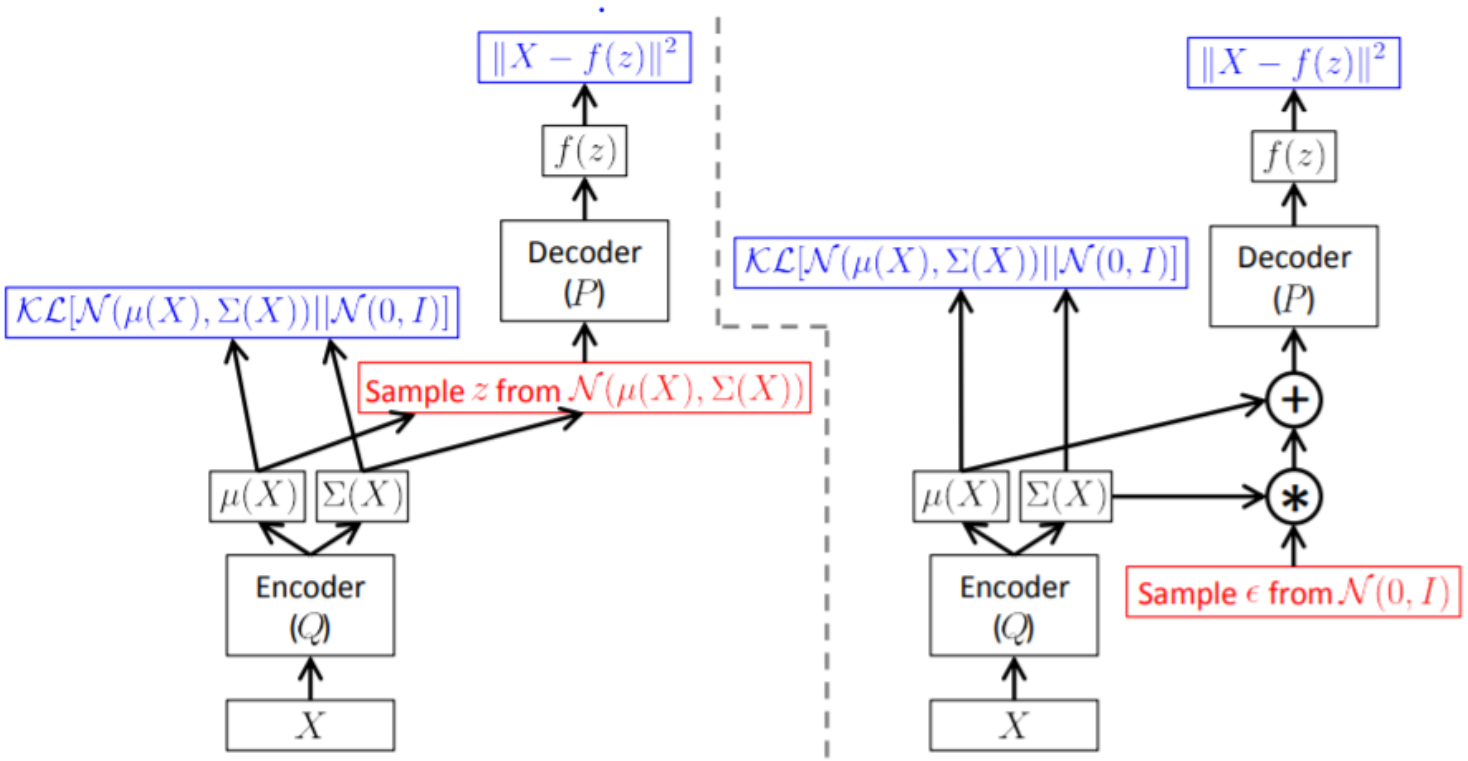
Автокодировщик, которому на вход подаются данные с шумом, а сравнение проводится с исходными данными.

$$\mathbf{f}_0(\mathbf{x}) = \mathbf{x} \xrightarrow{noise} \mathbf{f}_1(\mathbf{x}) = \tilde{\mathbf{x}} \xrightarrow{g_0^{1,2}(\tilde{\mathbf{x}})} \mathbf{f}_2(\mathbf{x}) \xrightarrow{g_0^{2,3}(\mathbf{u})} \mathbf{f}_3(\mathbf{x})$$

$\mathbf{x}, \tilde{\mathbf{x}} \in \mathrm{R}^n$; $\mathbf{W_h} \in \mathrm{R}^{h \times n}$; $\mathbf{W_n} \in \mathrm{R}^{n \times h}$; $\mathbf{u} \in \mathrm{R}^h$.
$\mathbf{g}_0^{0,1}(\mathbf{x}) = \sigma_h(W_h\tilde{\mathbf{x}}) : \mathrm{R}^n \to \mathrm{R}^h$.
$\mathbf{g}_0^{1,2}(\mathbf{u}) = \sigma_n(W_n\mathbf{u}) : \mathrm{R}^h \to \mathrm{R}^n$.
Модель задается формулой:

$$\mathbf{f}_2(\mathbf{x}) = \gamma_0^{2,3}\mathbf{g}_0^{2,3}(\gamma_0^{1,2}\mathbf{g}_0^{1,2}(\tilde{\mathbf{x}})).$$

Пример: VAE

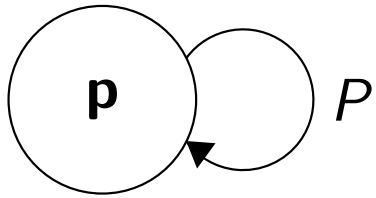


X — выборка из R^n;

$\mu(X) \in \mathrm{R}^n, \Sigma(X) \in \mathrm{R}^{n \times n}$ — выборочные среднее и ковариационная матрица.

На вход декодеру подается $z \in \mathrm{R}^m$.

Пример: марковская цепь

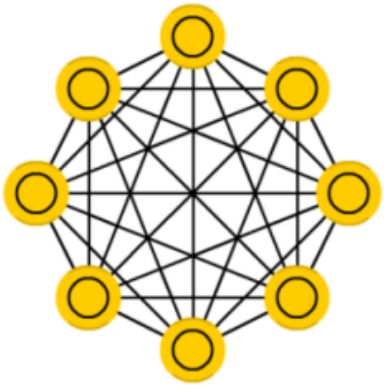


$\mathbf{p} \in \mathrm{R}^n$ — вектор вероятностей находиться в состояниях $\{1, 2, \ldots, n\}$;
$P \in \mathrm{R}^{n \times n}$ — матрица вероятностей перехода.

$$\mathbf{p}(t+1) = \mathbf{p}(t)P.$$

Как частный случай можно представить сеть Хопфилда.

Пример: сеть Хопфилда



Полносвязная сеть.

Каждый нейрон служит входным до обучения, скрытым во время него и выходным после.

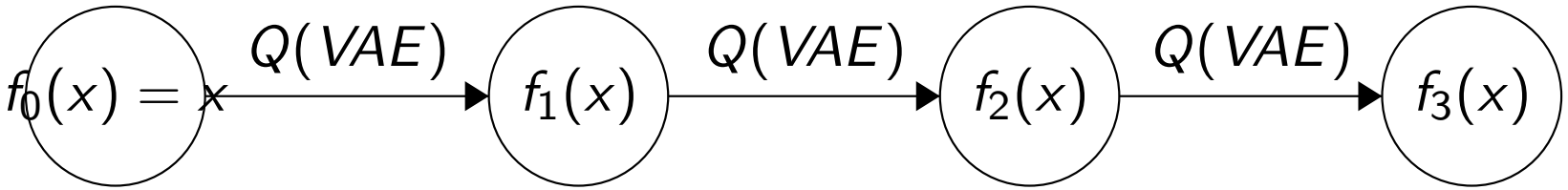
Пусть $S_i(t) \in \{-1, 1\}$ состояние i-ого нейрона в момент t.
Динамика состояния всех нейронов в сети из N нейронов:

$$\mathbf{S}(t+1) = \text{sign}\left(\mathbf{W}\mathbf{S}(t)\right),$$

где матрица $\mathbf{W} \in \mathbb{R}^{N \times N}$ —матрица весовых коэффициентов, описывающая взаимодействия нейронов.

$\mathbf{S}(t+1) = \text{sign}(W\mathbf{S}(t))$. Энергия: $E = -S^T WS$.
Стохастический частный случай — машина Больцмана, для которой согласно распределению Больцмана $p = \frac{1}{Z} e^{-E(s)}$ определяется вероятность нейрона оказаться в том или ином состоянии.

DBN

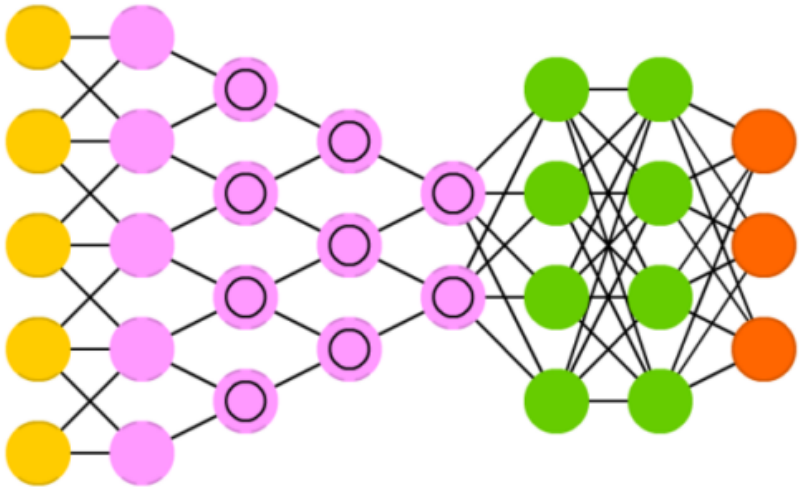


Пожалуй, самый яркий пример суперпозиции.

i-ый кодировщик отображает пространство $\mathrm{R}^{n_{i-1}}$ в R^{n_i}, $n_0 = n$.

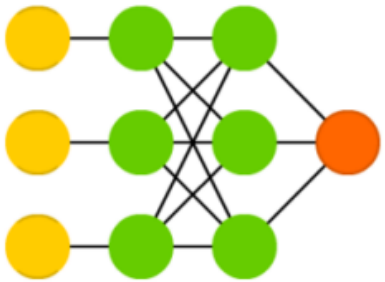
Суперпозиция: $\mathrm{R}^n \to \mathrm{R}^{n_i}$.

DCN



$$x \xrightarrow{Conv(x, c_0, c_1)} f_1(x) \xrightarrow{Pool} f_2(x) \xrightarrow{FFNN} f_3(x) \xrightarrow{Softmax} f_4(x)$$

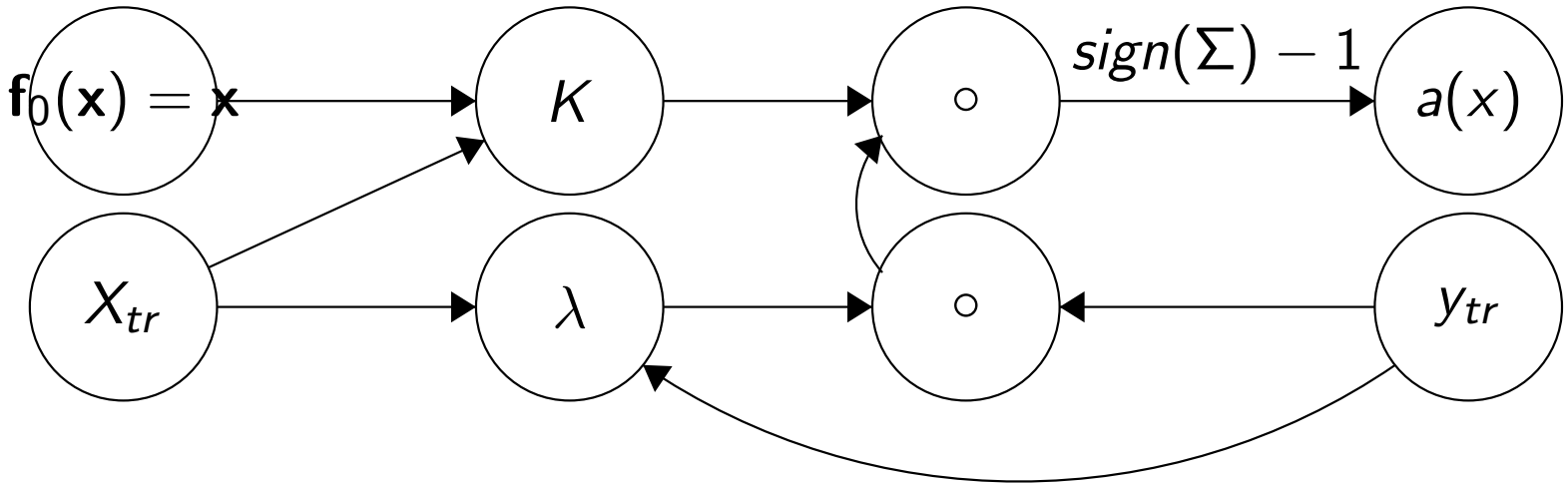
$$\mathrm{R}^{n \times n} \to \mathrm{R}^{m \times m} \to \mathrm{R}^{h}$$



Модель:

$$a(\mathbf{x}) = \text{sign}\left(\sum_{i=1}^{h} \lambda_i y_i K(\mathbf{x}_i, \mathbf{x}) + w_0\right).$$

$$K(u, v) = \tanh\left(k_0 + k_1\langle u, v\rangle\right) \to \sigma.$$
$$K(u, v) = \exp\left(-\beta\|v - u\|^2\right) \to \text{ RBF.}$$



$\mathbf{x} \in \mathrm{R}^n$; $X_{tr} \in \mathrm{R}^{n\times h}$; $y_{tr} \in \mathrm{R}^h$; $a(\mathbf{x}) \in \mathrm{R}$.
$K(x, X_{tr}) = [K(x, x_1), \ldots, K(x, x_h)]$;
$\lambda \in \mathrm{R}^h$ — вектор множителей Лагранжа;
$\circ$ — покомпонентное умножение.

Заключение

Результаты

- Предложено формальное определение модели, удобное для конструирования и оптимизации;
- Предложен способ описания модели машинного обучения;
- Построены описания некоторых распространенных моделей.