

Модели машинного обучения. Суперпозиции

Плетнев Никита Вячеславович

Московский физико-технический институт

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

- 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 \mathcal{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 нелинейных функций. Под структурой модели понимаются значения структурных параметров модели, т.е. величин, задающих вид итоговой суперпозиции.

Объект

пара (\mathbf{x}, y) , $\mathbf{x} \in X = \mathbb{R}^n$, $y \in Y$.

В случае задачи классификации y является распределением вероятностей принадлежности объекта $\mathbf{x} \in X$ множеству классов $\{1, \dots, Z\}$: $Y \subseteq [0, 1]^Z$, где Z — число классов.

В случае задачи регрессии Y является некоторым подмножеством вещественных чисел $Y \subseteq \mathbb{R}$.

\mathbf{x} — признаковое описание, y — метка объекта.

Модель \mathbf{f}

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

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

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 $K^{j,k}$: $\mathbf{g}^{j,k} = [\mathbf{g}_0^{j,k}, \dots, \mathbf{g}_{K^{j,k}}^{j,k}]$. Пусть также для каждой вершины v определена функция агрегации \mathbf{agg}_v . Граф (V, E) в совокупности со множеством векторов базовых функций $\{\mathbf{g}^{j,k}, (j, k) \in E\}$ и множеством функций агрегаций $\{\mathbf{agg}_v, v \in V\}$ называется *семейством моделей* \mathfrak{F} , если функция, задаваемая рекурсивно как

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

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

Примерами функций агрегации являются функции суммы и конкатенации векторов.

Слои, или подмодели

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

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

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

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

Конкатенация векторов $\gamma^{j,k}$.

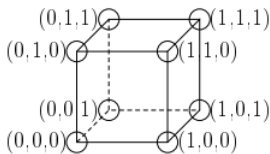
Множество всех возможных значений структуры Γ будем обозначать Γ . Векторы $\{\gamma^{j,k} \mid (j, k) \in E\}$ назовем *структурными параметрами модели*.

Параметризация множества моделей M

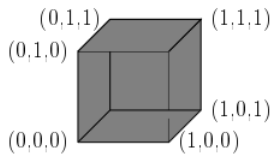
Семейство моделей \mathfrak{F} , такое что для каждой модели $\mathbf{f} \in M$ существует значение структуры модели Γ , при котором функция \mathbf{f} совпадает с функцией из определения семейства моделей.

Варианты множества структур Γ

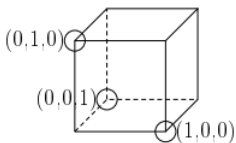
- ❶ Вершины булева куба: $\gamma^{j,k} \in \{0, 1\}^{K^{j,k}}$;
- ❷ Внутренность булева куба: $\gamma^{j,k} \in [0, 1]^{K^{j,k}}$;
- ❸ Вершины симплекса: $\gamma^{j,k} \in \overline{\Delta}^{K^{j,k}-1}$;
- ❹ Внутренность симплекса: $\gamma^{j,k} \in \Delta^{K^{j,k}-1}$.



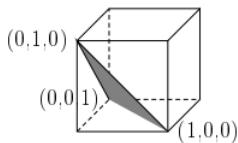
(а)



(б)



(в)



(г)

Рис. 1.2. Примеры ограничений для одного структурного параметра γ , $|\gamma| = 3$.
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.

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.

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

$$\mathbf{x} \in \mathbb{R}^n; \mathbf{w} \in \mathbb{R}^n.$$

$$\mathbf{g}_0^{0,1} : \mathbb{R}^n \rightarrow \mathbb{R}.$$

Модель задается формулой:

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

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

Пример: многослойный перцептрон



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

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

$$\mathbf{x} \in \mathbb{R}^n; \mathbf{W} \in \mathbb{R}^{h \times n}; \mathbf{u} \in \mathbb{R}^h; \mathbf{w} \in \mathbb{R}^h.$$

$$\mathbf{g}_0^{0,1}(\mathbf{x}) = \sigma_h(W\mathbf{x}) : \mathbb{R}^n \rightarrow \mathbb{R}^h.$$

$$\mathbf{g}_0^{1,2}(\mathbf{u}) = \sigma(\langle \mathbf{w}, \mathbf{u} \rangle) : \mathbb{R}^h \rightarrow \mathbb{R}.$$

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: сверточная нейросеть

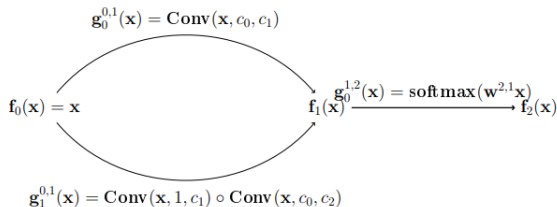


Рис. 1.1. Пример семейства моделей глубокого обучения: семейство описывает сверточную нейронную сеть.

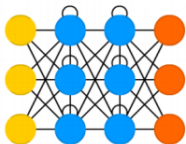
Пример семейства моделей, которое описывает сверточную нейронную сеть, представлена на Рис. 1.1. Семейство задает множество моделей с двумя операциями свертки с одинаковым размером фильтра c_0 и различным числом каналов c_1 и c_2 . Единичная свертка с c_1 каналами $\text{Conv}(\mathbf{x}, c_1, 1)$ требуется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

$$\mathbf{f} = \text{agg}_2 \left(\left\{ \gamma_0^{1,2} \mathbf{g}_0^{1,2} \left(\text{agg}_1 \left(\{ \gamma_0^{0,1} \mathbf{g}_0^{0,1}(\mathbf{x}), \gamma_1^{0,1} \mathbf{g}_1^{0,1}(\mathbf{x}) \} \right) \right) \right\} \right).$$

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

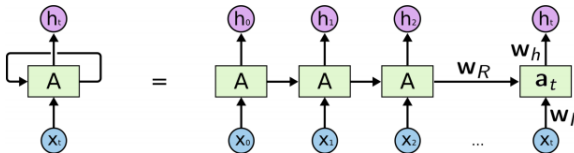
$$\mathbf{f} = \gamma_0^{1,2} \text{softmax} \left(\gamma_0^{0,1} \text{Conv}(\mathbf{x}, c_0, c_1)(\mathbf{x}) + \gamma_1^{0,1} \text{Conv}(\mathbf{x}, 1, c_1) \circ \text{Conv}(\mathbf{x}, c_0, c_2)(\mathbf{x}) \right).$$

Пример: рекуррентная нейросеть



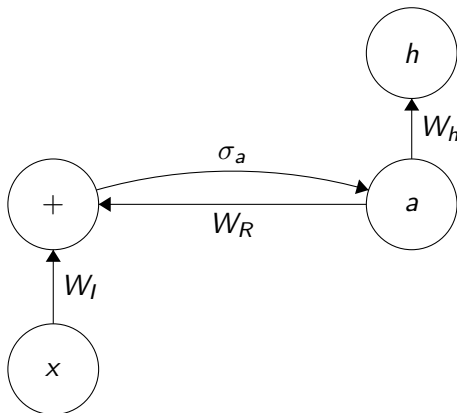
Нейроны получают информацию не только от предыдущего слоя, но и от самих себя в результате предыдущего прохода.

Развернем обратную связь одного нейрона:



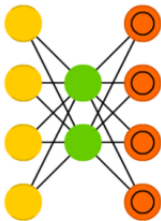
$$h_t = \sigma_h(W_h a_t) \quad a_t = \sigma_a(W_I x_t + W_R a_{t-1})$$

Пример: рекуррентная нейросеть



В этом случае все структурные параметры равны единице.
 $\mathbf{x} \in \mathbb{R}^n$; $\mathbf{W}_I \in \mathbb{R}^{a \times n}$; $\mathbf{W}_R \in \mathbb{R}^{a \times a}$; $\mathbf{W}_h \in \mathbb{R}^{h \times a}$; $\mathbf{a} \in \mathbb{R}^a$; $\mathbf{h} \in \mathbb{R}^h$.

Пример: автокодировщик



Другой способ использования FFNN.
Идея — автоматическое кодирование.

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

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

$$\mathbf{x} \in \mathbb{R}^n; \mathbf{W}_h \in \mathbb{R}^{h \times n}; \mathbf{W}_n \in \mathbb{R}^{n \times h}; \mathbf{u} \in \mathbb{R}^h.$$

$$\mathbf{g}_0^{0,1}(\mathbf{x}) = \sigma_h(\mathbf{W}_h \mathbf{x}) : \mathbb{R}^n \rightarrow \mathbb{R}^h.$$

$$\mathbf{g}_0^{1,2}(\mathbf{u}) = \sigma_n(\mathbf{W}_n \mathbf{u}) : \mathbb{R}^h \rightarrow \mathbb{R}^n.$$

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.

Результаты

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