

1 Аннотация

В работе рассматривается задача выбора структуры модели глубокого обучения. Модель — это вычислительный граф, т.е. граф, в котором ребрами выступают нелинейные функции, а вершинами — результаты действия функцией на выборку. Каждому ребру поставлено в соответствие множество нелинейных функций, такое что любая линейная комбинация этих функций определяет дифференцируемую функцию заданной сигнатуры. Структурой модели назовем веса линейной комбинации этих функций.

Для нахождения оптимальной структуры предлагается ввести вероятностную интерпретацию модели. Проводится градиентная оптимизация параметров и гиперпараметров модели на основе вариационного вывода. Оптимизация решает двухуровневую задачу: на первом уровне проводится оптимизация нижней оценки правдоподобия модели по вариационным параметрам модели. На втором уровне проводится оптимизация гиперпараметров модели. В качестве оптимизируемой функции для гиперпараметров модели предлагается обобщенная функция качества, позволяющая проводить оптимизацию в нескольких режимах: Add-Del, в режиме полного перебора, а также в режиме максимизации нижней оценки правдоподобия модели.

The paper presents the method of deep learning model structure selection. The model is a computation graph, i.e. a graph, which edges correspond to nonlinear functions and vertices correspond to the intermediate state of the dataset after application of these functions. Each edge has a set of available nonlinear function such that any linear combination of them is a differential function of the predetermined signature. The structure of the model is a vector of weights for linear combinations of these functions.

For the optimal model structure selection the author proposes a probabilistic interpretation of the model. The model parameters and hyperparameters are optimized using gradient methods based on variational inference. The optimization of the model is a bi-level optimization problem. The first level corresponds to the evidence lower bound optimization with regard to variational model parameter optimization. The second level corresponds to hyperparameter optimization. The author proposes a generalized loss function for this level, which allows to optimize the hyperparameters in different regimes: Add-Del, brute force optimization and evidence lower bound optimization.

2 Постановка задачи

Задана выборка

$$\mathcal{D} = \{(\mathbf{x}_i, y_i)\}, i = 1, \dots, m, \quad (1)$$

состоящая из множества пар «объект-метка»

$$\mathbf{x}_i \in \mathbf{X} \subset \mathbb{R}^n, \quad y_i \in \mathbf{y} \subset \mathbb{Y}.$$

Метка y объекта \mathbf{x} принадлежит либо множеству: $y \in \mathbb{Y} = \{1, \dots, Z\}$ в случае задачи классификации, где Z — число классов, либо некоторому подмножеству вещественных

чисел $y \in \mathbb{Y} \subseteq \mathbb{R}$ в случае задачи регрессии. Определим множество архитектур моделей глубокого обучения для дальнейшего выбора оптимальной.

Будем рассматривать модель как граф V, E , ребрами E которого являются функции из заданного множества функций \mathbf{O} , действующие на выборку, а вершинами V — промежуточные представления выборки под действием данных функций. Перейдем к формальному определению модели. Пусть задан граф V, E . Пусть для каждого ребра $\langle i, j \rangle \in E$ определено множество функций $\mathbf{o}(i, j)$. Граф V, E с множеством функций \mathbf{O} называется моделью, если функция, задаваемая рекурсивно как

$$f_i(\mathbf{x}) = \sum_{j \in \text{Adj}(v_i)} o(i, j)(f_j(\mathbf{x})),$$

является непрерывной дифференцируемой функцией из \mathbb{R}^n во множество \mathbb{Y} при любом $o(i, j)$, являющемся линейной комбинацией функций из множества $\mathbf{o}(i, j)$.

Пусть для каждого ребра i, j задан нормированный положительный вектор $\gamma_{i,j} \in \mathbb{R}^{|\mathbf{o}(i,j)|}$, определяющий веса функций из множества $\mathbf{o}(i, j)$. Будем считать, что вектор $\gamma_{i,j}$ распределен по распределению Gumbel-Softmax:

$$\gamma_{i,j} \sim \text{GS}(c, \mathbf{m}_{i,j}).$$

где c — вектор концентрации распределения, $\mathbf{m}_{i,j}$ — вектор средних. Обозначим за структуру модели $\mathbf{\Gamma}$ множество всех векторов γ . Заметим, что Gumbel-Softmax-распределение может быть заменено на распределение Дирихле. Однако в дальнейшем нам понадобится проводить градиентную оптимизацию по параметрам данного распределения. Наиболее простым вариантом для подобной оптимизации является Gumbel-Softmax.

Для более адекватной оптимизации положим априорное распределение параметров модели \mathbf{w} зависящим от структуры модели $\mathbf{\Gamma}$. Пусть задано соответствие S между каждым параметром $\mathbf{w} \in \mathbf{W}$ и параметром структуры γ_i . Положим распределение параметров \mathbf{w} нормальным с нулевым средним и диагональной ковариационной матрицей:

$$\mathbf{w} \sim \mathcal{N}(\mathbf{0}, \mathbf{A}^{-1} \cdot S).$$

Пусть также определено правдоподобие выборки $p(\mathbf{y}|\mathbf{X}, \mathbf{w}, \mathbf{\Gamma})$.

Определение Правдоподобием модели \mathbf{f} назовем следующее выражение:

$$p(\mathbf{y}|\mathbf{X}, \mathbf{A}, \mathbf{m}, c) = \int_{\mathbf{w}, \mathbf{\Gamma}} p(\mathbf{y}|\mathbf{X}, \mathbf{w}, \mathbf{\Gamma}) p(\mathbf{w}|\mathbf{A}) p(\mathbf{\Gamma}|\mathbf{m}, c) d\mathbf{w} d\mathbf{\Gamma}. \quad (2)$$

Пусть задано значение концентрации c . Требуется найти гиперпараметры модели \mathbf{A}, \mathbf{m} доставляющие максимум правдоподобия модели:

$$\arg \max_{\mathbf{A}, \mathbf{m}} p(\mathbf{y}|\mathbf{X}, \mathbf{A}, \mathbf{m}, c).$$

Утверждение (доказано). При $c \ll 0$ оптимизация (2) эквивалентна оптимизации дискретной оптимизации: $\gamma_{i,j} \in 2^{|\mathbf{o}(i,j)|}$.

3 Вариационная постановка задачи

В общем виде вычисление значения интеграла (2) является вычислительно сложной процедурой. В качестве приближенного значения интеграла будем использовать вариационную верхнюю оценку правдоподобия модели. Пусть заданы непрерывные параметрические распределения q_w, q_γ , аппроксимирующие апостериорные распределения $p(\mathbf{w}|\mathbf{y}, \mathbf{X}, \mathbf{A}, \mathbf{m}, c)$, $p(\Gamma|\mathbf{y}, \mathbf{X}, \mathbf{A}, \mathbf{m}, c)$. Тогда верно следующее выражение:

$$\log p(\mathbf{y}|\mathbf{X}, \mathbf{A}, \mathbf{m}, c) \leq \mathbb{E}_{q_w, q_\gamma} \log p(\mathbf{y}|\mathbf{X}, \mathbf{w}, \Gamma, \mathbf{A}, \mathbf{m}, c) - D_{KL}(q_\gamma||p(\Gamma)) - D_{KL}(q_w||p(\mathbf{w})). \quad (3)$$

Разница между верхней оценкой (3) и правдоподобием модели (2) определяется дивергенцией между апостериорными распределениями $p(\mathbf{w}|\mathbf{y}, \mathbf{X}, \mathbf{A}, \mathbf{m}, c)$, $p(\Gamma|\mathbf{y}, \mathbf{X}, \mathbf{A}, \mathbf{m}, c)$ и вариационными распределениями q_w, q_γ .

Сформулируем основные требования к оптимизационной задаче и оптимизируемым функционалам:

1. Оптимизируемые функции должны быть дифференцируемы.
2. Степень регуляризации структуры и параметров должна быть контролируемой.
3. Оптимизация должна приводить к максимуму вариационной оценки.
4. Оптимизация должна позволять калибровать количество эффективных параметров
5. Оптимизация должна позволять калибровать количество эффективных ребер.
6. Оптимизация должна позволять проводить полный перебор структуры.

Положим θ равным параметрам распределений q_w, q_γ . Положим $\mathbf{h} = [\mathbf{A}, \mathbf{m}]$.

Пусть L — приближенное значение вариационной оценки правдоподобия:

$$L = \beta \log p(\mathbf{y}|\hat{\mathbf{w}}, \hat{\Gamma}) - D_{KL}(q_\gamma||p(\Gamma)) - D_{KL}(q_w||p(\mathbf{w})),$$

где $\hat{\mathbf{w}} \sim q_w$, $\hat{\Gamma} \sim q_\gamma$, β — коэффициент, контролирующий степень регуляризации структуры и параметров.

Утверждение (доказано). Пусть β выражается как $\frac{m_0}{m}$, где m_0 — натуральное число. Тогда оптимизация функции L эквивалентна оптимизации вариационной нижней оценки правдоподобия для подвыборки \mathfrak{D} мощностью m_0 .

Пусть Q — валидационная функция:

$$Q(c, c_1, c_2, c_3, \mathbf{p}) = c_1 \log p(\mathbf{y}|\hat{\mathbf{w}}, \hat{\Gamma}) + c_2 [-D_{KL}(q_\gamma||p(\Gamma)) - D_{KL}(q_w||p(\mathbf{w}))] + \\ + c_3 \sum_{p_k \in \mathbf{p}} D_{KL}(q_\gamma||p_k),$$

где \mathbf{p} — заданные распределения на структурах, c_1, c_2, c_3 — коэффициенты.

Сформулируем задачу поиска оптимальной модели как двухуровневую задачу.

$$\hat{\mathbf{h}} = \arg \max_{\mathbf{h} \in \mathbb{R}^h} Q(T^n(\boldsymbol{\theta}_0, \mathbf{h})), \quad (4)$$

где T — оператор оптимизации, решающий задачу оптимизации:

$$L(T^n(\boldsymbol{\theta}_0, \mathbf{h})) \rightarrow \max.$$

Утверждение. Пусть $D_{KL}(q_w|p(\mathbf{w}|\mathbf{y}, \mathbf{X}, \mathbf{A}, \mathbf{m}, c)) = 0$, $D_{KL}(q_\gamma|p(\boldsymbol{\Gamma}|\mathbf{y}, \mathbf{X}, \mathbf{A}, \mathbf{m}, c)) = 0$, пусть $c_1 = 1, c_2 = 1, c_3 = 0$. Тогда оптимизация (4) эквивалентна оптимизации (2).

Определение (предварительно) Параметрической δ -сложностью модели назовем матожидание следующей величины:

$$C_p(\delta, \mathbf{w}) = \mathbb{E} \sum_{w \in \mathbf{w}} I(|w| > \delta).$$

Определение (предварительно) Структурной δ -сложностью модели назовем матожидание следующей величины:

$$C_s(\delta, \boldsymbol{\Gamma}) = \mathbb{E} \sum_{\gamma \in \boldsymbol{\Gamma}} \sum_{\gamma_i \in \gamma} I(\gamma_i > \delta).$$

Утверждение (предварительно). Пусть $c_1 = 1, c_3 = 0, c_2 > 0, c'_2 < c_2$. Пусть \mathbf{w}, \mathbf{w}' — параметры, полученные в результате соответствующих оптимизаций. Тогда $C_p(\delta, \mathbf{w}') \leq C_p(\delta, \mathbf{w})$.

Утверждение (доказано). Пусть $c_1 = 1, c_3 = 0$. Тогда $C_p(\delta, \mathbf{w}(c_2)) \rightarrow_{c_2 \rightarrow \infty} 0$.

Утверждение (предварительно). Пусть $c_1 = 1, c_3 = 0$. Максимум величины $C_p(\delta, \mathbf{w}(c_2))$ достигается при $c_2 = 0$.

Утверждение (доказано). Пусть $c_1 = c_2 = 1$. $C_s(\delta, \mathbf{w}') \rightarrow \min$ при $c \rightarrow 0$.

Утверждение (предварительно). Пусть $c_1 = c_2 = 1$. $C_s(\delta, \mathbf{w}') \rightarrow \max$ при $c \rightarrow \infty$.

Утверждение (предварительно, нужно развить). Пусть $c_3 > 0, c < 0$ и все $p_k \in \mathbf{p}$ отражают распределения на вершинах симплекса. Тогда оптимизация приведет к q_γ , сконцентрированному на одной из остальных вершин симплекса.

Утверждение (очень предварительно). Изменение c позволяет избежать ухода в локальный минимум.

Утверждение (очень предварительно). Изменение c_2 позволяет избежать ухода в локальный минимум.

Утверждение (очень предварительно). Взаимосвязь структуры и параметров в prig позволяет получить «хорошие» модели.

Утверждение (предварительно). Пусть $c_1 = c_2 = c_3 = 0$. Пусть $q_w \sim \mathcal{N}(\mathbf{0}, \sigma), \sigma \sim 0$. Тогда оптимизация эквивалентна обычной оптимизации параметров с l_2 -регуляризацией.

Далее будем рассматривать $q_w \sim \mathcal{N}(\mathbf{0}, \mathbf{A}_q^{-1}), \quad q_\gamma \sim \text{Gumbel-Softmax}(\mathbf{g}, \tau)$.

4 Вычислительный эксперимент

В качестве модельного эксперимента рассматривалась задача выбора модели линейной регрессии. Множество объектов \mathbf{X} было сгенерировано из трехмерного стандартного распределения:

$$\mathbf{X} \sim \mathcal{N}(0, \mathbf{I}), n = 3.$$

Множество меток было определено следующим правилом:

$$\mathbf{y} = \arg \max_{0,1} (\mathbf{X}_1 + \mathbf{X}_2),$$

третья компонента не участвовала в генерации ответа.

Рассматривались четыре возможные структуры:

1. $f_1 = \mathbf{w}_1 \mathbf{X}_1$ (модель — регрессия только по первому признаку),
2. $f_2 = \mathbf{w}_2 \mathbf{X}_2$ (модель — регрессия только по первому признаку),
3. $f_3 = \mathbf{w}_3 \mathbf{X}_3$ (модель — регрессия только по шумовому признаку),
4. $f_4 = \mathbf{w}_4 \mathbf{X} (---)$.

Ожидаемое поведение оптимизации:

1. При $c_1 = c_2 = 1, c \sim 0$ (Evidence с низкой температурой) будет произведен выбор структуры f_4 .
2. При $c_1 = c_2 = 1, c \gg 0$ (Evidence с высокой температурой) будет произведен выбор двух структур с одинаковым весом: f_1, f_2 .
3. При $c_1 = c_2 = 0, c_3 = 1, \mathbf{p} = [[0.0, 0.0, 1.0, 0.0]], c \sim 0$ (Поощряется выбор структуры с шумовой компонентой) будет произведен выбор структуры f_4 , при снижении параметра β выбор будет меняться в сторону f_3 .

Результаты

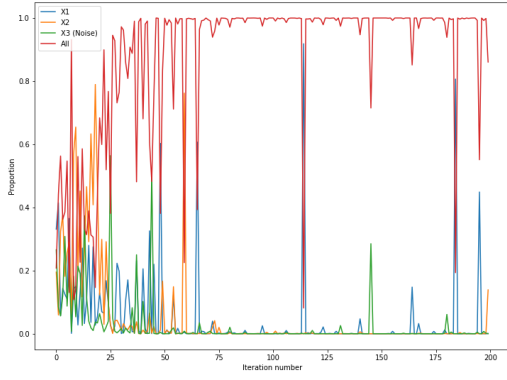


Figure 1: Evidence с низкой температурой

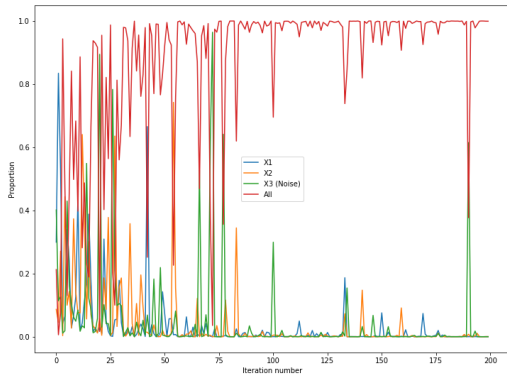


Figure 2: Evidence с высокой температурой

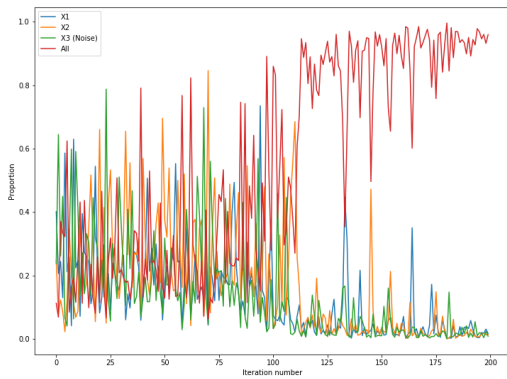


Figure 3: Evidence с высокой температурой, $\beta = 0.01$

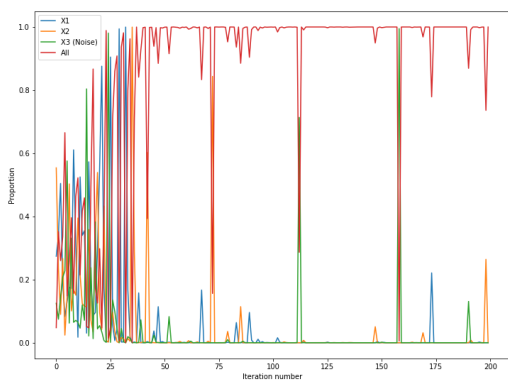


Figure 4: Поощрение выбора шумовой компоненты

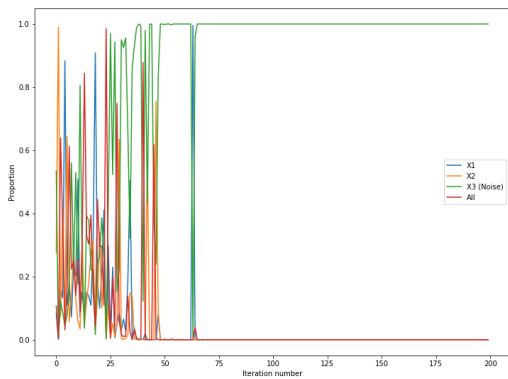


Figure 5: Поощрение выбора шумовой компоненты, $\beta = 0.01$