1 Аннотация

В работе рассматривается задача выбора структуры модели глубокого обучения. Модель — это вычислительный вероятностный 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 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, т.е. предположения о распределениях параметров и структуры модели. Проводится градиентная оптимизация параметров и гиперпараметров модели на основе байесовского вариационного вывода. Решается двухуровневая задача оптимизации: 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 правдоподобия. Показано, что данная функция позволяет проводить оптимизацию несколькими алгоритмами: последовательным добавлением и удалением параметров, полным перебором, а также максимизацией нижней оценки правдоподобия модели.

Проводится сравнение с эвристическими алгоритмами выбора структуры модели. Вычислительный эксперимент проводится на синтетических данных и выборке рукописных цифр MNIST.

Цель работы: предложить метод выбора модели субоптимальной сложности, позволяющий проводить выбор модели в нескольких режимах (ELBO, AddDel, полный перебор, оптимизация без регуляризации и с регуляризацией).

2 Постановка задачи

Задана выборка

$$\mathfrak{D} = \{(\mathbf{x}_i, y_i)\}, i = 1, \ldots, m, \tag{1}$$

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: $y \in \mathbb{Y} = \{1, \ldots, Z\}$ в случае задачи классификации, где Z — число классов, либо некоторому подмножеству вещественных чисел $y \in \mathbb{Y} \subseteq \mathbb{R}$ в случае задачи регрессии. Далее будем полагать, что объекты $\mathbf{x}$ являются реализацией некоторой случайно величины и порождены независимо.

Определим множество архитектур моделей глубокого обучения для дальнейшего выбора оптимальной. Будем рассматривать модель как граф V, E, ребрами E которого являются функции из заданого 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 $< i, j > \in E$ 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 выпуклой комбинацией функций из множества $\mathbf{o}(i, j)$.

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

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

где c — вектор концентрации распределения, $\mathbf{m}_{i,j}$ — вектор средних. Обозначим за структуру модели $\boldsymbol{\Gamma}$ множество всех векторов $\boldsymbol{\gamma}$.

Для более адекватной оптимизации положим априорное распределение параметров модели $\mathbf{w}$ зависящим от структуры модели $\boldsymbol{\Gamma}$. 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 квариационной матрицей:

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

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

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

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

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

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

Теорема. При $c << 0$ оптимизация (2) эквивалентна дискретной оптимизации: $\gamma_{ij} \in \{0, 1\}, ||\gamma_i|| = 1$.

Доказательство. При $c \to 0$ плотность вероятности сконцентрирована на вершинах симплекса (из оригинальной статьи). Т.к. функция вероятности ограничена, то по теореме Лебега допустим предельный переход к распределению на вершинах.

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 $p(\mathbf{w}|\mathbf{y}, \mathbf{X}, \mathbf{A}, \mathbf{m}, c)$, $p(\mathbf{\Gamma}|\mathbf{y}, \mathbf{X}, \mathbf{A}, \mathbf{m}, c)$. Тогда верно следующее выражение:

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

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

Определим основные величины, которые характеризуют сложность модели.
Определение Параметрической сложностью C_w модели назовем наименьшую дивергенцию вариационныъ параметров при условии априорного распределения параметров:

$$C_w = \arg\min_{\mathbf{A}} D_{\mathrm{KL}}\left(q|p\right).$$

(Примечание: кажется, здесь должны учитываться параметры и структура, т.к. в априорном распределении параметров зашита зависимость от структуры).

Определение Структурной сложностью C_γ модели назовем энтропию распределения структуры:

$$C_\gamma = \mathsf{E}_{q_\gamma} \log q_\gamma.$$

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 параметрическую сложность модели
5. Оптимизация должна позволять калибровать структурную сложность модели.
6. Оптимизация должна позволять проводить полный перебор структуры.

Сформулируем задачу как двухуровневую задачу оптимизации. Обозначим за $\boldsymbol{\theta}$ оптимизируемые на первом уровне величины. Обозначим за $\mathbf{h}$ величины, оптимизируемые на втором уровне. Положим $\boldsymbol{\theta}$ равным параметрам распределений q_w, q_γ. Положим $\mathbf{h} = [\mathbf{A}, \mathbf{m}]$.

Пусть L — приближенное значение вариационной оценки правдоподобия:

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

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

Теорема. Пусть $\beta \neq 1$. Тогда функция L сходится по вероятности к вариационной нижней оценке правдоподобия для подвыборки $\mathfrak{D}$ мощностью βm.

Доказательство. Рассмотрим произвольную подвыборку $\hat{\mathfrak{D}}$ мощностью m_0. Верхняя оценка правдоподобия модели для подвыборки имеет вид:

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

$$\log p(\hat{\mathbf{y}}|\hat{\mathbf{X}}, \mathbf{w}, \mathbf{\Gamma}, \mathbf{A}, \mathbf{m}, c) = \sum_i \log p(\hat{\mathbf{y}}_i|\hat{\mathbf{x}}_i, \mathbf{w}, \mathbf{\Gamma}, \mathbf{A}, \mathbf{m}, c) \to^p_{m\to\infty} m\mathsf{E}\log p(\mathbf{y}|\hat{\mathbf{x}}, \mathbf{w}, \mathbf{\Gamma}, \mathbf{A}, \mathbf{m}, c).$$

Формула нижней оценки получается подстановкой.

Пусть Q — валидационная функция:

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

$$+ c_3 \sum_{p_k \in \mathbf{p}} D_{KL}(q_\gamma||p_k),$$

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

Сформулируем задачу поиска оптимальной модели как двухуровневую задачу.

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

где T — оператор оптимизации, решающий задачу оптимизации:

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

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

Доказательство. При соблюдении условий теоремы неравенство вариационной оценки превращается в равенство.

Теорема. Пусть $c_1 = 1, c_3 = 0$. При $c_2 \to \infty$ параметрическая сложность $C_w \to 0$. **Доказательство (идея).** При $C_w \to \infty$ априорное распределение $p(\mathbf{w})$ будет стремиться к дельта-функции. Штраф за D_KL будет расти, параметры будут стремиться к нулю.

Теорема. Пусть $c_1 = 1, c_3 = 0, c_2 > 0, c_2' > c_2$ и логарифм распределения $\mathsf{E}_{q_w, q_\gamma} \log p(\mathbf{y}|\mathbf{X}, \mathbf{w}, \mathbf{\Gamma}, \mathbf{A}, \mathbf{m}$ является выпуклым. Пусть $q(\mathbf{w}), q(\mathbf{w})'$ — распределения, полученные в результате оптимизаций с различными коэффициентами c_2, c_2' соответственно. Тогда $C_w(q) \geq C_w(q')$.

Доказательство (не окончено). Заметим, что q и q' можно выразить следующим образом:

$$q = \arg \max_{\hat{q}: L(\hat{q}, \mathbf{A}, \mathbf{\Gamma}} Q(\mathbf{A}, \mathbf{\Gamma}, c_2),$$

$$q' = \arg \max_{\hat{q}: L(\hat{q}, \mathbf{A}, \mathbf{\Gamma}} Q(\mathbf{A}, \mathbf{\Gamma}, c_2').$$

Функция Q является выпуклой как сумма выпуклых функций. Отсюда справедливы следующие неравенства (по единственности точек экстремума):

$$\mathsf{E}_q \log p(\mathbf{y}|\mathbf{X}, \mathbf{w}, \mathbf{\Gamma}, \mathbf{A}, \mathbf{m}, c) - c_2 D_{\mathrm{KL}}(q||p) - \mathsf{E}_{q'} \log p(\mathbf{y}|\mathbf{X}, \mathbf{w}, \mathbf{\Gamma}, \mathbf{A}, \mathbf{m}, c) + c_2 D_{\mathrm{KL}}(q'||p) > 0,$$

$$\mathsf{E}_{q'} \log p(\mathbf{y}|\mathbf{X}, \mathbf{w}, \mathbf{\Gamma}, \mathbf{A}, \mathbf{m}, c) - c_2' D_{\mathrm{KL}}(q'||p) - \mathsf{E}_q \log p(\mathbf{y}|\mathbf{X}, \mathbf{w}, \mathbf{\Gamma}, \mathbf{A}, \mathbf{m}, c) + c_2 D_{\mathrm{KL}}(q||p) > 0.$$

Вычитая неравенства получим:

$$D_{\mathrm{KL}}(q||p) \geq D_{\mathrm{KL}}(q'||p'),$$

$$\mathsf{E}_{q'} \log p(\mathbf{y}|\mathbf{X}, \mathbf{w}, \mathbf{\Gamma}, \mathbf{A}, \mathbf{m}, c) \leq \mathsf{E}_q \log p(\mathbf{y}|\mathbf{X}, \mathbf{w}, \mathbf{\Gamma}, \mathbf{A}, \mathbf{m}, c).$$

Теорема Пусть при любых выпуклых комбинациях функций **O** модель является выпуклой. Пусть $c_1, c_2 \geq 0, c_3 \leq 0$. Пусть в качестве оценки правдоподобия выборки используется следующая:

$$\mathsf{E}_{q_w, q_\gamma} \log p(\mathbf{y}|\mathbf{X}, \mathbf{w}, \mathbf{\Gamma}, \mathbf{A}, \mathbf{m}, c) \sim \frac{1}{N} \sum_{k=1}^{N} \log p(\mathbf{y}|\mathbf{X}, \hat{\mathbf{w}}_i, \hat{\mathbf{\Gamma}}_i, \mathbf{A}, \mathbf{m}, c),$$

где $\hat{\mathbf{w}}_i, \hat{\mathbf{\Gamma}}_i$ зафиксированные получены под действием вариационной репараметризации с зафиксированной реализацией исходных случайных величин. Тогда L и Q являются выпуклыми функциями. **Доказательство** По сумме выпуклых функций.

Теорема (предварительно) При $c \to 0$ $C_\gamma \to 0$.

Теорема (предварительно) При $c \to \infty$ $C_\gamma \to \max$.

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 prior 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.

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. $f_4 = \mathbf{w}_4\mathbf{X}$ (модель — регрессия по всем признакам).

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 $c_1 = c_2 = 1, c >> 0$ (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

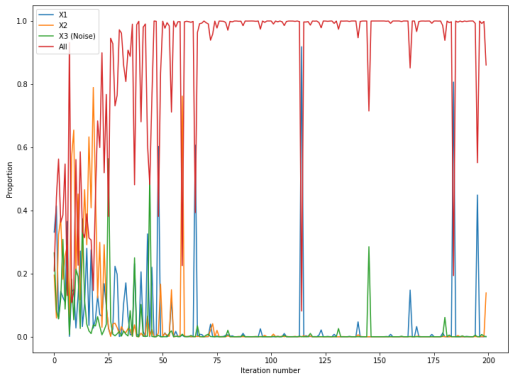

Figure 1: Evidence с низкой температурой

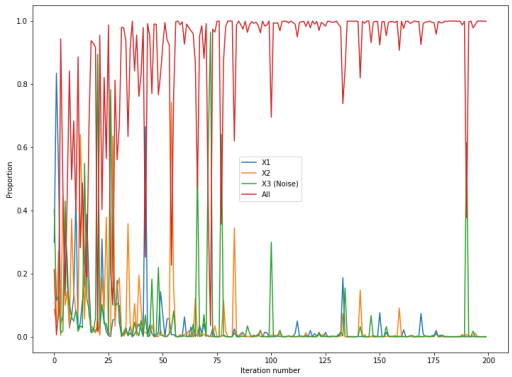

Figure 2: Evidence с высокой температурой

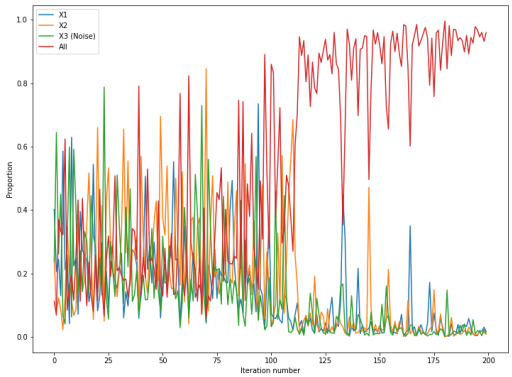

Figure 3: Evidence с высокой температурой, $\beta = 0.01$

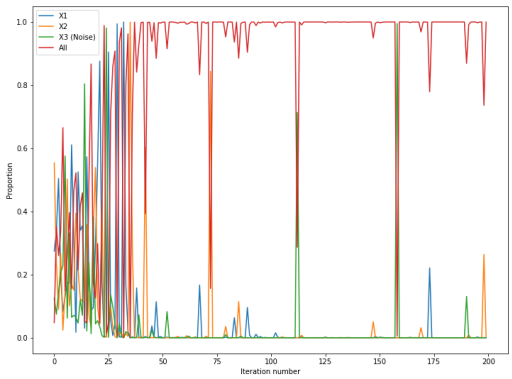

Figure 4: Поощрение выбора шумовой компоненты

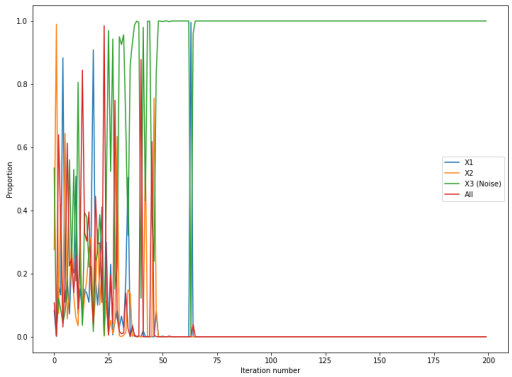

Figure 5: Поощрение выбора шумовой компоненты, $\beta = 0.01$