Байесовский выбор субоптимальной структуры модели глубокого обучения

О. Ю. Бахтеев

Диссертация на соискание ученой степени
кандидата физико-математических наук
05.13.17 — Теоретические основы информатики
Научный руководитель: д.ф.-м.н. В.В. Стрижов

Московский физико-технический институт
5 июня 2019 г.

Выбор структуры модели глубокого обучения

Цель: предложить метод выбора структуры модели глубокого обучения.

Задачи

1. Предложить критерии оптимальной и субоптимальной сложности модели глубокого обучения.
2. Предложить алгоритм построения модели субоптимальной сложности и оптимизации параметров.

Исследуемые проблемы

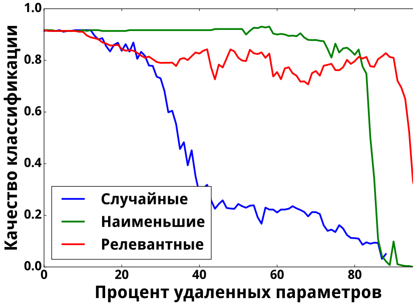
1. Большое число параметров и гиперпараметров модели, высокая вычислительная сложность оптимизации.
2. Многоэкстремальность и невыпуклость задачи оптимизации.

Методы исследования

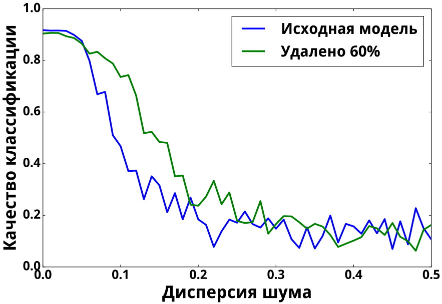
Рассматриваются графовое представление нейронной сети. Используются методы вариационного байесовского вывода. Для получения модели субоптимальной сложности используется метод автоматического определения релевантности параметров с использоваением градиентных методов оптимизации гиперпараметров и структурных параметров модели.

Проблема выбора оптимальной структуры

Правдоподобие моделей с избыточным числом параметров значимо не меняется при их удалении.



Избыточность параметров модели



Устойчивость модели

Глубокое обучение предполагает оптимизацию моделей с заведомо избыточной сложностью.

Модель глубокого обучения

Определение

Моделью $\mathbf{f}(\mathbf{w}, \mathbf{x})$ назовем дифференцируемую по параметрам $\mathbf{w}$ функцию из множества признаковых описаний объекта во множество меток:

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

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

Особенность задачи выбора модели *глубокого обучения* — значительное число параметров моделей приводит к неприменимости ряда методов оптимизации и выбора структуры модели (AIC, BIC, кросс-валидация).
Модель определяется параметрами $\mathbf{W}$ и структурой $\mathbf{\Gamma}$.
Структура задает набор суперпозиций, входящих в модель и выбирается согласно статистическим критериям сложности модели.
Эмпирические оценки статистической сложности модели:

1. число параметров;
2. число суперпозиций, из которых состоит модель.

Выбор структуры: двуслойная нейросеть

Модель $\mathbf{f}$ задана **структурой** $\mathbf{\Gamma} = [\gamma^{0,1}, \gamma^{1,2}]$.

$$\text{Модель: } \mathbf{f}(\mathbf{x}) = \mathbf{softmax}\left(\mathbf{f}_1(\mathbf{x})\mathbf{W}_0^{1,2}\right), \quad \mathbf{f}(\mathbf{x}) : \mathbb{R}^n \to [0,1]^{|\mathbb{Y}|}, \quad \mathbf{x} \in \mathbb{R}^n.$$

$$\mathbf{f}_1(\mathbf{x}) = \gamma_0^{0,1}\mathbf{g}_0^{0,1}(\mathbf{x}) + \gamma_1^{0,1}\mathbf{g}_1^{0,1}(\mathbf{x}),$$

где $\mathbf{w} = [\mathbf{W}_0^{0,1}, \mathbf{W}_1^{0,1}, \mathbf{W}_0^{1,2}]^{\mathsf{T}}$ — матрицы параметров, $\{\mathbf{g}_{0,1}^0, \mathbf{g}_{0,1}^1, \mathbf{g}_{1,2}^0\}$ — обобщенно-линейные функции скрытых слоев нейросети.

$$\begin{tikzcd}
\mathbf{f}_0(\mathbf{x}) = \mathbf{x} \arrow[r, bend left, "\gamma_0^{0,1}\mathbf{g}_0^{0,1}(\mathbf{x}) = \gamma_0^{0,1}\boldsymbol{\sigma}(\mathbf{x}\mathbf{W}_0^{0,1})"] \arrow[r, bend right, "\gamma_1^{0,1}\mathbf{g}_1^{0,1}(\mathbf{x}) = \gamma_1^{0,1}\boldsymbol{\sigma}(\mathbf{x}\mathbf{W}_1^{0,1})"'] & \mathbf{f}_1(\mathbf{x}) \arrow[r, "\gamma_0^{1,2}\mathbf{g}_0^{1,2}(\mathbf{x}) = \gamma_0^{1,2}\mathbf{softmax}(\mathbf{x}\mathbf{W}_0^{1,2})"] & \mathbf{f}_2(\mathbf{x})
\end{tikzcd}$$

Графовое представление модели глубокого обучения

Заданы:

1. ациклический граф (V, E);
2. для каждого ребра $(j, k) \in E$: вектор базовых дифференцируемых функций $\mathbf{g}^{j,k} = [\mathbf{g}^{j,k}_0, \dots, \mathbf{g}^{j,k}_{K^{j,k}}]$ мощности $K^{j,k}$;
3. для каждой вершины $v \in V$: дифференцируемая функция агрегации $\mathbf{agg}_v$.
4. Функция $\mathbf{f} = \mathbf{f}_{|V|-1}$, задаваемая по правилу

$$\mathbf{f}_v(\mathbf{w}, \mathbf{x}) = \mathbf{agg}_v\left(\{\langle \boldsymbol{\gamma}^{j,k}, \mathbf{g}^{j,k}\rangle \circ \mathbf{f}_j(\mathbf{x}) | j \in \mathrm{Adj}(v_k)\}\right), v \in \{1, \dots, |V|-1\}, \quad \mathbf{f}_0(\mathbf{x}) = \mathbf{x} \tag{1}$$

и являющаяся функцией из признакового пространства $\mathbb{X}$ в пространство меток $\mathbb{Y}$ при значениях векторов, $\boldsymbol{\gamma}^{j,k} \in [0,1]^{K^{j,k}}$.

Определение

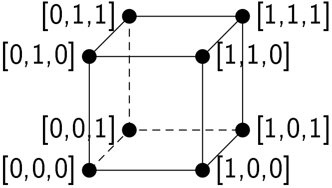
Граф (V, E) со множестом векторов базовых функций $\{\mathbf{g}^{j,k}, (j,k) \in E\}$ и функций агрегаций $\{\mathbf{agg}_v, v \in V\}$ назовем *параметрическим семейством моделей* $\mathfrak{F}$.

Утверждение

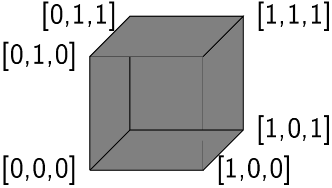
Для любого значения $\boldsymbol{\gamma}^{j,k} \in [0,1]^{K^{j,k}}$ функция $\mathbf{f} \in \mathfrak{F}$ является моделью.

Ограничения на структурные параметры

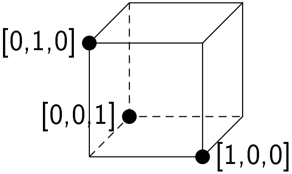
Примеры ограничений для одного структурного параметра $\gamma, |\gamma| = 3$.



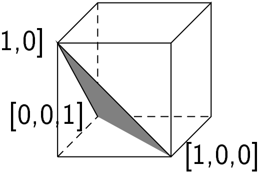
На вершинах куба



Внутри куба

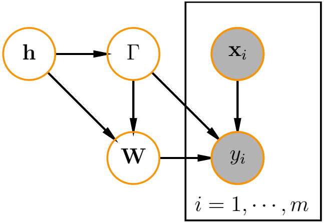


На вершинах симплекса



Внутри симплекса

Априорное распределение параметров



Определение

Априорным распределением параметров $\mathbf{w}$ и структуры $\mathbf{\Gamma}$ модели $\mathbf{f}$ назовем вероятностное распределение $p(\mathbf{W}, \mathbf{\Gamma}|\mathbf{h}) : \mathbb{W} \times \mathbb{\Gamma} \times \mathbb{H} \to \mathbb{R}^+$, где $\mathbb{W}$ — множество значений параметров модели, $\mathbb{\Gamma}$ — множество значений структуры модели.

Определение

Гиперпараметрами $\mathbf{h} \in \mathbb{H}$ модели назовем параметры распределения $p(\mathbf{w}, \mathbf{\Gamma}|\mathbf{h})$ (параметры распределения параметров модели $\mathbf{f}$).

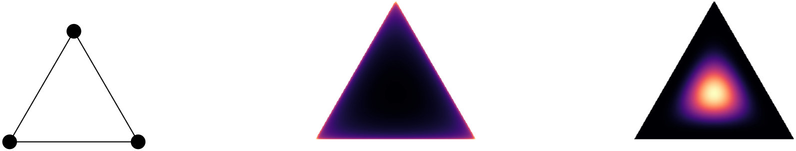
Модель $\mathbf{f}$ задается следующими величинами:

- **Параметры** $\mathbf{w} \in \mathbb{W}$ задают суперпозиции $\mathbf{f}_v$, из которых состоит модель $\mathbf{f}$.
- **Структурные параметры** $\mathbf{\Gamma} = \{\gamma^{j,k}\}_{(j,k)\in E} \in \mathbb{\Gamma}$ задают вклад суперпозиций $\mathbf{f}_v$ в модель $\mathbf{f}$.
- **Гиперпараметры** $\mathbf{h} \in \mathbb{H}$ задают распределение параметров и структурных параметров модели.
- **Метапараметры** $\boldsymbol{\lambda} \in \mathbb{\Lambda}$ задают вид оптимизации модели.

Априорное распределение на структуре модели

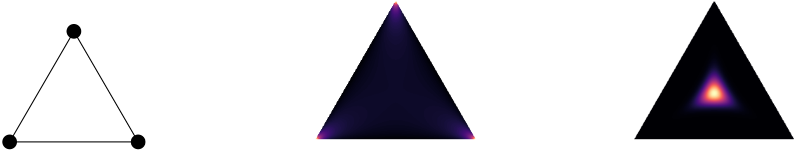
Каждая точка на симплексе задает модель.

Распределение Дирихле: $\boldsymbol{\Gamma} \sim \mathrm{Dir}(\mathbf{s}, \lambda_{\text{temp}})$



$\lambda_{\text{temp}} \to 0$ $\quad$ $\lambda_{\text{temp}} = 0.995$ $\quad$ $\lambda_{\text{temp}} = 5.0$

Распределение Гумбель-софтмакс: $\boldsymbol{\Gamma} \sim \mathrm{GS}(\mathbf{s}, \lambda_{\text{temp}})$

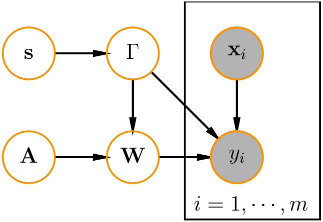


$\lambda_{\text{temp}} \to 0$ $\quad$ $\lambda_{\text{temp}} = 0.995$ $\quad$ $\lambda_{\text{temp}} = 5.0$

Байесовский выбор модели

Базовая модель

- **Параметры** модели: $\mathbf{w} \sim \mathcal{N}(0, \alpha^{-1})$,
- **Гиперпараметры** модели: $\mathbf{h} = [\alpha]$.



Предлагаемая модель

- **Параметры** модели: $\mathbf{w}_r^{j,k} \sim \mathcal{N}(0, \gamma_r^{j,k}(\mathbf{A}_r^{j,k})^{-1})$, $\mathbf{A}_r^{j,k}$ — диагональная матрица параметров, соответствующих базовых функций $\mathbf{g}_r^{j,k}$,
$\mathbf{A}_r^{j,k} \sim \text{inv-gamma}(\lambda_1, \lambda_2)$.
- **Структурные параметры** модели: $\mathbf{\Gamma} = \{\gamma^{j,k}, (j,k) \in E\}$,
$\gamma^{j,k} \sim \text{GS}(\mathbf{s}^{j,k}, \lambda_{\text{temp}})$.
- **Гиперпараметры** модели: $\mathbf{h} = [\text{diag}(\mathbf{A}), \mathbf{s}]$.
- **Метапараметры:** $\lambda_1, \lambda_2, \lambda_{\text{temp}}$.

Обоснованность как статистическая сложность

Статистическая сложность модели $\mathbf{f}$:

$$\mathrm{MDL}(\mathbf{y}, \mathbf{f}) = -\log\, p(\mathbf{h}) - \log\, \big(p(\mathbf{y}|\mathbf{X}, \mathbf{h})\delta\mathfrak{D})\big),$$

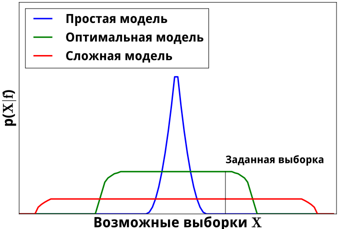
где $\delta\mathfrak{D}$ — допустимая точность передачи информации о выборке $\mathfrak{D}$.

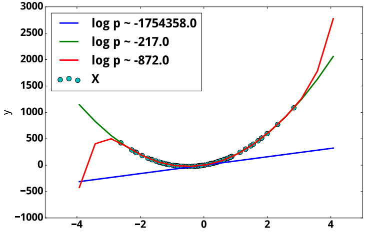
Выбор значений параметров $\mathbf{w}$ производится согласно **апостериорному распределению параметров** L:

$$L = \log p(\mathbf{w}|\mathbf{X}, \mathbf{y}, \mathbf{h}) \propto \log p(\mathbf{y}|\mathbf{X}, \mathbf{w}, \mathbf{h}) + \log p(\mathbf{w}|\mathbf{h}).$$

Выбор значений гиперпараметров производится в согласно **апостериорному распределению гиперпараметров** Q:

$$Q = \log p(\mathbf{h}|\mathbf{X}, \mathbf{y}) \propto \log p(\mathbf{h}) + \log \int\limits_{\mathbf{w}} p(\mathbf{y}|\mathbf{X}, \mathbf{w}) p(\mathbf{w}|\mathbf{h}) d\mathbf{w},$$





Вариационная нижняя оценка обоснованности

Интеграл обоснованности невычислим аналитически.

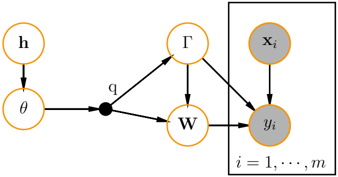
Обоснованность модели:

$$p(\mathbf{y}|\mathbf{X}, \lambda_{\text{temp}}) = \iint\limits_{\mathbf{w},\mathbf{\Gamma}} p(\mathbf{y}|\mathbf{X}, \mathbf{w}, \mathbf{\Gamma}) p(\mathbf{w}, \mathbf{\Gamma}|\lambda_{\text{temp}}) d\mathbf{w} d\mathbf{\Gamma}.$$

Определение

Вариационными параметрами модели $\boldsymbol{\theta} \in \mathbb{R}^u$ назовем параметры распределения q, приближающие апостериорное распределение параметров и структуры $p(\mathbf{w}, \mathbf{\Gamma}|\mathbf{X}, \mathbf{y}, \mathbf{h}, \lambda_{\text{temp}})$:

$$q \approx \frac{p(\mathbf{y}|\mathbf{X}, \mathbf{w}, \mathbf{\Gamma}) p(\mathbf{w}, \mathbf{\Gamma}|\mathbf{h}, \lambda_{\text{temp}})}{\iint\limits_{\mathbf{w}',\mathbf{\Gamma}'} p(\mathbf{y}|\mathbf{X}, \mathbf{w}', \mathbf{\Gamma}') p(\mathbf{w}', \mathbf{\Gamma}'|\mathbf{h}, \lambda_{\text{temp}}) d\mathbf{w}' d\mathbf{\Gamma}'}.$$



Получим нижнюю оценку интеграла правдоподобия.

$$\log p(\mathbf{y}|\mathbf{X}, \lambda_{\text{temp}}) \geq \mathsf{E}_q \log p(\mathbf{y}|\mathbf{X}, \mathbf{w}, \mathbf{\Gamma}) - D_{KL}(q(\mathbf{w}, \mathbf{\Gamma}) || p(\mathbf{w}, \mathbf{\Gamma}, \lambda_{\text{temp}})) = \log \hat{p}(\mathbf{y}|\mathbf{X}, \lambda_{\text{temp}}).$$

Оценка совпадает с интегралом правдоподобия при

$$D_{\text{KL}}(q(\mathbf{w}, \mathbf{\Gamma}) | (p(\mathbf{w}, \mathbf{\Gamma}|\mathbf{y}, \mathbf{X}, \lambda_{\text{temp}})) = 0.$$

Задача выбора модели

Зададим вариационное распределение $q = q_{\mathbf{w}} q_{\mathbf{\Gamma}}$ с параметрами $\boldsymbol{\theta}$, приближающие апостериорное распределение $p(\mathbf{w}, \mathbf{\Gamma} | \mathbf{X}, \mathbf{y}, \mathbf{h})$ параметров и структуры.

Определение

Функцией потерь $L(\boldsymbol{\theta} | \mathbf{h}, \mathbf{X}, \mathbf{y})$ назовем дифференцируемую функцию, качество модели на обучающей выборки при параметрах $\boldsymbol{\theta}$ распределения q.

Функцией валидации $Q(\mathbf{h} | \boldsymbol{\theta}, \mathbf{X}, \mathbf{y})$ назовем дифференцируемую функцию, качество модели при векторе $\boldsymbol{\theta}$, заданном неявно.

Задачей выбора модели $\mathbf{f}$ назовем двухуровневую задачу оптимизации:

$$\mathbf{h}^* = \arg\min_{\mathbf{h} \in \mathbb{H}} Q(\mathbf{h} | \boldsymbol{\theta}^*, \mathbf{X}, \mathbf{y}),$$

где $\boldsymbol{\theta}^*$ — решение задачи оптимизации

$$\boldsymbol{\theta}^* = \arg\min_{\boldsymbol{\theta} \in \mathbb{R}^{u}} L(\boldsymbol{\theta} | \mathbf{h}, \mathbf{X}, \mathbf{y}).$$

Обобщающая задача

Задачу выбора модели $\mathbf{h}^*, \boldsymbol{\theta}^*$ назовем обобщаюшей на множестве $U_\theta \times U_h \times U_\lambda \subset \mathbb{R}^u \times \mathbb{H} \times \mathbb{\Lambda}$, если выполнены условия:

1. Для каждого $\mathbf{h} \in U_h$ и каждого $\boldsymbol{\lambda} \in U_\lambda$ решение $\boldsymbol{\theta}^*$ определено однозначно.
2. *Условие максимизации правдоподобия выборки:* существует $\boldsymbol{\lambda} \in U_\lambda$ и $K_1 \in \mathbb{R}_+$, такие что для любых векторов гиперпараметров $\mathbf{h}_1, \mathbf{h}_2 \in U_h, Q(\mathbf{h}_1) - Q(\mathbf{h}_2) > K_1$: матожидания правдоподбия выборок: $\mathsf{E}_q \log\, p(\mathbf{y}|\mathbf{X}, \boldsymbol{\theta}_1, \lambda_{\text{temp}}) > \log \mathsf{E}_q\; p(\mathbf{y}|\mathbf{X}, \boldsymbol{\theta}_2, \lambda_{\text{temp}})$.
3. *Условие минимизации сложности модели:* существует $\boldsymbol{\lambda} \in U_\lambda$ и $K_2 \in \mathbb{R}_+$, такие что для любых векторов гиперпараметров $\mathbf{h}_1, \mathbf{h}_2 \in U_h, Q(\mathbf{h}_1) - Q(\mathbf{h}_2) > K_2$, $\mathsf{E}_q \log\, p(\mathbf{y}|\boldsymbol{\theta}_1, \lambda_{\text{temp}}) = \log \mathsf{E}_q\; p(\mathbf{y}|\boldsymbol{\theta}_2, \lambda_{\text{temp}})$, количество ненулевых параметров у первой модели меньше, чем у второй.
4. *Условие достижения максимума правдоподобия модели:* существует значение гиперпараметров $\boldsymbol{\lambda}$, такое что оптимизация задачи эквивалента оптимизации вариационной оценки правдоподобия модели:
$$\mathbf{h}^* = \arg\max p(\mathbf{y}|\mathbf{X}, \mathbf{h}', \lambda_{\text{temp}}), \quad \boldsymbol{\theta}^* = \arg\min D_{\text{KL}}(q|p(\mathbf{w}, \boldsymbol{\Gamma}|\mathbf{y}, \mathbf{X}, \lambda_{\text{temp}})).$$
5. *Условие многоэкстремальности:* Существует константа K_3, такая что для любых двух векторов $\mathbf{h}_1, \mathbf{h}_2$ и соответствующих векторов $\boldsymbol{\theta}_1^*, \boldsymbol{\theta}_2^* : D_{\text{KL}}(q_{\boldsymbol{\Gamma}_2}, q_{\boldsymbol{\Gamma}_1}) > K_3, D_{\text{KL}}(q_{\boldsymbol{\Gamma}_1}, q_{\boldsymbol{\Gamma}_2}) > K_3$: существуют значения гиперпараметров $\boldsymbol{\lambda}_1, \boldsymbol{\lambda}_2$, такие что $Q(\mathbf{h}_1, \lambda_1) > Q(\mathbf{h}_2, \lambda_1), Q(\mathbf{h}_1, \lambda_1) < Q(\mathbf{h}_2, \lambda_2)$.
6. Условие непрерывности: $\mathbf{h}^*, \boldsymbol{\theta}^*$ непрерывны по метапараметрам.

Анализ задач выбора моделей

Теорема

Следующие задачи выбора модели не являются обобщающими:

1. метод максимума правдоподобия: $\max_{\boldsymbol{\theta}} \mathsf{E}_q \log p(\mathbf{y}|\mathbf{X}, \boldsymbol{\theta}, \lambda_{\text{temp}})$;
2. метод максимума апостериорной вероятности $\max_{\boldsymbol{\theta}} \mathsf{E}_q \log p(\mathbf{y}|\mathbf{X}, \boldsymbol{\theta}) p(\boldsymbol{\theta}|\mathbf{h}, \lambda_{\text{temp}}) p(\mathbf{h}|\boldsymbol{\lambda})$;
3. метод максимума вариационной оценки правдоподобия модели $\max_{\mathbf{h}} \max_{\boldsymbol{\theta}} \mathsf{E}_q \log\ p(\mathbf{y}|\mathbf{X}, \mathbf{w}, \boldsymbol{\Gamma}) - \mathsf{D}_{KL}(q(\mathbf{w}, \boldsymbol{\Gamma})||p(\mathbf{w}, \boldsymbol{\Gamma}, \lambda_{\text{temp}})) p(\mathbf{h}|\boldsymbol{\lambda})$;
4. кросс-валидация $\max_{\mathbf{h}} \mathsf{E}_q \log p(\mathbf{y}_{\text{valid}}|\mathbf{X}_{\text{valid}}, \boldsymbol{\theta}^*, \lambda_{\text{temp}}) p(\mathbf{h}|\boldsymbol{\lambda})$, $\boldsymbol{\theta}^* = \arg\max_{\boldsymbol{\theta}} \mathsf{E}_q \log p(\mathbf{y}_{\text{train}}|\mathbf{X}_{\text{train}}, \boldsymbol{\theta}, \lambda_{\text{temp}}) p(\boldsymbol{\theta}|\mathbf{h})$.
5. AIC: $\max_{\boldsymbol{\theta}} \mathsf{E}_q \log p(\mathbf{y}|\mathbf{X}, \boldsymbol{\theta}, \lambda_{\text{temp}}) + |\theta_i : \theta_i \neq 0|$;
6. BIC: $\max_{\boldsymbol{\theta}} \mathsf{E}_q \log p(\mathbf{y}|\mathbf{X}, \boldsymbol{\theta}, \lambda_{\text{temp}}) + \log(m)|\theta_i : \theta_i \neq 0|$;
7. перебор структуры модели: $\max \boldsymbol{\Gamma}' \max_{\boldsymbol{\theta}} \mathsf{E}_q \log p(\mathbf{y}|\mathbf{X}, \boldsymbol{\theta}, \lambda_{\text{temp}}) \mathbb{I}(\boldsymbol{\Gamma} = \boldsymbol{\Gamma}')$.

Предлагаемая задача оптимизации

Теорема

Пусть функции потерь и валидации L, Q являются непрерывно-дифференцируемыми на некоторой области U. Тогда следующая задача является обобщающей на U.

$$\mathbf{h}^* = \arg\max_{\mathbf{h}} Q = \tag{Q^*}$$

$$= \lambda_{\text{train}} \mathsf{E}_{q^*} \log p(\mathbf{y}|\mathbf{X}, \mathbf{w}, \mathbf{\Gamma}, \mathbf{h}, \lambda_{\text{prior}}) -$$

$$-\lambda_{\text{prior}} \mathsf{D}_{KL}(p(\mathbf{w}, \mathbf{\Gamma}|\mathbf{h}, \lambda_{\text{temp}})||q^*(\mathbf{w}, \mathbf{\Gamma})) -$$

$$- \sum_{p' \in \mathbf{P}, \lambda \in \boldsymbol{\lambda}_{\text{comb}}} \lambda \mathsf{D}_{KL}(\mathbf{\Gamma}|p') + \log p(\mathbf{h}|\lambda_1, \lambda_2),$$

где

$$q^* = \arg\max_{q} L = \mathsf{E}_q \log p(\mathbf{y}|\mathbf{X}, \mathbf{w}, \mathbf{\Gamma}, \mathbf{A}^{-1}, \lambda_{\text{temp}}) \tag{L^*}$$

$$-\lambda_{\text{reg}} \mathsf{D}_{KL}(p(\mathbf{w}, \mathbf{\Gamma}|\mathbf{A}^{-1}, \mathbf{m}, \lambda_{\text{temp}})||q(\mathbf{w}), q(\mathbf{\Gamma})).$$

$\lambda_{\text{train}}, \lambda_{\text{prior}}, \lambda_{\text{temp}}, \boldsymbol{\lambda}_{\text{comb}}$ и параметры распределений $\mathbf{P}$ — метапараметры оптимизации.

Оптимизационная задача обобщает алгоритмы оптимизации: оптимизация правдоподобия, последовательное увеличение и снижение сложности модели, полный перебор структуры.

$\boldsymbol{\lambda}_{\text{comb},1} = [0; 0; 0]$.

$\boldsymbol{\lambda}_{\text{comb},1} = [1; 0; 0]$.

$\boldsymbol{\lambda}_{\text{comb},1} = [1; 1; 0]$.

Адекватность задачи оптимизации

Теорема

Пусть задано параметрическое множество вариационных распределений: $q(\boldsymbol{\theta})$.
Пусть $\lambda_{\text{train}} = \lambda_{\text{prior}} = \lambda_{\text{reg}} > 1, \lambda_{\text{comb}} = 0$. Тогда:

1. Задача оптимизации (Q^*) доставляет максимум апостериорной вероятности гиперпараметров с использованием вариационной оценки правдоподобия:
$$\log \hat{p}(\mathbf{y}|\mathbf{X}, \mathbf{h}, \lambda_{\text{temp}}) + \log p(\mathbf{h}|\lambda_1, \lambda_2) \to \max_{\mathbf{h}}.$$

2. Вариационное распределение q приближает апостериорное распределение $p(\mathbf{w}, \boldsymbol{\Gamma}|\mathbf{y}, \mathbf{X}, \mathbf{h}, \lambda_{\text{temp}})$ наилучшим образом:
$$D_{\text{KL}}(q||p(\mathbf{w}, \boldsymbol{\Gamma}|\mathbf{y}, \mathbf{X}, \mathbf{h}, \lambda_{\text{temp}})) \to \min_{\boldsymbol{\theta}}.$$

Пусть также распределение q декомпозируется на два независимых распределения для параметров $\mathbf{w}$ и структуры $\boldsymbol{\Gamma}$ модели $\mathbf{f}$:

$$q = q_{\mathbf{w}} q_{\boldsymbol{\Gamma}}, q_{\boldsymbol{\Gamma}} \approx p(\boldsymbol{\Gamma}|\mathbf{y}, \mathbf{X}, \mathbf{h}), q_{\mathbf{w}} \approx p(\mathbf{w}|\boldsymbol{\Gamma}, \mathbf{y}, \mathbf{X}, \mathbf{h}).$$

Тогда вариационные распределения $q_{\mathbf{w}}, q_{\boldsymbol{\Gamma}}$приближают апостериорные распределения $p(\boldsymbol{\Gamma}|\mathbf{y}, \mathbf{X}, \mathbf{h}, \lambda_{\text{temp}}), p(\mathbf{w}|\boldsymbol{\Gamma}, \mathbf{y}, \mathbf{X}, \mathbf{h}, \lambda_{\text{temp}})$ наилучшим образом:

$$D_{\text{KL}}(q_{\boldsymbol{\Gamma}}||p(\boldsymbol{\Gamma}|\mathbf{y}, \mathbf{X}, \mathbf{h}, \lambda_{\text{temp}})) \to \min, \quad D_{\text{KL}}(q_{\mathbf{w}}||p(\mathbf{w}|\mathbf{y}, \mathbf{X}, \mathbf{h})) \to \min.$$

Оператор оптимизации

Определение

Назовем *оператором оптимизации* T выбор вектора параметров $\boldsymbol{\theta}'$ по параметрам предыдущего шага $\boldsymbol{\theta}$.

Оператор стохастического градиентного спуска:

$$\hat{\boldsymbol{\theta}} = T \circ T \circ \dots \circ T(\boldsymbol{\theta}_0, \mathbf{A}^{-1}, \mathbf{m}) = T^\eta(\boldsymbol{\theta}_0, \mathbf{A}^{-1}, \mathbf{m}), \quad \text{где} T(\boldsymbol{\theta}, \mathbf{A}^{-1}, \mathbf{m}) =$$
$$= \boldsymbol{\theta} - \lambda_{\mathsf{lr}} \nabla L(\boldsymbol{\theta}, \mathbf{A}^{-1}, \mathbf{m})|_{\hat{\mathfrak{D}}},$$

λ_{lr} — длина шага градиентного спуска, $\boldsymbol{\theta}_0$ — начальное значение параметров $\boldsymbol{\theta}$, $\hat{\mathfrak{D}}$ — случайная подвыборка исходной выборки $\mathfrak{D}$.
Перепишем итоговую задачу оптимизации:

$$\mathbf{h}' = T^\eta\big(Q, \mathbf{h}, T^\eta(L, \boldsymbol{\theta}_0, \mathbf{h})\big),$$

где $\boldsymbol{\theta}_0$ — начальное значение $\boldsymbol{\theta}$.

Теорема

Пусть Q, L — локально выпуклы и непрерывны в некоторой области $U_W \times U_\Gamma \times U_H \times U_B \subset \mathbb{W} \times \mathbb{\Gamma} \times \mathbb{H} \times \mathbb{B}$, при этом $U_H \times U_B$ — компакт. Тогда решение задачи градиентной оптимизации стремится к локальному минимуму $\mathbf{h}^* \in U$ исходной задачи оптимизации $\big(Q^*\big)$ при $\eta \to \infty$, $\mathbf{h}^*$ является непрерывной функцией по метапараметрам модели.

Нижняя вариационная оценка правдоподобия на основе мультистарта

$$\log p(\mathbf{y}|\mathbf{X},\mathbf{h}) \geq \mathsf{E}_{q(\mathbf{W})}\log\ p(\mathbf{y},\mathbf{w}|\mathbf{X},\mathbf{h}) - \mathsf{E}_{q_{\mathbf{w}}}(-\log(q_{\mathbf{w}})).$$

Теорема [Бахтеев, 2016]

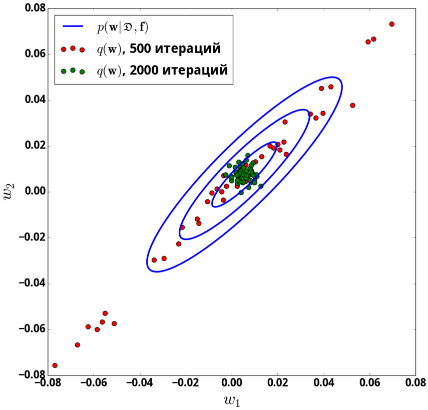
Пусть L — функция потерь, градиент которой — непрерывно-дифференцируемая функция с константой Липшица C.
Пусть $\boldsymbol{\theta} = [\mathbf{w}^1, \dots, \mathbf{w}^k]$ — начальные приближения оптимизации модели, λ — шаг градиентного спуска.
Тогда разность энтропий на смежных шагах оптимизации приближается следующим образом:

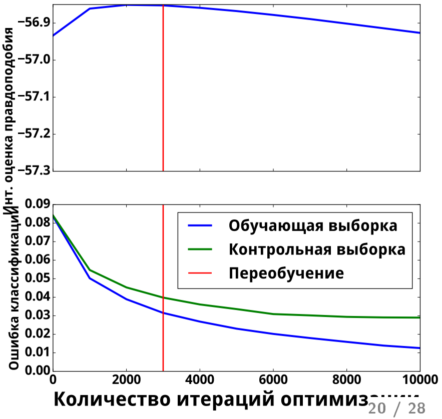
$$\mathsf{E}_{q_{\mathbf{w}}^{\tau}}(-\log(q_{\mathbf{w}}^{\tau})) - \mathsf{E}_{q_{\mathbf{w}}^{\tau-1}}(-\log(q_{\mathbf{w}}^{\tau-1})) \approx \frac{1}{k}\sum_{r=1}^{k}\left(\lambda Tr[\mathbf{H}(\mathbf{w}^r)] - \lambda^2 Tr[\mathbf{H}(\mathbf{w}^r)\mathbf{H}(\mathbf{w}^r)]\right),$$

где $\mathbf{H}$ — гессиан функции потерь L, $q_{\mathbf{w}}^{\tau}$ — распределение $q_{\mathbf{w}}^{\tau}$ в момент оптимизации τ.

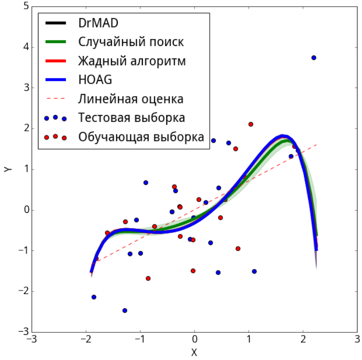
Градиентный спуск как вариационная оценка правдоподобия модели

Эмпирическая плотность, основанная на точках старта оптимизации — вариационное распределение.
Снижение вариационной оценки правдоподобия — начало переобучения.

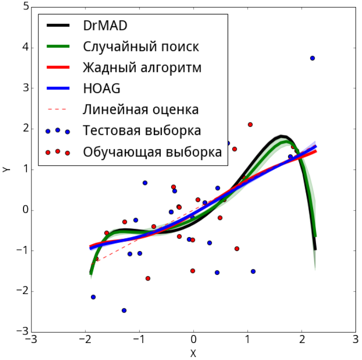




Оптимизация гиперпараметров: пример



Кросс-Валидация



Вариационная оценка

Анализ обобщающей задачи оптимизации

Теорема.

Пусть $\lambda_{\text{reg}} > 0, m \gg 0, \frac{m}{\lambda_{\text{reg}}} \in \mathbb{N}$.
Тогда оптимизация функции

$$L = \mathsf{E}_q \log p(\mathbf{y}|\mathbf{X}, \mathbf{w}, \mathbf{\Gamma}, \mathbf{A}^{-1}, \lambda_{\text{temp}}) -$$

$$-\lambda_{\text{reg}} \mathsf{D}_{KL}(p(\mathbf{w}, \mathbf{\Gamma}|\mathbf{A}^{-1}, \mathbf{m}, \lambda_{\text{temp}})||q(\mathbf{w}), q(\mathbf{\Gamma}))$$

эквивалентна минимизации ожидаемой дивергенции
$\mathsf{E}_{\hat{\mathbf{X}}, \hat{\mathbf{y}} \sim p(\mathbf{X}, \mathbf{y})} \mathsf{D}_{KL}(q||p(\mathbf{w}, \mathbf{\Gamma}|\hat{\mathbf{X}}, \hat{\mathbf{y}}))$, где $\hat{\mathbf{X}}, \hat{\mathbf{y}}$ — случайные подвыборки мощностью $\frac{m}{\lambda_{\text{reg}}}$ из генеральной совокупности.

Параметрическая сложность

Определение

Параметрической сложностью модели назовем минимальную дивергенцию между априорным и вариационным распределением:

$$C_p = \min_{\mathbf{h}} D_{\mathrm{KL}}(q||p(\mathbf{w}, \boldsymbol{\Gamma}|\lambda_{\mathrm{temp}})).$$

Вариационное удаление параметров модели

Будем удалять параметры с наибольшей относительной плотностью:

$$\rho(w) = \frac{q(0)}{q(w)} = \exp\left(-\frac{\mu^2}{\sigma^2}\right).$$

Теорема

При устремлении параметрической сложности модели к нулю относительная плотность параметров модели стремится к единице:

$$C_p \to 0 => \boldsymbol{\rho}(\mathbf{w}) \to 1.$$

Оптимизация параметрической сложности

Теорема

Пусть $\lambda_{\text{train}} = \lambda_{\text{reg}} = 1, \lambda_{\text{comb}} = 0$. Тогда предел оптимизации

$$\lim_{\lambda_{\text{prior}} \to \infty} \lim_{\eta \to \infty} T^{\eta}\big(Q, \mathbf{h}, T^{\eta}(L, \boldsymbol{\theta}_0, \mathbf{h})\big)$$

доставляет минимум параметрической сложности. Существует компактная область $\hat{U} \subset U(\mathbf{0})$, такая что для любой точки $\boldsymbol{\theta}_0 \in \hat{U}$ предел данной оптимизации доставляет нулевую параметрическую сложность: $C_p = 0$.

Теорема

Пусть $\lambda_{\text{train}} = 1, \lambda_{\text{comb}} = 0$. Пусть $\mathbf{f}_1, \mathbf{f}_2$ — результаты градиентной оптимизации при разных значениях гиперпараметров $\lambda^1_{\text{prior}}, \lambda^2_{\text{prior}}, \lambda^1_{\text{prior}} < \lambda^2_{\text{prior}}$, полученных при начальном значении вариационных параметров $\boldsymbol{\theta}_0$ и гиперпараметров $\mathbf{h}_0$. Пусть $\boldsymbol{\theta}_0, \mathbf{h}_0$ принадлежат области U, в которой соответствующие функции L и Q являются локально-выпуклыми. Тогда:

$$C_p(\mathbf{f}_1) - C_p(\mathbf{f}_2) \geq \lambda_{\text{reg}}(\lambda_{\text{reg}} - \lambda^1_{\text{prior}}) \text{sup}_{\boldsymbol{\theta}, \mathbf{h} \in U} |\nabla^2_{\boldsymbol{\theta}, \mathbf{h}} D_{KL}(q|p)(\nabla^2_{\boldsymbol{\theta}} L)^{-1} \nabla_{\boldsymbol{\theta}} D_{KL}(q|p))|.$$

Структурная сложность

Определение

Структурной сложностью C_s модели назовем энтропию структур $\mathbf{\Gamma}$, полученных из вариационного распределения q:

$$C_s = -\mathsf{E}_q \mathsf{E}_{\Gamma} \log p_{\mathbf{\Gamma}}.$$

Теорема

Пусть задано априорное распределение на структуре:

$$p(\gamma_{i,j}) = \lim_{\lambda_{\text{temp}} \to 0} \mathcal{GS}(\lambda_{\text{temp}}).$$

Пусть $\lambda_{\text{reg}} > 0, \lambda_{\text{train}} > 0, \lambda_{\text{prior}} > 0, \lambda_{\text{comb}} = 0$, $\mathbf{f}$ — глобальный оптимум задачи оптимизации. Тогда $C_s(\mathbf{f}) = 0$.

Пусть

$$p(\gamma_{i,j}) = \lim_{\lambda_{\text{temp}} \to \infty} \mathcal{GS}(\lambda_{\text{temp}}).$$

Тогда структурная сложность глобального оптимума $\mathbf{f}$ равняется максимуму:

$$C_s(\mathbf{f}) = \mathsf{E} \log \mathcal{U}.$$

Оптимизация структурной сложности

Теорема

Пусть $\lambda_{\text{train}} > 0$, $\boldsymbol{\theta}_1, \boldsymbol{\theta}_2$ — вариационные параметры, такие что $\boldsymbol{\theta}_1$ лежит внутри произведения симплексов структуры, $\boldsymbol{\theta}_2$ — на вершинах симплексов. Тогда

$$\lim_{\lambda_{\text{temp}} \to 0} \frac{L(\boldsymbol{\theta}_2)}{L(\boldsymbol{\theta}_1)} \to 0.$$

Теорема

Пусть $\lambda_{\text{train}} > 0$, $\boldsymbol{\theta}_1, \boldsymbol{\theta}_2$ — вариационные параметры, такие что $\boldsymbol{\theta}_1$ лежит внутри произведения симплексов структуры, $\boldsymbol{\theta}_2$ — в центре симплексов. Тогда

$$\lim_{\lambda_{\text{temp}} \to \infty} \frac{L(\boldsymbol{\theta}_2)}{L(\boldsymbol{\theta}_1)} \to 0.$$

Результаты, выносимые на защиту

1. Предложен метод выбора модели наиболее правдоподобной структуры, обобщающий ранее описанные алгоритмы оптимизации:
   - оптимизация правдоподобия;
   - последовательное увеличение сложности модели;
   - последовательное снижение сложности модели;
   - полный перебор вариантов структуры модели.
2. Предложен алгоритм оптимизации параметров, гиперпараметров и структурных параметров моделей глубокого обучения.
3. Проведено исследование свойств алгоритмов выбора модели при различных значениях мета-параметров.
4. Проведен вычислительный эксперимент, иллюстрирующий работу предложенного метода.

Список работ автора по теме диссертации

Публикации ВАК

1. Бахтеев О.Ю., Попова М.С., Стрижов В.В. Системы и средства глубокого обучения в задачах классификации. // Системы и средства информатики. 2016. № 26.2. С. 4-22.
2. Бахтеев О.Ю., Стрижов В.В. Выбор моделей глубокого обучения субоптимальной сложности. // Автоматика и телемеханика. 2018. №8. С. 129-147.
3. Огальцов А.В., Бахтеев О.Ю. Автоматическое извлечение метаданных из научных PDF-документов. // Информатика и её применения. 2018.
4. Смердов А.Н., Бахтеев О.Ю., Стрижов В.В. Выбор оптимальной модели рекуррентной сети в задачах поиска парафраза. // Информатика и ее применения. 2019.
5. Грабовой А.В., Бахтеев О.Ю., Стрижов В.В. Определение релевантности параметров нейросети. // Информатика и её применения. 2019.

Выступления с докладом

1. “Восстановление панельной матрицы и ранжирующей модели в разнородных шкалах”, Всероссийская конеренция «57-я научная конеренция МФТИ», 2014.
2. “A monolingual approach to detection of text reuse in Russian-English collection”, Международная конференция «Artificial Intelligence and Natural Language Conference», 2015.
3. “Выбор модели глубокого обучения субоптимальной сложности с использованием вариационной оценки правдоподобия”, Международная конференция «Интеллектуализация обработки информации», 2016.
4. “Author Masking using Sequence-to-Sequence Models”, Международная конференция «Conference and Labs of the Evaluation Forum», 2017.
5. “Градиентные методы оптимизации гиперпараметров моделей глубокого обучения”, Всероссийская конференция «Математические методы распознавания образов ММРО», 2017.
6. “Детектирование переводных заимствований в текстах научных статей из журналов, входящих в РИНЦ”, Всероссийская конференция «Математические методы распознавания образов ММРО», 2017.
7. “Байесовский выбор наиболее правдоподобной структуры модели глубокого обучения”, Международная конференция «Интеллектуализация обработки информации», 2018.