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
16 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

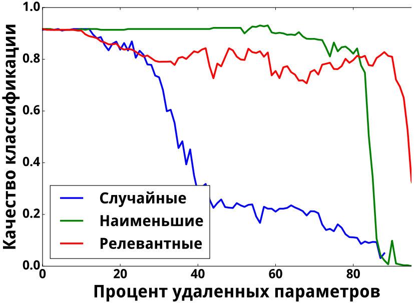
1. Большое число параметров и гиперпараметров модели, высокая вычислительная сложность оптимизации.
2. Многоэкстремальность и невыпуклость задачи оптимизации.

Методы исследования

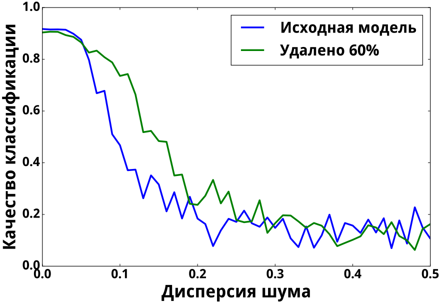
Рассматриваются графовое представление нейронной сети. Используются методы вариационного байесовского вывода. Для получения модели субоптимальной сложности используется метод автоматического определения релевантности параметров с использоваением градиентных методов оптимизации гиперпараметров и структурных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

Проблема выбора оптимальной структуры модели глубокого обучения

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.

Глубокое обучение

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 (AIC, BIC, кросс-валидация) приводит к неприменимости ряда методов оптимизации и выбора модели.

Эмпирические оценки сложности модели:

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

Выбор структуры: двуслойная нейросеть

Структурные параметры: $\mathbf{\Gamma} = [\gamma^{0,1}, \gamma^{1,2}]$.

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

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

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.

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

Графовое представление модели глубокого обучения

Определение

Пусть

1. задан ациклический граф (V, E);
2. для каждого ребра $(j, k) \in E$ определен вектор базовых липшецевых функций $\mathbf{g}^{j,k} = [\mathbf{g}_0^{j,k}, \dots, \mathbf{g}_{K^{j,k}}^{j,k}]$ мощности $K^{j,k}$;
3. для каждой вершины $v \in V$ определена липшецевая функция агрегации $\mathbf{agg}_v$.

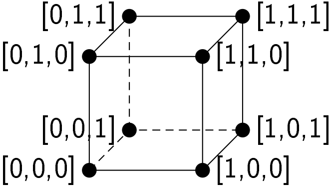
Граф (V, E) со множестом векторов базовых функций $\{\mathbf{g}^{j,k}, (j, k) \in E\}$ и функций агрегаций $\{\mathbf{agg}_v, v \in V\}$ задает *параметрическое семейство моделей* $\mathfrak{F}$ тогда и только тогда, когда функция $\mathbf{f} = \mathbf{f}_{|V|-1}$, задаваемая по правилу

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

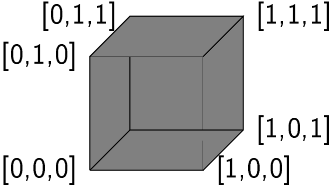
является дифференцируемой по параметрам $\mathbf{w}$ функцией из признакового пространства $\mathbb{X}$ в пространство меток $\mathbb{Y}$ при значениях векторов, $\gamma^{j,k} \in [0, 1]^{K^{j,k}}$.

Ограничения на структурные параметры

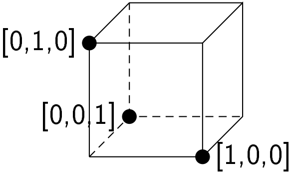
Примеры ограничений для одного структурного параметра $\gamma, |\gamma| = 3$.



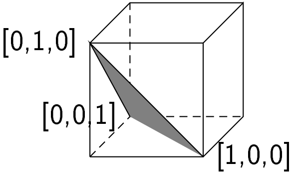
На вершинах куба



Внутри куба



На вершинах симплекса

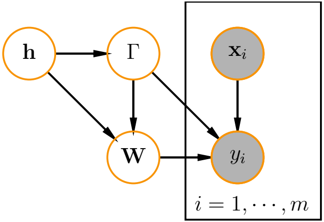


Внутри симплекса

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

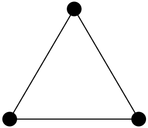
Гиперпараметрами $\mathbf{h} \in \mathbb{H}$ модели назовем параметры распределения $p(\mathbf{w}, \mathbf{\Gamma}|\mathbf{h})$ (параметры распределения параметров модели $\mathbf{f}$).

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

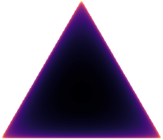
- **Параметры** $\mathbf{W}$ задают суперпозиции $\mathbf{f}_v$, из которых состоит модель $\mathbf{f}$.
- **Структурные параметры** $\mathbf{\Gamma}$ задают вклад суперпозиций $\mathbf{f}_v$ в модель $\mathbf{f}$.
- **Гиперпараметры** $\mathbf{h}$ задают распределение параметров и структурных параметров модели.
- **Метапараметры** $\boldsymbol{\beta}$ задают вид оптимизации модели.

Априорное распределение на структуре модели

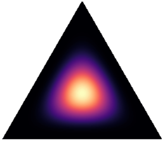
Распределение Дирихле: $\boldsymbol{\Gamma} \sim \mathrm{Dir}(\mathbf{s}, c_{\mathrm{temp}})$



$c_{\mathrm{temp}} \to 0$

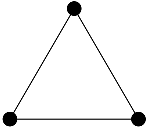


$c_{\mathrm{temp}} = 0.995$

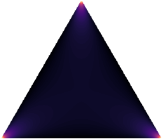


$c_{\mathrm{temp}} = 5.0$

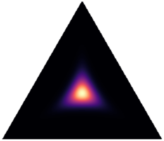
Распределение Гумбель-софтмакс: $\boldsymbol{\Gamma} \sim \mathrm{GS}(\mathbf{s}, c_{\mathrm{temp}})$



$c_{\mathrm{temp}} \to 0$



$c_{\mathrm{temp}} = 0.995$

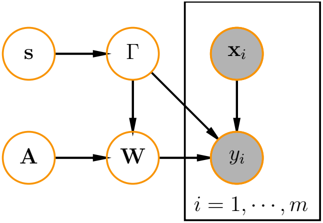


$c_{\mathrm{temp}} = 5.0$

Байесовский выбор модели

Bishop, 1998:

- **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 подмодели $\mathbf{g}_r^{j,k}$, $\mathbf{A}_r^{j,k} \sim \text{inv-gamma}(c_1, c_2)$.
- **Структурные параметры** модели: $\boldsymbol{\Gamma} = \{\gamma^{j,k}, (j,k) \in E\}$, $\gamma^{j,k} \sim \text{GS}(\mathbf{s}^{j,k}, c_{\text{temp}})$.
- **Гиперпараметры** модели: $\mathbf{h} = \{\mathbf{A}, \mathbf{s}\}$.
- **Метапараметры:** $c_1, c_2, c_{\text{temp}}$.

Правдоподобие как статистическая сложность

Статистическая сложность модели $\mathbf{f}$:

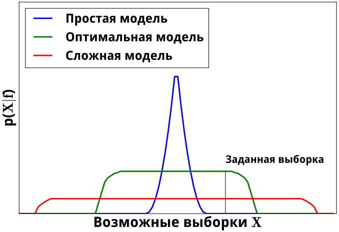
$$\text{MDL}(\mathbf{y}, \mathbf{f}) = -\log\, p(\mathbf{h}) - \log\, \left(p(\mathbf{y}|\mathbf{X}, \mathbf{h})\delta\mathfrak{D}\right),$$

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

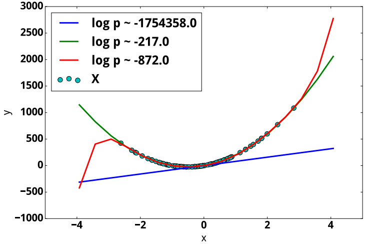
Правдоподобие модели:

$$Q(\mathbf{h}|\boldsymbol{\theta}^*, \mathbf{X}, \mathbf{y}) = \log p(\mathbf{h}|\mathbf{X}, \mathbf{y}) = \log p(\mathbf{h}) + \log \int_{\mathbf{W},\boldsymbol{\Gamma}} p(\mathbf{y}|\mathbf{X}, \mathbf{W}, \boldsymbol{\Gamma})p(\mathbf{W}, \boldsymbol{\Gamma}|\mathbf{h})d\mathbf{W}d\boldsymbol{\Gamma},$$

$$L(\boldsymbol{\theta}|\mathbf{h}, \mathbf{X}, \mathbf{y}) = \log p(\mathbf{W}, \boldsymbol{\Gamma}|\mathbf{X}, \mathbf{y}, \mathbf{h}) \propto \log p(\mathbf{y}|\mathbf{X}, \mathbf{W}, \boldsymbol{\Gamma}, \mathbf{h}) + \log p(\mathbf{W}, \boldsymbol{\Gamma}|\mathbf{h}).$$



Выбор модели по правдоподобию



Аппроксимация выборки полиномами

Вариационная нижняя оценка правдоподобия

Интеграл правдоподобия невычислим аналитически.

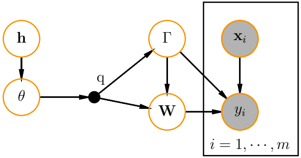
Правдоподобие модели:

$$p(\mathbf{y}|\mathbf{X}, c_{\text{temp}}) = \int_{\mathbf{W},\boldsymbol{\Gamma},c_{\text{temp}}} p(\mathbf{y}|\mathbf{X}, \mathbf{W}, \boldsymbol{\Gamma})p(\mathbf{W}, \boldsymbol{\Gamma}|c_{\text{temp}})d\mathbf{W}d\boldsymbol{\Gamma}.$$

Определение

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

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



Получим нижнюю оценку интеграла правдоподобия.

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

Полученная оценка совпадает с интегралом правдоподобия при

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

Задача выбора модели

Зададим вариационное распределение q 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 решение двухуровневой задачи 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

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

Выбор оптимальной модели

Требуется

Предложить метод поиска субоптимального решения задачи оптимизации, обобщающего различные алгоритмы оптимизации:

- Оптимизация правдоподобия.
- Последовательное увеличение и снижение сложности модели.
- Полный перебор вариантов структуры модели.

Общая задача оптимизации

Теорема (будет)

Следующая оптимизационная задача обобщает алгоритмы оптимизации: оптимизация правдоподобия, последовательное увеличение и снижение сложности модели, полный перебор вариантов структуры модели:

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

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

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

$$-c_{\text{comb}} \sum_{p' \in \mathbf{P}} \mathsf{D}_{KL}(\boldsymbol{\Gamma}|p') + \log p(\mathbf{h}|c_1, c_2),$$

где

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

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

$c_{\text{train}}, c_{\text{prior}}, c_{\text{temp}}, c_{\text{comb}}$ — метапараметры оптимизации.

Адекватность задачи оптимизации

Теорема

Пусть задано параметрическое множество вариационных распределений: $q \in \mathfrak{Q}$. Пусть $c_{\textbf{train}} = c_{\textbf{prior}} = c_{\textbf{reg}} > 1, c_{\textbf{comb}} = 0$. Тогда:

1. Задача оптимизации (Q^*) доставляет максимум апостериорной вероятности гиперпараметров с использованием вариационной оценки правдоподобия:

$$\log \hat{p}(\mathbf{y}|\mathbf{X}, \mathbf{h}, c_{\textbf{temp}}) p(\mathbf{h}|c_1, c_2) \to \max_{\mathbf{h}}.$$

2. Вариационное распределение q приближает апостериорное распределение $p(\mathbf{W}, \mathbf{\Gamma}|\mathbf{y}, \mathbf{X}, \mathbf{h}, c_{\textbf{temp}})$ наилучим образом среди множества распределений $\mathfrak{Q}$:

$$D_{\mathbf{KL}}(q||p(\mathbf{W}, \mathbf{\Gamma}|\mathbf{y}, \mathbf{X}, \mathbf{h}, c_{\textbf{temp}})) \to \min_{q \in \mathfrak{Q}}.$$

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

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

Тогда вариационные распределения $q_{\mathbf{W}}, q_{\mathbf{\Gamma}}$приближают апостериорные распределения $p(\mathbf{\Gamma}|\mathbf{y}, \mathbf{X}, \mathbf{h}, c_{\textbf{temp}}), p(\mathbf{W}|\mathbf{\Gamma}, \mathbf{y}, \mathbf{X}, \mathbf{h}, c_{\textbf{temp}})$ наилушчим образом:

$$D_{\mathbf{KL}}(q_{\mathbf{\Gamma}}||p(\mathbf{\Gamma}|\mathbf{y}, \mathbf{X}, \mathbf{h}, c_{\textbf{temp}})) \to \min, \quad D_{\mathbf{KL}}(q_{\mathbf{W}}||p(\mathbf{W}|\mathbf{y}, \mathbf{X}, \mathbf{h})) \to \min.$$

Оператор оптимизации

Определение

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

Оператор стохастического градиентного спуска:

$$\hat{\boldsymbol{\theta}} = T \circ T \circ \cdots \circ T(\boldsymbol{\theta}_0, \mathbf{A}^{-1}, \mathbf{m}) = T^{\eta}(\boldsymbol{\theta}_0, \mathbf{A}^{-1}, \mathbf{m}), \quad \text{где} T(\boldsymbol{\theta}, \mathbf{A}^{-1}, \mathbf{m}) =$$

$$= \boldsymbol{\theta} - \beta \nabla L(\boldsymbol{\theta}, \mathbf{A}^{-1}, \mathbf{m})|_{\hat{\mathfrak{D}}},$$

γ — 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 сильно выпуклы в некоторой области U. Тогда решение задачи градиентной оптимизации стремится к локальному минимуму $\mathbf{h}^* \in U$ исходной задачи оптимизации (Q^*) при $\eta \to \infty$.

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]

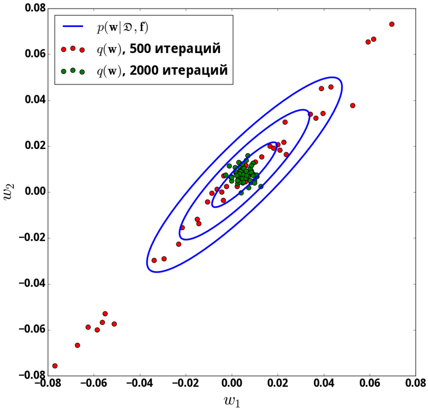
Пусть L — функция потерь, градиент которой — непрерывно-дифференцируемая функция с константой Липшица C.
Пусть $\boldsymbol{\theta} = [\mathbf{W}^1, \dots, \mathbf{W}^k]$ — начальные приближения оптимизации модели, β — шаг градиентного спуска.
Тогда разность энтропий на смежных шагах оптимизации приближается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

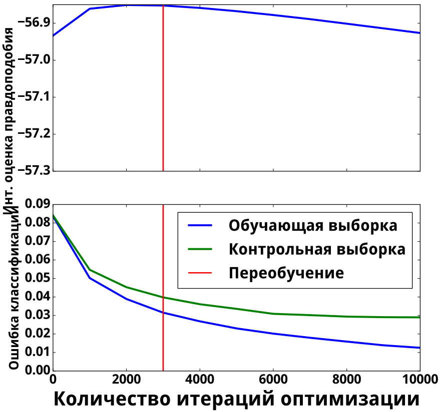
$$\mathsf{E}_{q^\tau_{\mathbf{W}}}(-\log(q^\tau_{\mathbf{W}})) - \mathsf{E}_{q^{\tau-1}_{\mathbf{W}}}(-\log(q^{\tau-1}_{\mathbf{W}})) \approx \frac{1}{k}\sum_{r=1}^{k}\big(\beta\, Tr[\mathbf{H}(\mathbf{W}^r)] - \beta^2\, Tr[\mathbf{H}(\mathbf{W}^r)\mathbf{H}(\mathbf{w}^r)]\big),$$

где $\mathbf{H}$ — гессиан функции потерь L, $q^\tau_{\mathbf{W}}$ — распределение $q^\tau_{\mathbf{W}}$ в момент оптимизации τ.

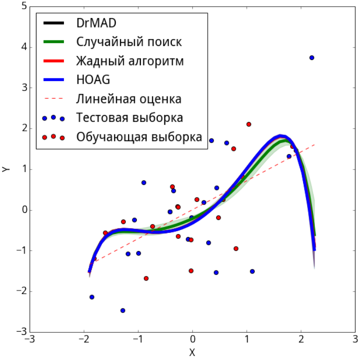
Градиентный спуск как вариационная оценка правдоподобия модели

Для вычисления правдоподобия был предложен ряд алгоритмов, основанных на стохастическом градиентном спуске.

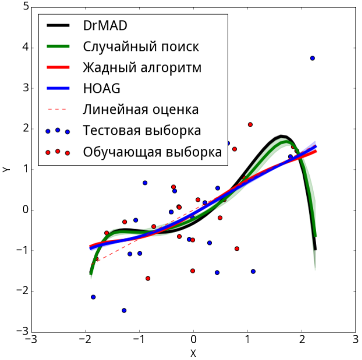




Оптимизация гиперпараметров: пример



Кросс-Валидация



Вариационная оценка

Оптимизация правдоподобия модели

Теорема.

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

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

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

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 $\frac{m}{c_{\text{reg}}}$ из генеральной совокупности.

Параметрическая сложность

Определение

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

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

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:

$$C_p \to 0 => \boldsymbol{\rho}(\mathbf{W}) \to \infty.$$

Оптимизация параметрическое сложности

Теорема

Пусть $c_{\text{train}} = c_{\text{prior}} = 1, c_{\text{comb}} = 0$. Существует область $U(\mathbf{0})$, такая что для любой точки $\theta_0 \in U$ оптимизация

$$\lim_{c_{\text{reg}} \to \infty} \lim_{\eta \to \infty} T^{\eta}\big(Q, \mathbf{h}, T^{\eta}(L, \theta_0, \mathbf{h})\big)$$

доставляет минимум параметрической сложности, $C_p = 0$.

Теорема

Пусть $c_{\text{train}} = 1, c_{\text{comb}} = 0$. Пусть $\mathbf{f}_1, \mathbf{f}_2$ — результаты градиентной оптимизации при разных значениях гиперпараметров $c^1_{\text{prior}}, c^2_{\text{prior}}, c^1_{\text{prior}} < c^2_{\text{prior}}$, полученных при начальном значении вариационных параметров θ и гиперпараметров $\mathbf{h}$. Пусть $\theta, \mathbf{h}$ принадлежат области U, в которой соответствующие функции L и Q являются локально-выпуклыми. Тогда модель $\mathbf{f}_1$ имеет параметрическую сложность, не меньшую чем у $\mathbf{f}_2$.

$$C_p(\mathbf{f}_1) \geq C_p(\mathbf{f}_2).$$

Структурная сложность

Определение

Структурной сложностью C_s модели назовем энтропию распределения структуры модели $\mathbf{\Gamma}$.

$$C_s = -\mathbb{E}_q \log q_{\mathbf{\Gamma}}.$$

Теорема

Пусть для каждого ребра (i, j) семейства моделей $\mathfrak{F}$ априорное распределение

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

Пусть $c_{\text{reg}} > 0, c_{\text{train}} > 0, c_{\text{prior}}, c_{\text{comb}} = 0$, $\mathbf{f}$ — глобавльный оптимум задачи оптимизации. Тогда структурная сложность модели $\mathbf{f}$ равняется нулю.

$$C_s(\mathbf{f}) = 0.$$

Оптимизация структурной сложности

Здесь еще одна теорема об отношении того, что на вершинах и вне веришн

Теорема [Бахтеев, 2018].

Пусть $\mathbf{f}_1 \in F(c_{\text{reg}}, c_{\text{train}}, c_{\text{prior}}, 0, \varnothing, c^1_{\text{temp}}), \mathbf{f}_2 \in \lim_{c^2_{\text{temp}} \to \infty} F(c_{\text{reg}}, c_{\text{train}}, c_{\text{prior}}, 0, \varnothing, c^2_{\text{temp}})$. Пусть вариационные параметры моделей f_1 и f_2 лежат в области U, в которой соответствующие функции L и Q являются локально-выпуклыми. Тогда разница структурных сложностей моделей ограничена выражением:

$$C_{\text{struct}}(\mathbf{f}_1) - C_{\text{struct}}(\mathbf{f}_2) \leq \mathsf{E}^1_q \log p(\mathbf{y}|\mathbf{X}, \mathbf{W}, \mathbf{\Gamma}.\mathbf{A}^{-1}, c^1_{\text{temp}}) - \mathsf{E}^2_q \log p(\mathbf{y}|\mathbf{X}, \mathbf{W}, \mathbf{\Gamma}, \mathbf{A}^{-1}).$$

Полный перебор

Пусть для каждого ребра (i, j) семейства моделей $\mathfrak{F}$ априорное распределение

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

Рассмотрим последовательность $\mathbf{P}$, состоящую из $N = \prod_{(j,k)\in E} K_{j,k}$ моделей, полученных в ходе оптимизаций вида:

$$f_1 \in F(c_{\text{reg}}, 0, 0, \varnothing, c_{\text{comb}}, c_{\text{temp}}),$$

$$f_2 \in F(c_{\text{reg}}, 0, 0, \{q_1(\boldsymbol{\Gamma})\}, c_{\text{comb}}, c_{\text{temp}}),$$

$$f_3 \in F(c_{\text{reg}}, 0, 0, \{q_1(\boldsymbol{\Gamma}), q_2(\boldsymbol{\Gamma})\}, c_{\text{comb}}, c_{\text{temp}}),$$

где $C_{\text{reg}} > 0, c_{\text{comb}} > 0.$

Теорема

Вариационные распределения $q_{\boldsymbol{\Gamma}}$ структур последовательности $\mathbf{P}$ вырождаются в распределения вида $\delta(\hat{\mathbf{m}})$, где $\hat{\mathbf{m}}$ — точка на декартовом произведении вершин симплексов структуры модели.
Последовательность соответствует полному перебору структуры $\boldsymbol{\Gamma}$.

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.