

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

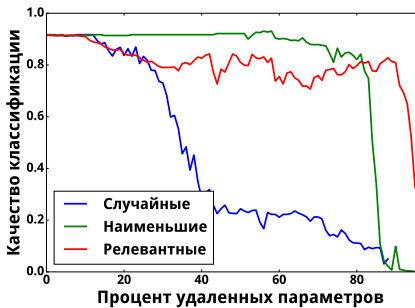
- 1 Большое число параметров и гиперпараметров модели, высокая вычислительная сложность оптимизации.
- 2 Многоэкстремальность и невыпуклость задачи оптимизации.

Методы исследования

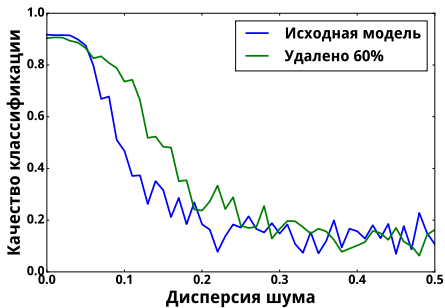
Рассматриваются графовое представление нейронной сети. Используются методы вариационного байесовского вывода. Для получения модели субоптимальной сложности используется метод автоматического определения релевантности параметров с использованием градиентных методов оптимизации гиперпараметров и структурных 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 признаков описаний объекта во множество меток:

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

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

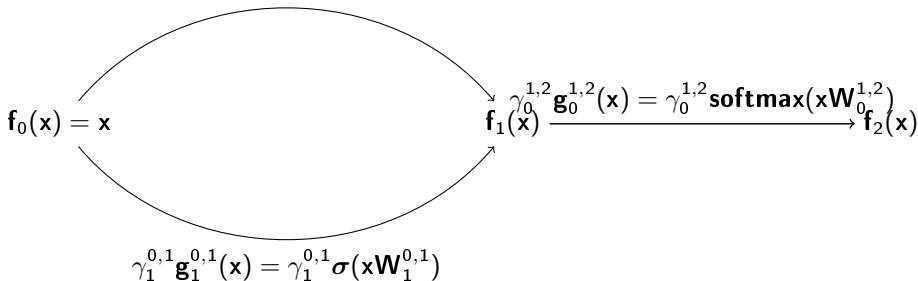
Модель \mathbf{f} задана структурой $\Gamma = [\gamma^{0,1}, \gamma^{1,2}]$.

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

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

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

$$\gamma_0^{0,1} \mathbf{g}_0^{0,1}(\mathbf{x}) = \gamma_0^{0,1} \sigma(\mathbf{x}\mathbf{W}_0^{0,1})$$



Графовое представление модели глубокого обучения

Заданы:

- ① ациклический граф (V, E) ;
- ② для каждого ребра $(j, k) \in E$: вектор базовых дифференцируемых функций $\mathbf{g}^{j,k} = [\mathbf{g}_0^{j,k}, \dots, \mathbf{g}_{K^{j,k}}^{j,k}]$ 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

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

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

Определение

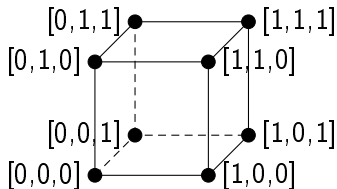
Граф (V, E) со множеством векторов базовых функций $\{\mathbf{g}^{j,k}, (j, k) \in E\}$ и функций агрегаций $\{\mathbf{agg}_v, v \in V\}$ назовем *параметрическим семейством моделей* \mathfrak{F} .

Утверждение

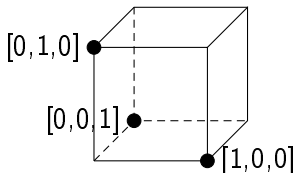
Для любого значения $\gamma^{j,k} \in [0, 1]^{K^{j,k}}$ функция $\mathbf{f} \in \mathfrak{F}$ является моделью.

Ограничения на структурные параметры

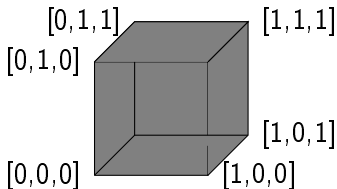
Примеры ограничений для одного структурного параметра γ , $|\gamma| = 3$.



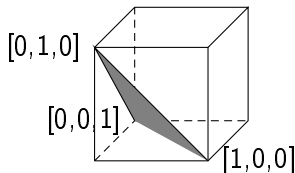
На вершинах куба



На вершинах симплекса



Внутри куба

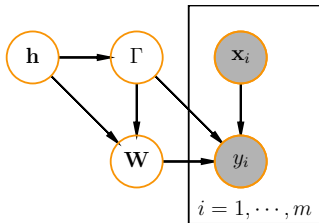


Внутри симплекса

Априорное распределение параметров

Определение

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



Определение

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

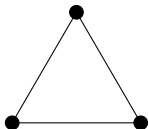
Модель \mathbf{f} задается следующими величинами:

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

Априорное распределение на структуре модели

Каждая точка на симплексе задает модель.

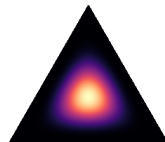
Распределение Дирихле: $\Gamma \sim \text{Dir}(\mathbf{s}, \lambda_{\text{temp}})$



$\lambda_{\text{temp}} \rightarrow 0$

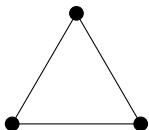


$\lambda_{\text{temp}} = 0.995$



$\lambda_{\text{temp}} = 5.0$

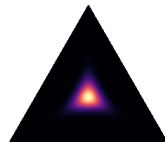
Распределение Гумбель-софтмакс: $\Gamma \sim \text{GS}(\mathbf{s}, \lambda_{\text{temp}})$



$\lambda_{\text{temp}} \rightarrow 0$



$\lambda_{\text{temp}} = 0.995$

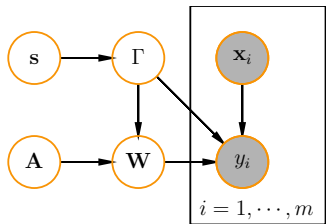


$\lambda_{\text{temp}} = 5.0$

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
 $\mathbf{g}_r^{j,k}$,
 $(\mathbf{A}_r^{j,k})^{-1} \sim \text{inv-gamma}(\lambda_1, \lambda_2)$.
- Структурные параметры модели:
 $\Gamma = \{\gamma^{j,k}, (j, k) \in E\}$,
 $\gamma^{j,k} \sim \text{GS}(s^{j,k}, \lambda_{\text{temp}})$.
- Гиперпараметры модели:
 $\mathbf{h} = [\text{diag}(\mathbf{A}), s]$.
- Метапараметры: $\lambda_1, \lambda_2, \lambda_{\text{temp}}$.

Обоснованность как статистическая сложность

Статистическая сложность модели f :

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

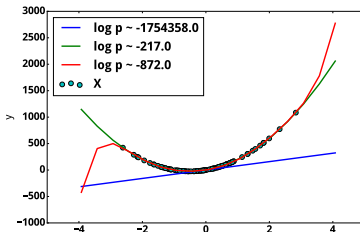
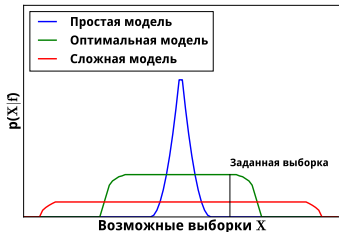
где $\delta\mathfrak{D}$ — допустимая точность передачи информации о выборке \mathfrak{D} .

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

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

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

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



Вариационная нижняя оценка обоснованности

Интеграл обоснованности невычислим аналитически.

Обоснованность модели:

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

Определение

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

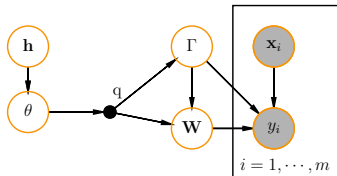
$$q \approx \frac{p(y|\mathbf{X}, \mathbf{w}, \Gamma) p(\mathbf{w}, \Gamma | \mathbf{h}, \lambda_{\text{temp}})}{\iint_{\mathbf{w}', \Gamma'} p(y|\mathbf{X}, \mathbf{w}', \Gamma') p(\mathbf{w}', \Gamma' | \mathbf{h}, \lambda_{\text{temp}}) d\mathbf{w}' d\Gamma'}.$$

Получим нижнюю оценку интеграла:

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

Оценка совпадает с интегралом обоснованности при

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



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 выборке 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:

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

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

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

Обобщающая задача

Задачу выбора модели \mathbf{h}^*, θ^* назовем обобщающей на множестве

$U_\theta \times U_h \times U_\lambda \subset \mathbb{R}^u \times \mathbb{H} \times \mathbb{A}$, если выполнены условия:

- ① Для каждого $\mathbf{h} \in U_h$ и каждого $\lambda \in U_\lambda$ решение θ^* определено однозначно.
- ② *Условие максимизации правдоподобия выборки:* существует $\lambda \in U_\lambda$ 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 правдоподобия выборок: $E_q \log p(\mathbf{y}|\mathbf{X}, \theta_1, \lambda_{\text{temp}}) > \log E_q p(\mathbf{y}|\mathbf{X}, \theta_2, \lambda_{\text{temp}})$.
- ③ *Условие минимизации сложности модели:* существует $\lambda \in U_\lambda$ и $K_2 \in \mathbb{R}_+$, такие что для любых векторов гиперпараметров $\mathbf{h}_1, \mathbf{h}_2 \in U_h$, $Q(\mathbf{h}_1) - Q(\mathbf{h}_2) > K_2$, $E_q \log p(\mathbf{y}|\theta_1, \lambda_{\text{temp}}) = \log E_q p(\mathbf{y}|\theta_2, \lambda_{\text{temp}})$, количество ненулевых параметров у первой модели меньше, чем у второй.
- ④ *Условие достижения максимума правдоподобия модели:* существует значение гиперпараметров λ , такое что оптимизация задачи эквивалента оптимизации вариационной оценки правдоподобия модели:
$$\mathbf{h}^* = \arg \max p(\mathbf{y}|\mathbf{X}, \mathbf{h}', \lambda_{\text{temp}}), \quad \theta^* = \arg \min D_{\text{KL}}(q|p(\mathbf{w}, \Gamma|\mathbf{y}, \mathbf{X}, \lambda_{\text{temp}})).$$
- ⑤ *Условие многоэкстремальности:* Существует константа K_3 , такая что для любых двух векторов $\mathbf{h}_1, \mathbf{h}_2$ и соответствующих векторов $\theta_1^*, \theta_2^* : D_{\text{KL}}(q_{\Gamma_2}, q_{\Gamma_1}) > K_3, D_{\text{KL}}(q_{\Gamma_1}, q_{\Gamma_2}) > K_3$: существуют значения гиперпараметров λ_1, λ_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}^*, θ^* непрерывны по метопараметрам.

Анализ задач выбора моделей

Теорема

Следующие задачи выбора модели не являются обобщающими:

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

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 .

$$\begin{aligned} \mathbf{h}^* &= \arg \max_{\mathbf{h}} Q = & (Q^*) \\ &= \lambda_{\text{train}} E_{q^*} \log p(\mathbf{y}|\mathbf{X}, \mathbf{w}, \mathbf{\Gamma}, \mathbf{h}, \lambda_{\text{prior}}) - \\ &\quad - \lambda_{\text{prior}} D_{\text{KL}}(p(\mathbf{w}, \mathbf{\Gamma}|\mathbf{h}, \lambda_{\text{temp}}) || q^*(\mathbf{w}, \mathbf{\Gamma})) - \\ &\quad - \sum_{\mathbf{p}' \in \mathbf{P}, \lambda \in \lambda_{\text{comb}}} \lambda D_{\text{KL}}(\mathbf{\Gamma}|\mathbf{p}') + \log p(\mathbf{h}|\lambda_1, \lambda_2), \end{aligned}$$

где

$$\begin{aligned} q^* &= \arg \max_q L = E_q \log p(\mathbf{y}|\mathbf{X}, \mathbf{w}, \mathbf{\Gamma}, \mathbf{A}^{-1}, \lambda_{\text{temp}}) & (L^*) \\ &\quad - \lambda_{\text{reg}} D_{\text{KL}}(p(\mathbf{w}, \mathbf{\Gamma}|\mathbf{A}^{-1}, \mathbf{m}, \lambda_{\text{temp}}) || q(\mathbf{w}), q(\mathbf{\Gamma})). \end{aligned}$$

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

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



$$\lambda_{\text{comb}} = [0; 0; 0].$$



$$\lambda_{\text{comb}} = [1; 0; 0].$$



$$\lambda_{\text{comb}} = [1; 1; 0].$$

Адекватность задачи оптимизации

Теорема

Пусть задано параметрическое множество вариационных распределений: $q(\theta)$.

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:

$$\log \hat{p}(\mathbf{y}|\mathbf{X}, \mathbf{h}, \lambda_{\text{temp}}) + \log p(\mathbf{h}|\lambda_1, \lambda_2) \rightarrow \max_{\mathbf{h}}.$$

- 2 Вариационное распределение q приближает апостериорное распределение $p(\mathbf{w}, \Gamma|\mathbf{y}, \mathbf{X}, \mathbf{h}, \lambda_{\text{temp}})$ наилучшим образом:

$$D_{\text{KL}}(q||p(\mathbf{w}, \Gamma|\mathbf{y}, \mathbf{X}, \mathbf{h}, \lambda_{\text{temp}})) \rightarrow \min_{\theta}.$$

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

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

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

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

Оператор оптимизации

Определение

Назовем *оператором оптимизации* T выбор вектора параметров θ' по параметрам предыдущего шага θ .

Оператор стохастического градиентного спуска:

$$\hat{\theta} = T \circ T \circ \dots \circ T(\theta_0, \mathbf{A}^{-1}, \mathbf{m}) = T^\eta(\theta_0, \mathbf{A}^{-1}, \mathbf{m}), \quad \text{где } T(\theta, \mathbf{A}^{-1}, \mathbf{m}) = \\ = \theta - \lambda_{lr} \nabla L(\theta, \mathbf{A}^{-1}, \mathbf{m})|_{\hat{\mathcal{D}}},$$

λ_{lr} — длина шага градиентного спуска, θ_0 — начальное значение параметров θ , $\hat{\mathcal{D}}$ — случайная подвыборка исходной выборки \mathcal{D} .

Перепишем итоговую задачу оптимизации:

$$\mathbf{h}' = T^\eta(Q, \mathbf{h}, T^\eta(L, \theta_0, \mathbf{h})),$$

где θ_0 — начальное значение θ .

Теорема

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

Нижняя вариационная оценка правдоподобия на основе мультистарта

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

Теорема [Бахтеев, 2016]

Пусть L — функция потерь, градиент которой — непрерывно-дифференцируемая функция с константой Липшица C .

Пусть $\theta = [\mathbf{w}^1, \dots, \mathbf{w}^k]$ — начальные приближения оптимизации модели, λ_{lr} — шаг градиентного спуска.

Тогда разность энтропий на смежных шагах оптимизации приближается следующим образом:

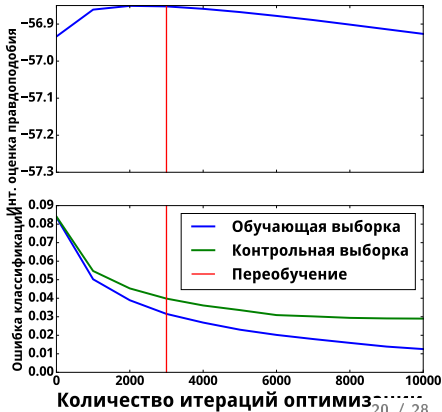
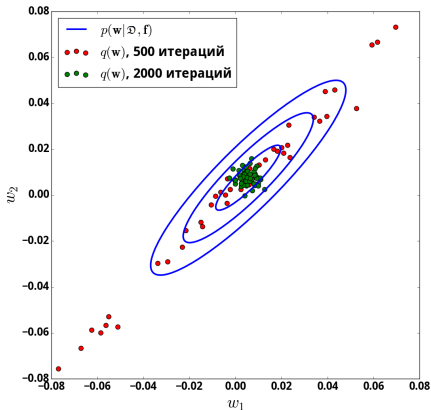
$$\mathbb{E}_{q_{\mathbf{w}}^{\tau}}(-\log(q_{\mathbf{w}}^{\tau})) - \mathbb{E}_{q_{\mathbf{w}}^{\tau-1}}(-\log(q_{\mathbf{w}}^{\tau-1})) \approx \frac{1}{k} \sum_{r=1}^k (\lambda_{lr} \text{Tr}[\mathbf{H}(\mathbf{w}^r)] - \lambda_{lr}^2 \text{Tr}[\mathbf{H}(\mathbf{w}^r)\mathbf{H}(\mathbf{w}^r)]),$$

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

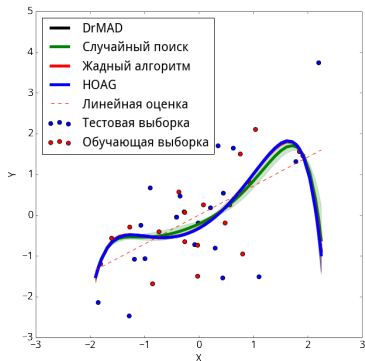
Градиентный спуск как вариационная оценка правдоподобия модели

Эмпирическая плотность, основанная на точках старта оптимизации — вариационное распределение.

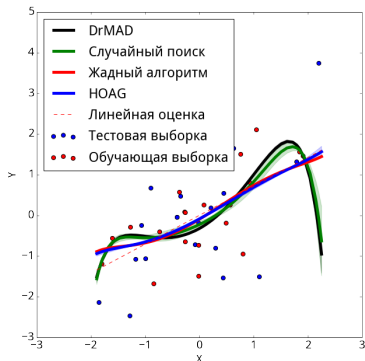
Снижение вариационной оценки правдоподобия — начало переобучения.



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

$$L = \mathbb{E}_q \log p(y|\mathbf{X}, \mathbf{w}, \mathbf{\Gamma}, \mathbf{A}^{-1}, \lambda_{\text{temp}}) - \\ - \lambda_{\text{reg}} D_{KL}(p(\mathbf{w}, \mathbf{\Gamma} | \mathbf{A}^{-1}, \mathbf{m}, \lambda_{\text{temp}}) || q(\mathbf{w}), q(\mathbf{\Gamma}))$$

эквивалентна минимизации ожидаемой дивергенции

$\mathbb{E}_{\hat{\mathbf{X}}, \hat{\mathbf{y}} \sim p(\mathbf{X}, \mathbf{y})} D_{KL}(q || p(\mathbf{w}, \mathbf{\Gamma} | \hat{\mathbf{X}}, \hat{\mathbf{y}}))$, 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:

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

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:

$$C_p \rightarrow 0 \Rightarrow \rho(\mathbf{w}) \rightarrow 1.$$

Оптимизация параметрической сложности

Теорема

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

$$\lim_{\lambda_{\text{prior}} \rightarrow \infty} \lim_{\eta \rightarrow \infty} T^\eta(Q, \mathbf{h}, T^\eta(L, \theta_0, \mathbf{h}))$$

доставляет минимум параметрической сложности. Существует компактная область U , такая что для любой точки $\theta_0 \in U$ 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 $\lambda_{\text{prior}}^1, \lambda_{\text{prior}}^2, \lambda_{\text{prior}}^1 < \lambda_{\text{prior}}^2$, полученных при начальном значении вариационных параметров θ_0 и гиперпараметров \mathbf{h}_0 . Пусть θ_0, \mathbf{h}_0 принадлежат области U , в которой соответствующие функции L и Q являются локально-выпуклыми. Тогда:

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

Структурная сложность

Определение

Структурной сложностью C_s модели назовем энтропию структур Γ , полученных из вариационного распределения q :

$$C_s = -E_q E_{\Gamma} \log p_{\Gamma}.$$

Теорема

Пусть задано априорное распределение на структуре:

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

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

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

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

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

Оптимизация структурной сложности

Теорема

Пусть $\lambda_{\text{train}} > 0$, θ_1, θ_2 — вариационные параметры, такие что θ_1 лежит внутри произведения симплексов структуры, θ_2 — на вершинах симплексов. Тогда

$$\lim_{\lambda_{\text{temp}} \rightarrow 0} \frac{L(\theta_2)}{L(\theta_1)} \rightarrow 0.$$

Теорема

Пусть $\lambda_{\text{train}} > 0$, θ_1, θ_2 — вариационные параметры, такие что θ_1 лежит внутри произведения симплексов структуры, θ_2 — в центре симплексов. Тогда

$$\lim_{\lambda_{\text{temp}} \rightarrow \infty} \frac{L(\theta_2)}{L(\theta_1)} \rightarrow 0.$$

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 конференция «57-я научная конференция 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.