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
28 ноября 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

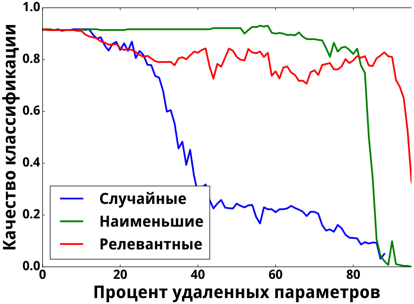
1. Большое число параметров и гиперпараметров модели, высокая вычислительная сложность оптимизации.
2. Многоэкстремальность и невыпуклость задачи оптимизации.

Методы исследования

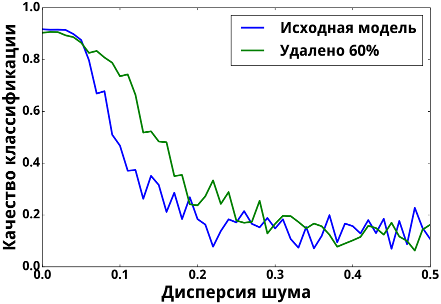
Рассматривается графовое представление нейронной сети. Используются методы вариационного байесовского вывода. Для получения модели субоптимальной сложности используется метод автоматического определения релевантности параметров с использоваением градиентных методов оптимизации гиперпараметров и структурных 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** $\boldsymbol{\Gamma} = [\gamma^{0,1}, \gamma^{1,2}]$.

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

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

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

$$\begin{tikzcd}
\mathbf{f}_0(\mathbf{x}) = \mathbf{x} \arrow[r, bend left, "\gamma_0^{0,1}\mathbf{g}_0^{0,1}(\mathbf{x}) = \gamma_0^{0,1}\boldsymbol{\sigma}\left((\mathbf{w}_0^{0,1})^\mathsf{T}\mathbf{x}\right)"] \arrow[r, bend right, "\gamma_1^{0,1}\mathbf{g}_1^{0,1}(\mathbf{x}) = \gamma_1^{0,1}\boldsymbol{\sigma}\left((\mathbf{w}_1^{0,1})^\mathsf{T}\mathbf{x}\right)"'] & \mathbf{f}_1(\mathbf{x}) \arrow[r, "\gamma_0^{1,2}\mathbf{g}_0^{1,2}(\mathbf{x}) = \gamma_0^{1,2}\textbf{softmax}\left((\mathbf{w}_0^{1,2})^\mathsf{T}\mathbf{x}\right)"] & \mathbf{f}_2(\mathbf{x})
\end{tikzcd}$$

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

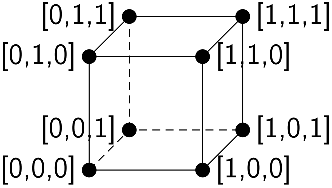
Граф (V, E) со множестом векторов базовых функций $\{\mathbf{g}^{j,k}, (j, k) \in E\}$ и функций агрегаций $\{\mathbf{agg}_v, v \in V\}$ назовем *параметрическим семейством моделей* $\mathfrak{F}$.

Утверждение

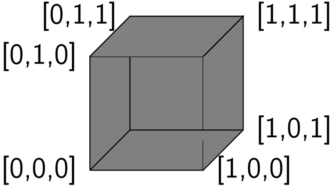
Для любого значения $\boldsymbol{\gamma}^{j,k} \in [0,1]^{K^{j,k}}$ функция $\mathbf{f} \in \mathfrak{F}$ является моделью.

Ограничения на структурные параметры

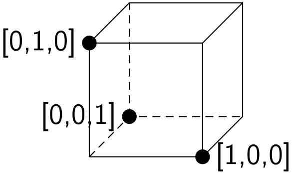
Примеры ограничений для одного структурного параметра $\gamma, |\gamma| = 3$.



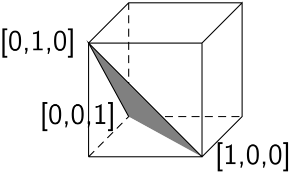
На вершинах куба



Внутри куба



На вершинах симплекса

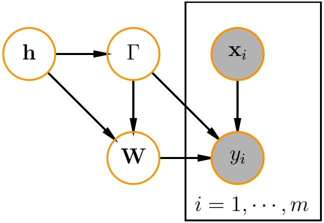


Внутри симплекса

Априорное распределение параметров

Определение

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



Определение

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

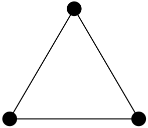
Модель $\mathbf{f}$ задается следующими величинами:

- **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.

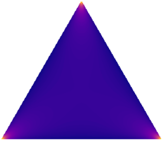
Априорное распределение на структуре модели

Каждая точка на симплексе задает модель.

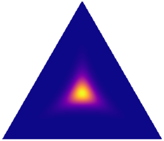
Распределение Гумбель-софтмакс: $\boldsymbol{\Gamma} \sim \mathrm{GS}(\mathbf{s}, \lambda_{\text{temp}})$



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

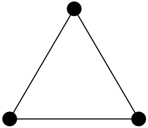


$\lambda_{\text{temp}} = 0.995$

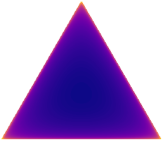


$\lambda_{\text{temp}} = 5.0$

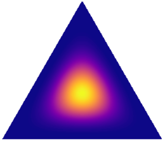
Распределение Дирихле: $\boldsymbol{\Gamma} \sim \mathrm{Dir}(\mathbf{s}, \lambda_{\text{temp}})$



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



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

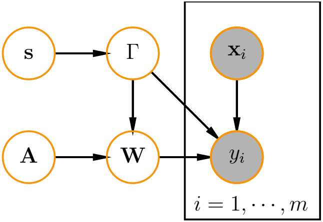


$\lambda_{\text{temp}} = 5.0$

Байесовский выбор модели

Базовая модель:

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



Предлагаемая модель:

- **параметры** модели $\mathbf{w}_r^{j,k} \sim \mathcal{N}(0, (\gamma_r^{j,k})^2 (\mathbf{A}_r^{j,k})^{-1})$, $\mathbf{A}_r^{j,k}$ — 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 $\mathbf{\Gamma} = \{\gamma^{j,k}, (j,k) \in E\}$, $\gamma^{j,k} \sim \text{GS}(\mathbf{s}^{j,k}, \lambda_{\text{temp}})$,
- **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 $\mathbf{f}$:

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

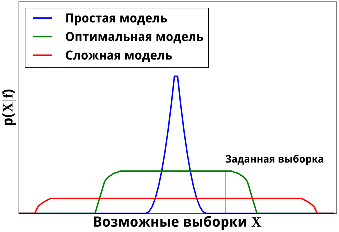
где $\delta\mathfrak{D}$ — допустимая точность передачи информации о выборке $\mathfrak{D}$.

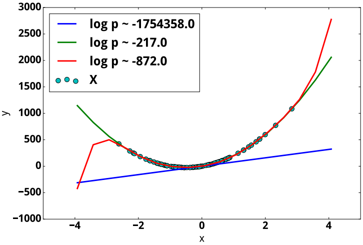
Оптимизация параметров $\mathbf{w}$ производится согласно **апостериорному распределению параметров**:

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

Оптимизация гиперпараметров производится в согласно **апостериорному распределению гиперпараметров**:

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





Вариационная нижняя оценка обоснованности

Интеграл обоснованности невычислим аналитически.

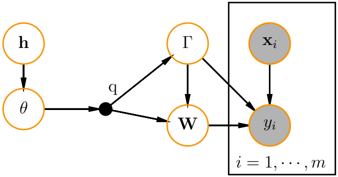
Обоснованность модели:

$$p(\mathbf{y}|\mathbf{X}, \mathbf{h}, \boldsymbol{\lambda}) = \iint\limits_{\mathbf{w}, \boldsymbol{\Gamma}} p(\mathbf{y}|\mathbf{X}, \mathbf{w}, \boldsymbol{\Gamma}) p(\mathbf{w}, \boldsymbol{\Gamma}|\mathbf{h}, \boldsymbol{\lambda}) d\mathbf{w} d\boldsymbol{\Gamma}.$$

Определение

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

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



Получим нижнюю оценку $\log \hat{p}(\mathbf{y}|\mathbf{X}, \mathbf{h}, \boldsymbol{\lambda})$ интеграла

$$\log\, p(\mathbf{y}|\mathbf{X}, \mathbf{h}, \boldsymbol{\lambda}) \geq \mathsf{E}_q \log\, p(\mathbf{y}|\mathbf{X}, \mathbf{w}, \boldsymbol{\Gamma}, \mathbf{h}, \boldsymbol{\lambda}) - \mathsf{D}_{KL}\big(q(\mathbf{w}, \boldsymbol{\Gamma}) || p(\mathbf{w}, \boldsymbol{\Gamma}|\mathbf{h}, \boldsymbol{\lambda})\big).$$

Она совпадает с интегралом обоснованности при

$$D_{\mathrm{KL}}\big(q(\mathbf{w}, \boldsymbol{\Gamma}) | p\,(\mathbf{w}, \boldsymbol{\Gamma}|\mathbf{y}, \mathbf{X}, \boldsymbol{\lambda}, \mathbf{h})\big) = 0.$$

Задача выбора модели

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

Определение

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

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

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

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

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

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

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. Область параметров, гиперпараметров и метапараметров не является пустым или точкой.
2. Для каждого $\mathbf{h} \in U_h$ и каждого $\boldsymbol{\lambda} \in U_\lambda$ решение $\boldsymbol{\theta}^*$ определено однозначно.
3. **Критерий непрерывности:** L, Q непрерывны по метапараметрам.
4. **Критерий перехода между структурами:** существует константа $K_3 > 0$, такая, что для произвольных локальных оптимумов $\mathbf{h}_1, \mathbf{h}_2$ задачи оптимизации Q, полученных при метапараметрах $\boldsymbol{\lambda}$ и удовлетворяющих неравенствам

$$D_{\text{KL}}\left(p(\boldsymbol{\Gamma}|\mathbf{h}_1, \boldsymbol{\lambda})|p(\boldsymbol{\Gamma}|\mathbf{h}_1, \boldsymbol{\lambda})\right) > K_3, D_{\text{KL}}\left(p(\boldsymbol{\Gamma}|\mathbf{h}_1, \boldsymbol{\lambda})|p(\boldsymbol{\Gamma}|\mathbf{h}_2, \boldsymbol{\lambda})\right) > K_3,$$

$$Q(\mathbf{h}_1|\boldsymbol{\lambda}) > Q(\mathbf{h}_2|\boldsymbol{\lambda}),$$

существует значение метапараметров $\boldsymbol{\lambda}' \neq \boldsymbol{\lambda}$, такое, что

   1. соответствие между вариационными параметрами $\boldsymbol{\theta}^*(\mathbf{h}_1), \boldsymbol{\theta}^*(\mathbf{h}_2)$ сохраняется при $\boldsymbol{\lambda}'$,
   2. выполняется неравенство $Q(\mathbf{h}_1|\boldsymbol{\lambda}') < Q(\mathbf{h}_2|\boldsymbol{\lambda}')$.

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:

5. **Критерий максимизации правдоподобия выборки:** существует $\boldsymbol{\lambda} \in U_\lambda$ и $K_1 \in \mathbb{R}_+$, такие что для любых векторов гиперпараметров $\mathbf{h}_1, \mathbf{h}_2 \in U_h, Q(\mathbf{h}_1) - Q(\mathbf{h}_2) > K_1$: выполнено: $\mathsf{E}_{q(\mathbf{w},\boldsymbol{\Gamma}|\boldsymbol{\theta}^*(\mathbf{h_1}))}\log p(\mathbf{y}|\mathbf{X},\mathbf{w},\boldsymbol{\Gamma}) > \mathsf{E}_{q(\mathbf{w},\boldsymbol{\Gamma}|\boldsymbol{\theta}^*(\mathbf{h_2}))}\log p(\mathbf{y}|\mathbf{X},\mathbf{w},\boldsymbol{\Gamma})$.
6. **Критерий минимизации параметрической сложности модели:** существует $\boldsymbol{\lambda} \in U_\lambda$ и $K_2 \in \mathbb{R}_+$, такие что для любых векторов гиперпараметров $\mathbf{h}_1, \mathbf{h}_2 \in U_h, Q(\mathbf{h}_1) - Q(\mathbf{h}_2) > K_2$, сложность первой модели меньше, чем второй.
7. **Критерий максимизации обоснованности модели:** существует значение гиперпараметров $\boldsymbol{\lambda}$, такое что оптимизация задачи эквивалента оптимизации вариационной оценки обоснованности модели:

$$\mathbf{h}^* \propto \arg\max \mathsf{E}_{q(\mathbf{w},\boldsymbol{\Gamma}|\boldsymbol{\theta})} \log p(\mathbf{y}|\mathbf{X},\mathbf{w},\boldsymbol{\Gamma}) - D_{\mathrm{KL}}\left(q(\mathbf{w},\boldsymbol{\Gamma}|\boldsymbol{\theta})||p(\mathbf{w},\boldsymbol{\Gamma}|\mathbf{h},\boldsymbol{\lambda})\right) + \log p(\mathbf{h}|\boldsymbol{\lambda}),$$

$$\boldsymbol{\theta}^* = \arg\min D_{\mathrm{KL}}(q|p(\mathbf{w},\boldsymbol{\Gamma}|\mathbf{y},\mathbf{X},\mathbf{h},\boldsymbol{\lambda})).$$

Анализ задач выбора моделей

Теорема [Бахтеев, 2019]

Следующие задачи выбора модели не являются обобщающими:

1. критерий максимума правдоподобия: $\max_{\boldsymbol{\theta}} \mathsf{E}_q \log p(\mathbf{y}|\mathbf{X}, \boldsymbol{\theta}, \lambda_{\text{temp}}, \mathbf{f})$;
2. критерий максимума апостериорной вероятности $\max_{\boldsymbol{\theta}} \mathsf{E}_q \log p(\mathbf{y}|\mathbf{X}, \boldsymbol{\theta}, \mathbf{f}) p(\boldsymbol{\theta}|\mathbf{h}, \lambda_{\text{temp}})$;
3. метод максимума вариационной оценки обоснованности модели $\max_{\mathbf{h}} \max_{\boldsymbol{\theta}} \mathsf{E}_q \log\ p(\mathbf{y}|\mathbf{X}, \mathbf{w}, \boldsymbol{\Gamma}, \mathbf{f}) - \mathsf{D}_{KL}\big(q(\mathbf{w}, \boldsymbol{\Gamma})||p(\mathbf{w}, \boldsymbol{\Gamma}, \lambda_{\text{temp}})\big) + \log\ p(\mathbf{h}|\mathbf{f})$;
4. кросс-валидация $\max_{\mathbf{h}} \mathsf{E}_q \log p(\mathbf{y}_{\text{valid}}|\mathbf{X}_{\text{valid}}, \boldsymbol{\theta}^*, \lambda_{\text{temp}}, \mathbf{f})$, $\boldsymbol{\theta}^* = \arg\max_{\boldsymbol{\theta}} \mathsf{E}_q \log p(\mathbf{y}_{\text{train}}|\mathbf{X}_{\text{train}}, \boldsymbol{\theta}, \lambda_{\text{temp}}, \mathbf{f}) p(\boldsymbol{\theta}|\mathbf{h})$.
5. AIC: $\max_{\boldsymbol{\theta}} \mathsf{E}_q \log p(\mathbf{y}|\mathbf{X}, \boldsymbol{\theta}, \lambda_{\text{temp}}, \mathbf{f}) - |\theta_i : \mathsf{D}_{\mathsf{KL}}\ (q(w_i)|p(w_i|\boldsymbol{\Gamma}, \mathbf{h}, \boldsymbol{\lambda}) < \lambda|$;
6. BIC: $\max_{\boldsymbol{\theta}} \mathsf{E}_q \log p(\mathbf{y}|\mathbf{X}, \boldsymbol{\theta}, \lambda_{\text{temp}}, \mathbf{f}) - \frac{1}{2}\log(|\mathbb{W}||\theta_i : \mathsf{D}_{\mathsf{KL}}\ (q(w_i)|p(w_i|\boldsymbol{\Gamma}, \mathbf{h}, \boldsymbol{\lambda}) < \lambda|$;
7. перебор структуры модели: $\max \boldsymbol{\Gamma}' \max_{\boldsymbol{\theta}} \mathsf{E}_q \log p(\mathbf{y}|\mathbf{X}, \boldsymbol{\theta}, \lambda_{\text{temp}}, \mathbf{f}) \mathbb{I}(q(\boldsymbol{\Gamma}\Gamma = p')$, где p' — распределение на структуре (метапараметр).

Предлагаемая задача оптимизации

Теорема [Бахтеев, 2018]

Тогда следующая задача является обобщающей:

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

$$= \lambda^{\mathbf{Q}}_{\mathbf{likelihood}} \mathsf{E}_{q(\mathbf{w},\boldsymbol{\Gamma}|\boldsymbol{\theta}^*)} \log p(\mathbf{y}|\mathbf{X},\mathbf{w},\boldsymbol{\Gamma},\mathbf{h},\boldsymbol{\lambda}) -$$

$$-{}^{\text{prior}}_{\mathbf{Q}} \mathrm{D}_{KL}\big(q(\mathbf{w},\boldsymbol{\Gamma}|\boldsymbol{\theta}^*)||p(\mathbf{w},\boldsymbol{\Gamma}|\mathbf{h},\boldsymbol{\lambda})\big) -$$

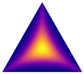
$$- \sum_{p' \in \mathfrak{P}, \lambda \in \boldsymbol{\lambda}^{\text{struct}}_{\mathbf{Q}}} \lambda \mathrm{D}_{KL}(p(\boldsymbol{\Gamma}|\mathbf{h},\boldsymbol{\lambda})|p') + \log p(\mathbf{h}|\boldsymbol{\lambda}),$$

где

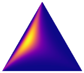
$$\boldsymbol{\theta}^* = \arg\max_{\boldsymbol{\theta}} L = \mathsf{E}_q \log p(\mathbf{y}|\mathbf{X},\mathbf{w},\boldsymbol{\Gamma},\mathbf{h},\boldsymbol{\lambda})$$

$$-{}^{\text{prior}}_{\mathbf{L}} \mathrm{D}_{KL}\big(q^*(\mathbf{w},\boldsymbol{\Gamma})||p(\mathbf{w},\boldsymbol{\Gamma}|\mathbf{h},\boldsymbol{\lambda})\big).$$

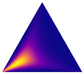
Оптимизационная задача обобщает алгоритмы оптимизации: оптимизация правдоподобия и обоснованности, последовательное увеличение и снижение сложности модели, полный перебор структуры.



$\boldsymbol{\lambda}^{Q}_{\text{struct}} = [0;0;0].$



$\boldsymbol{\lambda}^{Q}_{\text{struct}} = [1;0;0].$



$\boldsymbol{\lambda}^{Q}_{\text{struct}} = [1;1;0].$

Адекватность задачи оптимизации

Теорема, [Бахтеев, 2018]

Пусть задано параметрическое множество вариационных распределений: $q(\boldsymbol{\theta})$.
Пусть $\lambda^{L}_{\text{likelihood}} = \lambda^{L}_{\text{prior}} = \lambda^{Q}_{\text{prior}} = 1, \boldsymbol{\lambda}^{Q}_{\text{struct}} = \mathbf{0}$. Тогда:

1. Предлагаемая задача оптимизации доставляет максимум апостериорной вероятности гиперпараметров с использованием вариационной оценки обоснованности: $$\log \hat{p}(\mathbf{y}|\mathbf{X}, \mathbf{h}, \boldsymbol{\lambda}) + \log p(\mathbf{h}|\boldsymbol{\lambda}) \to \max_{\mathbf{h}}.$$
2. Вариационное распределение q приближает апостериорное распределение $p(\mathbf{w}, \boldsymbol{\Gamma}|\mathbf{y}, \mathbf{X}, \mathbf{h}, \boldsymbol{\lambda}, \mathbf{f})$ наилучшим образом: $$D_{\text{KL}}(q||p(\mathbf{w}, \boldsymbol{\Gamma}|\mathbf{y}, \mathbf{X}, \mathbf{h}, \boldsymbol{\lambda})) \to \min_{\boldsymbol{\theta}}.$$

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

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

Если существуют значения вариационных параметров, такие что
$q(\mathbf{w}) = p(\mathbf{w}|\boldsymbol{\Gamma}, \mathbf{h}, \boldsymbol{\lambda})$, $q(\boldsymbol{\Gamma}) = p(\boldsymbol{\Gamma}|\mathbf{h}, \boldsymbol{\lambda})$, то решение задачи оптимизации для функции L доставляет эти значения.

Оператор оптимизации

Определение

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

Оператор стохастического градиентного спуска:

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

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:

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

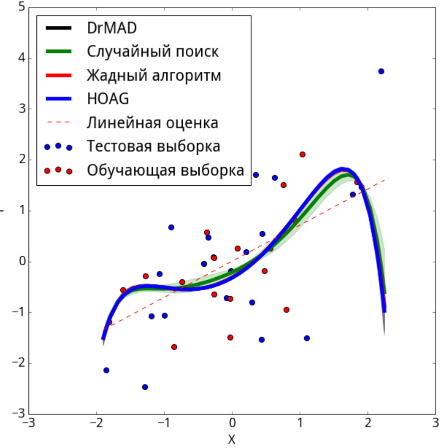
где $\boldsymbol{\theta}_0$ — начальное значение $\boldsymbol{\theta}$.

Теорема, [Бахтеев, 2019]

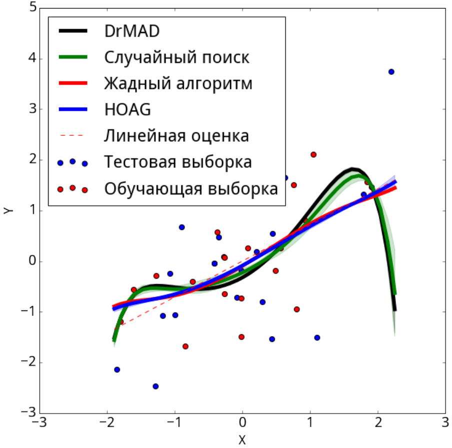
Пусть $\frac{\lambda^{\boldsymbol{Q}}_{\text{prior}}}{\lambda^{\boldsymbol{Q}}_{\text{likelihood}}} = \lambda^L_{\text{prior}}$. Тогда задача оптимизации представима в виде одноуровневой задач.

Оптимизация гиперпараметров: пример

Исследованы градиентные методы оптимизации гиперпараметров.

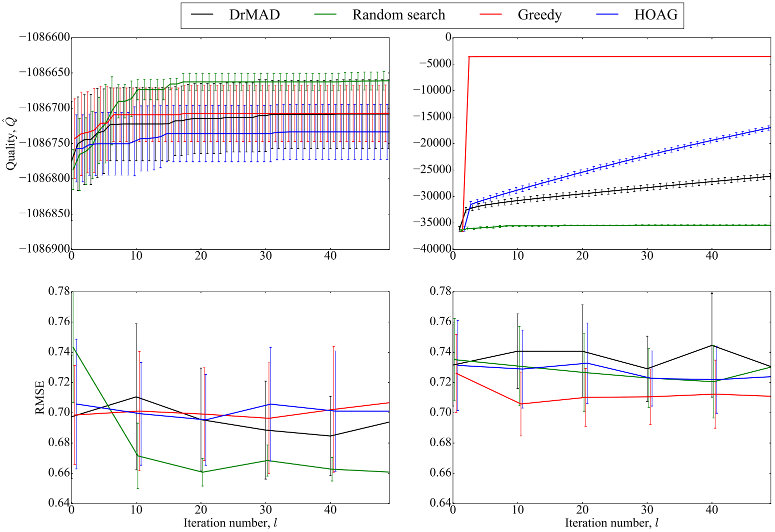


Кросс-Валидация

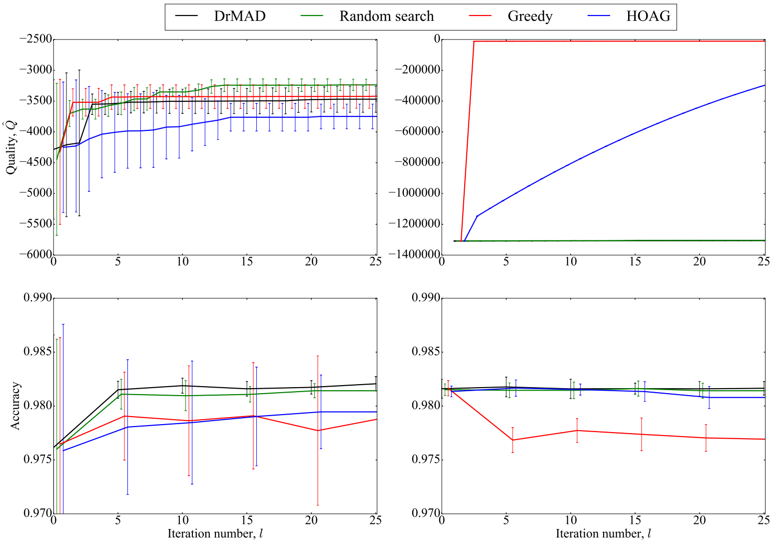


Вариационная оценка

Эксперименты: WISDM



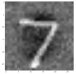
Эксперименты: MNIST



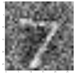
Эксперименты: MNIST

Добавление гауссового шума $\mathcal{N}(\mathbf{0}, \sigma^2 \mathbf{I})$:

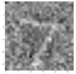
Без шума



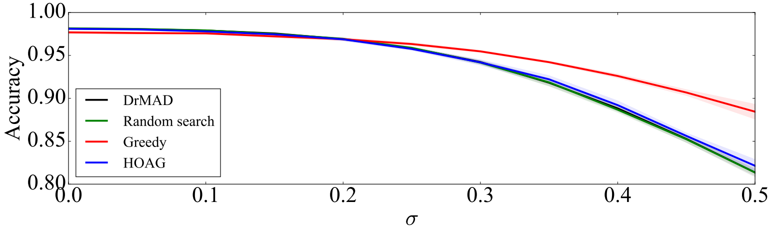
$\sigma = 0.1$



$\sigma = 0.25$



$\sigma = 0.5$



Нижняя вариационная оценка обоснованности на основе мультистарта

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

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_{lr} — шаг градиентного спуска.
Тогда разность энтропий на смежных шагах оптимизации приближается следующим образом:

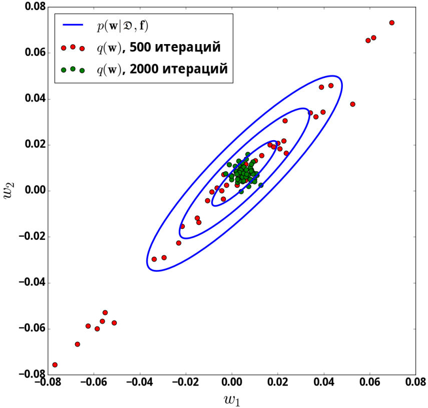
$$\mathsf{E}_{q_{\mathbf{w}}^{\tau}}(-\log(q_{\mathbf{w}}^{\tau})) - \mathsf{E}_{q_{\mathbf{w}}^{\tau-1}}(-\log(q_{\mathbf{w}}^{\tau-1})) \approx \frac{1}{k}\sum_{r=1}^{k}\left(\lambda_{\mathrm{lr}} Tr[\mathbf{H}(\mathbf{w}^r)] - \lambda_{\mathrm{lr}}^2 Tr[\mathbf{H}(\mathbf{w}^r)\mathbf{H}(\mathbf{w}^r)]\right),$$

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

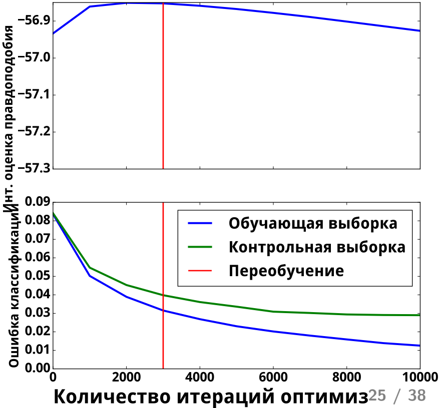
Градиентный спуск как вариационная оценка обоснованности модели

Эмпирическое распределение на точках старта оптимизации — вариационное распределение.

Градиентный спуск не оптимизирует оценку обоснованности.



Снижение вариационной оценки обоснованности — начало переобучения.



Анализ обобщающей задачи оптимизации

Теорема, [Бахтеев, 2018]

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

$$L = \mathsf{E}_q \log p(\mathbf{y}|\mathbf{X}, \mathbf{w}, \mathbf{\Gamma}) - \lambda_{\text{prior}}^{L} \mathsf{D}_{KL}(q||p(\mathbf{w}, \mathbf{\Gamma}|\mathbf{h}, \boldsymbol{\lambda}))$$

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

Определение

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

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

Теорема, [Бахтеев, 2018]

Пусть $\boldsymbol{\lambda}_{\text{struct}}^{Q} = \mathbf{0}$. Пусть $\boldsymbol{\theta}_1, \boldsymbol{\theta}_2, \mathbf{h}_1, \mathbf{h}_2$ — результаты оптимизации при разных значениях гиперпараметров $\lambda_{\text{prior}_1}^{\text{Q}}, \lambda_{\text{prior}_2}^{\text{Q}}, \lambda_{\text{prior}_1}^{\text{Q}} > \lambda_{\text{prior}_2}^{\text{Q}}$ на компакте U. Пусть функция $Q(\mathbf{h}|\mathbf{y}, \mathbf{X}, \boldsymbol{\theta}, \boldsymbol{\lambda})$ является вогнутой на U при $\lambda_{\text{prior}_2}^{\text{Q}}$. Тогда:

$$C_p(\boldsymbol{\theta}_1|U_{\mathbf{h}}, \boldsymbol{\lambda}_1) - C_p(\boldsymbol{\theta}_2|U_{\mathbf{h}}, \boldsymbol{\lambda}_2) < \frac{\lambda_{\text{prior}}^{\text{L}}}{\lambda_{\text{prior}_2}^{\text{Q}}}(\lambda_{\text{prior}_2}^{\text{Q}} - \lambda_{\text{prior}}^{\text{L}})C,$$

где C — некоторая константа.

Анализ параметрической сложности

Определение

Относительной вариационной плотностью назовем отношение:

$$\rho(w|\mathbf{\Gamma}, \boldsymbol{\theta}_{\mathbf{w}}, \mathbf{h}, \boldsymbol{\lambda}) = \frac{q_{\mathbf{w}}(\text{mode } p(\mathbf{w}|\mathbf{\Gamma}, \mathbf{h}, \boldsymbol{\lambda}))}{q_{\mathbf{w}}(\text{mode } q_{\mathbf{w}})}.$$

Теорема, [Бахтеев, 2018]

Пусть заданы компактные множества $U_{\mathbf{h}} \subset \mathbb{H}, U_{\boldsymbol{\theta}_{\mathbf{w}}} \subset \mathbb{O}_{\mathbf{w}}, U_{\boldsymbol{\theta}_{\mathbf{\Gamma}}} \subset \mathbb{O}_{\mathbf{\Gamma}}$, вариационное и априорное распределение $q_{\mathbf{w}}(\mathbf{w}|\mathbf{\Gamma}, \boldsymbol{\theta}_{\mathbf{w}})$, $p(\mathbf{w}|\mathbf{\Gamma}, \mathbf{h}, \boldsymbol{\lambda})$ являются абсолютно непрерывным и унимодальным на $U_{\boldsymbol{\theta}}$ с совпадающей модой и матожиданием. Пусть мода и матожидание априорного распределения не зависят от гиперпараметров $\mathbf{h}$ и структуры $\mathbf{\Gamma}$.
Пусть задана бесконечная последовательность векторов вариационных параметров $\boldsymbol{\theta}[1], \boldsymbol{\theta}[2], \dots, \boldsymbol{\theta}[i], \dots \in U_{\boldsymbol{\theta}}$, такая, что $\lim_{i \to \infty} C_p(\boldsymbol{\theta}[i]|U_{\mathbf{h}}, \boldsymbol{\lambda}) = 0$. Тогда:

$$\lim_{i \to \infty} \mathsf{E}_{q_{\mathbf{\Gamma}}(\mathbf{\Gamma}|\boldsymbol{\theta}_{\mathbf{\Gamma}}[i])} \rho(\mathbf{w}|\mathbf{\Gamma}, \boldsymbol{\theta}_{\mathbf{w}}[i], \mathbf{h}[i], \boldsymbol{\lambda})^{-1} = 1, \mathbf{h}[i] = \arg\min D_{\mathrm{KL}}\big(q(\mathbf{w}, \mathbf{\Gamma}|\boldsymbol{\theta}_i)||p(\mathbf{w}, \mathbf{\Gamma}|\mathbf{h}, \boldsymbol{\lambda})\big).$$

Результаты, выносимые на защиту

1. Предложен метод байесовского выбора субоптимальной структуры модели глубокого обучения с использованием автоматического определения релевантности параметров.
2. Предложены критерии оптимальной и субоптимальной сложности модели глубокого обучения.
3. Предложен метод графового описания моделей глубокого обучения.
4. Предложено обобщение задачи оптимизации структуры модели, включающее ранее описанные методы выбора модели:
    - оптимизация обоснованности;
    - последовательное увеличение сложности модели;
    - последовательное снижение сложности модели;
    - полный перебор вариантов структуры модели.
5. Предложен метод оптимизации вариационной оценки обоснованности на основе мультистарта оптимизации модели.
6. Предложен алгоритм оптимизации параметров, гиперпараметров и структурных параметров моделей глубокого обучения.
7. Исследованы свойства оптимизационной задачи при различных значениях метапараметров. Рассмотрены ее асимптотические свойства.

Список работ автора по теме диссертации

Публикации ВАК

1. Bakhteev, O., Kuznetsova, R., Romanov, A. and Khritankov, A. A monolingual approach to detection of text reuse in Russian-English collection // In 2015 Artificial Intelligence and Natural Language and Information Extraction, Social Media and Web Search FRUCT Conference (AINL-ISMW FRUCT) (pp. 3-10). IEEE. 2015.
2. 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.
3. Romanov, A., Kuznetsova, R., Bakhteev, O. and Khritankov, A. Machine-Translated Text Detection in a Collection of Russian Scientific Papers. // Computational Linguistics and Intellectual Technologies. 2016.
4. Bakhteev, O. and Khazov, A., Author Masking using Sequence-to-Sequence Models // In CLEF (Working Notes). 2017.
5. Бахтеев О.Ю., Стрижов В.В. Выбор моделей глубокого обучения субоптимальной сложности // Автоматика и телемеханика. 2018. №8. С. 129-147.
6. Огальцов А.В., Бахтеев О.Ю. Автоматическое извлечение метаданных из научных PDF-документов // Информатика и её применения. 2018.
7. Смердов А.Н., Бахтеев О.Ю., Стрижов В.В. Выбор оптимальной модели рекуррентной сети в задачах поиска парафраза // Информатика и ее применения. 2019.
8. Грабовой А.В., Бахтеев О.Ю., Стрижов В.В. Определение релевантности параметров нейросети // Информатика и её применения. 2019.
9. Bakhteev O., Strijov V. Comprehensive analysis of gradient-based hyperparameter optimization algorithms // Annals of Operations Research. 2019.

Выступления с докладом

1. "Восстановление панельной матрицы и ранжирующей модели в разнородных шкалах", Всероссийская конеренция «57-я научная конеренция МФТИ», 2014.
2. "Выбор модели глубокого обучения субоптимальной сложности с использованием вариационной оценки правдоподобия", Международная конференция «Интеллектуализация обработки информации», 2016.
3. "Градиентные методы оптимизации гиперпараметров моделей глубокого обучения", Всероссийская конференция «Математические методы распознавания образов ММРО», 2017.
4. "Детектирование переводных заимствований в текстах научных статей из журналов, входящих в РИНЦ", Всероссийская конференция «Математические методы распознавания образов ММРО», 2017.
5. "Байесовский выбор наиболее правдоподобной структуры модели глубокого обучения", Международная конференция «Интеллектуализация обработки информации», 2018.

Иллюстративный пример

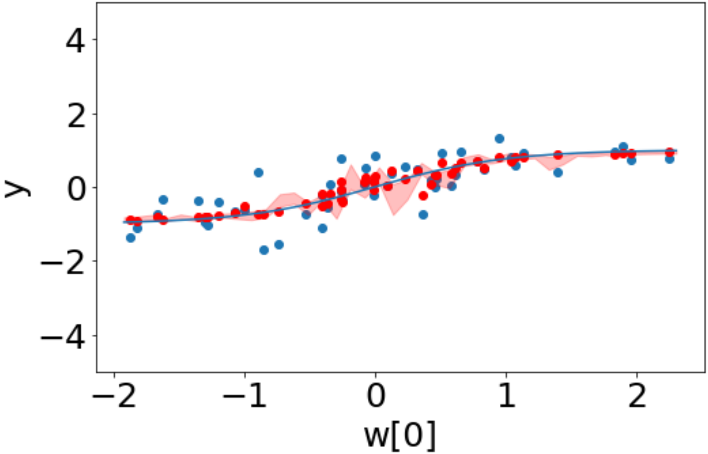
Цель эксперимента: проверка анализируемых свойств предлагаемой модели оптимизации.

Модель $\mathbf{f}$ — ансамбль трех моделей:

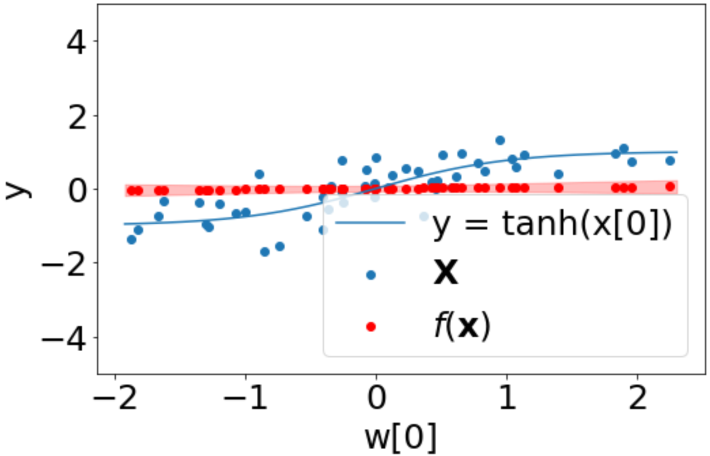
1. $\mathbf{g}_0^{0,1} = tanh(wx)$;
2. $\mathbf{g}_1^{0,1} = tanh(\mathbf{w}^\mathsf{T}[x, x^2, \dots, x^{10}])$;
3. $\mathbf{g}_2^{0,1} = w$.

Рассматривалось поведение оптимизируемой функции при калибровке $\lambda_{\text{temp}}, \lambda_{\text{prior}}^{\text{L}}$.

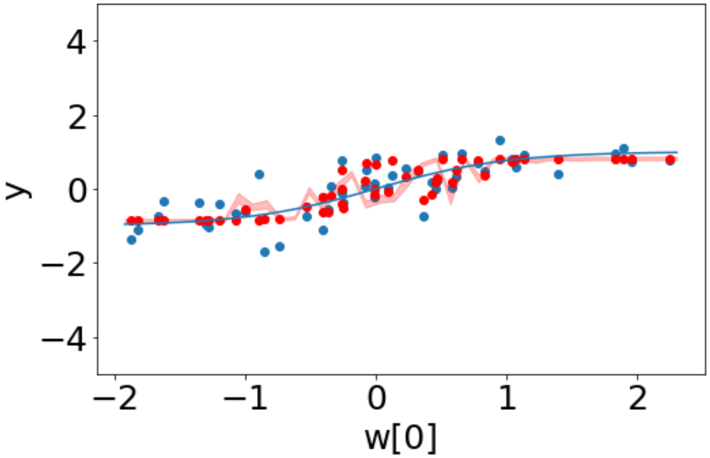
Эксперимент при $\lambda_{\text{temp}} \ll 1, \lambda_{\text{prior}}^{\text{L}} \ll 1$



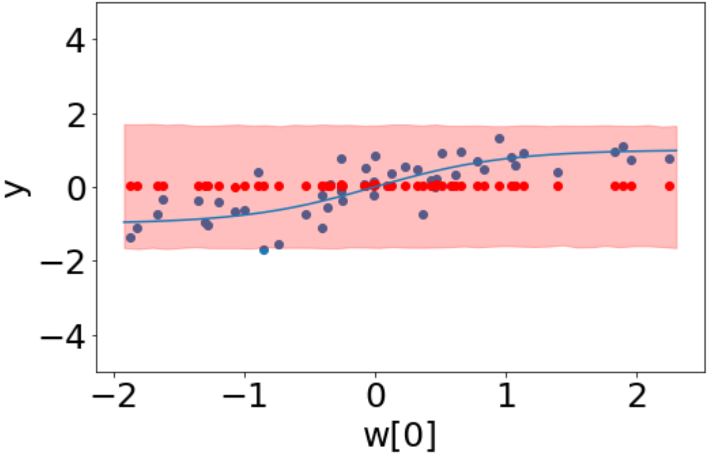
Эксперимент при $\lambda_{\mathbf{temp}} \ll 1, \lambda_{\mathbf{prior}}^{\mathbf{L}} \gg 1$



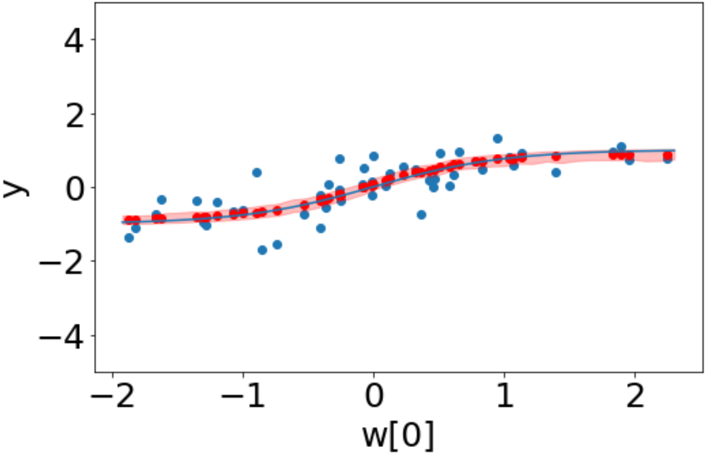
Эксперимент при $\lambda_{\text{temp}} \gg 1, \lambda^{\mathbf{L}}_{\text{prior}} \ll 1$



Эксперимент при $\lambda_{\mathbf{temp}} \gg 1, \lambda^{\mathbf{L}}_{\mathbf{prior}} \gg 1$



Эксперимент при $\lambda_{\textsf{temp}} = 1, \lambda^{\textsf{L}}_{\textsf{prior}} = 1$



Эксперимент по выбору моделей

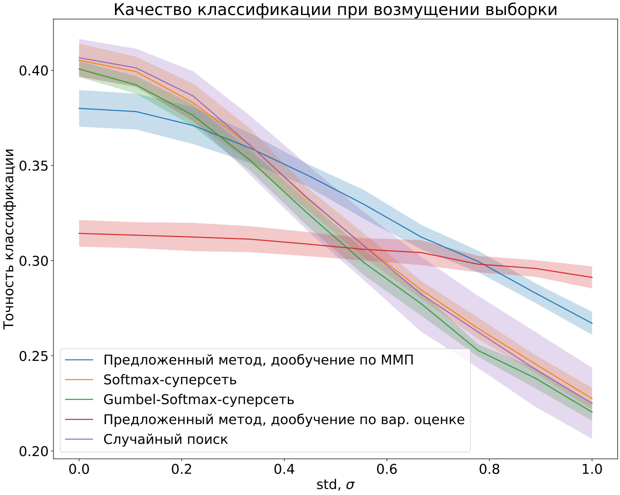
	Алгоритмы					
Выборка $\mathfrak{D}$	ML	ML+VAL	MAP+VAL	Var+SGD	Var+Lang	Var+Norm
Результаты, RMSE/Accuracy						
Boston, 1 слой	8,1 ± 2,0	5,9 ± 0,7	5,2 ± 0,6	$\mathbf{3,7 \pm 0,2}$	6,7 ± 0,7	5,0 ± 0,4
Boston, 3 слоя	7,1 ± 1,3	4,3 ± 0,1	4,4 ± 0,4	$\mathbf{3,2 \pm 0,06}$	4,6 ± 0,4	6,8 ± 1,6
Protein	5,1 ± 0,0	5,1 ± 0,0	5,1 ± 0,0	5,1 ± 0,0	5,1 ± 0,0	$\mathbf{5,0 \pm 0,1}$
MSD	12,2 ± 0,0	$\mathbf{10,9 \pm 0,1}$	$\mathbf{10,9 \pm 0,1}$	12,2 ± 0,0	12,9 ± 0,0	19,6 ± 3,6
MNIST	0,985 ± 0,002	0,984 ± 0,002	$\mathbf{0,986 \pm 0,002}$	0,914 ± 0,005	0,979 ± 0,003	0,971 ± 0,001
Результаты, RMSE/Accuracy$_{\mathbf{0,5}}$						
Boston, 1 слой	43,9 ± 9,4	18,6 ± 2,0	15,8 ± 2,3	$\mathbf{11,9 \pm 1,1}$	20,3 ± 3,1	18,2 ± 3,3
Boston, 3 слоя	23,4 ± 4,9	18,7 ± 2,8	18,3 ± 3,0	**9,0 ± 0,7**	14,5 ± 2,6	15,2 ± 2,7
Protein	19,5 ± 0,3	18,5 ± 0,5	18,6 ± 0,3	$\mathbf{16,7 \pm 0,3}$	19,3 ± 0,6	19,7 ± 3,7
MSD	178,3 ± 0,8	$\mathbf{121,3 \pm 4,5}$	123,7 ± 2,5	175,8 ± 1,0	203,8 ± 1,4	292,0 ± 2,0
MNIST	0,931 ± 0,004	0,929 ± 0,006	$\mathbf{0,934 \pm 0,007}$	0,857 ± 0,007	0,919 ± 0,008	0,916 ± 0,004
Результаты, RMSE/Accuracy$_{\mathbf{1,0}}$						
Boston, 1 слой	120,9 ± 33,4	42,5 ± 6,3	32,5 ± 6,0	$\mathbf{25,7 \pm 3,2}$	42,4 ± 5,7	41,3 ± 6,3
Boston, 3 слоя	46,1 ± 15,8	40,5 ± 5,3	38,6 ± 8,0	**16,5 ± 2,5**	30,4 ± 7,9	26,2 ± 6,9
Protein	37,0 ± 0,8	34,4 ± 1,1	35,0 ± 1,0	$\mathbf{30,6 \pm 0,6}$	36,6 ± 1,1	35,0 ± 8,1
MSD	319,6 ± 1,4	$\mathbf{217,5 \pm 8,2}$	221,9 ± 4,2	314,8 ± 1,8	363,7 ± 1,9	521,6 ± 3,1
MNIST	$\mathbf{0,814 \pm 0,010}$	0,808 ± 0,010	0,812 ± 0,008	0,772 ± 0,010	0,802 ± 0,009	0,800 ± 0,009

Выбор структуры модели: число параметров в модели

Метод выбора структуры	Число параметров
Случайный поиск	1.3M $\pm$ 0.8M
Softmax-суперсеть	3M $\pm$ 0.0M
Gumbel-Softmax-суперсеть	1.5M $\pm$ 0.4M
Предложенный метод	0.2M $\pm$ 0.1M

Результаты были округлены до десятой миллиона.

Выбор структуры модели: устойчивость моделей при добавлении шума в выборку



Выбор структуры модели: устойчивость моделей при добавлении шума в параметры