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 работы:

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.

Основные проблемы:

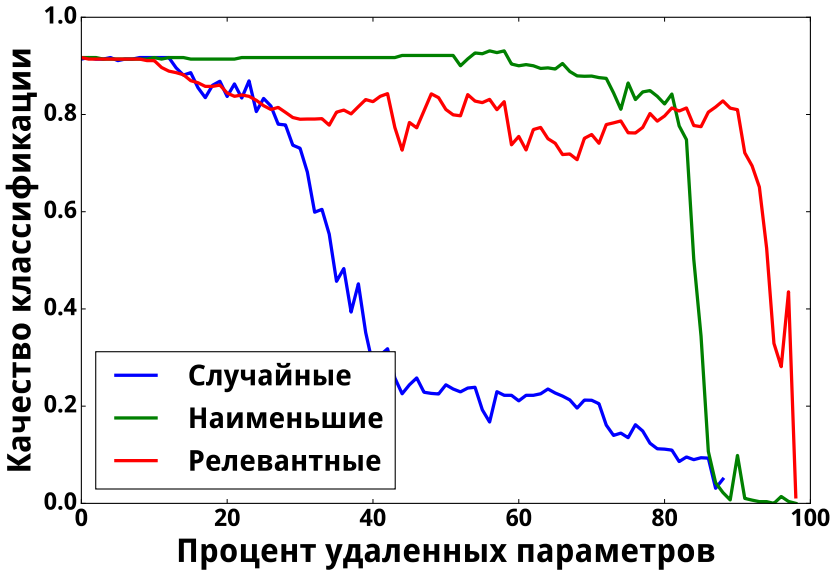
1. Большое число параметров и гиперпараметров.
2. Многоэкстремальность и невыпуклость задачи оптимизации параметров и гиперпараметров модели.
3. Высокая вычислительная сложность оптимизации.

Методы исследования.

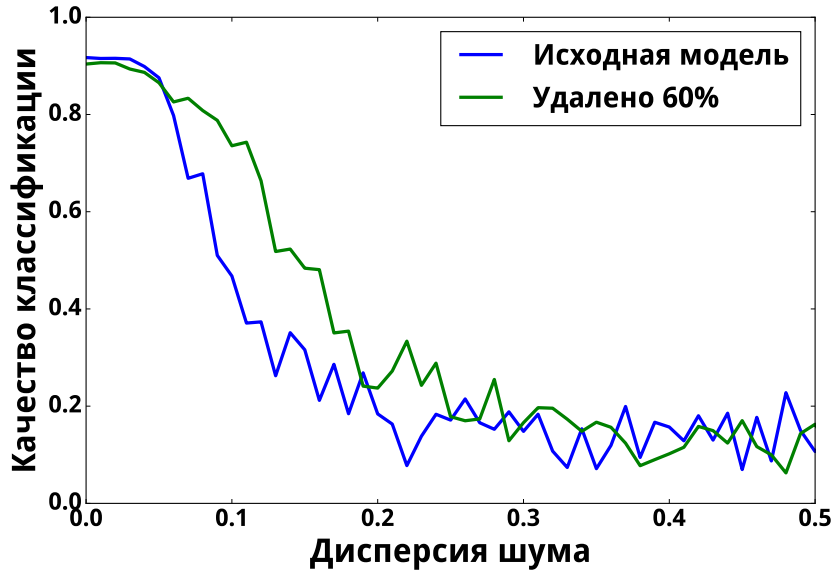
Используются методы вариационного байесовского вывода. Рассматриваются графовое представление нейронной сети. Для получения вариационных оценок правдоподобия модели используется метод, основанный на градиентном спуске. В качестве метода получения модели субоптимальной сложности используется метод автоматического определения релевантности параметров с использоваением градиентных методов оптимизации гиперпараметров.

Проблема выбора оптимальной структуры модели глубокого обучения

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



Избыточность параметров модели



Неустойчивость модели

Глубокое обучение предполагает оптимизацию моделей с заведомо избыточной сложностью.

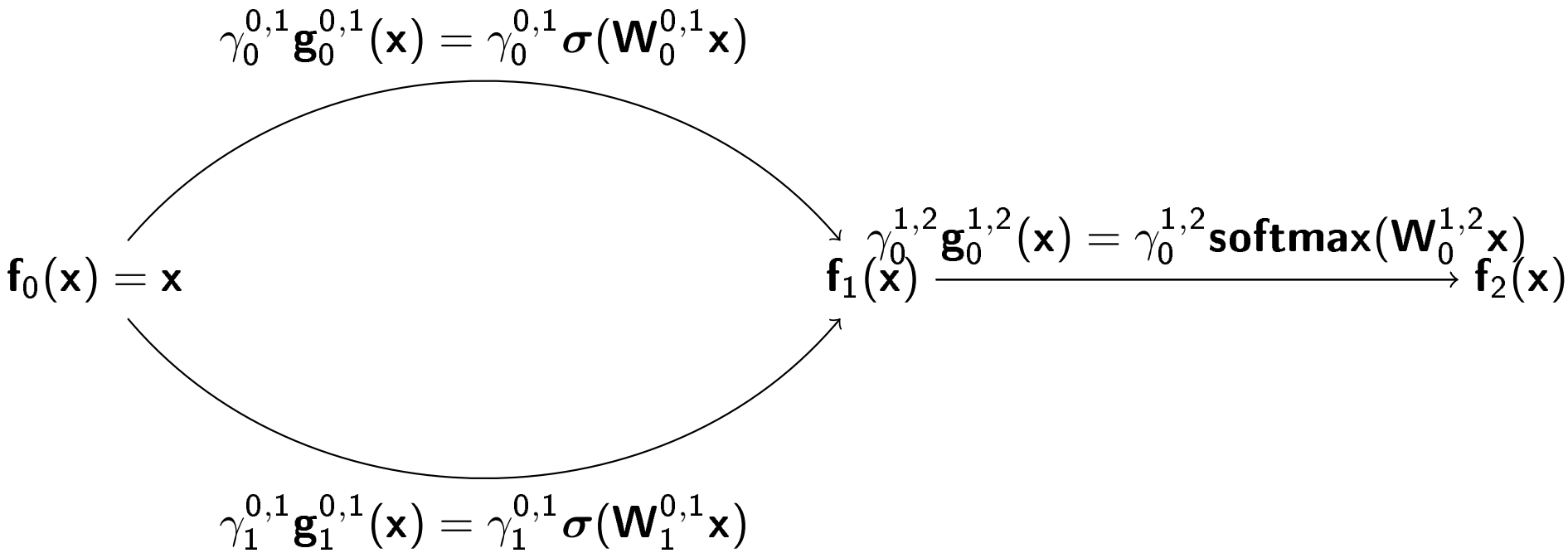
Выбор структуры: двуслойная нейросеть

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

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

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

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.



Графовое представление модели глубокого обучения

Определение

Пусть:

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

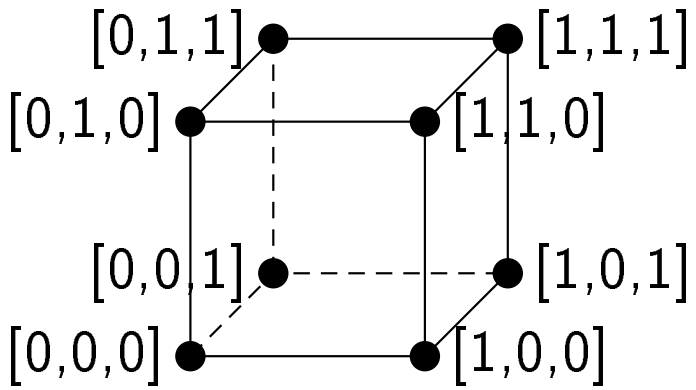
Граф (V, E) в совокупности со множестом векторов базовых функций $\{\mathbf{g}^{j,k}, (j,k) \in E\}$ и множеством функций агрегаций $\{\mathbf{agg}_v, v \in V\}$ задает *параметрическое семейство моделей* $\mathfrak{F}$, если функция $\mathbf{f} = \mathbf{f}_{|V|-1}$, задаваемая по правилу

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

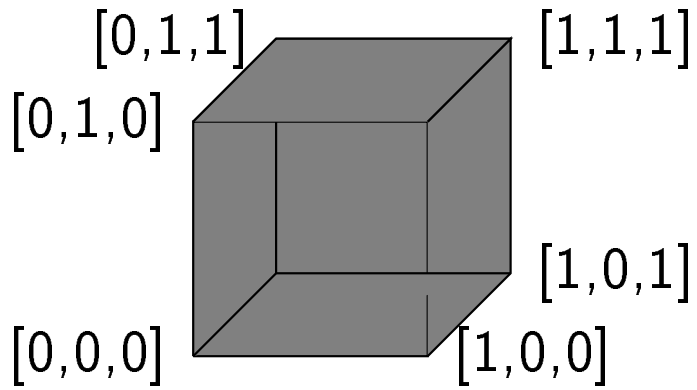
является дифференцируемой по параметрам $\mathbf{w}$ функцией из признакового пространства $\mathbb{X}$ в пространство меток $\mathbb{Y}$ при значениях векторов, $\boldsymbol{\gamma}^{j,k} \in [0,1]^{K^{j,k}}$.

Ограничения на структурные параметры

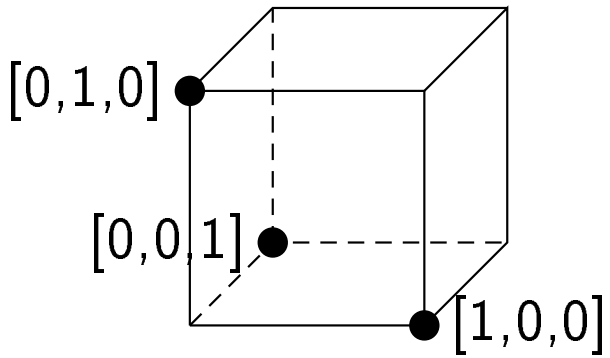
Примеры ограничений для одного структурного параметра $\gamma, |\gamma| = 3$.



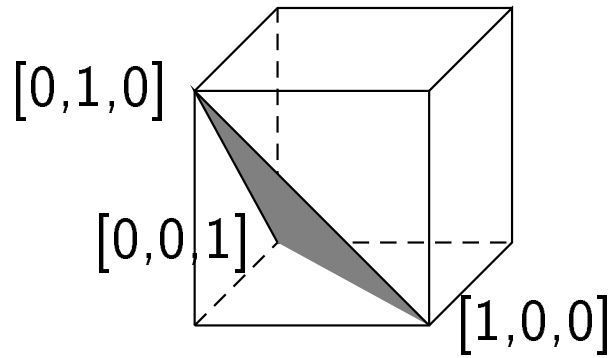
Структурный параметр лежит на вершинах куба



Структурный параметр лежит внутри куба



Структурный параметр лежит на вершинах симплекса

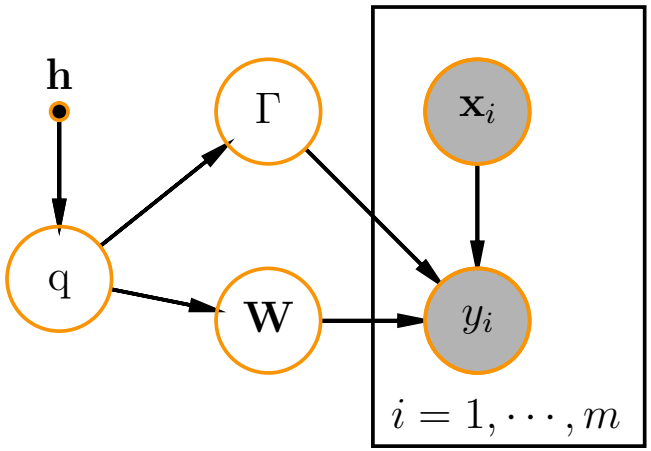


Структурный параметр лежит внутри симплекса

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.



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

Вариационными параметрами модели $\boldsymbol{\theta} \in \mathbb{R}^u$ назовем параметры распределения q, приближающие апостериорное распределение параметров и структур:

$$q \approx p(\mathbf{W}, \mathbf{\Gamma}|\mathbf{X}, \mathbf{y}, \mathbf{h}) = \frac{p(\mathbf{y}|\mathbf{X}, \mathbf{W}, \mathbf{\Gamma}, \mathbf{h})p(\mathbf{W}, \mathbf{\Gamma}|\mathbf{h})}{\int\int_{\mathbf{W}', \mathbf{\Gamma}'} p(\mathbf{y}|\mathbf{X}, \mathbf{W}', \mathbf{\Gamma}', \mathbf{h})p(\mathbf{W}', \mathbf{\Gamma}'|\mathbf{h})d\mathbf{W}'d\mathbf{\Gamma}'}.$$

Оптимизационная задача

Определение. Пусть задано вариационное распределение q с параметрами $\boldsymbol{\theta}$, приближающие апостериорное распределение $p(\mathbf{W}, \boldsymbol{\Gamma}|\mathbf{X}, \mathbf{y}, \mathbf{h})$ параметров и структуры.

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

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

Выбором 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

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

Правдоподобие как статистическая сложность

Статистическая сложность модели $\mathbf{f}$:

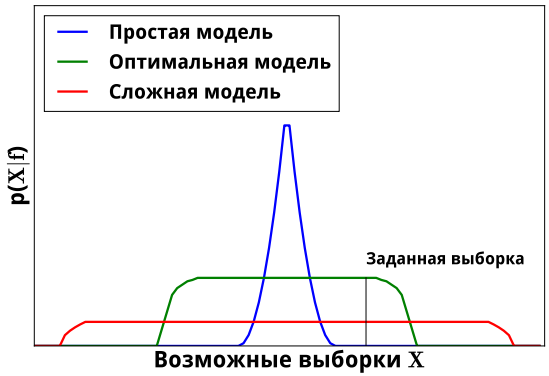
$$\mathrm{MDL}(\mathbf{y}, \mathbf{f}) = -\log\, p(\mathbf{h}) - \log\, \big(p(\mathbf{y}|\mathbf{X}, \mathbf{h})\delta\mathfrak{D}\big)\big),$$

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

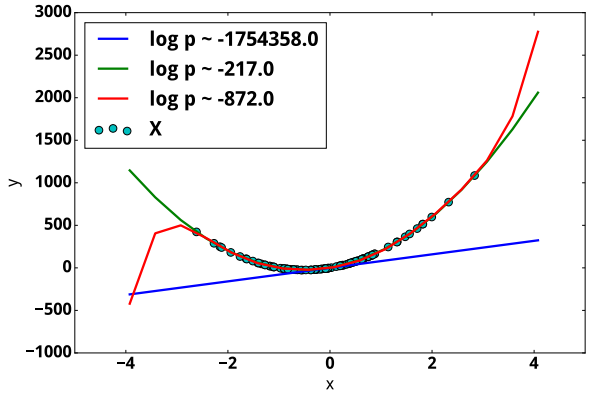
Правдоподобие модели:

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

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



Выбор модели по правдоподобию



Аппроксимация выборки полиномами

Выбор оптимальной модели

Основные проблемы выбора оптимальной модели

- Интеграл правдоподобия $p(\mathbf{y}|\mathbf{X},\mathbf{h})$ невычислим аналитически.
- Задача его оптимизации многоэкстремальна и невыпукла.

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.

Вариационная нижняя оценка правдоподобия

Интеграл правдоподобия невычислим аналитически.
Правдоподобие модели:

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

Пусть $q(\mathbf{W},\boldsymbol{\Gamma}) = q_{\mathbf{W}}(\mathbf{W}) q_{\boldsymbol{\Gamma}}(\boldsymbol{\Gamma})$ — непрерывное распределение, аппроксимирующее апостериорное распределение $p(\mathbf{W},\boldsymbol{\Gamma}|\mathbf{y},\mathbf{X})$.
Получим нижнюю оценку интеграла правдоподобия.

$$\log p(\mathbf{y}|\mathbf{X}) \geq \mathrm{E}_q \log\, p(\mathbf{y}|\mathbf{X},\mathbf{W},\boldsymbol{\Gamma}) - \mathrm{D}_{KL}(p(\mathbf{w},\boldsymbol{\Gamma})||q(\mathbf{W},\boldsymbol{\Gamma})) = \log \hat{p}_{q_{\mathbf{W}} q_{\boldsymbol{\Gamma}}}(\mathbf{y}|\mathbf{X}).$$

Полученная оценка совпадает с интегралом правдоподобия при

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

Общая задача оптимизации

Теорема (будет)

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

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

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

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

$$- c_{\text{comb}} \sum_{p' \in \mathbf{P}} \mathsf{D}_{KL}(\mathbf{\Gamma}|p'),$$

где

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

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

Нижняя вариационная оценка правдоподобия на основе мультистарта

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

Теорема [Бахтеев, 2016]

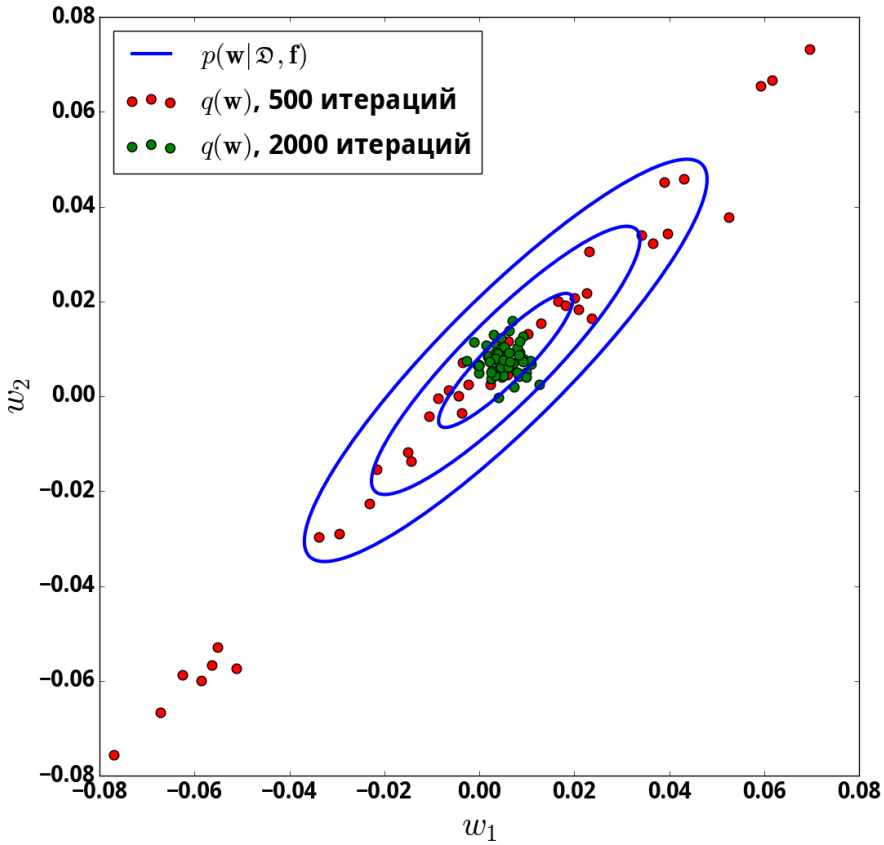
Пусть L — функция потерь, градиент которой — непрерывно-дифференцируемая функция с константой Липшица C.
Пусть $\boldsymbol{\theta} = [\mathbf{W}^1, \ldots, \mathbf{W}^k]$ — начальные приближения оптимизации модели, β — шаг градиентного спуска.
Тогда разность энтропий на смежных шагах оптимизации приближается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

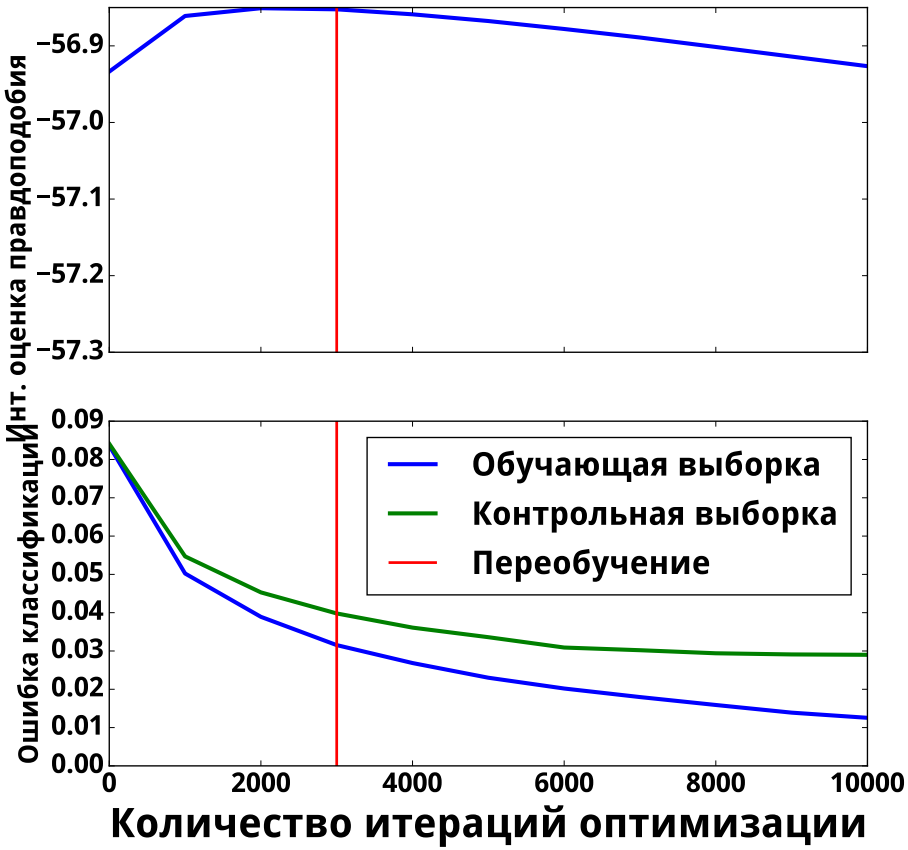
$$\mathsf{E}_{q^\tau_\mathsf{W}}(-\log(q^\tau_\mathsf{W})) - \mathsf{E}_{q^{\tau-1}_\mathsf{W}}(-\log(q^{\tau-1}_\mathsf{W})) \approx \frac{1}{k}\sum_{r=1}^{k}\left(\beta\, Tr[\mathbf{H}(\mathbf{W}^r)] - \beta^2\, Tr[\mathbf{H}(\mathbf{W}^r)\mathbf{H}(\mathbf{w}^r)]\right),$$

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

Градиентный спуск как вариационная оценка правдоподобия модели

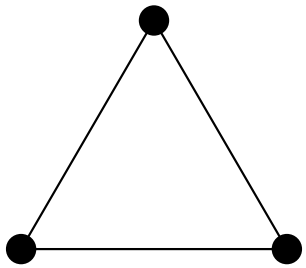
Для вычисления правдоподобия был предложен ряд алгоритмов, основанных на стохастическом градиентном спуске.



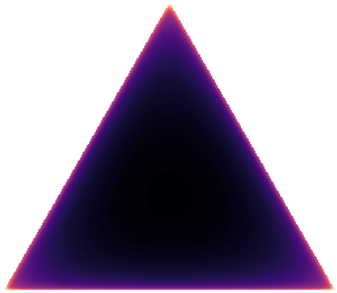


Априорное распределение на структуре модели

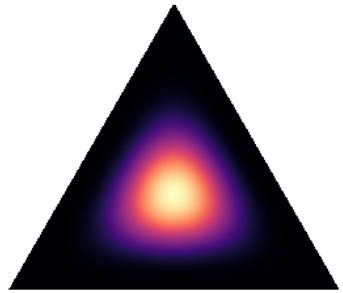
Распределение Дирихле



$\tau \to 0$

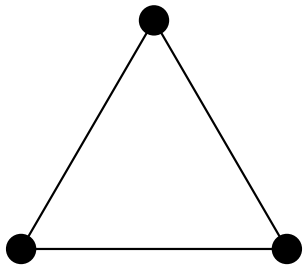


$\tau = 0.995$

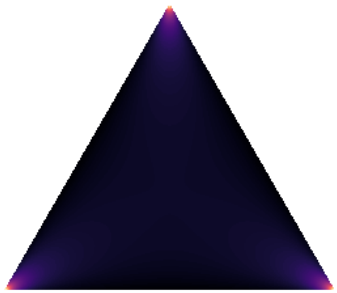


$\tau = 5.0$

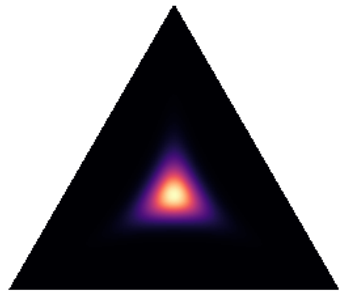
Распределение Гумбель-софтмакс



$\tau \to 0$



$\tau = 0.995$



$\tau = 5.0$

Оператор оптимизации

Определение

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

Оператор стохастического градиентного спуска:

$$\hat{\theta} = T \circ T \circ \cdots \circ T(\theta_0, \mathbf{A}^{-1}, \mathbf{m}) = T^{\eta}(\theta_0, \mathbf{A}^{-1}, \mathbf{m}), \quad \text{где} T(\theta, \mathbf{A}^{-1}, \mathbf{m}) =$$

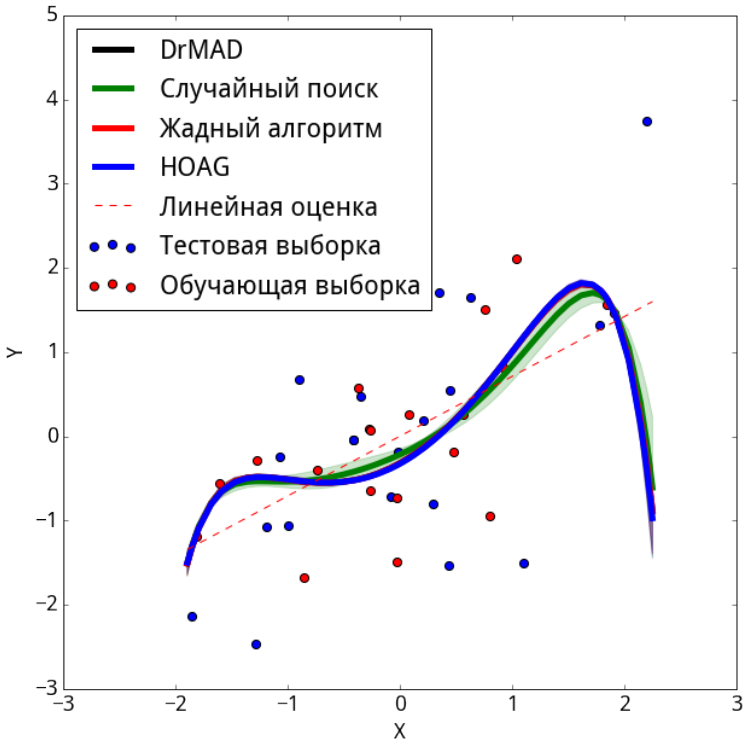
$$= \theta - \beta \nabla L(\theta, \mathbf{A}^{-1}, \mathbf{m})|_{\hat{\mathfrak{D}}},$$

γ — длина шага градиентного спуска, θ_0 — начальное значение параметров θ, $\hat{\mathfrak{D}}$ — случайная подвыборка исходной выборки $\mathfrak{D}$.
Перепишем итоговую задачу оптимизации:

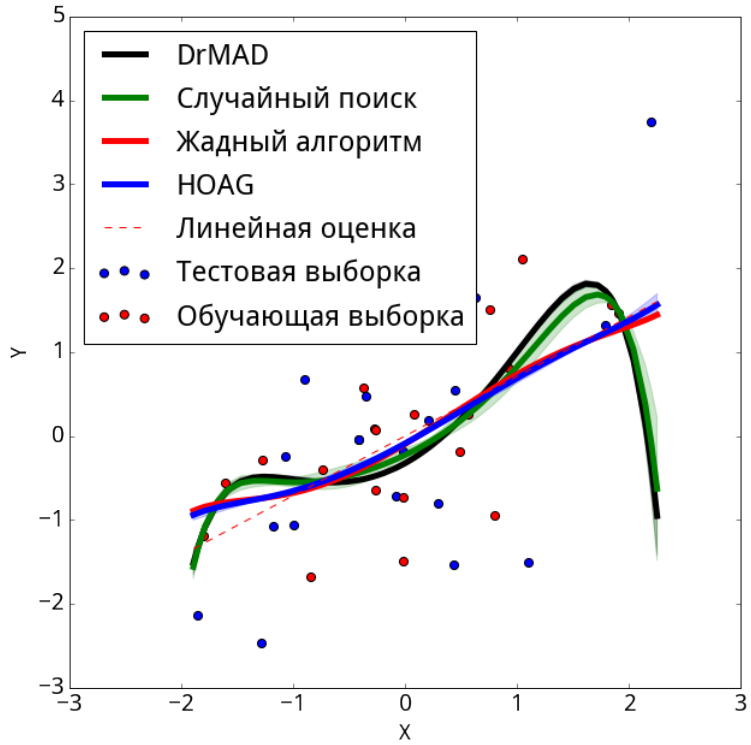
$$\mathbf{h}^* = \arg\max_{\mathbf{h}} Q\left(T^{\eta}(\theta_0, \mathbf{A}^{-1}, \mathbf{m})\right),$$

где θ_0 — начальное значение θ.

Оптимизация гиперпараметров: пример



Кросс-Валидация



Вариационная оценка

Оптимизация правдоподобия модели

Теорема [Бахтеев, 2018].

Пусть существуют параметры распределения $q(\mathbf{W}, \mathbf{\Gamma})$, такие что $\mathrm{D}_{\mathrm{KL}}(q(\mathbf{W}, \mathbf{\Gamma})|p(\mathbf{W}, \mathbf{\Gamma}|\mathbf{y}, \mathbf{X}, \mathbf{A}, \mathbf{m}, c_{\mathrm{temp}})) = 0$.
Тогда двухуровневая задача оптимизация эквивалентна задаче оптимизации правдоподобия модели:

$$\arg\max_{\mathbf{A},\mathbf{m}} p(\mathbf{y}|\mathbf{X}, \mathbf{A}, \mathbf{m}, c_{\mathrm{temp}})$$

при $c_{\mathrm{reg}} = c_{\mathrm{prior}} = c_{\mathrm{train}} > 0, c_{\mathrm{comb}} = 0$.

Параметрическая сложность

Обозначим за $F(c_{\text{reg}}, c_{\text{train}}, c_{\text{prior}}, c_{\text{comb}}, \mathbf{P}, c_{\text{temp}})$ множество экстремумов функции L при решении задачи двухуровневой оптимизации.

Теорема [Бахтеев, 2018].

Пусть $\mathbf{f} \in F(1, 1, c_{\text{prior}}, 0, \varnothing, c_{\text{temp}})$. При устремлении c_{prior} к бесконечности параметрическая сложность модели $\mathbf{f}$ устремляется к нулю.

$$\lim_{c_{\text{prior}} \to \infty} C_{\text{param}}(\mathbf{f}) = 0.$$

Теорема [Бахтеев, 2018].

Пусть $\mathbf{f}_1 \in F(1, 1, c^1_{\text{prior}}, 0, \varnothing, c_{\text{temp}}), \mathbf{f}_2 \in F(1, 1, c^2_{\text{prior}}, 0, \varnothing, c_{\text{temp}})$, $c^1_{\text{prior}} < c^2_{\text{prior}}$. Пусть вариационные параметры моделей $\mathbf{f}_1$ и $\mathbf{f}_2$ лежат в области U, в которой соответствующие функции L и Q являются локально-выпуклыми. Тогда модель $\mathbf{f}_1$ имеет параметрическую сложность, не меньшую чем у $\mathbf{f}_2$.

$$C_{\text{param}}(\mathbf{f}_1) \geq C_{\text{param}}(\mathbf{f}_2).$$

Структурная сложность

Теорема [Бахтеев, 2018].

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 $c_{\text{reg}} > 0, c_{\text{train}} > 0, c_{\text{prior}} > 0$. Пусть $\mathbf{f} \in F(c_{\text{reg}}, c_{\text{train}}, c_{\text{prior}}, 0, \varnothing, c_{\text{temp}})$. Тогда структурная сложность модели $\mathbf{f}$ равняется нулю.

$$C_{\text{struct}}(\mathbf{f}) = 0.$$

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

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

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

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

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

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

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

Теорема

Вариационные распределения $q_{\mathbf{\Gamma}}$ 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 $\mathbf{\Gamma}$.

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.