Байесовский выбор наиболее правдоподобной структуры модели глубокого обучения

О. Ю. Бахтеев

Научный руководитель: д.ф.-м.н. В.В. Стрижов
Московский физико-технический институт (государственный университет)

Интеллектуализация обработки информации
ИОИ-2018
11.10.2018

Выбор структуры модели глубокого обучения

Цель работы:
Развитие теории байесовского выбора модели и исследование свойств методов выбора моделей глубокого обучения.

Задачи:

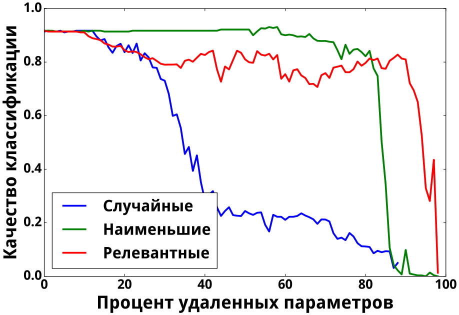
- Предложить алгоритм совместной оптимизации параметров, гиперпараметров и структурных параметров моделей глубокого обучения.
- Предложить метод выбора модели наиболее правдоподобной структуры.
- Исследовать свойства оптимизационных алгоритмов выбора модели.

Основные проблемы

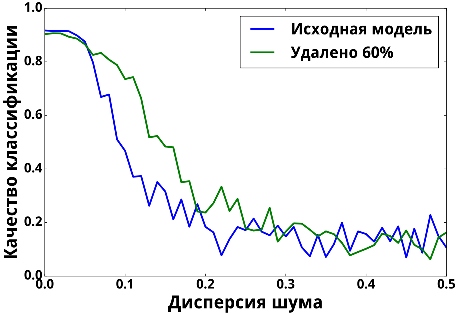
- Многоэкстремальность задачи оптимизации параметров модели.
- Вычислительная сложность оптимизации.
- Большое число параметров и гиперпараметров.

Проблемы оптимизации моделей глубокого обучения

Правдоподобие моделей с избыточным количеством параметров не меняется при удалении параметров.



Избыточность параметров модели



Неустойчивость модели

Глубокое обучение предполагает оптимизацию моделей с заведомо избыточной сложностью.

Задача выбора структуры модели

Однослойная нейросеть:

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

$$\mathbf{f}_1(\mathbf{x}) = \gamma_{0,1}^1 \mathbf{g}_{0,1}^1(\mathbf{x}) + \dots + \gamma_{0,1}^K \mathbf{g}_{0,1}^K(\mathbf{x}) = \gamma_{0,1}^1 \boldsymbol{\sigma}(\mathbf{W}_1^{\mathsf{T}}\mathbf{x}) + \dots + \gamma_{0,1}^K \boldsymbol{\sigma}(\mathbf{W}_K^{\mathsf{T}}\mathbf{x}),$$

где $\mathbf{W} = [\mathbf{W}_0, \mathbf{W}_1, \dots, \mathbf{W}_K]^{\mathsf{T}}$ — матрицы параметров, $\{\mathbf{g}_{0,1}^i\}_{i=1}^K$ — базовые функции скрытого слоя нейросети.

Структурные параметры: $\boldsymbol{\Gamma} = [\gamma_{0,1}]$.
Структура модели задается вершиной K-мерного симплекса.

Задача выбора структуры модели: два скрытых слоя

Двухслойная нейросеть:

$$\mathbf{f}(\mathbf{x}) = \text{softmax}\left(\mathbf{W}^{\mathsf{T}}\mathbf{f}_2(\mathbf{x})\right), \quad f(\mathbf{x}) : \mathbb{R}^n \to [0,1]^{|\mathbb{Y}|}, \quad \mathbf{x} \in \mathbb{R}^n.$$

$$\mathbf{f}_2(\mathbf{x}) = \gamma_{1,2}^1 \mathbf{g}_{1,2}^1(\mathbf{f}_1(\mathbf{x})) + \dots + \gamma_{1,2}^K \mathbf{g}_{1,2}^K(\mathbf{f}_1(\mathbf{x})) = \gamma_{1,2}^1 \boldsymbol{\sigma}(\mathbf{W}_{K+1}^{\mathsf{T}}\mathbf{f}_1(\mathbf{x})) + \dots + \gamma_{1,2}^K \boldsymbol{\sigma}(\mathbf{W}_{2K}^{\mathsf{T}}\mathbf{f}_1(\mathbf{x})),$$

$$\mathbf{f}_1(\mathbf{x}) = \gamma_{0,1}^1 \mathbf{g}_{0,1}^1(\mathbf{x}) + \dots + \gamma_{0,1}^K \mathbf{g}_{0,1}^K(\mathbf{x}) = \gamma_{0,1}^1 \boldsymbol{\sigma}(\mathbf{W}_1^{\mathsf{T}}\mathbf{x}) + \dots + \gamma_{0,1}^K \boldsymbol{\sigma}(\mathbf{W}_K^{\mathsf{T}}\mathbf{x}),$$

где $\mathbf{W} = [\mathbf{W}_0, \mathbf{W}_1, \dots, \mathbf{W}_{2K}]^{\mathsf{T}}$ — матрицы параметров, $\{\mathbf{g}_{0,1}^i, \mathbf{g}_{1,2}^i\}_{i=1}^K$ — базовые функции скрытых слоев нейросети.

Структурные параметры: $\boldsymbol{\Gamma} = [\gamma_{0,1}, \gamma_{1,2}]$.
Структура модели задается вершинами двух K-мерных симплексов.

Исследование основывается на следующих работах

- Graves A. Practical variational inference for neural networks // Advances in Neural Information Processing Systems. – 2011.
- Maclaurin D., Duvenaud D., Adams R. Gradient-based hyperparameter optimization through reversible learning // International Conference on Machine Learning. – 2015.
- Hanxiao L. et al., DARTS: Differentiable Architecture Search // arXiv preprint:1806.09055. - 2018.
- О. Ю. Бахтеев, В. В. Стрижов. Выбор моделей глубокого обучения субоптимальной сложности // Автоматика и телемеханика, 2018.

Графовое представление модели глубокого обучения

Определение

Задан граф V, E.
Для каждого ребра $(j,k) \in E$ определен вектор **базовых функций** $\mathbf{g}_{j,k}$ мощностью $K_{j,k}$.
Граф V, E со множеством функций $\{\mathbf{g}_{j,k}\}_{(j,k)\in E}$ называется **семейством моделей**, если функция, задаваемая рекурсивно как

$$\mathbf{f}_j(\mathbf{x}) = \sum_{k \in \mathrm{Adj}(v_j)} \langle \gamma_{j,k}, \mathbf{g}_{j,k} \rangle (\mathbf{f}_k(\mathbf{x})), \quad \mathbf{f}_0(\mathbf{x}) = \mathbf{x},$$

является дифференцируемой по параметрам $\mathbf{W}$ функцией из $\mathbb{R}^n$ во множество $\mathbb{Y}$ при любых значениях векторов γ.

Модель задается параметрами подмоделей $\{\mathbf{f}_j\}_{j=1}^{|V|}$ и структурными параметрами γ.

Параметры модели W — конкатенация параметров всех подмоделей $\{\mathbf{f}_j\}_{j=1}^{|V|}$.
Структура модели Г — конкатенация структурных параметров γ.

Эксплуатационные критерии качества модели

Точность $S(\mathbf{W}, \boldsymbol{\Gamma})$ модели $\mathbf{f}(\mathbf{x})$ — величина ошибки на контрольной выборке.

Устойчивость $\eta(\mathbf{W})$ модели — число обусловленности матрицы ковариации $\mathbf{A}$:

$$\eta(\mathbf{W}) = \frac{\lambda_{\text{max}}}{\lambda_{\text{min}}} \quad \text{при гипотезе } \mathbf{W} \sim \mathcal{N}(\mathbf{0}, \mathbf{A}^{-1}),$$

λ_{max} — максимальное, а λ_{min} — минимальное собственные числа матрицы $\mathbf{A}$.

Статистические критерии качества модели

Параметрическая сложность — наименьшая дивергенция между априорным распределением параметров и апостериорным распределением параметров:

$$C_{\text{param}} = \min_{\mathbf{A},\mathbf{m}} \mathrm{D}_{\mathrm{KL}}\left(p(\mathbf{W}, \boldsymbol{\Gamma}|\mathbf{y}, \mathbf{X})||p(\mathbf{W}, \boldsymbol{\Gamma}||\mathbf{A}, \mathbf{m})\right).$$

где $\mathbf{m}$ — гиперпараметры априорного распределения структуры модели.
Структурная сложность модели — энтропия апостериорного распределения структуры модели:

$$C_{\text{struct}} = -\mathsf{E}_p \log p(\boldsymbol{\Gamma}|\mathbf{y}, \mathbf{X}).$$

Правдоподобие как статистическая сложность

Предлагается метод выбора модели, учитывающий все перечисленные критерии качества.

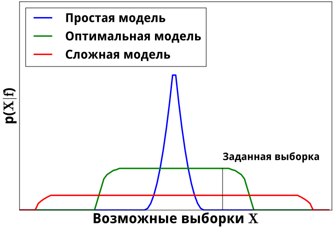
Статистическая сложность модели $\mathbf{f}$:

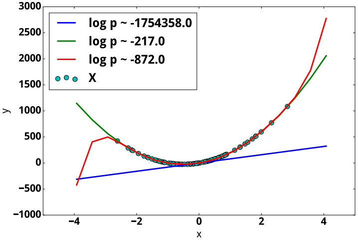
$$\mathrm{MDL}(\mathbf{y}, \mathbf{f}) = -\log\, p(\mathbf{f}) - \log\, \big(p(\mathbf{y}|\mathbf{X}, \mathbf{f})\delta\mathfrak{D}\big),$$

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

Правдоподобия модели:

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





Выбор оптимальной модели

Основные проблемы выбора оптимальной модели

- Интеграл правдоподобия $p(\mathbf{y}|\mathbf{X}, \mathbf{f})$ (далее — $p(\mathbf{y}|\mathbf{X})$) 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:

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

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

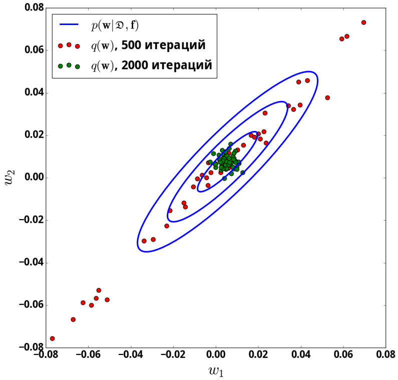
$$\log p(\mathbf{y}|\mathbf{X}) \geq \mathsf{E}_q \log p(\mathbf{y}|\mathbf{X},\mathbf{W},\mathbf{\Gamma}) - \mathsf{D}_{KL}(p(\mathbf{w},\mathbf{\Gamma})||q(\mathbf{W},\mathbf{\Gamma})) = \log \hat{p}_{q_{\mathbf{W}} q_{\mathbf{\Gamma}}}(\mathbf{y}|\mathbf{X}).$$

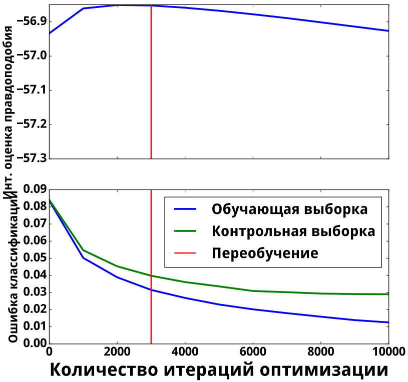
Полученная оценка совпадает с интегралом правдоподобия при

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

Градиентный спуск как вариационная оценка правдоподобия модели

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





Выбор вариационного распределения q

Вариационное распределение параметров $q_{\mathbf{W}}$:

$$q_{\mathbf{W}} = \mathcal{N}(\boldsymbol{\mu}_q, \mathbf{A}_q^{-1}).$$

Вариационное распределение структуры $q_{\boldsymbol{\Gamma}}$:

$$q_{\boldsymbol{\Gamma}}(\mathbf{m}_q, c_{\text{temp}}) = \prod_{(j,k)\in E} q_{\gamma_{j,k}}(c_{\text{temp}}, \mathbf{m}_q^{j,k}) \sim \mathcal{GS}(c_{\text{temp}}, \mathbf{m}_q^{j,k}), \quad |\mathbf{m}_q^{j,k}| = K^{j,k}$$

Свойства:

- $\lim_{c_{\text{temp}}\to\infty} \mathcal{GS}(c_{\text{temp}}) = \mathcal{U}(\Delta^{K^{j,k}-1})$.
- При $c_{\text{temp}} \to 0$ распределение вырождается в дискретное распределение.
- Существует вычислительно устойчивый метод вычисления градиента по параметрам распределения от реализации случайной величины.

Оптимизация параметров вариационного распределения

Параметры вариационного распределения $q(\mathbf{W}, \mathbf{\Gamma}) = q_{\mathbf{W}}(\mathbf{W})q_{\mathbf{\Gamma}}(\mathbf{\Gamma})$ оптимизируем:

$$L = \mathsf{E}_q \log p(\mathbf{y}|\mathbf{X}, \mathbf{W}, \mathbf{\Gamma}.\mathbf{A}^{-1}, c_{\text{temp}}) - c_{\text{reg}} \mathrm{D}_{KL}\left(p(\mathbf{w}, \mathbf{\Gamma}|\mathbf{A}^{-1}, \mathbf{m}, c_{\text{temp}})||q(\mathbf{W}), q(\mathbf{\Gamma})\right) \to \max_{\mathbf{A}_q, \boldsymbol{\mu}_q, \mathbf{m}_q}.$$

Теорема [Бахтеев, 2018].

Пусть $c_{\text{reg}} > 0$. Тогда функция L сходится по вероятности к вариационной нижней оценке логарифма правдоподобия $\log p(\mathbf{y}|\mathbf{x})$ для случайной подвыборки $\mathfrak{D}$ мощностью $c_{\text{reg}}m$:

$$L \to^p c_{\text{reg}} m \log \hat{p}_{q_{\mathbf{W}} q_{\mathbf{\Gamma}}}(\mathbf{y}|\mathbf{X}).$$

Теорема [Бахтеев, 2018].

Для любых значений ковариационных матриц $\mathbf{A}, \mathbf{A}_q$, любого вектора $\boldsymbol{\mu}_q$ и любой точки $\mathbf{m}$ из декартового произведения симплексов структуры $\mathbf{\Gamma}$ существует такая точка $\mathbf{m}_q^1$ на вершинах симплексов структуры $\mathbf{\Gamma}$, что для любой точки $\mathbf{m}_q^2$ внутри симплексов справедливо выражение:

$$\lim_{c_{\text{temp}} \to 0} \frac{\log \hat{p}_{q_{\mathbf{W}} q_{\mathbf{\Gamma}}^2}(\mathbf{y}|\mathbf{X})}{\log \hat{p}_{q_{\mathbf{W}} q_{\mathbf{\Gamma}}}(\mathbf{y}|\mathbf{X})} = -\infty, \quad \text{где} q_{\mathbf{\Gamma}}^1 = q_{\mathbf{\Gamma}}(\mathbf{m}_q^1, c_{\text{temp}}), \quad q_{\mathbf{\Gamma}}^2 = q_{\mathbf{\Gamma}}^1(\mathbf{m}_q^2, c_{\text{temp}})).$$

Оптимизация параметров априорного распределения

Гиперпараметры $\mathbf{A}, \mathbf{m}$ оптимизируем:

$$Q = c_{\text{train}} \mathsf{E}_q \log p(\mathbf{y}|\mathbf{X}, \mathbf{W}, \mathbf{\Gamma}.\mathbf{A}^{-1}, c_{\text{prior}}) - c_{\text{prior}} \mathsf{D}_{KL}(p(\mathbf{W}, \mathbf{\Gamma}|\mathbf{A}^{-1}, \mathbf{m}, c_{\text{temp}})||q(\mathbf{W}, \mathbf{\Gamma})) -$$

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

где $\mathbf{P}$ — множество (возможно пустое) распределений на структуре модели.

- c_{train} — коэффициент правдоподобия выборки;
- c_{prior} — коэффициент регуляризации модели;
- c_{comb} — коэффициент перебора структуры.

Общая задача оптимизации

Общая задача оптимизации — двухуровневая:

$$\hat{\mathbf{A}}, \hat{\mathbf{m}} = \arg\max_{\mathbf{A},\mathbf{m}} Q =$$

$$= c_{\text{train}} \mathsf{E}_{\hat{q}} \log\, p(\mathbf{y}|\mathbf{X}, \mathbf{W}, \mathbf{\Gamma}.\mathbf{A}^{-1}, c_{\text{prior}}) - c_{\text{prior}} \mathrm{D}_{KL}(p(\mathbf{W}, \mathbf{\Gamma}|\mathbf{A}^{-1}, \mathbf{m}, c_{\text{temp}})||\hat{q}(\mathbf{W}, \mathbf{\Gamma})) -$$

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

где

$$\hat{q} = \arg\max_{q} L = \mathsf{E}_q \log\, p(\mathbf{y}|\mathbf{X}, \mathbf{W}, \mathbf{\Gamma}.\mathbf{A}^{-1}, c_{\text{temp}}) - c_{\text{reg}} \mathrm{D}_{KL}(p(\mathbf{w}, \mathbf{\Gamma}|\mathbf{A}^{-1}, \mathbf{m}, c_{\text{temp}})||q(\mathbf{W}), q(\mathbf{\Gamma}))$$

Оператор оптимизации

Обозначим за $\mathbf{h}$ гиперпараметры $\mathbf{A}, \mathbf{m}$.
Обозначим за $\boldsymbol{\theta}$ параметры распределений $q_{\mathbf{w}}, q_{\mathbf{\Gamma}}$.

Определение

Оператором T назовем оператор стохастического градиентного спуска, производящий η шагов оптимизации:

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

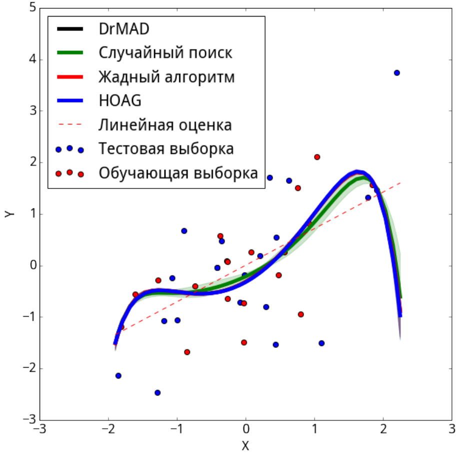
γ — длина шага градиентного спуска, $\boldsymbol{\theta}_0$ — начальное значение параметров $\boldsymbol{\theta}$, $\hat{\mathfrak{D}}$ — случайная подвыборка исходной выборки $\mathfrak{D}$.

Перепишем итоговую задачу оптимизации:

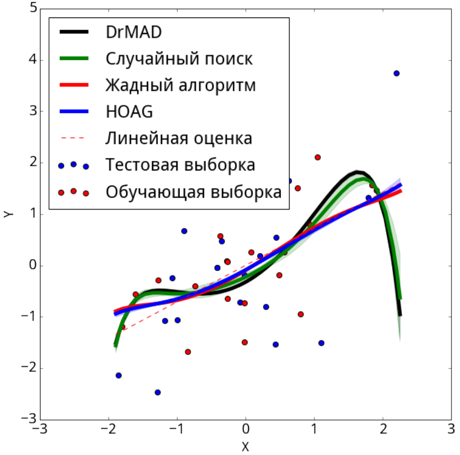
$$\hat{\mathbf{h}} = \arg\max_{\mathbf{h}} Q\left(T^{\eta}(\boldsymbol{\theta}_0, \mathbf{A}^{-1}, \mathbf{m})\right),$$

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

Оптимизация гиперпараметров: пример



Кросс-Валидация



Вариационная оценка

Оптимизация правдоподобия модели

Теорема [Бахтеев, 2018].

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

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

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

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 $\mathbf{\Gamma}$.

Заключение

- Предложен алгоритм оптимизации параметров, гиперпараметров и структурных параметров моделей глубокого обучения.
- Предложен метод выбора модели наиболее правдоподобной структуры, обобщающий различные алгоритмы оптимизации:
  - оптимизация правдоподобия;
  - последовательное увеличение сложности модели;
  - последовательное снижение сложности модели;
  - полный перебор вариантов структуры модели.
- Проведено исследование свойства оптимизационных алгоритмов выбора модели.