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

Основные проблемы

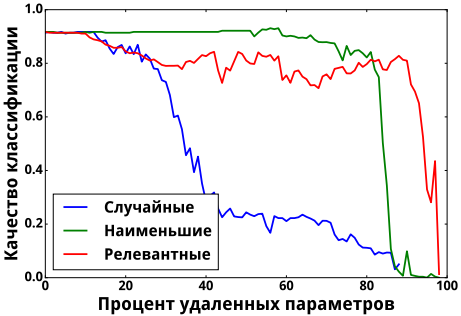
- Многоэкстремальность задачи оптимизации параметров модели.
- Вычислительная сложность оптимизации.
- Большое число параметров и гиперпараметров.

Исследование основывается на следующих работах

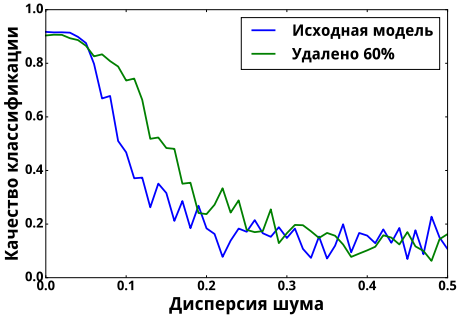
- Graves A. Practical variational inference for neural networks //Advances in Neural Information Processing Systems. – 2011
- Maclaurin D., Duvenaud D., Adams R. Gradient-based hyperparameter optimization through reversible learning //International Conference on Machine Learning. – 2015
- Luketina J. et al. Scalable gradient-based tuning of continuous regularization hyperparameters //International Conference on Machine Learning. - 2016
- J. Fu et al., DrMAD: Distilling Reverse-Mode Automatic Differentiation for Optimizing Hyperparameters of Deep Neural Networks // IJCAI - 2016
- Pedregosa F. Hyperparameter optimization with approximate gradient //International Conference on Machine Learning. – 2016. –

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:

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

$$f_1(\mathbf{x}) = \gamma_{0,1}^1 \mathbf{g}_{0,1}^1(\mathbf{x}) + \cdots + \gamma_{0,1}^K \mathbf{g}_{0,1}^K(\mathbf{x}) = \gamma_{0,1}^1 \boldsymbol{\sigma}$$

где $\mathbf{W}_1, \ldots, \mathbf{W}_K$ — матрицы параметров, $\{\mathbf{g}_{0,1}^i\}_{i=1}^K$ — базовые функции для скрытого слоя нейросети.

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

Задача выбора структуры модели: два скрытых слоя

Двухслойная нейросеть:

$$\mathbf{f}(\mathbf{x}) = \text{softmax}\left(\mathbf{W}^{\mathsf{T}}\mathbf{f}_2(\mathbf{x})\right), \quad f(\mathbf{x}) : \mathbb{R}^n \to 2^{|\mathbb{Y}|}, \quad \mathbf{x} \in \mathbb{R}^n.$$

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

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

где $\mathbf{W}_1, \ldots, \mathbf{W}_{2K}$ — матрицы параметров, $\{\mathbf{g}^i_{0,1}, \mathbf{g}^i_{1,2}\}_{i=1}^K$ — базовые функции для скрытых слоев нейросети.

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

Картинка: связь параметров со структурой.

Семейство моделей

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 $\mathfrak{G} = \{\mathbf{g}_{j,k}\}_{(j,k)\in E}$ называется **семейством моделей**, если функция, задаваемая рекурсивно как

$$\mathbf{f}_j(\mathbf{x}) = \sum_{k\in \mathrm{Adj}(v_j)} < \boldsymbol{\gamma}_{j,k}, \mathbf{g}_{j,k} > (f_k(\mathbf{x})), \quad \mathbf{f}_0(\mathbf{x}) = \mathbf{x},$$

является непрерывной дифференцируемой функцией из $\mathbb{R}^n$ во множество $\mathbb{Y}$ при любых значениях векторов $\boldsymbol{\gamma}$.
Модель определяется параметрами подмоделей f_j и структурными параметрами $\boldsymbol{\gamma}$.
Обозначим за вектор **параметров модели W** конкатенацию параметров всех подмоделей $\{f_j\}_{j=1}^{|V|}$.
Обозначим за вектор **параметров модели Г** конкатенацию структурных параметров $\boldsymbol{\gamma}$.

Критерии качества модели

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

Устойчивость модели $\mathbf{f}(\mathbf{x})$ — число обусловленности матрицы $\mathbf{A}$:

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

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

Статистические критерии качества модели

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 $\mathfrak{D}$.

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

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

где $\mathbf{m}$ — гиперпараметры априорного распределения структуры модели.

TODO: надо ли про bits-back?

Структурная сложность модели — энтропия апостериорного распределения структуры модели:

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

В данной работе предлагается метод оптимизации модели, учитывающий все перечисленные критерии качества модели.

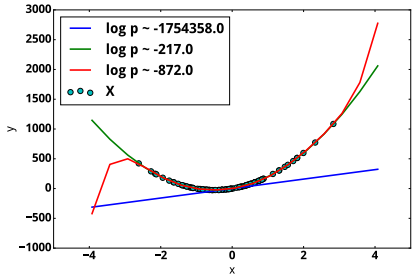
Правдоподобие модели

Пусть заданы априорное распределение параметров и структуры $p(\mathbf{W}, \mathbf{\Gamma})$.

Модель $\mathbf{f}$ оптимальна, если достигается максимум **правдоподобия модели**:

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

Схема выбора модели по правдоподобию



Пример: полиномы

Статистические критерии качества модели

Выбор оптимальной модели

Основные проблемы выбора оптимальной модели

- Интеграл правдоподобия невычислим аналитически.
- Задача оптимизации многоэкстремальна и невыпукла.

Требуется

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

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

Вариационная нижняя оценка правдоподобия

Интеграл правдоподобия невычислим аналитически.
Правдоподобие модели:

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

Получим нижнюю оценку интеграла правдоподобия.
Пусть q — непрерывное распределение.

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

Полученная оценка совпадает с интегралом правдоподобия при

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

Распределение на структуре

Пусть для каждого ребра (j,k) задан нормированный положительный вектор $\gamma_{j,k} \in \mathbb{R}_+^{|K_{j,k}|}$, определяющий веса базовых функций из $\mathbf{g}(j,k)$.

Будем считать, что вектор $\gamma_{j,k}$ распределен по распределению Gumbel-Softmax(GS):

$$p(\gamma) = (K_{j,k}-1)! {c_{\text{temp}}}^{K_{j,k}-1} \left(\prod_{h=1}^{K_{j,k}-1} \alpha_h \gamma_h^{-c_{\text{temp}}-1} \right) \left(\sum_{h=1}^{u} \alpha_h \gamma_h^{-c_{\text{temp}}} \right),$$

где $\mathbf{m} = [\alpha_1, \ldots, \alpha_{K_{j,k}}]$ — параметры сдвига распределения, c_{temp} — температура распределения.

Вариационный вывод: распределение структурных параметров

Для реализации h-й компоненты случайной величины $\boldsymbol{\gamma}$ справедлива следующая формула:

$$\hat{\boldsymbol{\gamma}}^h = \exp\left(\log\left(\alpha_h + \mathrm{Gum}_h\right) c_{\mathrm{temp}}^{-1}\right) \sum_{l=1}^{K_{j,k}} \exp\left(\log\left(\alpha_l + \mathrm{Gum}_l\right) c_{\mathrm{temp}}^{-1}\right),$$

где $\mathrm{Gum} \sim -\log(-\log \mathcal{U}(0,1))$.
Для всех элементов структуры $\boldsymbol{\gamma}_{j,k}$ положим:

$$p(\boldsymbol{\gamma}_{j,k}) = GS([\alpha_1, \ldots, \alpha_{K_{j,k}}, c_{\mathrm{temp}}]), \quad q(\boldsymbol{\gamma}_{j,k}) = GS([\hat{\alpha}_1, \ldots, \hat{\alpha}_{K_{j,k}}, \hat{c}_{\mathrm{temp}}]),$$

где $\hat{\alpha}_1, \ldots, \hat{\alpha}_{K_{j,k}}, \hat{c}_{\mathrm{temp}}$ — параметры вариационного распределения.

Вариационный вывод: распределение параметров

Пусть $\mathbf{W} \sim \mathcal{N}(\mathbf{0}, \mathbf{A}^{-1})$ и структура модели $\mathbf{\Gamma}$ определена однозначно.
Пусть $q_{\mathbf{W}} = \mathcal{N}(\boldsymbol{\mu}_q, \mathbf{A}_q^{-1}), \quad \boldsymbol{\theta} = [\boldsymbol{\mu}_q, \mathbf{A}_q^{-1}]$.
Тогда вариационная оценка имеет вид:

$$\int_{\mathbf{W}} q(\mathbf{W}) \log p(\mathfrak{D}, \mathbf{W}, \mathbf{A}^{-1}) d\mathbf{W} - D_{\mathrm{KL}}\left(q_{\mathbf{W}}(\mathbf{W}) || p(\mathbf{W}|\mathbf{A}^{-1})\right) \simeq$$

$$\sum_{i=1}^{m} \log p(\mathbf{x}_i|\mathbf{W}_i) - D_{\mathrm{KL}}\left(q_{\mathbf{W}}(\mathbf{W}) || p(\mathbf{W}|\mathbf{A}^{-1})\right) = -L(\boldsymbol{\theta}, \mathbf{A}^{-1}, \mathfrak{D}),$$

где $\mathbf{W}_i \sim q_{\mathbf{W}}$.
Дивергенция $D_{\mathrm{KL}}\left(q_{\mathbf{W}}(\mathbf{w}) || p(\mathbf{W}|\mathbf{A}^{-1})\right)$ вычисляется аналитически:

$$D_{\mathrm{KL}}\left(q_{\mathbf{W}}(\mathbf{W}) || p(\mathbf{w}|\mathbf{A}^{-1})\right) = \frac{1}{2}\left(\mathrm{tr}(\mathbf{A}\mathbf{A}_q^{-1}) + \boldsymbol{\mu}_q^{\mathsf{T}}\mathbf{A}\boldsymbol{\mu}_q - n + \ln |\mathbf{A}^{-1}| - \ln |\mathbf{A}_q^{-1}|\right).$$

Вариационная оценка на основе мультистарта

$$\log p(\mathbf{y}|\mathbf{X},\mathbf{A}) \geq \mathrm{E}_{q(\mathbf{W})}\log\, p(\mathbf{y},\mathbf{W}|\mathbf{X},\mathbf{A}^{-1}) - \mathrm{E}_{q_\mathbf{W}}(-\log(q_\mathbf{W})).$$

Теорема [Бахтеев, 2016]. Пусть L — функция потерь, градиент которой — непрерывно-дифференцируемая функция с константой Липшица C. Пусть $\boldsymbol{\theta} = [\mathbf{W}^1, \ldots, \mathbf{W}^k]$ — начальные приближения оптимизации модели. Пусть β — шаг градиентного спуска, такой что:

- $\beta < \frac{1}{C}$,
- $\beta^{(-1)} > \max_{r\in\{1,\ldots,k\}} \lambda_{\max}(\mathbf{H}(\mathbf{W}^r))$.

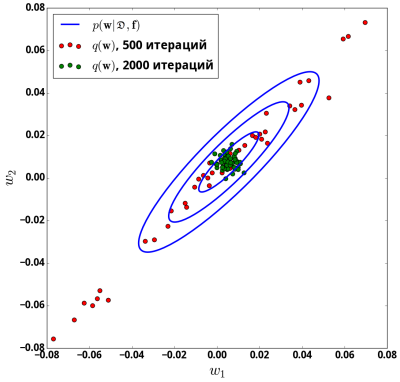
Тогда

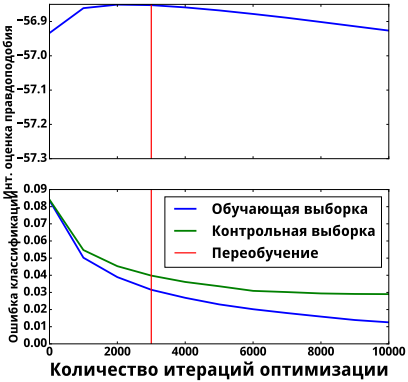
$$\mathrm{E}_{q^\tau_\mathbf{W}}(-\log(q^\tau_\mathbf{W})) - \mathrm{E}_{q^{\tau-1}_\mathbf{W}}(-\log(q^{\tau-1}_\mathbf{W})) \sim \frac{1}{k}\sum_{r=1}^{k}\left(\beta\, Tr[\mathbf{H}(\mathbf{W}^r)] - \beta^2\, Tr[\mathbf{H}(\mathbf{W}^r)\mathbf{H}(\mathbf{w}^r)]\right) + o_{\beta\to 0}(1),$$

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

Вариационная оценка с использованием градиентного спуска

Максимизация вариационной оценки эквивалентна минимизации $\mathrm{D}_{\mathrm{KL}}(q(\mathbf{W})||p(\mathbf{W}|\mathfrak{D}, \mathbf{A}^{-1}))$.
Градиентный спуск не минимизирует $\mathrm{D}_{\mathrm{KL}}(q(\mathbf{W})||p(\mathbf{W}|\mathfrak{D}, \mathbf{A}^{-1}))$.





Оптимизация параметров вариационного распределения

Оптимизацию параметров вариационного распределения $q(\mathbf{W}, \mathbf{\Gamma}) = q_{\mathbf{W}}(\mathbf{W}) q_{\mathbf{\Gamma}}(\mathbf{\Gamma})$ будем проводить по следующему функционалу:

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

Теорема

Пусть $c_{\text{reg}} > 0$. Тогда функция L сходится по вероятности к вариационной нижней оценке правдоподобия для подвыборки $\mathfrak{D}$ мощностью $c_{\text{reg}} m$:

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

Теорема

Для любых значений $\mathbf{A}, \mathbf{m}$ и вариационных параметров $q_{\mathbf{W}}$ существует точка на вершинах симплексов структуры, определяющая распределение $q_{\mathbf{\Gamma}}$, что для любой точки внутри симплексов, определяющей распределение $q'_{\mathbf{\Gamma}}$ справедливо выражение:

$$\lim_{c_{\text{temp}} \to 0} \frac{\log \hat{p}'_q(\mathbf{y}|\mathbf{X})}{\log \hat{p}_q(\mathbf{y}|\mathbf{X})} = -\infty.$$

Оптимизация параметров априорного распределения

Оптимизацию параметров априорного распределения будем проводить по следующему функционалу:

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

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

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

Общая задача оптимизации

Общая задача оптимизации — двухуровневая:

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

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

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

где

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

Оператор оптимизации

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

Определение

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

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

где

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

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

Оптимизация гиперпараметров

Перепишем итоговую задачу оптимизации:

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

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

Утверждение, Luketina et al., 2016

Пусть функции L и Q являются дважды дифференцируемыми и выпуклыми. Пусть гессиан функции функции L аппроксимируется единичной матрицей:

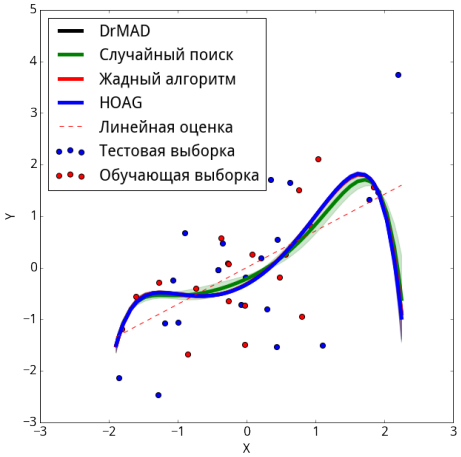
$$\mathbf{H}(L, \boldsymbol{\theta}) \approx \mathbf{I}.$$

Тогда допустима следующая оптимизация гиперпараметров:

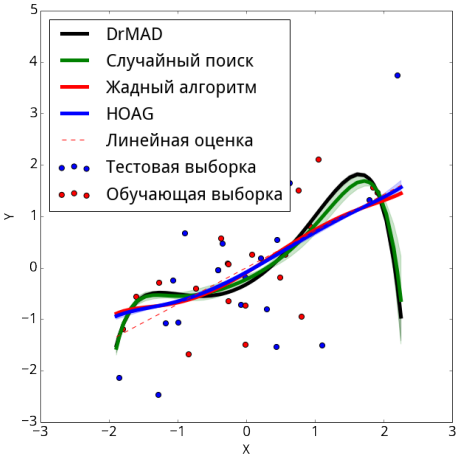
$$\mathbf{h}' = \mathbf{h} - \beta^h \nabla_{\mathbf{h}} Q(T(\boldsymbol{\theta}), \mathbf{h}),$$

где β^h — шаг оптимизации гиперпараметров.

Оптимизация гиперпараметров: пример



Кросс-Валидация



Вариационная оценка

Оптимизация правдоподобия модели

Теорема

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

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.

Параметрическая сложность

TODO: тип сходимости

Теорема

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

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

Теорема

Пусть $\mathbf{f}_1 \in F(1, 1, c_{\text{prior}}, 0, \{\}, c_{\text{temp}}), \mathbf{f}_2 \in F(1, 1, c_{\text{prior}}, 0, \{\}, c'_{\text{temp}})$, $c_{\text{prior}} < c'_{\text{prior}}$.
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 большую чем у $\mathbf{f}_2$.

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

Структурная сложность

Теорема

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

$$p(\gamma_{i,j}) = GS(\alpha_1, \ldots, \alpha_{K_{i,j}}, c_{\text{temp}}).$$

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

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

Теорема

Пусть $\mathbf{f}_1 \in F(c_{\text{reg}}, c_{\text{train}}, c_{\text{prior}}, 0, \{\}, c_{\text{temp}}), \mathbf{f}_2 = \in \lim_{c_{\text{temp}'} \to \infty} F(c_{\text{reg}}, c_{\text{train}}|, c_{\text{prior}}, 0, \{\}, c_{\text{temp}'})$. 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:

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

Полный перебор

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

$$p(\gamma_{i,j}) = GS(\alpha_1, \ldots, \alpha_{K_{i,j}}, c_{\text{temp}}).$$

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

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

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

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

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

Теорема

Вариационные распределения структур $q_{\mathbf{\Gamma}}$ последовательности вырождаются в распределения вида $\delta(\hat{\Gamma})$, где $\hat{\Gamma}$ — точка на декартовом произведении вершин симплексов структуры модели. Вариационные распределения последовательности проходят все возможные комбинации структур модели.

Заключение

- Предложен алгоритм оптимизации параметров, гиперпараметров и структурных параметров моделей глубокого обучения.
- Предложен метод выбора модели наиболее правдоподобной структуры.
- Проведено исследование свойства оптимизационных алгоритмов выбора модели.