

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

Задача выбора структуры модели

Рассматривается задача регрессии или классификации:

$$f(\mathbf{x}) : \mathbb{R}^n \rightarrow \mathbb{Y}, \quad \mathbf{x} \in \mathbf{X}, |\mathbf{X}| = m.$$

Рассмотрим задачу выбора структуры для однослойной нейросети в задаче классификации:

$$f(\mathbf{x}) = \text{softmax}(\mathbf{W}^T \mathbf{f}_1(\mathbf{x})),$$

$$\mathbf{f}_1(\mathbf{x}) = \gamma_{0,1}^1 \mathbf{g}_{0,1}^1(\mathbf{x}) + \cdots + \gamma_{0,1}^K \mathbf{g}_{0,1}^K(\mathbf{x}) = \gamma_{0,1}^1 \tanh(\mathbf{W}_1^T \mathbf{x}) + \cdots + \gamma_{0,1}^K \tanh(\mathbf{W}_K^T \mathbf{x}),$$

где $\mathbf{W}_1, \dots, \mathbf{W}_K$ — матрицы одинаковой размерности с различным количеством нулевых строк, $\{\mathbf{g}_{0,1}^i\}_{i=1}^K$ — базовые функции для скрытого слоя нейросети.

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

Задача выбора структуры модели: два скрытых слоя

Рассмотрим задачу выбора структуры для однослойной нейросети в задаче классификации:

$$f(\mathbf{x}) = \text{softmax}(\mathbf{W}^T \mathbf{f}_2(\mathbf{x})),$$

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

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

где $\mathbf{W}_1, \dots, \mathbf{W}_{2K}$ — матрицы одинаковой размерности с различным количеством нулевых строк, $\{\mathbf{g}_{0,1}^i, \mathbf{g}_{1,2}^i\}_{i=1}^K$ — базовые функции для скрытого слоя нейросети.

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

Пример для более сложной структуры. Сложная нейросеть.

Выбор структуры модели: базовые определения

- **Модель глубокого обучения** — суперпозиция дифференцируемых функций.
- Качество модели определяется **параметрами** модели.
- Оптимизация параметров определяется **гиперпараметрами и структурными параметрами** модели.
- **Гиперпараметры** — параметры распределения параметров модели.
- **Структурные параметры** — параметры, определяющие структуры модели.

Выбор структуры модели глубокого обучения

Цель работы:

Развитие теории исследования метода байесовского выбора модели глубокого обучения

Задачи:

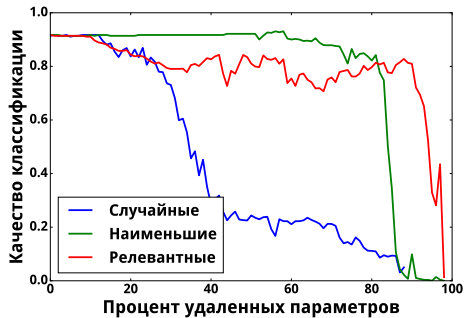
- Предложить алгоритм оптимизации параметров, гиперпараметров и структурных параметров моделей глубокого обучения.
- Предложить метод выбора наиболее правдоподобной структурной модели.
- Исследовать свойства оптимизационных алгоритмов выбора модели.

Основные проблемы

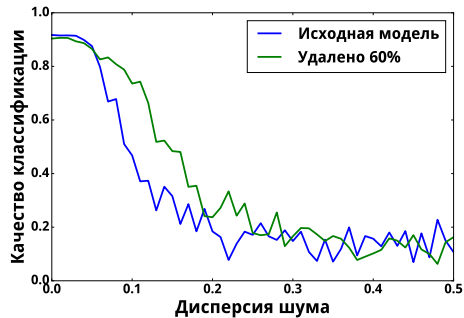
- Многоэкстремальность задачи оптимизации параметров модели.
- Вычислительная сложность оптимизации.
- Большое количество параметров и гиперпараметров.

Проблемы обучения сетей

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



Избыточность параметров модели



Неустойчивость модели

Правдоподобие модели

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

Модель f оптимальна, если достигается максимум **правдоподобия модели**:

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

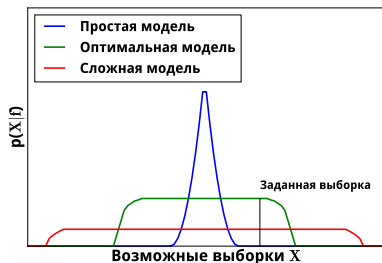
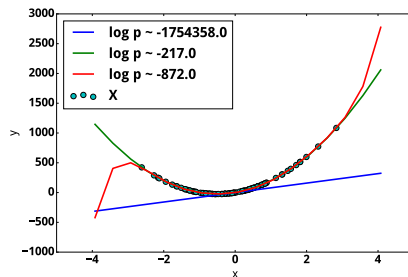


Схема выбора модели по правдоподобию



Пример: полиномы

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, позволяющего проводить оптимизацию в различных режимах:

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

Вариационная нижняя оценка правдоподобия

Правдоподобие модели:

$$p(y|\mathbf{X}, \mathbf{A}, \mathbf{m}, c_{\text{temp}}) = \int_{\mathbf{W}, \mathbf{\Gamma}} p(y|\mathbf{X}, \mathbf{W}, \mathbf{\Gamma}) p(\mathbf{W}|\mathbf{A}, \mathbf{\Gamma}) p(\mathbf{\Gamma}|\mathbf{m}, c_{\text{temp}}) d\mathbf{W} d\mathbf{\Gamma}.$$

Пусть q — непрерывное распределение.

$$\begin{aligned} \log p(y|\mathbf{X}, \mathbf{A}, \mathbf{m}, c_{\text{temp}}) &\geq \\ &\geq \int q(\mathbf{W}) \log p(y|\mathbf{X}, \mathbf{W}, \mathbf{\Gamma} \cdot \mathbf{A}^{-1}, c_{\text{temp}}) d\mathbf{W} d\mathbf{\Gamma} - \\ &\quad - \int q(\mathbf{w}) \log \frac{p(\mathbf{w}, \mathbf{\Gamma} | \mathbf{A}^{-1}, \mathbf{m}, c_{\text{temp}})}{q(\mathbf{W}, \mathbf{\Gamma})} d\mathbf{W} d\mathbf{\Gamma} = \\ &= \mathbb{E}_q \log p(y|\mathbf{X}, \mathbf{W}, \mathbf{\Gamma} \cdot \mathbf{A}^{-1}, c_{\text{temp}}) - D_{KL}(p(\mathbf{w}, \mathbf{\Gamma} | \mathbf{A}^{-1}, \mathbf{m}, c_{\text{temp}}) || q(\mathbf{W}, \mathbf{\Gamma})). \end{aligned}$$

Постановка задачи

TODO: ослабить? Как писать заголовки для этих слайдов?

TODO: Что мы считаем моделью?

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 **моделью**, если функция, задаваемая рекурсивно как

$$f_j(\mathbf{x}) = \sum_{k \in \text{Adj}(v_j)} \langle \gamma_{j,k}, \mathbf{g}_{j,k} \rangle (f_k(\mathbf{x})), \quad f_0(\mathbf{x}) = \mathbf{x},$$

является непрерывной дифференцируемой функцией из \mathbb{R}^n во множество \mathbb{Y} при любых значениях векторов γ .

Обозначим за вектор **параметров модели** \mathbf{W} конкатенацию параметров всех подмоделей $\{f_j\}_{j=1}^{|V|}$.

Распределение на структуре

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

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

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

где $\alpha_1, \dots, \alpha_h$ — параметры сдвига распределения, c_{temp} — температура распределения.

Обозначим за **структуру** модели $\mathbf{\Gamma}$ множество всех векторов $\gamma_{j,k}$.

Обозначим за \mathbf{m} параметры сдвига всех распределений, соответствующих структуре $\mathbf{\Gamma}$.

Вариационный вывод: распределение структурных параметров

Для каждого элемента структуры γ зададим распределение весов базовых функций по распределению Gumbel-Softmax с параметрами $\hat{\alpha}_1, \dots, \hat{\alpha}_L, c$, где параметр c — общий для всех весов.

Для реализации h -й компоненты случайной величины γ справедлива следующая формула:

$$\hat{\gamma}^h = \exp(\log(\alpha_h + \text{Gum}_h) c^{-1}) \sum_{h=1}^L \exp(\log(\alpha_l + \text{Gum}_l) c^{-1}),$$

где $\text{Gum} \sim -\log(-\log \mathcal{U}(0, 1))$.

Вариационный вывод: распределение параметров

Пусть $\mathbf{W} \sim \mathcal{N}(\mathbf{0}, \mathbf{A}^{-1})$ и структура модели Γ определена однозначно.

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

Тогда вариационная оценка имеет вид:

$$\int_{\mathbf{W}} q(\mathbf{W}) \log p(\mathcal{D}, \mathbf{W}, \mathbf{A}^{-1}) d\mathbf{W} - D_{\text{KL}}(q_{\mathbf{W}}(\mathbf{W}) || p(\mathbf{W} | \mathbf{A}^{-1})) \simeq$$
$$\sum_{i=1}^m \log p(\mathbf{x}_i | \mathbf{W}_i) - D_{\text{KL}}(q_{\mathbf{W}}(\mathbf{W}) || p(\mathbf{W} | \mathbf{A}^{-1})) = -L(\boldsymbol{\theta}, \mathbf{A}^{-1}, \mathcal{D}),$$

где $\mathbf{W}_i \sim q_{\mathbf{W}}$.

Дивергенция $D_{\text{KL}}(q_{\mathbf{W}}(\mathbf{w}) || p(\mathbf{W} | \mathbf{A}^{-1}))$ вычисляется аналитически:

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

Вариационная оценка на основе мултистарта

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

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

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

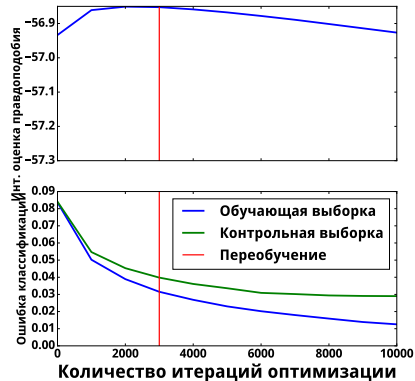
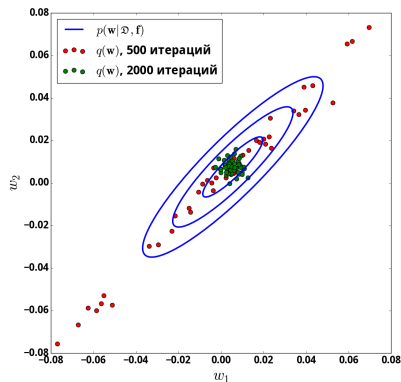
Тогда

$$E_{q_{\mathbf{W}}^{\tau}}(-\log(q_{\mathbf{W}}^{\tau})) - E_{q_{\mathbf{W}}^{\tau-1}}(-\log(q_{\mathbf{W}}^{\tau-1})) \sim \frac{1}{k} \sum_{r=1}^k (\beta \text{Tr}[\mathbf{H}(\mathbf{W}^r)] - \beta^2 \text{Tr}[\mathbf{H}(\mathbf{W}^r)\mathbf{H}(\mathbf{w}^r)]) + o_{\beta \rightarrow 0}(1),$$

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

Вариационная оценка с использованием градиентного спуска

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



Оптимизация параметров вариационного распределения

Оптимизацию параметров вариационного распределения будем проводить по следующему функционалу:

$$L = E_q \log p(y|X, W, \Gamma, A^{-1}, c_{\text{temp}}) - c_{\text{reg}} D_{KL}(p(W, \Gamma | A^{-1}, m, c_{\text{temp}}) || q(W), q(\Gamma)) \rightarrow \max$$

Теорема

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

$$L \xrightarrow{P} c_{\text{reg}} m \int q(W, \Gamma) \log \frac{p(y, W, \Gamma | X, A, m, c_{\text{temp}})}{q(W, \Gamma)} dW d\Gamma$$

Теорема

Для любых значений A, m и вариационных параметров q_W существует точка на вершинах симплексов структуры, определяющая распределение q_Γ , что для любой точки внутри симплексов, определяющей распределение q'_Γ справедливо выражение:

$$\lim_{c_{\text{temp}} \rightarrow 0} \frac{ELBO'}{ELBO} = -\infty.$$

Оптимизация параметров априорного распределения

Оптимизацию параметров априорного распределения будем проводить по следующему функционалу:

$$Q = c_{\text{train}} E_q \log p(y|X, \mathbf{W}, \mathbf{\Gamma}, \mathbf{A}^{-1}, c_{\text{prior}}) - c_{\text{prior}} D_{KL}(p(\mathbf{W}, \mathbf{\Gamma} | \mathbf{A}^{-1}, \mathbf{m}, c_{\text{temp}}) || q(\mathbf{W}, \mathbf{\Gamma})) - \\ - c_{\text{comb}} \sum_{p' \in \mathbf{P}} D_{KL}(\mathbf{\Gamma} | p') \rightarrow \max,$$

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

Общая задача оптимизации

Общая задача оптимизации — двухуровневая:

$$\begin{aligned}\hat{\mathbf{A}}, \hat{\mathbf{m}} &= \arg \max_{\mathbf{A}, \mathbf{m}} Q = \\ &= c_{\text{train}} E_{\hat{q}} \log p(\mathbf{y} | \mathbf{X}, \mathbf{W}, \mathbf{\Gamma}, \mathbf{A}^{-1}, c_{\text{prior}}) - c_{\text{prior}} D_{KL}(p(\mathbf{W}, \mathbf{\Gamma} | \mathbf{A}^{-1}, \mathbf{m}, c_{\text{temp}}) || \hat{q}(\mathbf{W}, \mathbf{\Gamma})) - \\ &\quad - c_{\text{comb}} \sum_{p' \in \mathbf{P}} D_{KL}(\mathbf{\Gamma} | p'),\end{aligned}$$

при

$$\hat{q} = \arg \max_q L = E_q \log p(\mathbf{y} | \mathbf{X}, \mathbf{W}, \mathbf{\Gamma}, \mathbf{A}^{-1}, c_{\text{temp}}) - c_{\text{reg}} D_{KL}(p(\mathbf{w}, \mathbf{\Gamma} | \mathbf{A}^{-1}, \mathbf{m}, c_{\text{temp}}) || q(\mathbf{W}), q(\mathbf{\Gamma}))$$

Оператор оптимизации

Обозначим за \mathbf{h} гиперпараметры \mathbf{A}, \mathbf{m} .

Обозначим за θ параметры распределений $q_{\mathbf{W}}, q_{\Gamma}$.

Определение

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

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

где

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

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

начальное значение параметров θ .

Оптимизация гиперпараметров

Перепишем итоговую задачу оптимизации:

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

где $\boldsymbol{\theta}_0$ — н

Утверждение, Luketina et al., 2016

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

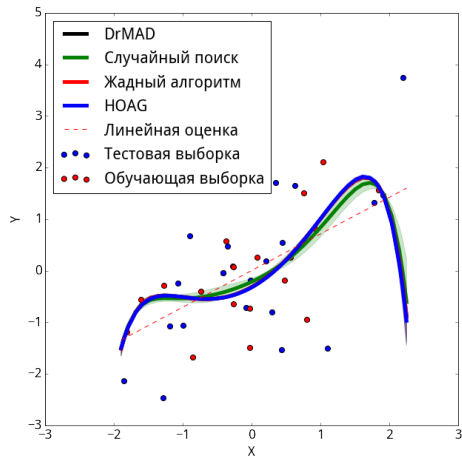
$$\mathbf{H}(L, \boldsymbol{\theta}) \approx \mathbf{I}.$$

Тогда допустима следующая оптимизация гиперпараметров:

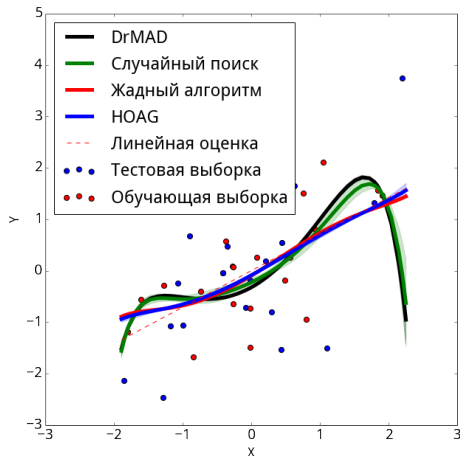
$$\mathbf{h}' = \mathbf{h} - \beta^h \nabla_{\mathbf{h}} Q(T(\boldsymbol{\theta}), \mathbf{h}),$$

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 $D_{\text{KL}}(q(\mathbf{W}, \mathbf{\Gamma})|p(\mathbf{W}, \mathbf{\Gamma}|\mathbf{y}, \mathbf{X}, \mathbf{A}, \mathbf{m}, c_{\text{temp}})) = 0$.

Тогда двухуровневая задача оптимизация эквивалентна исходной задаче оптимизации правдоподобия модели:

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

при $c_{\text{reg}} = c_{\text{prior}} = c_{\text{train}} = 1, c_{\text{comb}} = 0$.

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

Назовем **параметрической сложностью** модели наименьшую дивергенцию между априорным распределением параметров и вариационным распределением параметров:

$$C_{\text{param}} = \min_{\mathbf{A}, \mathbf{m}} D_{\text{KL}}(q(\mathbf{W}, \mathbf{\Gamma}) || p(\mathbf{W}, \mathbf{\Gamma} | \mathbf{A}, \mathbf{m}, c_{\text{temp}})).$$

TODO: надо ли про bits-back?

Теорема

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

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

Теорема

Пусть $f_1 \in F(1, 1, c_{\text{prior}}, 0, \{\}, c_{\text{temp}})$, $f_2 \in F(1, 1, c_{\text{prior}}, 0, \{\}, c'_{\text{temp}})$, $c_{\text{prior}} < c'_{\text{prior}}$. Пусть вариационные параметры моделей f_1 и f_2 лежат в области U , в которой соответствующие функции L и Q являются локально-выпуклыми. Тогда модель f_1 имеет параметрическую сложность, не большую чем у f_2 .

$$C_{\text{param}}(f_1) \leq C_{\text{param}}(f_2).$$

Структурная сложность

Назовем **структурной сложностью модели** энтропию вариационного распределения структуры модели:

$$C_{\text{struct}} = -E_{q_{\Gamma}} \log q_{\Gamma}(\Gamma).$$

Теорема

Пусть $f \in F(c_{\text{reg}}, c_{\text{train}}, c_{\text{prior}}, 0, \{\}, c_{\text{temp}}) | \lim_{c_{\text{temp}} \rightarrow 0}$.

Сложность модели f равняется нулю.

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

Теорема

Пусть $f_1 \in F(c_{\text{reg}}, c_{\text{train}}, c_{\text{prior}}, 0, \{\}, c_{\text{temp}})$, $f_2 \in F(c_{\text{reg}}, c_{\text{train}} | \lim_{c'_{\text{temp}} \rightarrow \infty}, c_{\text{prior}}, 0, \{\}, c_{\text{temp}'})$. 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:

$$C_{\text{struct}}(f_1) - C_{\text{struct}}(f_2) \leq E_q \log p(y|X, W, \Gamma \cdot A^{-1}, c_{\text{temp}}) - E'_q \log p(y|X, W, \Gamma \cdot A^{-1}, c_{\text{temp}'}).$$

Полный перебор

Пусть $p(\Gamma) = \lim_{c_{\text{temp}} \rightarrow 0} GS$.

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

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

$$f_2 \in F(1, 0, 0, \{q_1(\Gamma)\}, 0, c_{\text{temp}}), q_1 \in f_1,$$

$$f_3 \in F(1, 0, 0, \{q_1(\Gamma), q_2(\Gamma)\}, 0, c_{\text{temp}}), q_2 \in f_2.$$

TODO: формализовать что такое q_1, q_2

Теорема

Вариационные распределения структур q_{Γ} последовательности

$\mathfrak{F} = \{\lim_{c_{\text{temp}} \rightarrow 0} f_1, \dots, \lim_{c_{\text{temp}} \rightarrow 0} f_N\}$ вырождаются в распределения вида $\delta(\hat{\Gamma})$, где $\hat{\Gamma}$ — точка на декартовом произведении вершин симплексов структуры модели.

Вариационные распределения последовательности \mathfrak{F} проходят все возможные комбинации структур модели.

TODO: пояснить про дискретность структуры?

Заключение

???

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. –