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)

ИОИ-2018
11.10.2018

Задача выбора структуры модели

Рассматривается задача регресси или классификации:

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

Рассмотрим задачу выбора структуры для однослойной нейросети в задаче классификации:

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

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

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

Структура модели γ определяется вершиной булевого K-мерного куба.

Выбор структуры модели глубокого обучения

- **Цель работы:** разработка метода построения наиболее правдоподобной структуры модели глубокого обучения.
- **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

Задачи

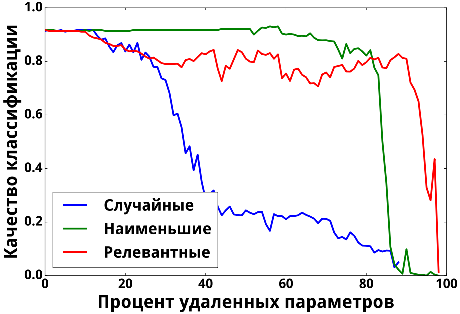
- Исследовать методы построения моделей глубокого обучения.
- Предложить критерии оптимльной и субоптимальной сложности модели глубокого обучения.
- Предложить метод построения модели субоптимальной сложности.

Основные проблемы

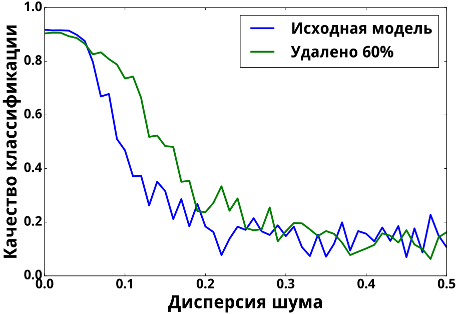
- Многоэкстремальность задачи оптимизации параметров модели.
- Вычислительная сложность оптимизации.
- Большое количество параметров и гиперпараметров.

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 $p(\mathbf{W}, \mathbf{\Gamma})$.

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

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

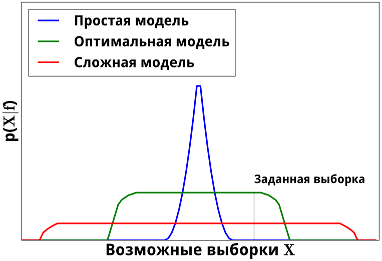
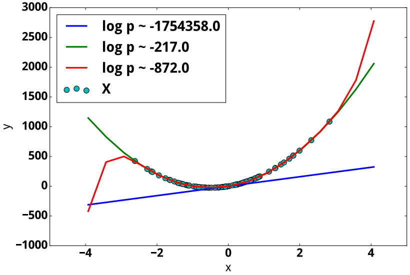


Схема выбора модели по правдоподобию



Пример: полиномы

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, позвяоляюещго проводить оптимизацию в различных режимах:

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

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

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

является непрерывной дифференцируемой функцией из $\mathbb{R}^n$ во множество $\mathbb{Y}$ при любых значениях векторов γ.
Обозначим за вектор **параметров модели 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 $\alpha_1, \ldots, \alpha_h$ — 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 $\mathbf{\Gamma}$.

Задача оптимизации

Пусть параметры распределения распределены нормально:

$$p(\mathbf{W}) \sim \mathcal{N}(\mathbf{0}, \mathbf{A}^{-1}).$$

Требуется найти гиперпараметры модели $\mathbf{A}, \mathbf{m}$ доставляющие максимум правдоподобия модели:

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

TODO: насколько это формально?

Теорема

При устремлении c_{temp} к нулю задача становится эквивалентна дискретной задаче оптимизации:

$$\arg\max_{\mathbf{A},\{\gamma_{j,k}\in\Delta^{K_{j,k}-1},(j,k)\in E\}} p(\mathbf{y}|\mathbf{X}, \mathbf{A}, \mathbf{m}, c_{\text{temp}}) \text{ при } c_{\text{temp}} \to 0,$$

где $\Delta^{K_{j,k}-1}$ — множество векторов, соответствующих вершинам $(K_{j,k}-1)$-сипмлекса.

Вариационная нижняя оценка правдоподобия

Правдоподобие модели:

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

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

Пусть q — непрерывное распределение.

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

$$\geq \int q(\mathbf{W}) \log p(\mathbf{y}|\mathbf{X}, \mathbf{W}, \boldsymbol{\Gamma}.\mathbf{A}^{-1}, c_{\text{temp}}) d\mathbf{W} d\boldsymbol{\Gamma} -$$

$$- \int q(\mathbf{w}) \log \frac{p(\mathbf{w}, \boldsymbol{\Gamma}|\mathbf{A}^{-1}, \mathbf{m}, c_{\text{temp}})}{q(\mathbf{W}, \boldsymbol{\Gamma})} d\mathbf{W} d\boldsymbol{\Gamma} =$$

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

Вариационный вывод: распределение параметров

Пусть структура модели 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

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

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. Пусть β — шаг градиентного спуска, такой что:

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

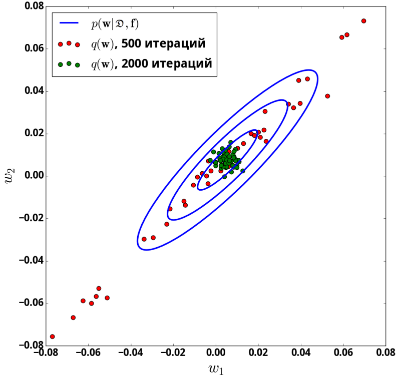
Тогда

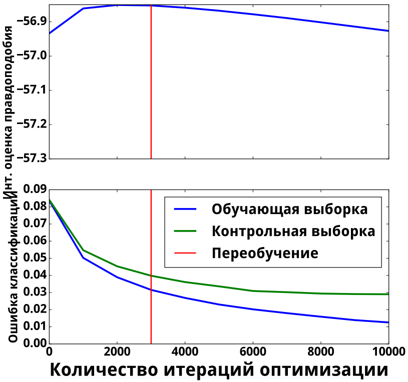
$$\mathsf{E}_{q^\tau_\mathbf{W}}(-\log(q^\tau_\mathbf{W})) - \mathsf{E}_{q^{\tau-1}_\mathbf{W}}(-\log(q^{\tau-1}_\mathbf{W})) \sim \frac{1}{k} \sum_{r=1}^{k} \big(\beta\, Tr[\mathbf{H}(\mathbf{W}^r)] - \beta^2\, Tr[\mathbf{H}(\mathbf{W}^r)\mathbf{H}(\mathbf{w}^r)]\big) + o_{\beta \to 0}(1),$$

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 $\mathrm{D_{KL}}(q(\mathbf{W})||p(\mathbf{W}|\mathfrak{D}, \mathbf{A}^{-1}))$.
Градиентный спуск не минимизирует $\mathrm{D_{KL}}(q(\mathbf{W})||p(\mathbf{W}|\mathfrak{D}, \mathbf{A}^{-1}))$.





Вариационный вывод: распределение структурных параметров

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

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

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

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

Оптимизация параметров вариационного распределения

Оптимизацию параметров вариационного распределения 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 $c_{\text{reg}} m$:

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

Оптимизация параметров априорного распределения

Оптимизацию параметров априорного распределения будем проводить по следующему функционалу:

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

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

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

Общая задача оптимизации

Общая задача оптимизации — двухуровневая:

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

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

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

при

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

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:

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

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 $\mathfrak{D}$.

ачальное значение параметров $\boldsymbol{\theta}$.

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 функции L можно аппроксимировать единичной матрицей:

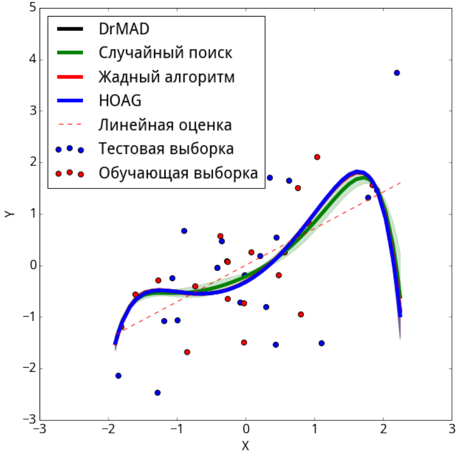
$$\mathbf{H}(L, \boldsymbol{\theta}) \approx \mathbf{I}.$$

Тогда допустима следующая оптимизация гиперпараметров:

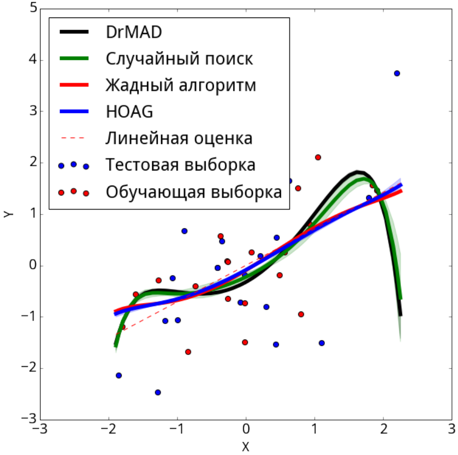
$$\mathbf{h}' = \mathbf{h} - \beta^h \nabla_{\mathbf{h}} Q(T(\boldsymbol{\theta}), \mathbf{h}),$$

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 $\mathrm{D}_{\mathrm{KL}}(q(\mathbf{W}, \mathbf{\Gamma}) | p(\mathbf{W}, \mathbf{\Gamma} | \mathbf{y}, \mathbf{X}, \mathbf{A}, \mathbf{m}, c_{\text{temp}})) = 0$.
Тогда двухуровневая задача оптимизация эквивалентна 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 распредеделением параметров:

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

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.

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

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:

$$C_{\text{struct}} = -\mathsf{E}_{q_\Gamma} \log q_\Gamma(\mathbf{\Gamma}).$$

Теорема

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

$$\lim_{c_{\text{temp}} \to \infty} C_{\text{struct}}(f) = 0$$

Теорема

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

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

Полный перебор

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

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

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

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

TODO: формализовать что такое q_1, q_2

Теорема

Вариационные распределения структур $q_{\mathbf{\Gamma}}$ последовательности $\mathfrak{F} = \{\lim_{c_{\text{temp}}\to 0} f_1, \dots, \lim_{c_{\text{temp}}\to 0} f_N\}$ вырождаются в дельта-функции 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. –