Метаоптимизация и структура

Бахтеев Олег

МФТИ

12.11.2019

В предыдущих сериях: гиперпараметры

Определение

Априорным распределением $p(\mathbf{w}|\mathbf{h})$ параметров модели назовем вероятностное распределение, соответствующее предположениям о распределении параметров модели.

Определение

Гиперпараметрами $\mathbf{h} \in \mathbb{H}$ модели назовем параметры априорного распределения (параметры распределения параметров модели).

В предыдущих сериях: гиперпараметры

Задана дифференцируемая по параметрам модель, приближающая зависимую переменную y:

$$\mathbf{f} : \mathbb{R}^n \to \mathbb{Y}, \quad \mathbf{w} \in \mathbb{R}^u.$$

Функция $\mathbf{f}$ задает правдоподобие выборки $\log p(\mathbf{y}|\mathbf{X}, \mathbf{w})$.
Пусть также задано априорное распределение параметров $p(\mathbf{w}|\mathbf{h})$.

Пример:

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

где $\mathbf{A}^{-1} = \text{diag}[\alpha_1, \dots, \alpha_u]^{-1}$ — матрица ковариаций диагонального вида, определяемая *гиперпараметрами* $[\alpha_1, \dots, \alpha_u] = \mathbf{h}$.

В предыдущих сериях: гиперпараметры

Пусть $\boldsymbol{\theta} \in \mathbb{R}^{\boldsymbol{s}}$ — множество всех оптимизируемых параметров.
$L(\boldsymbol{\theta}, \mathbf{h})$ — дифференцируемая функция потерь по которой производится оптимизация функции $\mathbf{f}$.
$Q(\boldsymbol{\theta}, \mathbf{h})$ — дифференцируемая функция определяющая итоговое качество модели $\mathbf{f}$ и приближающая интеграл.
Требуется найти параметры $\boldsymbol{\theta}^*$ и гиперпараметры $\mathbf{h}^*$ модели, доставляющие минимум следующему функционалу:

$$\mathbf{h}^* = \arg\max_{\mathbf{h} \in \mathbb{H}} Q(\boldsymbol{\theta}^*(\mathbf{h}), \mathbf{h}),$$

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

В предыдущих сериях: вариационная оценка

Вариационная оценка Evidence, Evidence lower bound — метод нахождения приближенного значения аналитически невычислимого распределения $p(\mathbf{w}|\mathfrak{D}, \mathbf{h})$ распределением $q(\mathbf{w}) \in \mathfrak{Q}$. Получение вариационной нижней оценки обычно сводится к задаче минимизации

$$\mathrm{KL}(q(\mathbf{w})||p(\mathbf{w}|\mathfrak{D})) = -\int_{\mathbf{w}} q(\mathbf{w}) \log \frac{p(\mathbf{w}|\mathfrak{D})}{q(\mathbf{w})} d\mathbf{w} = \color{blue}{\mathsf{E}_{\mathbf{w}} \log p(\mathfrak{D}|\mathbf{w})} - \color{red}{\mathrm{KL}(q(\mathbf{w})||p(\mathbf{w}|\mathbf{h}))}.$$

Частным случаем вариационного распределения можно считать распределение параметров модели под действием оптимизации.

Метапараметры

Wikipedia

A parameter that controls the value of one or more others.

Определение

Метапараметрами $\boldsymbol{\lambda}$ модели назовем параметры оптимизации.

Чаще всего метапараметры назначаются экспертно и не подлежат оптимизации в ходе решения задачи выбора модели.

Что можно считать метапараметрами:

- параметры оператора оптимизации;
- параметры задачи оптимизации;
- структуру модели;
- функции активации слоев сети;
- вид априорного распределения и функции правдоподобия.

A neural network that embeds its own meta-levels

Предлагается разделить подмодели внутри модели сети по назначениям:

- "Normal" model: обучение и вывод.
- Evaluation model: оценка качества Q.
- Analyzing model: анализ параметров модели.
- Modifiyng model: модификация параметров.

Представлен градиентный алгоритм оптимизации нейронной сети.

Learning to learn by gradient descent by gradient descent

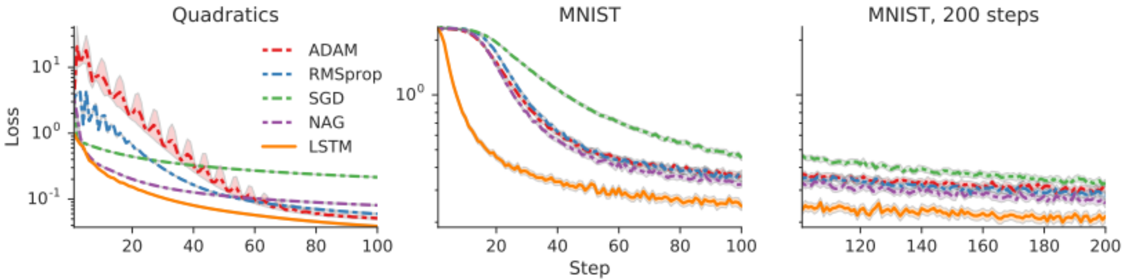
Идея: рассматривать оператор оптимизации T как дифференцируемую функцию:

$$T(\boldsymbol{\theta}) = \text{LSTM}(\boldsymbol{\theta}).$$

Оптимизационная задача:

$$\sum_{t=t_0}^{t_\eta} L\left(T^t(\boldsymbol{\theta}_{t_0})\right) \to \max.$$

LSTM имеет небольшое число параметров и делит параметры между всеми метапараметрами оператора.



Optimal Brain Damage

Рассматривается задача удаления неинформативных параметров.

Идея метода: Разложим функцию потерь в ряд Тейлора в окрестности максимума $\boldsymbol{\theta}^*$:

$$L(\boldsymbol{\theta}^* + \Delta\boldsymbol{\theta}) - L(\boldsymbol{\theta}^*) = -\frac{1}{2}\boldsymbol{\theta}^{\mathsf{T}}\mathbf{H}\boldsymbol{\theta} + o(||\Delta\boldsymbol{\theta}||^3),$$

где $\mathbf{H}$ — гессиан функции $-L$.

Для простоты вычисления будем полагать гессиан диагональным. Задача удаления параметров сводится к рассмотрению задач условной оптимизации вида:

$$L(\boldsymbol{\theta}^* + \Delta\boldsymbol{\theta}) \to \max$$

при

$$\theta_i^* + \Delta\theta_i = 0.$$

Показатель информативности параметра:

$$\frac{\theta_i^2}{2[\mathbf{H}^{-1}]_{i,i}}.$$

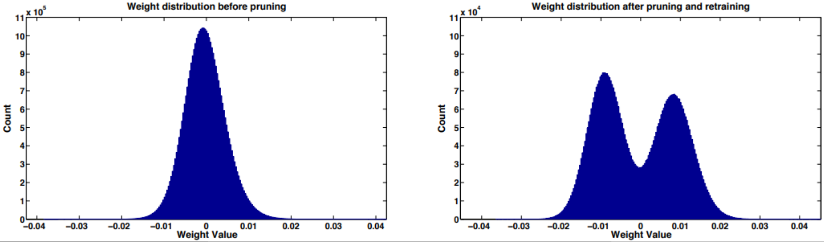
Learning both Weights and Connections for Efficient Neural Networks

Идея подхода:

1. Оптимизируем модель;
2. Удаляем наименьшие по модулю параметры;
3. Запускаем оптимизацию заново.

Почти очевидные факты, которые подтверждаются в статье:

- L_2 лучше для прунинга, чем L_1 в случае, если после прунинга идет оптимизация.
- Оптимизацию лучше производить из предыдущего оптимума, чем из случайной точки.
- После прунинга распределение параметров становится мультимодальным.

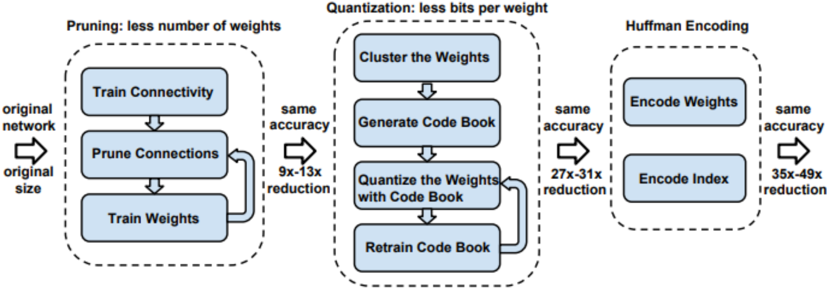


Deep Compression

Идея подхода:

1. Удаляем ненужные параметры модели, аналогично предыдущем подходу.
2. Кластеризуем параметры (K-means на каждом слое).
3. Производим повторную оптимизацию на центроидах.
4. Кодируем индексы параметров с использованием кодов Хаффмана.

Результат: уменьшение размеров модели в 40 раз, ускорение в 3 раза.



Graves, 2011

$$\mathrm{MDL}(\mathbf{f}, \mathfrak{D}) = L(\mathbf{f}) + L(\mathfrak{D}|\mathbf{f}),$$

где $\mathbf{f}$ — модель, $\mathfrak{D}$ — выборка, L — длина описания в битах.

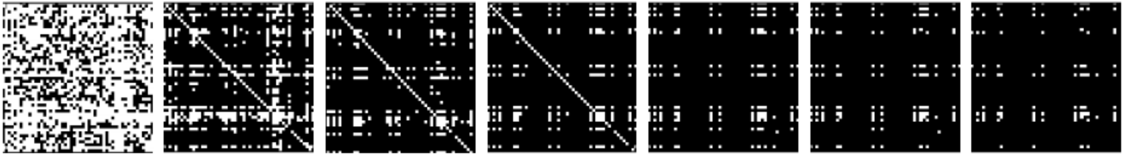
$$\mathrm{MDL}(\mathbf{f}, \mathfrak{D}) \sim L(\mathbf{f}) + L(\mathbf{w}^*|\mathbf{f}) + L(\mathfrak{D}|\mathbf{w}^*, \mathbf{f}),$$

$\mathbf{w}^*$ — оптимальные параметры модели.

$$L = \sum_{\mathbf{x},\mathbf{y}} \log p(\mathbf{y}|\mathbf{x}, \hat{\mathbf{w}}) + \frac{1}{2}\left(\mathrm{tr}(\mathbf{A}_q) + \boldsymbol{\mu}_q^{\mathsf{T}}\mathbf{A}^{-1}\boldsymbol{\mu}_q - \ln|\mathbf{A}_q|\right).$$

Прунинг параметра w_i определяется относительной плотностью:

$$\lambda = \frac{q(\mathbf{0})}{q(\boldsymbol{\mu}_{i,q})} = \exp\left(-\frac{\mu_i^2}{2\sigma_i^2}\right).$$



$\lambda = 0.01$ $\lambda = 0.05$ $\lambda = 0.1$ $\lambda = 0.2$ $\lambda = 0.5$ $\lambda = 1$ $\lambda = 2$

Bayesian Compression for Deep Learning

Модель 1:

$$\mathbf{z} \sim p(\mathbf{z}); \mathbf{w} \sim \mathcal{N}(0; \mathbf{z}^2),$$

где z поставлено в соответствие группе параметров (пример: нейронам).
Априорное распределение $p(z) \propto \frac{1}{|z|^2}$.
Вариационное распределение: $\mathbf{z} \sim \mathcal{N}(\boldsymbol{\mu}_z, \boldsymbol{\sigma}_z), \mathbf{w} \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\sigma})$
Критерий удаления параметров:

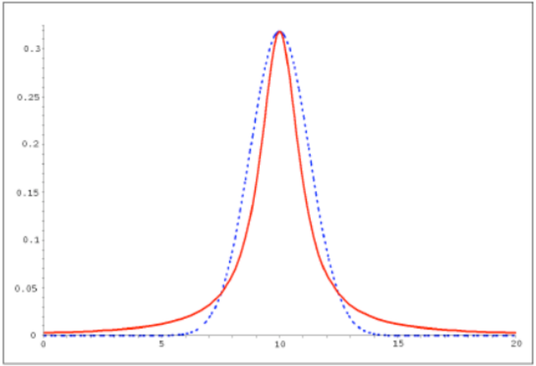
$$\log \sigma_z - \log \mu_z > \lambda.$$

Bayesian Compression for Deep Learning

Модель 2: Априорное распределение:

$$s \sim C^{+}(0, \lambda_0), z_i \sim C^{+}(0, 1), \hat{w_{i,j}} \sim \mathcal{N}(0, 1), w_{i,j} = s z_i \hat{w_{i,j}}.$$

Вариационное распределение для z_i: $\mathbf{z}_i \sim \mathcal{L}\wr\}\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\sigma})\mathcal{L}\wr\}\mathcal{N}(\boldsymbol{\mu}_i, \boldsymbol{\sigma}_i)$



[Comparing the Cauchy and Gaussian (Normal) density functions, F. Masci]

Learning the structure of deep sparse graphical models

Рассматривается глубокая генеративная модель.
Структура (т.е. связи параметров) сэмплируются в соответствии с распределением Индийского буфета. Интерпретация распределения: *В ресторане находится конечное количество клиентов и бесконечное количество блюд. j-й клиент берет блюдо k с вероятностью:*

$$\frac{\eta_k}{j+\beta-1},$$

где η_k *количество выборов этого блюда предыдщими клиентами (популярность блюда), а также несколько новых блюд в соответствии с рапределением Пуассона с параметром* $\frac{\alpha\beta}{j+\beta-1}$.
Порождение параметров и структуры происходит с помощью MCMC.
Главный вывод: структуру можно рассматривать как случайную величину и применять вероятностные методы.

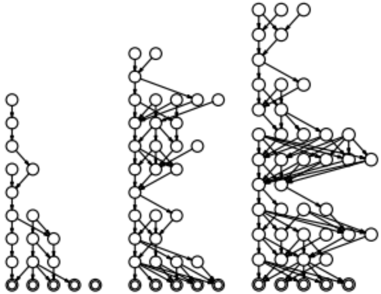
Learning the structure of deep sparse graphical models

(a) $\alpha = 1,\ \beta = 1$

(b) $\alpha = 1,\ \beta = \frac{1}{2}$

(c) $\alpha = \frac{1}{2},\ \beta = 1$

(d) $\alpha = 1,\ \beta = 2$



(e) $\alpha = \frac{3}{2},\ \beta = 1$

Вид вариационной оценки

Обозначим структуру $\mathbf{\Gamma}$. Пусть априорное распределение параметров зависит от структуры:

$$\mathbf{w} \sim p(\mathbf{w}|\mathbf{\Gamma}, \mathbf{h}), \quad \mathbf{w} \sim^q q(\mathbf{w}|\mathbf{\Gamma})$$

Тогда вариационная оценка имеет вид:

$$\mathsf{E}_q \mathrm{log} p(\mathbf{y}|\mathbf{x}, \mathbf{w}, \mathbf{\Gamma}) - \mathrm{KL}\left(q(\mathbf{\Gamma})||p(\mathbf{\Gamma}|\mathbf{h})\right) - \mathsf{E}_{\mathbf{\Gamma}\sim q(\mathbf{\Gamma})}\mathrm{KL}\left(q(\mathbf{w}|\mathbf{\Gamma})||p(\mathbf{w}|\mathbf{\Gamma}, \mathbf{h})\right).$$

Параметрическая сложность

Относительной вариационной плотностью назовем отношение:

$$\rho(w|\mathbf{h}) = \frac{q_{\mathbf{w}}(\text{mode } p(\mathbf{w}|\boldsymbol{\Gamma}, \mathbf{h}, \boldsymbol{\lambda}))}{q_{\mathbf{w}}(\text{mode } q_{\mathbf{w}})}, \quad \boldsymbol{\rho} = \prod_{w \in \mathbf{w}} \rho(w|\boldsymbol{\Gamma}, \boldsymbol{\theta}_{\mathbf{w}}, \mathbf{h}, \boldsymbol{\lambda}).$$

Определение

Параметрической сложностью модели назовем минимальную дивергенцию между априорным и вариационным распределением:

$$C_p = \min_{\mathbf{h} \in U} D_{\text{KL}}(q||p(\mathbf{w}, \boldsymbol{\Gamma}|\mathbf{h}, \lambda_{\text{temp}}, \mathbf{f})),$$

где U — некоторый компакт.

Утверждение

Пусть задана последовательность $\boldsymbol{\theta}[1], \boldsymbol{\theta}[2], \dots,$ такая, что $\lim_{i \to \infty} C_p(\boldsymbol{\theta}[i]) = 0$. Тогда:

$$\lim_{i \to \infty} \mathsf{E}_{q(\boldsymbol{\Gamma})} \boldsymbol{\rho}^{-1}(\mathbf{w}|\mathbf{h}[i]) = 1,$$

где $\mathbf{h} = \arg\min_{\mathbf{h} \in U} D_{\text{KL}}(q||p(\mathbf{w}, \boldsymbol{\Gamma}|\mathbf{h}, \lambda_{\text{temp}}, \mathbf{f}))$

Обобщающая задача оптимизации

Какие требования можно выдвинуть к “хорошей” функции оптимизации?

1. При некоторых значениях метапараметров функция должна приближать метод максимального правдоподобия.
2. При некоторых значениях метапараметров функция должна штрафовать излишне сложные модели.
3. При некоторых значениях метапараметров функция должна приближать обоснованность модели.
4. При некоторых значениях метапараметров функция должна позволять переходить между оптимальными структурами модели.
5. Функции потерь и валидации должны быть непрерывны по метапараметрам.
6. Область определения функции должна быть нетривиальна.

Критерии перехода между структурами

критерий 1

Существует нетривиальная константа K и метапараметры $\boldsymbol{\lambda}$, такие что для любой пары локальных оптимумов $\mathbf{h}_1, \mathbf{h}_2$ и соответствующих им вариационных параметров $\boldsymbol{\theta}(\mathbf{h}_1), \boldsymbol{\theta}(\mathbf{h}_2)$, таких что $\mathrm{KL}(q(\boldsymbol{\Gamma}|\boldsymbol{\theta}_1)||q(\boldsymbol{\Gamma}|\boldsymbol{\theta}_2)) > K, \mathrm{KL}(q(\boldsymbol{\Gamma}|\boldsymbol{\theta}_2)||q(\boldsymbol{\Gamma}|\boldsymbol{\theta}_1)) > K$ и $Q(\mathbf{h}_1|\boldsymbol{\lambda}) > Q(\mathbf{h}_2|\boldsymbol{\lambda})$, существует значение $\boldsymbol{\lambda}'$, такое что:

1. соответствия $\boldsymbol{\theta}(\mathbf{h}_1), \boldsymbol{\theta}(\mathbf{h}_2)$ сохраняются;
2. $Q(\mathbf{h}_1|\boldsymbol{\lambda}') < Q(\mathbf{h}_2|\boldsymbol{\lambda}')$.

Критерий 2

Существует нетривиальная константа K и метапараметры $\boldsymbol{\lambda}$, такие что для любой пары локальных оптимумов $\mathbf{h}_1, \mathbf{h}_2$ и соответствующих им вариационных параметров $\boldsymbol{\theta}(\mathbf{h}_1), \boldsymbol{\theta}(\mathbf{h}_2)$, таких что $\mathrm{KL}(p(\boldsymbol{\Gamma}|\mathbf{h}_1)||p(\boldsymbol{\Gamma}|\mathbf{h}_2)) > K, \mathrm{KL}(p(\boldsymbol{\Gamma}|\mathbf{h}_2)||p(\boldsymbol{\Gamma}|\mathbf{h}_1)) > K$ и $Q(\mathbf{h}_1|\boldsymbol{\lambda}) > Q(\mathbf{h}_2|\boldsymbol{\lambda})$, существует значение $\boldsymbol{\lambda}'$, такое что:

1. соответствия $\boldsymbol{\theta}(\mathbf{h}_1), \boldsymbol{\theta}(\mathbf{h}_2)$ сохраняются;
2. $Q(\mathbf{h}_1|\boldsymbol{\lambda}') < Q(\mathbf{h}_2|\boldsymbol{\lambda}')$.

Для $L = Q$ — вариационной нижней оценки критерий 2 не выполняется.

ДЗ: выбор задания

Дедлайн: 20 ноября, 0 часов.

```
from zlib import crc32

theory = crc32('фамилия кириллицей'.lower().encode('utf-8'))%5+1

practice = crc32('фамилия латиницей'.lower().encode('utf-8'))%3+1
```

Задания заливаются на github:
https://github.com/Intelligent-Systems-Phystech/model_selection/фамилия латиницей

ДЗ: теория

Формат: tex + pdf. Задание 1: доказать вид вариационной функции при структуре $\mathbf{\Gamma}$ (расписать дивергенцию).

ДЗ: теория

Формат: tex + pdf. Задание 2: доказать утверждение

Утверждение

Пусть

1. Заданы компактные множества $U_{\mathbf{h}} \subset \mathbb{H}$, $U_{\theta_{\mathbf{w}}} \subset \Theta_{\mathbf{w}}$, $U_{\theta_{\Gamma}} \subset \Theta_{\Gamma}$.
2. Вариационное распределение $q_{\mathbf{w}}(\mathbf{w}|\Gamma, \theta_{\mathbf{w}})$ является абсолютно непрерывным и унимодальным на U_{θ}. Его мода и матожидание совпадают:
$$\text{mode } q_{\mathbf{w}}(\mathbf{w}|\Gamma, \theta_{\mathbf{w}}) = \mathsf{E}_{q_{\mathbf{w}}(\mathbf{w}|\Gamma, \theta_{\mathbf{w}})}\mathbf{w}.$$
3. Априорное распределение $p(\mathbf{w}|\Gamma, \mathbf{h}, \lambda)$ является абсолютно непрерывным и унимодальным на $U_{\mathbf{h}}$. Его мода и матожидание совпадают и не зависят от гиперпараметров $\mathbf{h}$ на $U_{\mathbf{h}}$ и структуры Γ на $U_{\theta_{\Gamma}}$: $\mathsf{E}_{p(\mathbf{w}|\Gamma,\mathbf{h},\lambda)}\, \mathbf{w} = \text{mode } p(\mathbf{w}|\Gamma_1, \mathbf{h}_1, \lambda) = \text{mode } p(\mathbf{w}|\Gamma_1, \mathbf{h}_2, \lambda) = \mathbf{m}$
   для любых $\mathbf{h}_1, \mathbf{h}_2 \in U_{\mathbf{h}}, \Gamma_1, \Gamma_2 \in U_{\Gamma}$.
4. Параметры модели $\mathbf{w}$ имеют конечные вторые моменты по маргинальным распределениям: $\int_{\Gamma} q_{\Gamma}(\Gamma|\theta_{\Gamma}) q_{\mathbf{w}}(\mathbf{w}|\Gamma, \theta_{\mathbf{w}}) d\Gamma, \quad \int_{\Gamma} q_{\Gamma}(\Gamma|\theta_{\Gamma}) p(\mathbf{w}|\Gamma, \mathbf{h}, \lambda) d\Gamma$ при любых $\theta_{\mathbf{w}} \in U_{\theta_{\mathbf{w}}}, \theta_{\Gamma} \in U_{\theta_{\Gamma}}, \mathbf{h} \in U_{\mathbf{h}}$.
5. Вариационное распределение $q_{\mathbf{w}}(\mathbf{w}|\Gamma, \theta_{\mathbf{w}})$ является липшицевым по $\mathbf{w}$.
6. Значение $q_{\mathbf{w}}(\mathbf{w}|\Gamma, \theta_{\mathbf{w}})$ не равно нулю при любых $\theta \in U_{\theta}, \Gamma \in \mathbb{\Gamma}$.
7. Точная нижняя грань $q_{\mathbf{w}}(\mathbf{m}|\Gamma, \theta_{\mathbf{w}})$ не равна нулю при $\theta_{\mathbf{w}} \in U_{\theta_{\mathbf{w}}}$ и $\Gamma \in \mathbb{\Gamma}$:
$$\inf_{\Gamma \in \mathbb{\Gamma}, \theta_{\mathbf{w}} \in U_{\theta_{\mathbf{w}}}} q_{\mathbf{w}}(\mathbf{m}|\Gamma, \theta_{\mathbf{w}}) > 0.$$

Тогда
$$\left|\mathsf{E}_{q_{\Gamma}(\Gamma|\theta_{\Gamma})} \rho(\mathbf{w}|\Gamma, \theta_{\mathbf{w}}, \mathbf{h}, \lambda)^{-1} - 1\right| \leq \text{Const} \iint_{\Gamma, \mathbf{w}} |\mathbf{w}| \cdot |q_{\mathbf{w}}(\mathbf{w}|\Gamma, \theta_{\mathbf{w}}) - p(\mathbf{w}|\Gamma, \mathbf{h}, \lambda)| q_{\Gamma}(\Gamma|\theta_{\Gamma}) d\mathbf{w} d\Gamma.$$

ДЗ: теория

Формат: tex + pdf. Задание 3: доказать утверждение

Утверждение

Пусть

1. Вариационное распределение $q_{\mathbf{w}}(\mathbf{w}|\boldsymbol{\Gamma}, \boldsymbol{\theta}_{\mathbf{w}})$ и априорное распределение $p(\mathbf{w}|\boldsymbol{\Gamma}, \mathbf{h}, \boldsymbol{\lambda})$ являются абсолютно непрерывными.
2. Решение задачи
$$\mathbf{h}^* = \mathop{\arg\min}_{\mathbf{h} \in U_{\mathbf{h}}} D_{\mathrm{KL}}\big(q(\mathbf{w}, \boldsymbol{\Gamma}|\boldsymbol{\theta}) || p(\mathbf{w}, \boldsymbol{\Gamma}|\mathbf{h}, \boldsymbol{\lambda})\big) \tag{1}$$
единственно для любого $\boldsymbol{\theta} \in U_{\boldsymbol{\theta}}$.
3. Задана бесконечная последовательность векторов вариационных параметров $\boldsymbol{\theta}[1], \boldsymbol{\theta}[2], \dots, \boldsymbol{\theta}[i], \dots \in U_{\boldsymbol{\theta}}$, такая что $\lim_{i \to \infty} C_p(\boldsymbol{\theta}[i]|U_{\mathbf{h}}, \boldsymbol{\lambda}) = 0$.

Тогда следующее выражение стремится к нулю:

$$\iint_{\mathbf{w}, \Gamma} |p(\mathbf{w}|\boldsymbol{\Gamma}, \mathbf{h}[i], \boldsymbol{\lambda}) - q_{\mathbf{w}}(\mathbf{w}|\boldsymbol{\Gamma}, \boldsymbol{\theta}_{\mathbf{w}}[i])| q_{\boldsymbol{\Gamma}}(\boldsymbol{\Gamma}|\boldsymbol{\theta}_{\boldsymbol{\Gamma}}[i]) d\Gamma d\mathbf{w},$$

где $\boldsymbol{\theta}[i] = [\boldsymbol{\theta}_{\mathbf{w}}[i], \boldsymbol{\theta}_{\boldsymbol{\Gamma}}[i]]$, $\mathbf{h}[i]$ — решение задачи (1) для $\boldsymbol{\theta}[i]$.

(Воспользоваться неравенством Пинскера)

ДЗ: теория

Формат: tex + pdf. Задание 4: доказать утверждение

Утверждение

Пусть выполнены условия предыдущих двух утврежденений (из д.з.). Тогда справедливо следующее выражение:

$$\lim_{i \to \infty} \mathsf{E}_{q_{\boldsymbol{\Gamma}}(\boldsymbol{\Gamma}|\boldsymbol{\theta}_{\boldsymbol{\Gamma}}[i])} \rho(\mathbf{w}|\Gamma, \boldsymbol{\theta}_{\mathbf{w}}[i], \mathbf{h}[i], \boldsymbol{\lambda})^{-1} = 1.$$

(Вариант рассуждений:
https://math.stackexchange.com/questions/112786/convergence-in-law-and-uniformly-integrability)

ДЗ: теория

Задание 5: доказать утверждение

Утверждение

Пусть $q_{\boldsymbol{\Gamma}}$ — абсолютно непрерывное распределение с дифференцируемой плотностью, такой что:

1. Градиент плотности $\nabla_{\boldsymbol{\theta}_{\boldsymbol{\Gamma}}} q(\boldsymbol{\Gamma}|\boldsymbol{\theta}_{\boldsymbol{\Gamma}})$ является ненулевым почти всюду.
2. Выражение $\nabla_{\boldsymbol{\theta}_{\boldsymbol{\Gamma}}} q(\boldsymbol{\Gamma}|\boldsymbol{\theta}_{\boldsymbol{\Gamma}}) \log p(\boldsymbol{\Gamma}|\mathbf{h}, \boldsymbol{\lambda})$ ограничено на $U_{\boldsymbol{\theta}}$ абсолютно непрерывной случайной величиной, не зависящей от $\boldsymbol{\Gamma}$, с конечным первым моментом.
3. Не существует значений метапараметров $\boldsymbol{\lambda}_1, \boldsymbol{\lambda}_2$, таких что:

$$p(\boldsymbol{\Gamma}|\mathbf{h}, \boldsymbol{\lambda}_1) = p(\boldsymbol{\Gamma}|\mathbf{h}, \boldsymbol{\lambda}_2)$$

   для всех $\boldsymbol{\Gamma}$.

Тогда оптимизация вариационной оценки не удовлетворяет критерию 2 перехода между структурами.

ДЗ: практика

Формат: ipynb. Реализовать пример удаления параметров для логистической регрессии на MNIST и сравнить качество со случайным удалением параметров (ось X — процент удаленных параметров):

1. Удаление по модулю с использованием L_1, L_2 - регуляризаций. С переобучением после прунинга и без;
2. С использованием вариационного вывода (Graves, 2011);
3. Optimal Brain Damage.

При оценивании будут учитываться аккуратность кода ноутбуков и наглядность примера.

Используемые материалы

1. Schmidhuber, Jürgen. "A neural network that embeds its own meta-levels." IEEE International Conference on Neural Networks. IEEE, 1993.
2. Andrychowicz, Marcin, et al. "Learning to learn by gradient descent by gradient descent." Advances in neural information processing systems. 2016.
3. LeCun, Yann, John S. Denker, and Sara A. Solla. "Optimal brain damage." Advances in neural information processing systems. 1990.
4. Han, Song, et al. "Learning both weights and connections for efficient neural network." Advances in neural information processing systems. 2015.
5. Han, Song, Huizi Mao, and William J. Dally. "Deep compression: Compressing deep neural networks with pruning, trained quantization and huffman coding." arXiv preprint arXiv:1510.00149 (2015).
6. Graves, Alex. "Practical variational inference for neural networks." Advances in neural information processing systems. 2011.
7. Louizos, Christos, Karen Ullrich, and Max Welling. "Bayesian compression for deep learning." Advances in Neural Information Processing Systems. 2017.
8. Adams, Ryan, Hanna Wallach, and Zoubin Ghahramani. "Learning the structure of deep sparse graphical models." Proceedings of the thirteenth international conference on artificial intelligence and statistics. 2010.
9. http://web.ipac.caltech.edu/staff/fmasci/home/mystats/CauchyVsGaussian.pdf
10. Грабовой АВ, Бахтеев ОЮ, Стрижов ВВ. Определение релевантности параметров нейросети. Информатика и её применения. 2019;13(2):62-70.