Оптимизация гиперпараметров

Бахтеев Олег

МФТИ

23.10.2019

Что такое 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).

Постановка задачи

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

Постановка задачи

Пусть $\boldsymbol{\theta} \in \mathbb{R}^s$ — 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:

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

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

Байесовский вывод

Пусть $\boldsymbol{\theta} = [\mathbf{w}]^{\mathsf{T}}$.
Первый уровень:

$$\boldsymbol{\theta}^* = \arg\max\big(-L(\boldsymbol{\theta}, \mathbf{h})\big) = p(\mathbf{w}|\mathbf{X}, \mathbf{y}, \mathbf{h}) = \frac{p(\mathbf{y}|\mathbf{X}, \mathbf{w})p(\mathbf{w}|\mathbf{h})}{p(\mathbf{y}|\mathbf{X}, \mathbf{h})}.$$

Второй уровень:

$$p(\mathbf{h}|\mathbf{X}, \mathbf{y}) \propto p(\mathbf{y}|\mathbf{X}, \mathbf{h})p(\mathbf{h}),$$

Полагая распределение параметров $p(\mathbf{h})$ равномерным на некоторой большой окрестности, получим задачу оптимизации гиперпараметров:

$$Q(\boldsymbol{\theta}, \mathbf{h}) = p(\mathbf{y}|\mathbf{X}, \mathbf{h}) = \int_{\mathbf{w}\in\mathbb{R}^u} p(\mathbf{y}|\mathbf{X}, \mathbf{w})p(\mathbf{w}|\mathbf{h}) \to \max_{\mathbf{h}\in\mathbb{H}}.$$

Кросс-валидация

Разобьем выборку $\mathfrak{D}$ на k равных частей:

$$\mathfrak{D} = \mathfrak{D}_1 \sqcup \cdots \sqcup \mathfrak{D}_k.$$

Запустим k оптимизаций модели, каждую на своей части выборки. Положим $\boldsymbol{\theta} = [\mathbf{w}_1, \ldots, \mathbf{w}_k]$, где $\mathbf{w}_1, \ldots, \mathbf{w}_k$ — параметры модели при оптимизации k.
Пусть L — функция потерь:

$$L(\boldsymbol{\theta}, \mathbf{h}) = -\frac{1}{k} \sum_{q=1}^{k} \Big(\frac{k}{k-1} \log p(\mathbf{y} \setminus \mathbf{y}_q | \mathbf{X} \setminus \mathbf{X}_q, \mathbf{w}_q) + \log p(\mathbf{w}_q | \mathbf{h}) \Big). \tag{1}$$

Пусть Q — функция качества модели:

$$Q(\boldsymbol{\theta}, \mathbf{h}) = \frac{1}{k} \sum_{q=1}^{k} k \log p(\mathbf{y}_q | \mathbf{X}_q, \mathbf{w}_q).$$

Вариационная нижняя оценка

Пусть $L = -Q$:

$$\log\, p(\mathbf{y}|\mathbf{X}, \mathbf{A}) \geq \sum_{\mathbf{x},y} \log\, p(y|\mathbf{x}, \hat{\mathbf{w}}) - D_{\mathrm{KL}}\big(q(\mathbf{w})||p(\mathbf{w}|\mathbf{A})\big) = -L(\boldsymbol{\theta}, \mathbf{A}^{-1}) = Q(\boldsymbol{\theta}, \mathbf{A}^{-1}),$$

где q — нормальное распределение с диагональной матрицей ковариаций:

$$q \sim \mathcal{N}(\boldsymbol{\mu}_q, \mathbf{A}_q^{-1}),$$

$$D_{\mathrm{KL}}\big(q(\mathbf{w})||p(\mathbf{w}|\mathbf{f})\big) = \frac{1}{2}\big(\mathrm{Tr}[\mathbf{A}\mathbf{A}_q^{-1}] + (\boldsymbol{\mu} - \boldsymbol{\mu}_q)^{\mathrm{T}}\mathbf{A}(\boldsymbol{\mu} - \boldsymbol{\mu}_q) - u + \ln\, |\mathbf{A}^{-1}| - \ln\, |\mathbf{A}_q^{-1}|\big).$$

В качестве оптимизируемых параметров $\boldsymbol{\theta}$ выступают параметры распределения q:

$$\boldsymbol{\theta} = [\alpha_1, \ldots, \alpha_u, \mu_1, \ldots, \mu_u].$$

Evidence vs Кросс-валидация

Оценка Evidece:

$$\log p(\mathfrak{D}|\mathbf{f}) = \log p(\mathfrak{D}_1|\mathbf{f}) + \log p(\mathfrak{D}_2|\mathfrak{D}_1, \mathbf{f}) + \cdots + \log p(\mathfrak{D}_n|\mathfrak{D}_1, \ldots, \mathfrak{D}_{n-1}, \mathbf{f}).$$

Оценка leave-one-out:

$$\text{LOU} = \text{E}\log p(\mathfrak{D}_n|\mathfrak{D}_1, \ldots, \mathfrak{D}_{n-1}, \mathbf{f}).$$

Кросс-валидация использует среднее значение последнего члена $p(\mathfrak{D}_n|\mathfrak{D}_1, \ldots, \mathfrak{D}_{n-1}, \mathbf{f})$ для оценки сложности.
Evidence учитывает **полную** сложность описания заданной выборки, определяющую предсказательную способность модели с самого начала.

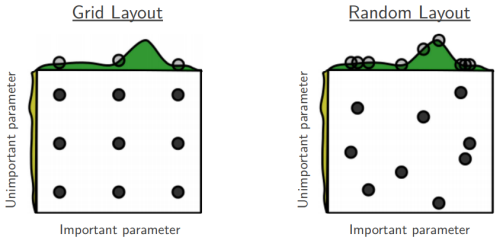
Базовые методы оптимизации гиперпараметров

Варианты:

- Поиск по решетке;
- Случайный поиск.

Оба метода страдают от проклятия размерности.

Случайный поиск может быть более эффективным, если пространство гиперпараметров вырождено.



Bergstra et al., 2012

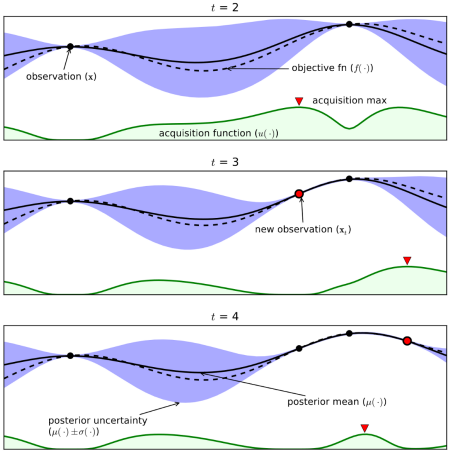
Гауссовый процесс

Идея:
Будем моделировать $Q(\theta(\mathbf{h})^*, \mathbf{h})$ гауссовым процессом, зависящим от $\mathbf{h}$.

Плюсы:

- Гибкость модели.
- Дешевле, чем обучения модели.

Минусы: кубическая сложность по количеству гиперпараметров.



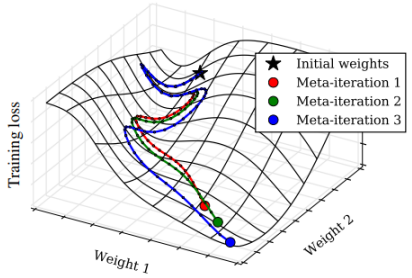
Shahriari et. al, 2016. Пример работы гауссового процесса.

Градиентные методы

Идея: Будем производить оптимизацию вдоль всей траектории оптимизации параметров.
Плюсы:

- Оптимизация гиперпараметров будет учитывать оптимизацию параметров.
- Сложность меняется незначительно от количества гиперпараметров.

Минусы: вычилсительно дорого.



Maclaurin et. al, 2015. Пример работы.

Формальная постановка задачи: градиентная оптимизация

Определение

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

$$\hat{\boldsymbol{\theta}} = T \circ T \circ \cdots \circ T(\boldsymbol{\theta}_0, \mathbf{h}) = T^{\eta}(\boldsymbol{\theta}_0, \mathbf{h}), \tag{2}$$

где

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

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:

$$\mathbf{h}^* = \arg\max_{\mathbf{h} \in \mathbb{H}} Q(T^{\eta}(\boldsymbol{\theta}_0, \mathbf{h})),$$

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

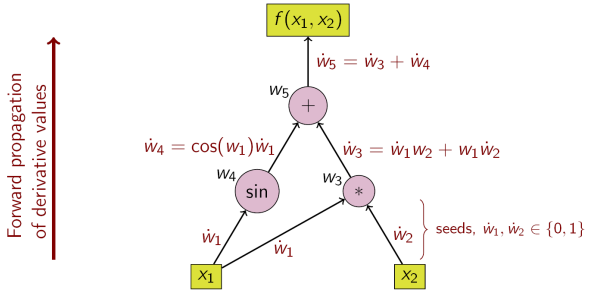
Forward-mode differentiation

Идея дифференцирования: применение формулы:

$$\frac{\partial y}{\partial x} = \frac{\partial y}{\partial w_{n-1}} \frac{\partial w_{n-1}}{\partial x} = \frac{\partial y}{\partial w_{n-1}} \left(\frac{\partial w_{n-1}}{\partial w_{n-2}} \frac{\partial w_{n-2}}{x} \partial x \right) = \dots .$$

Пример (wiki):

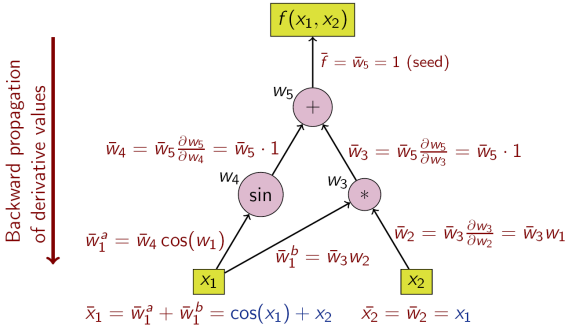
$$x_1 x_2 + \sin(x_1)$$



Reverse-mode differentiation

Идея дифференцирования: применение формулы:

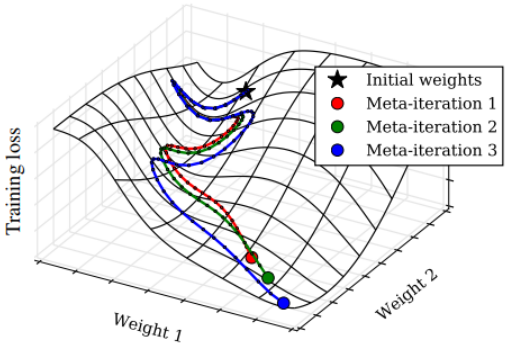
$$\frac{\partial y}{\partial x} = \frac{\partial y}{\partial w_1}\frac{\partial w_1}{\partial x} = \left(\frac{\partial y}{\partial w_2}\frac{\partial w_2}{\partial w_1}\right)\frac{\partial w_1}{\partial x} = \dots.$$



RMAD, Maclaurin et. al, 2015

Алгоритм RMAD основывается на Reverse-mode differentiation.

1. Провести η шагов оптимизации с моментом γ: $\boldsymbol{\theta} = T(\boldsymbol{\theta}_0, \mathbf{h})$.
2. Положим $\hat{\nabla}\mathbf{h} = \nabla_{\mathbf{h}} Q(\boldsymbol{\theta}, \mathbf{h})$.
3. Положим $d\mathbf{v} = \mathbf{0}$.
4. Для $\tau = \eta \dots 1$ повторить:
5. Вычислить $\boldsymbol{\theta}^{\tau-1}$.
6. Вычислить градиент на шаге $\tau - 1$, используя RMD.



DrMAD

Алгоритм DrMad — упрощенный RMAD. Вводится предположение о линейности траектории обновления параметров $\boldsymbol{\theta}$.

1. Провести η шагов оптимизации с моментом γ: $\boldsymbol{\theta} = T(\boldsymbol{\theta}_0, \mathbf{h})$.
2. Положим $\hat{\nabla}\mathbf{h} = \nabla_{\mathbf{h}} Q(\boldsymbol{\theta}, \mathbf{h})$.
3. Положим $d\mathbf{v} = \mathbf{0}$.
4. Для $\tau = \eta \dots 1$ повторить:
5. Вычислить $\boldsymbol{\theta}^{\tau-1}$.
6. Вычислить градиент на шаге $\tau - 1$, используя RMD.

1. Провести η шагов оптимизации с моментом γ: $\boldsymbol{\theta} = T(\boldsymbol{\theta}_0, \mathbf{h})$.
2. Положим $\hat{\nabla}\mathbf{h} = \nabla_{\mathbf{h}} Q(\boldsymbol{\theta}, \mathbf{h})$.
3. Положим $d\mathbf{v} = \mathbf{0}$.
4. Для $\tau = \eta \dots 1$ повторить:
5. $\boldsymbol{\theta}^{\tau-1} = \boldsymbol{\theta}_0 + \frac{\tau-1}{\eta}\boldsymbol{\theta}^{\eta}$.
6. Вычислить градиент на шаге $\tau - 1$, используя RMD.

Аналитическая формула оптимизации параметров

Утверждение (Pedregosa, 2016)

Пусть L — дифференцируемая функция, такая что все стационарные точки L являются локальными минимумами. Пусть также гессиан $\mathbf{H}^{-1}$ функции потерь L является обратимым в каждой стационарной точке.
Тогда

$$\nabla_{\mathbf{h}} Q(T(\theta_0, \mathbf{h}), \mathbf{h}) = \nabla_{\mathbf{h}} Q(\theta^\eta, \mathbf{h}) - \nabla_{\mathbf{h}} \nabla_{\theta} L(\theta^\eta, \mathbf{h})^{\mathsf{T}} \mathbf{H}^{-1} \nabla_{\theta} Q(\theta^\eta, \mathbf{h}).$$

Жадная оптимизация гиперпараметров

На каждом шаге оптимизации параметров $\boldsymbol{\theta}$:

$$\mathbf{h}' = \mathbf{h} - \beta_{\mathbf{h}} \nabla_{\mathbf{h}} Q\big(T(\boldsymbol{\theta}, \mathbf{h}), \mathbf{h}\big) = \mathbf{h} - \beta_{\mathbf{h}} \nabla_{\mathbf{h}} Q\big(\boldsymbol{\theta} - \beta \nabla L(\boldsymbol{\theta}, \mathbf{h}), \mathbf{h}\big)\big),$$

где $\beta_{\mathbf{h}}$ — длина шага оптимизации гиперпараметров.

- Можно рассматривать как упрощение алгоритма RMAD, использующее только один элемент истории обновления параметров.
- Является приближением к решению аналитической формуле в случае $\mathbf{H}^{-1} \sim \mathbf{I}$.

HOAG

Численное приближение аналитической формулы:

$$\nabla_{\mathbf{h}} Q(\boldsymbol{\theta}^{\eta}, \mathbf{h}) - \nabla_{\mathbf{h}} \nabla_{\boldsymbol{\theta}} L(\boldsymbol{\theta}^{\eta}, \mathbf{h})^{\mathrm{T}} \mathbf{H}^{-1} \nabla_{\boldsymbol{\theta}} Q(\boldsymbol{\theta}^{\eta}, \mathbf{h}).$$

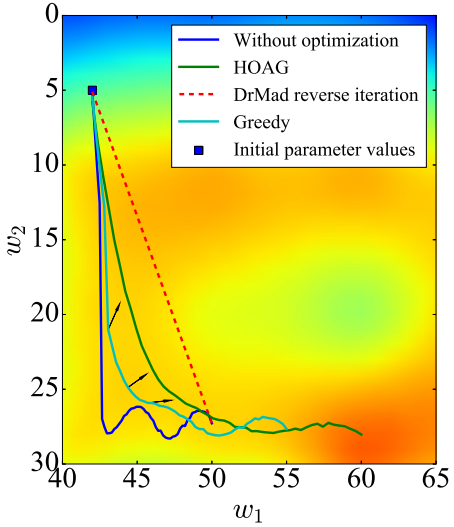
1. Провести η шагов оптимизации: $\boldsymbol{\theta} = T(\boldsymbol{\theta}_0, \mathbf{h})$.
2. Решить линейную систему для вектора $\boldsymbol{\lambda}$: $\mathbf{H}(\boldsymbol{\theta})\boldsymbol{\lambda} = \nabla_{\boldsymbol{\theta}} Q(\boldsymbol{\theta}, \mathbf{h})$.
3. Приближенное значение градиентов гиперпараметра вычисляется как: $\hat{\nabla}_{\mathbf{h}} Q = \nabla_{\mathbf{h}} Q(\boldsymbol{\theta}, \mathbf{h}) - \nabla_{\boldsymbol{\theta}, \mathbf{h}} L(\boldsymbol{\theta}, \mathbf{h})^{T} \boldsymbol{\lambda}$.

Итоговое правило обновления:

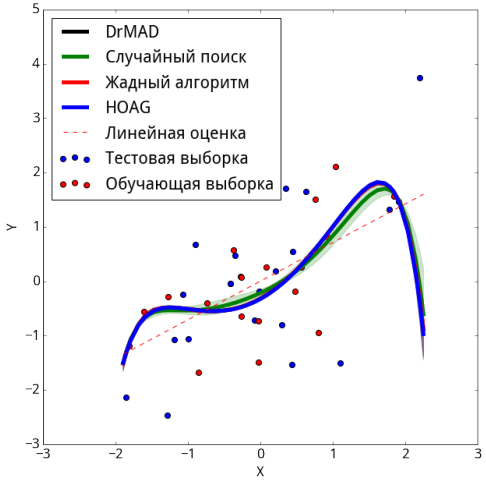
$$\mathbf{h}' = \mathbf{h} - \gamma_{\mathbf{h}} \hat{\nabla}_{\mathbf{h}} Q.$$

Сравнение алгоритмов

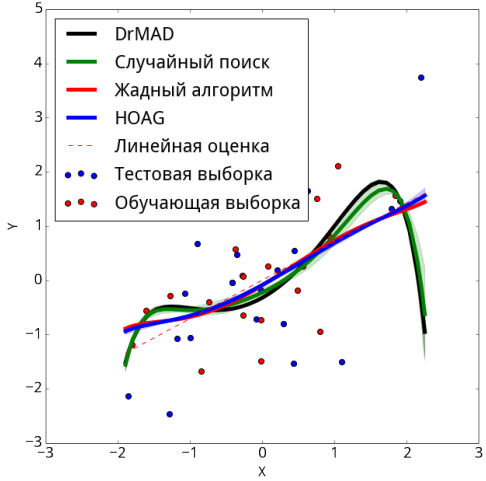
Алгоритм	+	-
Random search	Легко реализовать	Проклятие размерности
Жадная оптимизация	Оптимизация проводится внутри цикла оптимизации параметров. Легко реализовать	Жадность, неоптимальность.
HOAG	Быстрая сходимость.	Качество результатов зависит от решения линейного уравнения $\mathbf{H}(\boldsymbol{\theta})\boldsymbol{\lambda} = \nabla_{\boldsymbol{\theta}} Q(\boldsymbol{\theta}, \mathbf{h})$.
DrMAD	Учитывает особенности оператора оптимизации. Можно использовать для оптимизации мета-параметров.	Неустойчив при больших значениях длины градиентного шага $\gamma_{\mathbf{h}}$. Качество оптимизации зависит от кривизны траектории обновления параметров.


Without optimization
HOAG
DrMad reverse iteration
Greedy
Initial parameter values
w_1
w_2
0
5
10
15
20
25
30
40
45
50
55
60
65

Эксперименты: полиномы

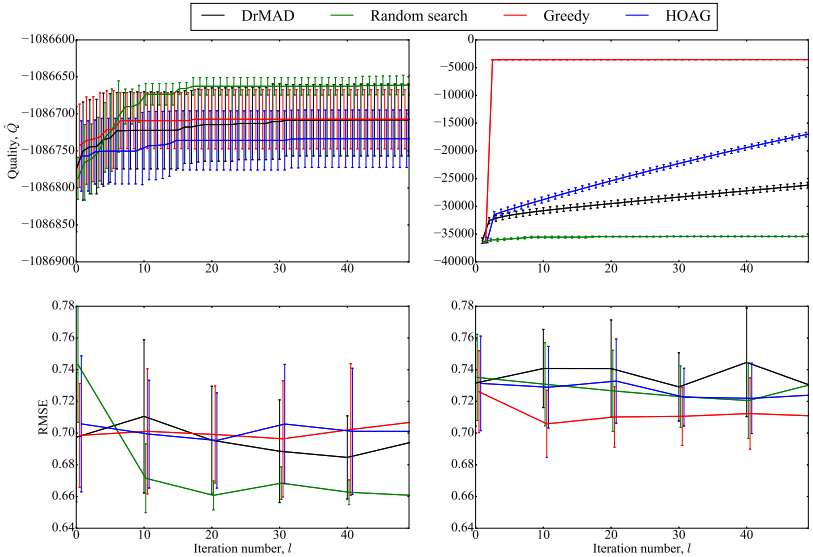


Кросс-валидация

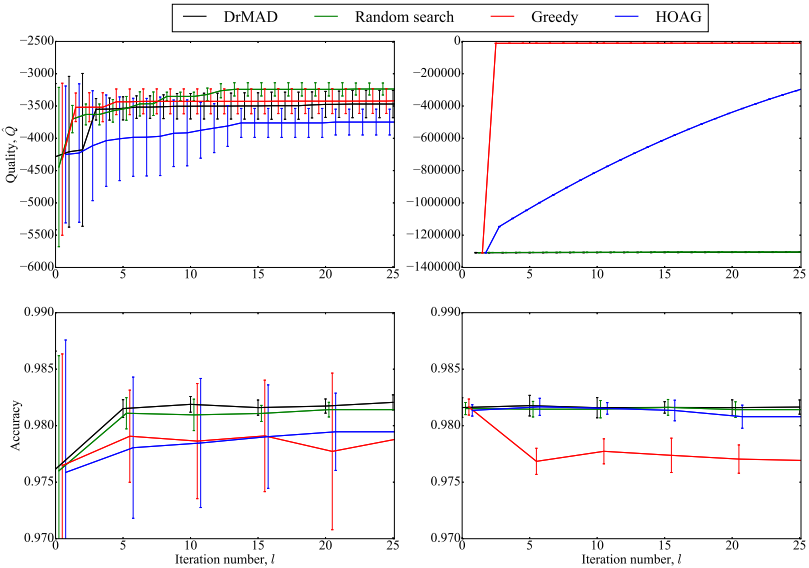


Evidence

Эксперименты: WISDM



Эксперименты: MNIST



Эксперименты: MNIST

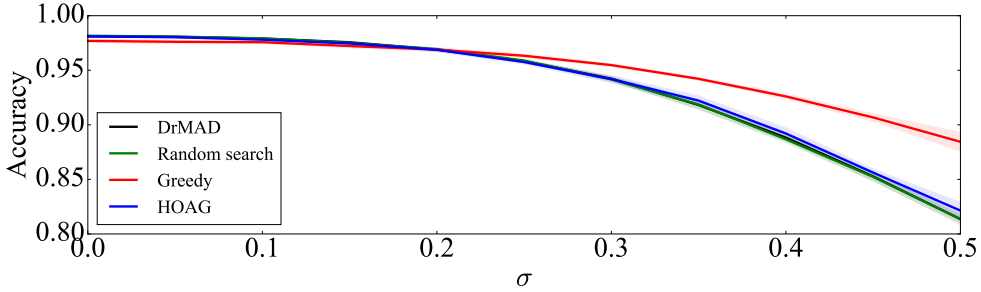
Добавление гауссового шума $\mathcal{N}(\mathbf{0}, \sigma^2\mathbf{I})$:

Без шума

$\sigma = 0.1$

$\sigma = 0.25$

$\sigma = 0.5$



ДЗ: выбор задания

Дедлайн: 30 октября, 0 часов.

```
 from zlib import crc32

theory = crc32('фамилия кириллицей'.lower().encode('utf-8'))%2+1

practice = crc32('фамилия латиницей'.lower().encode('utf-8'))%2+1
```

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

ДЗ: теория

Формат: tex + pdf.

1. Доказать утверждение (Pedregosa, 2016);
   - Воспользоваться (Pedregosa, 2016);
2. Расписать с комментариями RMAD для SGD без момента;
   - Воспользоваться (Maclaurin et al., 2015).

ДЗ: практика

Формат: ipynb. Реализовать пример оптимизации гиперпараметров на небольшой выборке с ошибкой на валидации в качестве функции Q.
Количество гиперпараметров: не менее 20.
Рассмотреть алгоритмы: случайный поиск, гауссовый процесс (библиотечная реализация) и:

1. HOAG;
2. Жадный алгоритм.

При оценивании будут учитываться аккуратность кода ноутбуков и наглядность примера.

Пример должен быть выполнен на **простых** игрушечных синтетических данных.

Используемые материалы

1. David J. C. MacKay, Information Theory, Inference & Learning Algorithms, 2003
2. Christopher Bishop, Pattern Recognition and Machine Learning, 2006
3. Bergstra et al., Random Search for Hyper-Parameter Optimization, 2012
4. Dougal Maclaurin et. al, Gradient-based Hyperparameter Optimization through Reversible Learning, 2015
5. Jelena Luketina et. al, Scalable Gradient-Based Tuning of Continuous Regularization Hyperparameters, 2016
6. Jie Fu et. al, DrMAD: Distilling Reverse-Mode Automatic Differentiation for Optimizing Hyperparameters of Deep Neural Networks, 2016
7. Fabian Pedregosa, Hyperparameter optimization with approximate gradient, 2016
8. Bobak Shahriari et. al, Taking the Human Out of the Loop: A Review of Bayesian Optimization, 2016
9. Bakhteev, Strijov, Comprehensive analysis of gradient-based hyperparameter optimization algorithms, 2018
10. Feurer et al, AUTOML: METHODS, SYSTEMS, CHALLENGES