

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 :

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

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 $[\alpha_1, \dots, \alpha_u] = \mathbf{h}$.

Постановка задачи

Пусть $\theta \in \mathbb{R}^s$ — множество всех оптимизируемых параметров.

$L(\theta, \mathbf{h})$ — дифференцируемая функция потерь по которой производится оптимизация функции \mathbf{f} .

$Q(\theta, \mathbf{h})$ — дифференцируемая функция определяющая итоговое качество модели \mathbf{f} и приближающая интеграл.

Требуется найти параметры θ^* и гиперпараметры \mathbf{h}^* модели, доставляющие минимум следующему функционалу:

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

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

Байесовский вывод

Пусть $\theta = [\mathbf{w}]^T$.

Первый уровень:

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

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:

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

Кросс-валидация

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

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

Запустим k оптимизаций модели, каждую на своей части выборки. Положим $\theta = [\mathbf{w}_1, \dots, \mathbf{w}_k]$, где $\mathbf{w}_1, \dots, \mathbf{w}_k$ — параметры модели при оптимизации k .

Пусть L — функция потерь:

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

Пусть Q — функция качества модели:

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

Вариационная нижняя оценка

Пусть $L = -Q$:

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

где q — нормальное распределение с диагональной матрицей ковариаций:

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

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

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

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

Evidence vs Кросс-валидация

Оценка Evidence:

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

Оценка leave-one-out:

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

Кросс-валидация использует среднее значение последнего члена $p(\mathcal{D}_n|\mathcal{D}_1, \dots, \mathcal{D}_{n-1}, \mathbf{f})$ для оценки сложности.

Evidence учитывает **полную** сложность описания заданной выборки, определяющую предсказательную способность модели с самого начала.

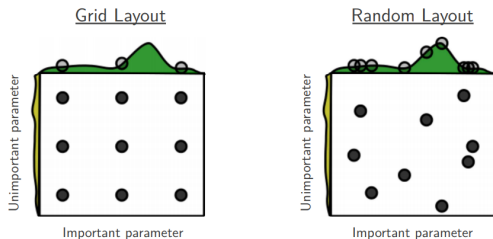
Базовые методы оптимизации гиперпараметров

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

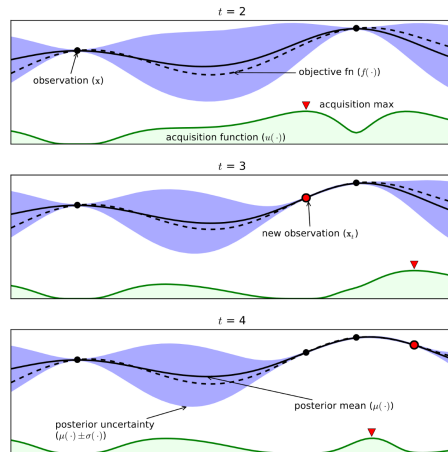
Идея:

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.

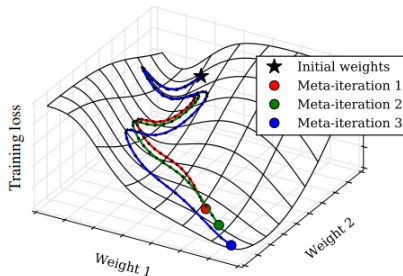
Градиентные методы

Идея: Будем производить оптимизацию вдоль всей траектории оптимизации параметров.

Плюсы:

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

Минусы: вычислительно дорого.



Maclaurin et. al, 2015. Пример работы.

Формальная постановка задачи: градиентная оптимизация

Определение

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

$$\hat{\theta} = T \circ T \circ \dots \circ T(\theta_0, \mathbf{h}) = T^\eta(\theta_0, \mathbf{h}), \quad (2)$$

где

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

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

Перепишем итоговую задачу оптимизации:

$$\mathbf{h}^* = \arg \max_{\mathbf{h} \in \mathbb{H}} Q(T^\eta(\theta_0, \mathbf{h})),$$

где θ_0 — начальное значение параметров θ .

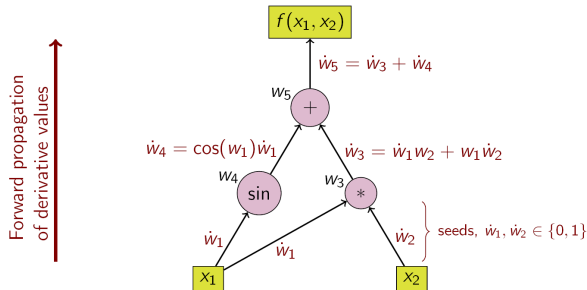
Forward-mode differentiation

Идея дифференцирования: применение формулы:

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

Пример (wiki):

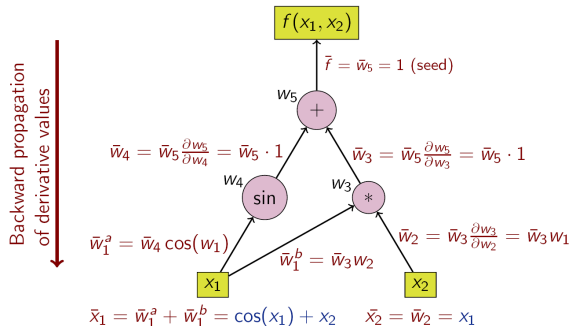
$$x_1 x_2 + \sin(x_1)$$



Reverse-mode differentiation

Идея дифференцирования: применение формулы:

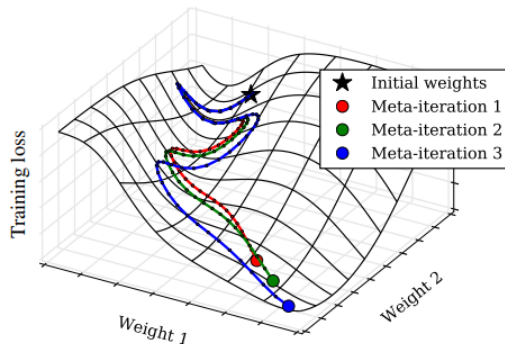
$$\frac{\partial y}{\partial x} = \frac{\partial y}{\partial w_1} \frac{\partial w_1}{\partial x} = \left(\frac{\partial y}{\partial w_2} \frac{\partial w_2}{\partial w_1} \right) \frac{\partial w_1}{\partial x} = \dots$$



RMAD, Maclaurin et. al, 2015

Алгоритм RMAD основывается на Reverse-mode differentiation.

- ① Провести η шагов оптимизации с моментом γ : $\theta = T(\theta_0, \mathbf{h})$.
- ② Положим $\hat{\nabla} \mathbf{h} = \nabla_{\mathbf{h}} Q(\theta, \mathbf{h})$.
- ③ Положим $d\mathbf{v} = \mathbf{0}$.
- ④ Для $\tau = \eta \dots 1$ повторить:
- ⑤ Вычислить $\theta^{\tau-1}$.
- ⑥ Вычислить градиент на шаге $\tau - 1$, используя RMD.



DrMAD

Алгоритм DrMad — упрощенный RMAD. Вводится предположение о линейности траектории обновления параметров θ .

- | | |
|---|--|
| <ul style="list-style-type: none">① Провести η шагов оптимизации с моментом γ: $\theta = T(\theta_0, \mathbf{h})$.② Положим $\hat{\nabla} \mathbf{h} = \nabla_{\mathbf{h}} Q(\theta, \mathbf{h})$.③ Положим $d\mathbf{v} = \mathbf{0}$.④ Для $\tau = \eta \dots 1$ повторить:⑤ Вычислить $\theta^{\tau-1}$.⑥ Вычислить градиент на шаге $\tau - 1$, используя RMD. | <ul style="list-style-type: none">① Провести η шагов оптимизации с моментом γ: $\theta = T(\theta_0, \mathbf{h})$.② Положим $\hat{\nabla} \mathbf{h} = \nabla_{\mathbf{h}} Q(\theta, \mathbf{h})$.③ Положим $d\mathbf{v} = \mathbf{0}$.④ Для $\tau = \eta \dots 1$ повторить:⑤ $\theta^{\tau-1} = \theta_0 + \frac{\tau-1}{\eta} \theta^\eta$.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

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

Жадная оптимизация гиперпараметров

На каждом шаге оптимизации параметров θ :

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

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:

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

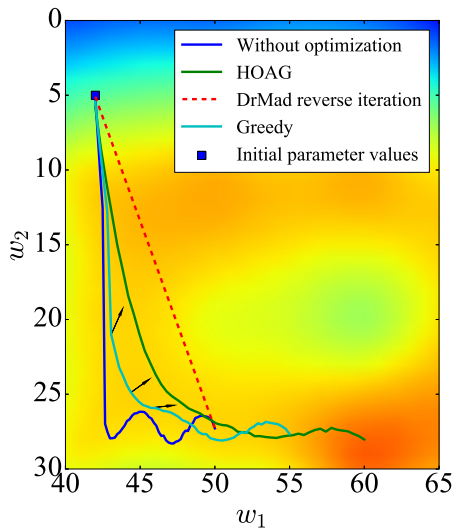
- ① Провести η шагов оптимизации: $\boldsymbol{\theta} = T(\boldsymbol{\theta}_0, \mathbf{h})$.
- ② Решить линейную систему для вектора $\boldsymbol{\lambda}$: $\mathbf{H}(\boldsymbol{\theta})\boldsymbol{\lambda} = \nabla_{\boldsymbol{\theta}} Q(\boldsymbol{\theta}, \mathbf{h})$.
- ③ Приближенное значение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

Итоговое правило обновления:

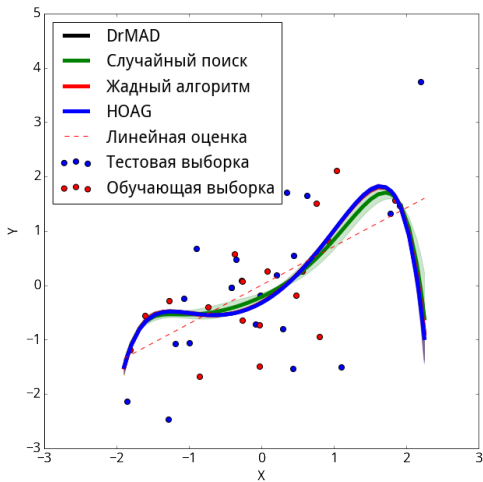
$$\mathbf{h}' = \mathbf{h} - \gamma_{\mathbf{h}} \hat{\nabla}_{\mathbf{h}} Q.$$

Сравнение алгоритмов

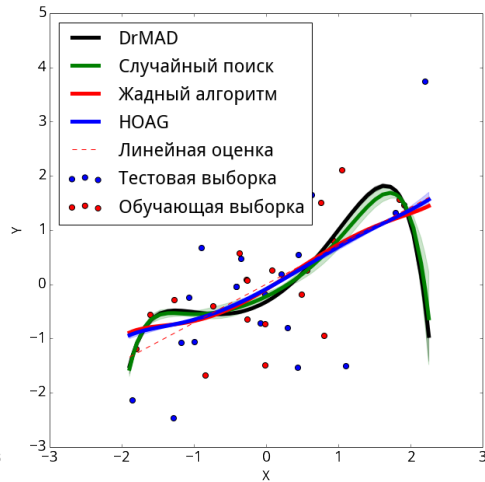
Алгоритм	+	-
Random search	Легко реализовать	Проклятие размерности
Жадная оптимизация	Оптимизация проводится внутри цикла оптимизации параметров. Легко реализовать	Жадность, неоптимальность.
HOAG	Быстрая сходимость.	Качество результатов зависит от решения линейного уравнения $\mathbf{H}(\theta)\lambda = \nabla_{\theta} Q(\theta, \mathbf{h})$.
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.



Эксперименты: полиномы

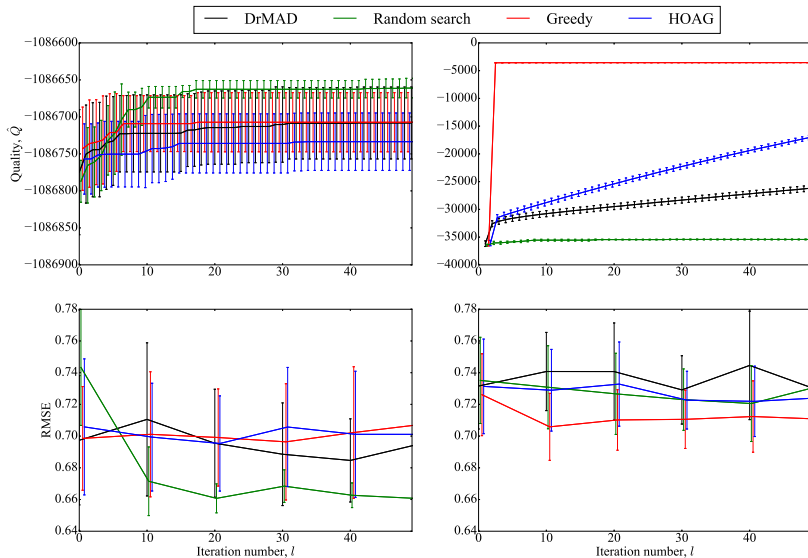


Кросс-валидация

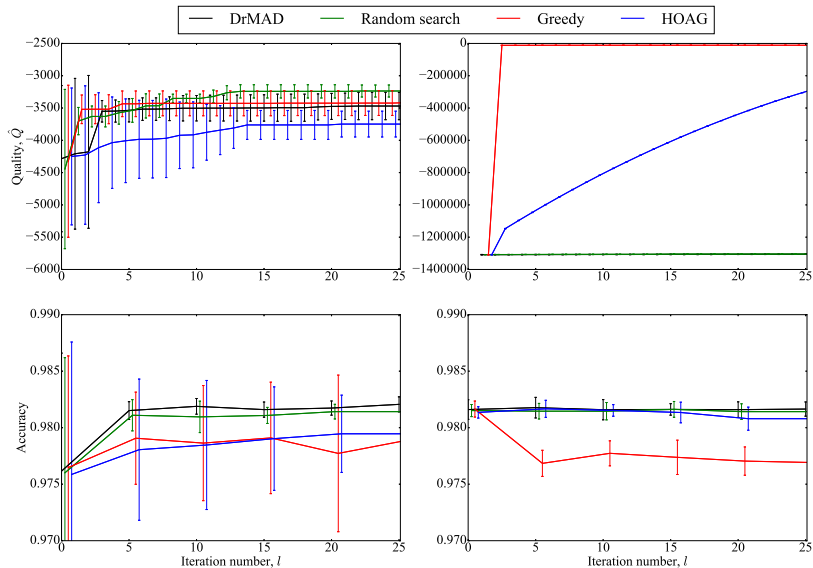


Evidence

Эксперименты: WISDM



Эксперименты: MNIST



Эксперименты: MNIST

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



Без шума



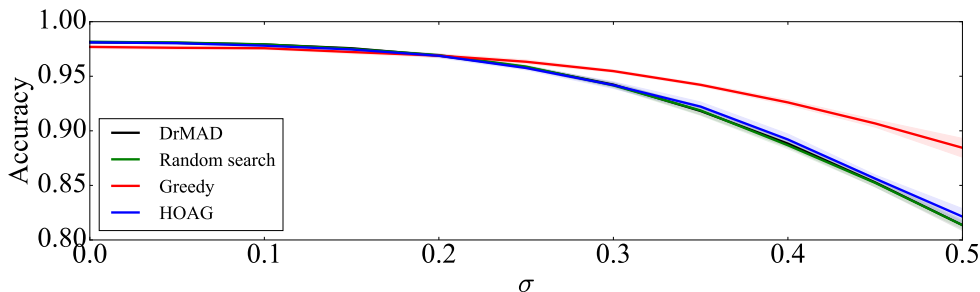
$\sigma = 0.1$



$\sigma = 0.25$



$\sigma = 0.5$



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: ірупв. 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 НОАГ;
- 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