Автоматическое определение релевантности параметров нейросети

Грабовой Андрей Валериевич

Московский физико-технический институт

ИОИ-12, г. Гаэта, Италия, 2018

Цель работы

Иследуются

Методы автоматического прореживания нейросетей.

Требуется

Предложить алгоритм нахождения оптимальной структуры модели на основе алгоритмов прореживания.

Проблемы

- Вычислительная сложность оптимизации,
- Невозможность получения адекватной статистической оценки параметров.

Литература

- *LeCun Y., Denker J. , Solla S.*
  Optimal Brain Damage // Advances in Neural Information Processing Systems, 1989. Vol. 2. P. 598–605.
- *Graves A.*
  Practical Variational Inference for Neural Networks // Advances in Neural Information Processing Systems, 2011. P. 2348–2356.
- *Bishop C.*
  Pattern Recognition and Machine Learning. — Berlin: Springer, 2006. 758 p.
- *Neychev R., Katrutsa A., Strijov V.*
  Robust selection of multicollinear features in forecasting // Factory Laboratory, 2016. Vol. 82. No 2. P. 68–74.

Постановка задачи

$$\mathfrak{D} = \{\mathbf{x}_i, y_i\},\ i = 1, ..., N,$$

где $\mathbf{x}_i \in \mathbb{R}^m$, $y_i \in \{1, \ldots, Y\}$, Y — число классов.

$$f(\mathbf{x}, \mathbf{w}) : \mathbb{R}^m \times \mathbb{R}^n \to \{1, \ldots, Y\},$$

где $\mathbf{w} \in \mathbb{R}^n$ — пространство параметров модели

$$f(\mathbf{x}, \mathbf{w}) = \text{softmax}\big(f_1(f_2(...(f_l(\mathbf{x}, \mathbf{w}))\big),$$

где $f_i(\mathbf{x}, \mathbf{w}) = \sigma(\mathbf{w}^\mathsf{T}\mathbf{x})$, $i \in \{1 \ldots l\}$, l — число слоев нейросети, σ некоторая функция активации.

$$\mathbb{W}_\mathcal{A} = \{\mathbf{w} \in \mathbb{R}^n \mid w_j \neq 0,\ j \in \mathcal{A}\},$$

где $\mathcal{A} \subset \mathcal{J} = \{1, ..., n\}$ — множество активных индексов

Постановка задачи

Правдоподобие выборки:

$$\mathcal{L}_{\mathfrak{D}}(\mathfrak{D}, \mathcal{A}, \mathbf{w}) = \log p(\mathfrak{D}|\mathcal{A}, \mathbf{w}),$$

где $p(\mathfrak{D}|\mathcal{A}, \mathbf{w})$ — апостериорная вероятность $\mathfrak{D}$ при заданых $\mathbf{w}, \mathcal{A}$

Правдоподобие модели:

$$\mathcal{L}_{\mathcal{A}}(\mathfrak{D}, \mathcal{A}) = \log p(\mathfrak{D}|\mathcal{A}) = \log \int_{\mathbf{w} \in \mathbb{W}_{\mathcal{J}}} p(\mathfrak{D}|\mathbf{w}) p(\mathbf{w}|\mathcal{A}) d\mathbf{w},$$

где $p(\mathbf{w}|\mathcal{A})$ — априорная вероятность $\mathbf{w}$ в пространстве $\mathbb{W}_{\mathcal{A}}$

Постановка задачи

$$\mathcal{L}_{\mathcal{A}}(\mathfrak{D},\mathcal{A}) = \log p(\mathfrak{D}|\mathcal{A}) = \log \int_{\mathbf{w}\in\mathbb{W}_{\mathcal{J}}} p(\mathfrak{D}|\mathbf{w})p(\mathbf{w}|\mathcal{A})d\mathbf{w} =$$

$$= \int_{\mathbf{w}\in\mathbb{W}_{\mathcal{J}}} q(\mathbf{w}) \log \frac{p(\mathfrak{D},\mathbf{w}|\mathcal{A})}{q(\mathbf{w})} d\mathbf{w} - \int_{\mathbf{w}\in\mathbb{W}_{\mathcal{J}}} q(\mathbf{w}) \log \frac{p(\mathbf{w}|\mathfrak{D},\mathcal{A})}{q(\mathbf{w})} d\mathbf{w} \approx$$

$$\approx \int_{\mathbf{w}\in\mathbb{W}_{\mathcal{J}}} q(\mathbf{w}) \log \frac{p(\mathfrak{D},\mathbf{w}|\mathcal{A})}{q(\mathbf{w})} d\mathbf{w} =$$

$$= \int_{\mathbf{w}\in\mathbb{W}_{\mathcal{J}}} q(\mathbf{w}) \log \frac{p(\mathbf{w}|\mathcal{A})}{q(\mathbf{w})} d\mathbf{w} + \int_{\mathbf{w}\in\mathbb{W}_{\mathcal{J}}} q(\mathbf{w}) \log p(\mathfrak{D}|\mathcal{A},\mathbf{w}) d\mathbf{w} =$$

$$= \mathcal{L}_{\mathbf{w}}(\mathfrak{D},\mathcal{A},\mathbf{w}) + \mathcal{L}_{E}(\mathfrak{D},\mathcal{A}),$$

где $q(\mathbf{w})$ — распределение апроксимирующее неизвестное апостериорное распределение $p(\mathbf{w}|\mathfrak{D},\mathcal{A})$

Постановка задачи

Задача оптимизации:

$$\hat{\mathbf{w}} = \underset{\mathcal{A}\subset\mathcal{J},\ \mathbf{w}\in\mathbb{W}_{\mathcal{A}}}{\arg\min} -\mathcal{L}_{\mathcal{A}}(\mathfrak{D}, \mathcal{A}, \mathbf{w}) =$$

$$= \underset{\mathcal{A}\subset\mathcal{J},\ \mathbf{w}\in\mathbb{W}_{\mathcal{A}}}{\arg\min} D_{KL}\big(q(\mathbf{w})||p(\mathbf{w}|\mathcal{A})\big) - \mathcal{L}_{\mathfrak{D}}(\mathfrak{D}, \mathcal{A}, \mathbf{w})$$

Некоторые методы прореживания нейросетей

Случайное удаление параметров:

$\xi \sim \mathcal{U}(\mathcal{A})$ — индекс наименее релевантного параметра.

Оптимальное прореживание:

$\delta\mathcal{L} = \sum_{j\in\mathcal{A}} g_j \delta w_j + \frac{1}{2}\sum_{i,j\in\mathcal{A}} h_{ij}\delta w_i \delta w_j + O(||\delta\mathbf{w}||^3)$

Релеватность параметров определяется как рост ошибки вызванной удалением w_j:

$\xi = \arg\min_{j\in\mathcal{A}} h_{jj}\frac{w_j^2}{2}$ — индекс наименее релевантного параметра.

Вариационная оценка:

$\xi = \arg\max_{j\in\mathcal{A}} \frac{p_j(\mathbf{w}|\mathcal{A})(0)}{p_j(\mathbf{w}|\mathcal{A})(\mu_j)}$ — индекс наименее релевантного параметра.

Метод Белсли

Рассмотри:

$$\hat{\mathbf{w}} = \underset{\mathcal{A}\subset\mathcal{J},\ \mathbf{w}\in\mathbb{W}_{\mathcal{A}}}{\arg\min} -\mathcal{L}_{\mathfrak{D}}(\mathfrak{D}, \mathcal{A}, \mathbf{w})$$

Пусть:

$\mathbf{A}_{\text{ps}}$ — матрица ковариационная матрица вектора $\hat{\mathbf{w}}$

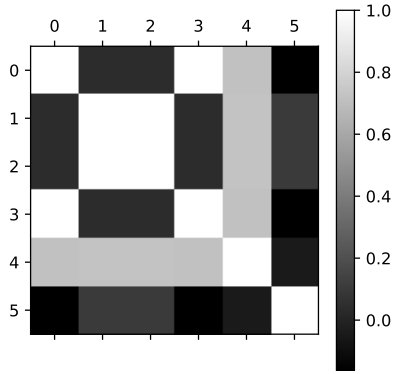
$$\mathbf{A}_{\text{ps}} = \mathbf{U}\mathbf{\Lambda}\mathbf{V}^{\mathsf{T}} \Rightarrow \eta_j = \frac{\lambda_j}{\max(\mathbf{\Lambda})}$$

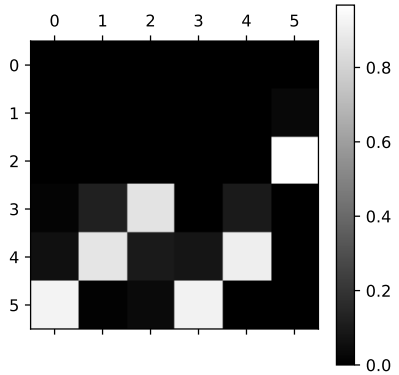
$$\xi = \underset{j\in\mathcal{A}}{\arg\max}\,\eta_j \qquad q_{ij} = \frac{u_{ij}^2/\lambda_{jj}}{\sum_{j=1}^{n} u_{ij}^2/\lambda_{jj}}$$

$q_{\xi j}$ — максимальные значения отвечают наиболее зависимым параметрам

Илюстрация метода Белсли

$$\hat{\mathbf{w}} = \begin{bmatrix} \sin(x) \\ \cos(x) \\ 2+\cos(x) \\ 2+\sin(x) \\ \cos(x)+\sin(x) \\ x \end{bmatrix}, \quad x \in [0.0, 0.02, ..., 20.0]$$





η_0	η_1	η_2	η_3	η_4	η_5
1.0	1.5	3.3	$2 \cdot 10^{15}$	$8 \cdot 10^{15}$	$1 \cdot 10^{16}$

Рис.: Илюстрация метода Белсли

Экспериментальные данные

Таблица: Описание выборок

Выборка	Тип задачи	Размер выборки	Число признаков
Wine	класификация	178	13
Boston Housing	регресия	506	13
Synthetic data	регресия	10000	100

Синтетические данные

Этап первый:

$$\mathbf{w}_{\text{synthetic}} \sim \mathcal{N}(\mathbf{m}_{\text{synthetic}}, \mathbf{A}_{\text{synthetic}})$$

$$\mathbf{m}_{\text{synthetic}} = \begin{bmatrix} 1.0 \\ 0.0025 \\ \dots \\ 0.0025 \end{bmatrix} \quad \mathbf{A}_{\text{synthetic}} = \begin{bmatrix} 1.0 & 10^{-3} & \cdots & 10^{-3} & 10^{-3} \\ 10^{-3} & 1.0 & \cdots & 0.95 & 0.95 \\ \dots & \dots & \dots & \dots & \dots \\ 10^{-3} & 0.95 & \cdots & 0.95 & 1.0 \end{bmatrix}$$

Этап второй:

$$\mathfrak{D}_{\text{synthetic}} = \{(\mathbf{x}_i, y_i) | \mathbf{x}_i \sim \mathcal{N}(\mathbf{1}, \mathbf{I}),\ y_i = x_{i0},\ i = 1...10000\}$$

Результаты эксперимента

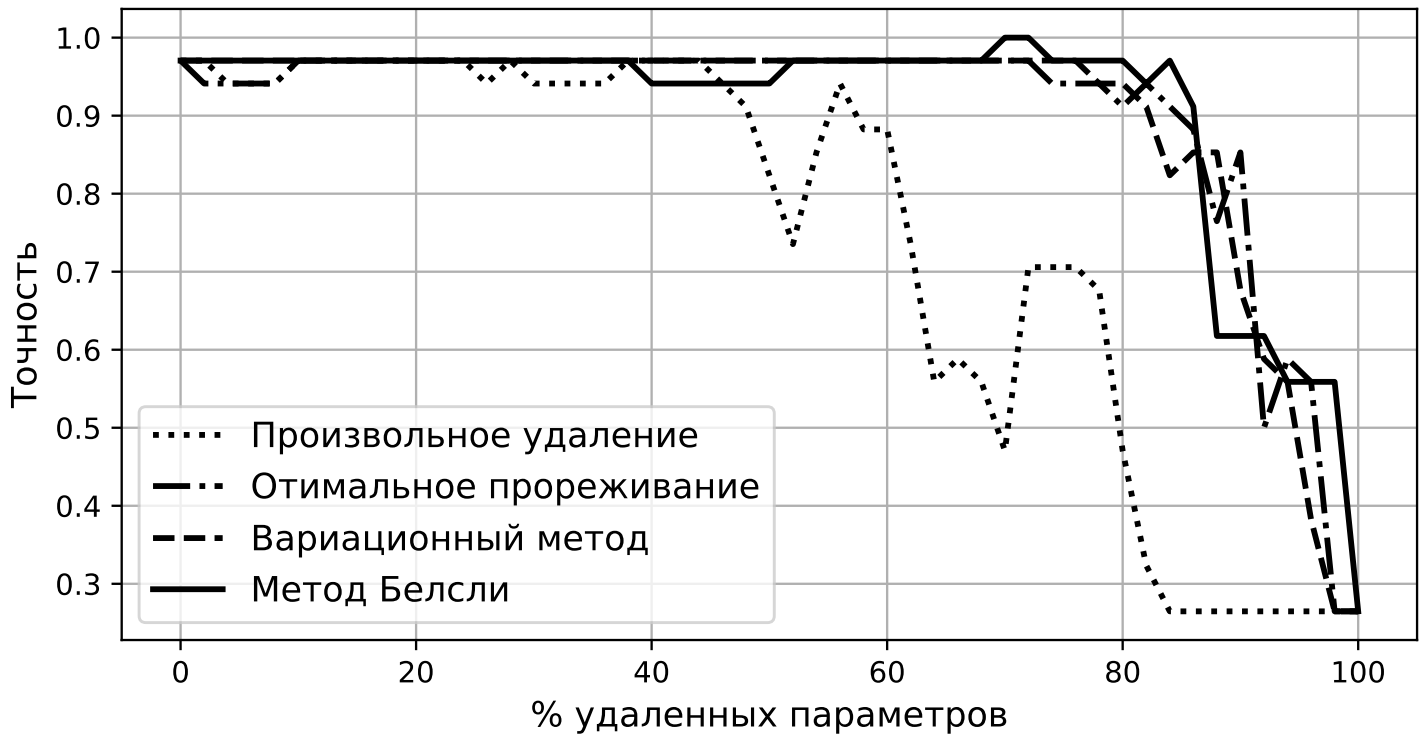


Рис.: Качество прогноза при удаление параметров на выборке Wine

Результаты эксперимента

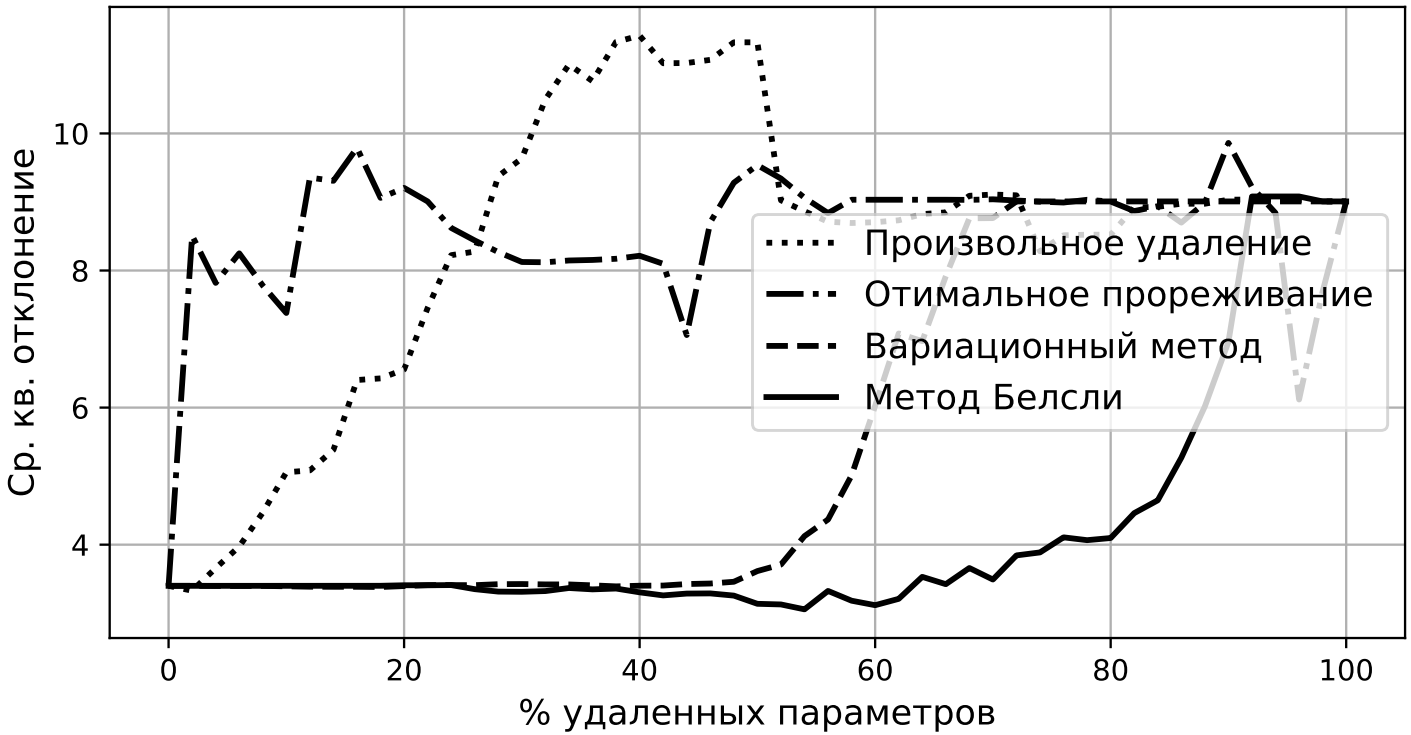


Рис.: Качество прогноза при удаление параметров на выборке Boston

Результаты эксперимента

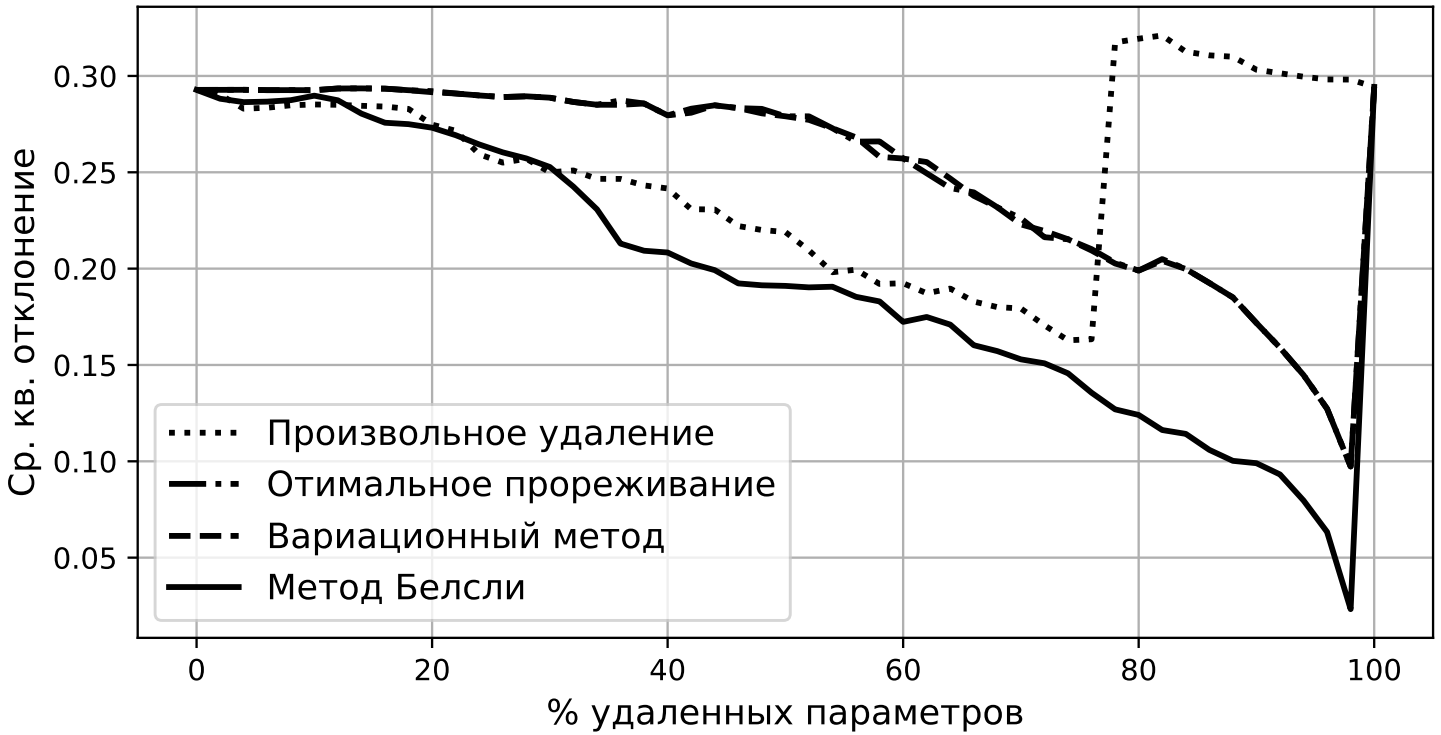
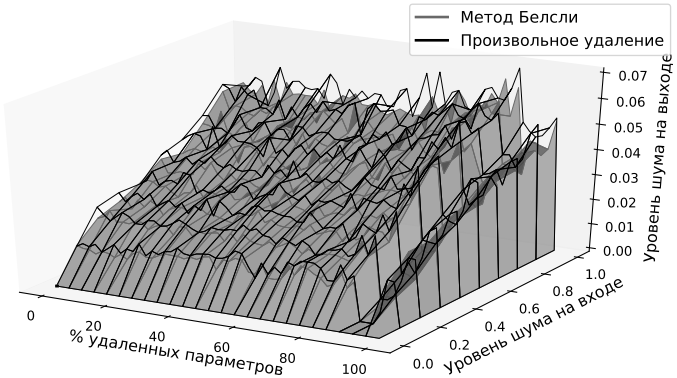
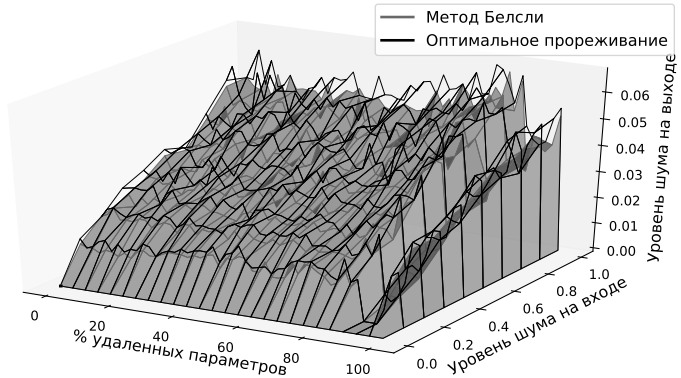


Рис.: Качество прогноза при удаление параметров на синтетических данных

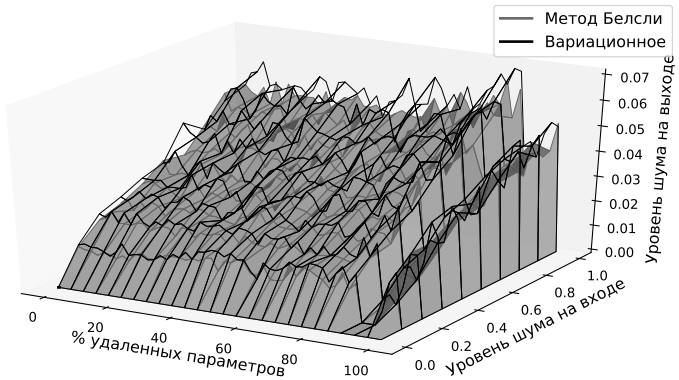
Результаты эксперимента



(a) Произвольное удаление параметров



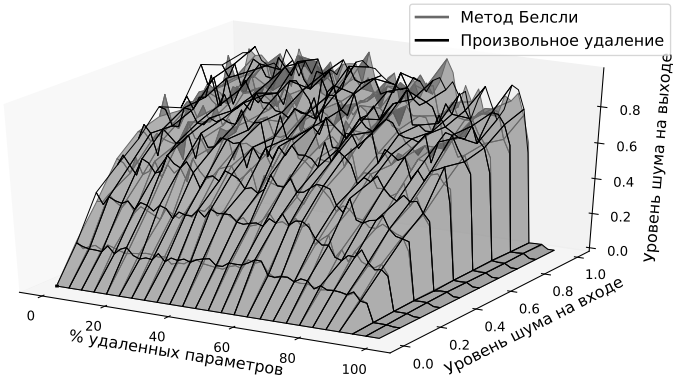
(b) Оптимальное прореживание



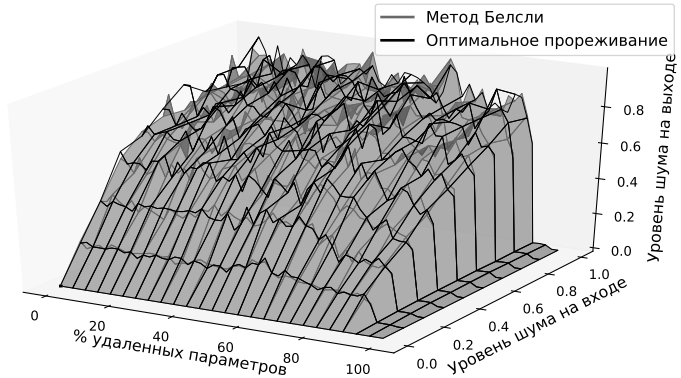
(c) Вариационный метод

Рис.: Влияние шума в начальных данных на шум выхода нейросети на выборке Wine

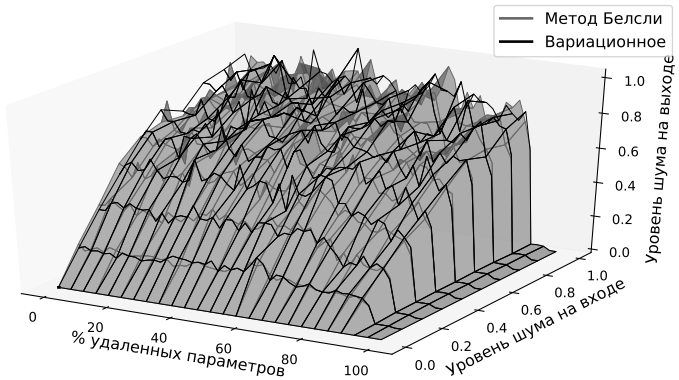
Результаты эксперимента



(a) Произвольное удаление параметров



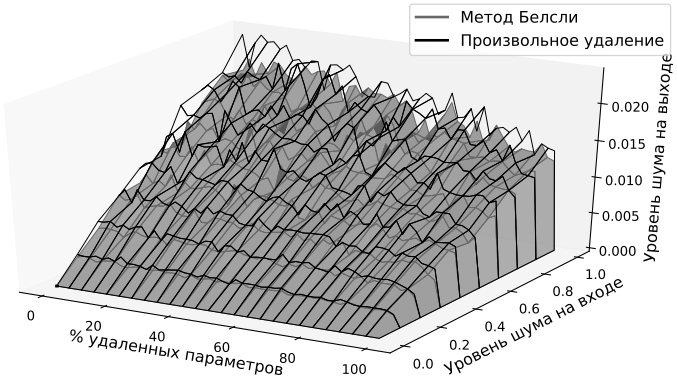
(b) Оптимальное прореживание



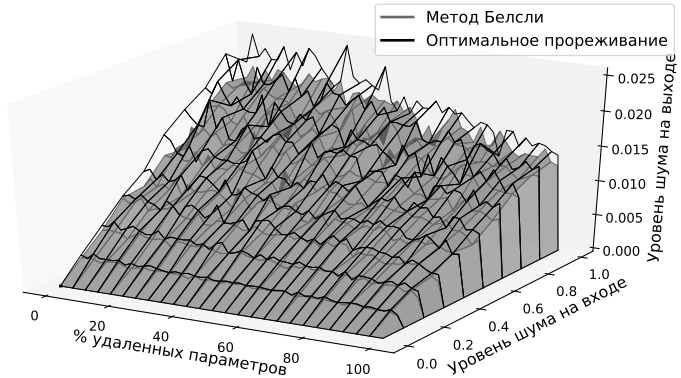
(c) Вариационный метод

Рис.: Влияние шума в начальных данных на шум выхода нейросети на выборке Boston

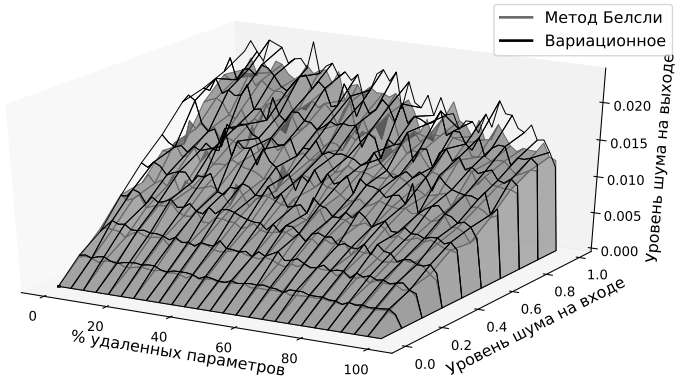
Результаты эксперимента



(a) Произвольное удаление параметров



(b) Оптимальное прореживание



(c) Вариационный метод

Рис.: Влияние шума в начальных данных на шум выхода нейросети на синтетических данных

Спасибо за внимание!