# ESTIMATION OF THE RELEVANCE OF THE NEURAL NETWORK PARAMETERS*

**A. V. Grabovoy[1], O. Yu. Bakhteev[2], V. V. Strijov[3]**

**Abstract**: This paper investigates a method for optimizing the structure of a neural network. It assumes that the number of neural network parameters can be reduced without significant loss of quality and without significant increase in the variance of the loss function. The paper proposes a method for automatic estimation of the relevance of parameters to prune a neural network. This method analyzes the covariance matrix of the posteriori distribution of the model parameters and removes the least relevant and multicorrelate parameters. It uses the Belsly method to search for multicorrelation in the neural network. The proposed method was tested on the Boston Housing data set, the Wine data set, and synthetic data.

**Keywords**: neural network; hyperparameters optimisation; Belsly method; relevance of parameters; neural network pruning

*This research was supported by RFBR, project 19-07-0875, and by Government of the Russian Federation, agreement 05.Y09.21.0018.

[1]Moscow Institute of Physics and Technology, grabovoy.av@phystech.edu

[2]Moscow Institute of Physics and Technology, bakhteev@phystech.edu

[3]Dorodnicyn Computing Center, Russian Academy of Sciences, strijov@ccas.ru

# References


[1] Sutskever, I., O. Vinyals, and Q. Le. 2014. Sequence to Sequence Learning with Neural Networks. *Advances in Neural Information Processing Systems*. Quebec. 2:3104–3112.

[2] Maclaurin, D., D. Duvenaud. and R. Adams. 2015. Gradient-based Hyperparameter Optimization Through Reversible Learning. *Proceedings of the 32th International Conference on Machine Learning*. Lille. 37:2113–2122.

[3] Luketina, J., M. Berglund, T. Raiko, and K. Greff. 2016. Scalable Gradient-based Tuning of Continuous Regularization Hyperparameters. *Proceedings of the 33th International Conference on Machine Learning*. New York. 48:2952–2960.

[4] Molchanov, D., A. Ashukha, and D. Vetrov. 2017. Variational Dropout Sparsifies Deep Neural Networks. *Proceedings of the 34th International Conference on Machine Learning*. Sydney. 70:2498–2507.

[5] Neal, A., and M. Radford. 1995. Bayesian Learning for Neural Networks. Toronto, Canada: University of Toronto. PhD Thesis. 195 p.

[6] LeCun, Y., J. Denker, and S. Solla. 1989. Optimal Brain Damage. *Advances in Neural Information Processing Systems*. Denver. 2:598–605.

[7] Louizos, C., K. Ullrich, and M. Welling. 2017. Bayesian Compression for Deep Learning. *Advances in Neural Information Processing Systems*. California. 3288–3298.

[8] Graves, A. 2011. Practical Variational Inference for Neural Networks. *Advances in Neural Information Processing Systems*. Granada. 2348–2356.

[9] Neychev, R., A. Katrutsa, and V. Strijov. 2016. Robust selection of multicollinear features in forecasting. *Factory Laboratory*. 82(3):68–74.

[10] Harrison, D., and D. Rubinfeld. Hedonic prices and the demand for clean air, 1991. Available at: `https://www.cs.toronto.edu/~delve/data/boston/bostonDetail.html`.

[11] Aeberhard, S. Wine Data Set, 1991. Available at: `http://archive.ics.uci.edu/ml/datasets/Wine`.

[12] Bishop, C. 2006. *Pattern Recognition and Machine Learning*. Berlin: Springer. 758 p.

# ОПРЕДЕЛЕНИЕ РЕЛЕВАНТНОСТИ ПАРАМЕТРОВ НЕЙРОСЕТИ*

**А. В. Грабовой**[1], **О. Ю. Бахтеев**[2], **В. В. Стрижов**[3]

**Аннотация:** Работа посвящена оптимизации структуры нейронной сети. Предполагается, что число параметров нейросети можно существенно снизить без значимой потери качества и значимого повышения дисперсии функции ошибки. Предлагается метод прореживания параметров нейронной сети, основанный на автоматическом определении релевантности параметров. Для определения релевантности параметров предлагается проанализировать ковариационую матрицу апостериорного распределения параметров и удалить из нейросети мультикоррелирующие параметры. Для определения мультикорреляции используется метод Белсли. Для анализа качества представленного алгоритма проводятся эксперименты на выборке Boston Housing, а также на синтетических данных.

**Ключевые слова**: нейронные сети; оптимизация гиперпараметров; метод Белсли; релевантность параметров; прореживание нейронной сети

**DOI**: 00.00000/00000000000000

*Работа выполнена при поддержке РФФИ (проект 19-07-0875) и правительства РФ (соглашение 05.Y09.21.0018).

[1]Московский физико-технический институт, grabovoy.av@phystech.edu

[2]Московский физико-технический институт, bakhteev@phystech.edu

[3]Вычислительный центр им. А. А. Дородницына ФИЦ ИУ РАН, strijov@ccas.ru

# 1 Введение

Решается задача выбора оптимальной структуры нейронной сети. В силу высокой вычислительной сложности, время оптимизации нейронных сетей может занимать до нескольких дней [1]. Поэтому построение и выбор оптимальной структуры нейронной сети также является вычислительно сложной процедурой, которая значимо влияет на итоговое качество модели. Использование избыточно сложных моделей с избыточным числом неинформативных параметров является препятствием для использования глубоких сетей на мобильных устройствах в режиме реального времени.

Существует ряд подходов к построению оптимальной сети. В работах [2, 3] предлагается использовать модель градиентного спуска для оптимизации сети. В [4] используются байесовские методы [5] оптимизации параметров нейронных сетей. Другим методом поиска оптимальной структуры является прореживание избыточно сложной модели [6, 7, 8]. В работе [6] предлагается удалять наименее релевантные параметры на основе значений первой и второй производных функции ошибки.

Данная работа посвящена прореживанию структуры сети. Предлагается удалять наименее релевантные параметры модели. Под релевантностью [6] подразумевается то, насколько параметр влияет на функцию ошибки. Малая релевантность указывает на то, что удаление этого параметра не влечет значимого изменения функции ошибки. Метод предлагает построение исходной избыточной сложности нейросети с большим числом избыточных параметров. Для определения релевантности параметров предлагается оптимизацировать параметры и гиперпараметры в единой процедуре. Для удаления параметров предлагается использовать метод Белсли [9] .

Проверка и анализ метода проводится на выборке Boston Housing [10], Wine [11] и синтетических данных. Результат сравнивается с моделью, полученной при помощи базовых алгоритмов.

# 2 Постановка задачи

Задана выборка

$$\mathfrak{D} = \{\mathbf{x}_i, y_i\},\ i = 1, ..., N, \tag{2.1}$$

где $\mathbf{x}_i \in \mathbb{R}^m$, $y_i \in \{1, \ldots, Y\}$, $Y$ — число классов. Рассмотрим модель $f(\mathbf{x}, \mathbf{w})$ : $\mathbb{R}^m \times \mathbb{R}^n \to \{1, \ldots, Y\}$, где $\mathbf{w} \in \mathbb{R}^n$ — пространство параметров модели,

$$f(\mathbf{x}, \mathbf{w}) = \text{softmax}\big(f_1(f_2(...(f_l(\mathbf{x}, \mathbf{w})\big), \tag{2.2}$$

где $f_k(\mathbf{x}, \mathbf{w}) = \tanh(\mathbf{w}^\mathsf{T}\mathbf{x})$, $l$ — число слоев нейронной сети, $k \in \{1 \ldots l\}$. Параметр $w_j$ модели $f$ называется активным, если $w_j \neq 0$. Множество индексов активных параметров обозначим $\mathcal{A} \subset \mathcal{J} = \{1, ..., n\}$. Задано пространство параметров модели:

$$\mathbb{W}_\mathcal{A} = \{\mathbf{w} \in \mathbb{R}^n \mid w_j \neq 0,\ j \in \mathcal{A}\}, \tag{2.3}$$

Для модели $f$ с множеством индексов активных параметров $\mathcal{A}$ и соответствующего ей вектора параметров $\mathbf{w} \in \mathbb{W}_\mathcal{A}$ определим логарифмическую функцию правдоподобия выборки:

$$\mathcal{L}_\mathfrak{D}(\mathfrak{D}, \mathcal{A}, \mathbf{w}) = \log p(\mathfrak{D}|\mathcal{A}, \mathbf{w}), \tag{2.4}$$

где $p(\mathfrak{D}|\mathcal{A}, \mathbf{w})$ — апостериорная вероятность выборки $\mathfrak{D}$ при заданных $\mathbf{w}, \mathcal{A}$. Оптимальные значения $\mathbf{w}, \mathcal{A}$ находятся из минимизации $-\mathcal{L}_\mathcal{A}(\mathfrak{D}, \mathcal{A})$ — логарифма правдоподобия модели:

$$\mathcal{L}_\mathcal{A}(\mathfrak{D}, \mathcal{A}) = \log p(\mathfrak{D}|\mathcal{A}) = \log \int_{\mathbf{w} \in \mathbb{W}_\mathcal{J}} p(\mathfrak{D}|\mathbf{w}) p(\mathbf{w}|\mathcal{A}) d\mathbf{w}, \tag{2.5}$$

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

Так как вычисление интеграла (2.5) является вычислительно сложной задачей, рассмотрим вариационный подход [12] для решения этой задачи. Пусть задано распределение $q$:

$$q(\mathbf{w}) \sim \mathcal{N}(\mathbf{m}, \mathbf{A}_{\text{ps}}^{-1}), \tag{2.6}$$

где $\mathbf{m}, \mathbf{A}_{\text{ps}}^{-1}$ — вектор средних и матрица ковариации, аппроксимирующее неизвестное апостериорное распределение $p(\mathbf{w}|\mathfrak{D}, \mathcal{A})$, полученное при априорном предположении:

$$p(\mathbf{w}|\mathcal{A}) \sim \mathcal{N}(\boldsymbol{\mu}, \mathbf{A}_{\text{pr}}^{-1}), \tag{2.7}$$

где $\boldsymbol{\mu}, \mathbf{A}_{\text{pr}}^{-1}$ — вектор средних и матрица ковариации априорного распределения.

Приблизим интеграл (2.5) методом, предложеном в [12]:

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

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

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

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

$$= \mathcal{L}_{\mathbf{w}}(\mathfrak{D},\mathcal{A},\mathbf{w}) + \mathcal{L}_{E}(\mathfrak{D},\mathcal{A}). \tag{2.8}$$

Первое слагаемое формулы (2.8) — это сложность модели. Оно определяется расстоянием Кульбака-Лейблера:

$$\mathcal{L}_{\mathbf{w}}(\mathfrak{D},\mathcal{A},\mathbf{w}) = -D_{KL}\big(q(\mathbf{w})||p(\mathbf{w}|\mathcal{A})\big). \tag{2.9}$$

Второе слагаемое формулы (2.8) является матожиданием правдоподобия выборки $\mathcal{L}_{\mathfrak{D}}(\mathfrak{D},\mathcal{A},\mathbf{w})$. В данной работе оно является функцией ошибки:

$$\mathcal{L}_{E}(\mathfrak{D},\mathcal{A}) = \mathsf{E}_{\mathbf{w}\sim q}\mathcal{L}_{\mathfrak{D}}(\mathbf{y},\mathfrak{D},\mathcal{A},\mathbf{w}). \tag{2.10}$$

Требуется найти параметры, доставляющие минимум суммарному функционалу потерь $\mathcal{L}_{\mathcal{A}}(\mathfrak{D},\mathcal{A},\mathbf{w})$ из (2.8):

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

$$= \underset{\mathcal{A}\subset\mathcal{J},\ \mathbf{w}\in\mathbb{W}_{\mathcal{A}}}{\arg\min} D_{KL}\big(q(\mathbf{w})||p(\mathbf{w}|\mathcal{A})\big) - \mathcal{L}_{\mathfrak{D}}(\mathfrak{D},\mathcal{A},\mathbf{w}). \tag{2.11}$$

# 3 Базовые методы прореживания нейросети

## 3.1 Случайное удаление

Метод случайного удаления заключается в том, что случайным образом удаляется некоторый параметр $w_\xi$ из множества активных параметров сети. Индекс параметра $\xi$ из равномерного распределения случайная величина, предположительно доставляющая оптимум в (2.11).

$$\xi \sim \mathcal{U}(\mathcal{A}). \tag{3.1.1}$$

## 3.2 Оптимальное прореживание

Метод оптимального прореживания [6] использует вторую производную целевой функции (2.4) по параметрам для определения нерелевантных параметров. Рассмотрим функцию потерь $\mathcal{L}$ (2.4) разложенную в ряд Тейлора в некоторой окрестности вектора параметров $\mathbf{w}$:

$$\delta\mathcal{L} = \sum_{j\in\mathcal{A}} g_j \delta w_j + \frac{1}{2}\sum_{i,j\in\mathcal{A}} h_{ij}\delta w_i \delta w_j + O(||\delta\mathbf{w}||^3), \tag{3.2.1}$$

где $\delta w_j$ — компоненты вектора $\delta\mathbf{w}$, $g_j$ — компоненты вектора градиента $\nabla\mathcal{L}$, а $h_{ij}$ — компоненты гессиана $\mathbf{H}$:

$$g_j = \frac{\partial\mathcal{L}}{\partial w_j}, \qquad h_{ij} = \frac{\partial^2\mathcal{L}}{\partial w_i \partial w_j}. \tag{3.2.2}$$

Задача является вычислительно сложной в силу высокой размерности матрицы $\mathbf{H}$. Введем предположение [6], о том что удаление нескольких параметров приводит к такому же изменению функции потерь $\mathcal{L}$, как и суммарное изменение при индивидуальном удалении:

$$\delta\mathcal{L} = \sum_{j\in\mathcal{A}} \delta\mathcal{L}_j, \tag{3.2.3}$$

где $\mathcal{A}$ — множество активных параметров, $\delta\mathcal{L}_j$ — изменение функции потерь при удалении одного параметра $\mathbf{w}_j$.

В силу данного предположения будем рассматривать только диагональные элементы матрицы $\mathbf{H}$. После введенного предположения, выражение (3.2.1) принимает вид

$$\delta\mathcal{L} = \frac{1}{2}\sum_{j\in\mathcal{A}} h_{jj}\delta w_j^2, \tag{3.2.4}$$

Получаем следующую задачу оптимизации:

$$\xi = \arg\min_{j\in\mathcal{A}} h_{jj}\frac{w_j^2}{2}, \tag{3.2.5}$$

где $\xi$ — индекс наименее релевантного, удаляемого параметра, предположительно доставляющая оптимум в (2.11).

## 3.3 Удаление неинформативных параметров с помощью вариационного вывода

Для удаления параметров в работе [8] предлагается удалить параметры, которые имеют максимальное отношение плотности $p(\mathbf{w}|\mathcal{A})$ априорной вероятности в нуле к плотности вероятности априорной вероятности в математическом ожидании $\mu_j$ параметра $w_j$.
Для гауссовского распределения с диагональной матрицей ковариации получаем:

$$p_j(\mathbf{w}|\mathcal{A})(w) = \frac{1}{\sqrt{2\sigma_j^2}} \exp(-\frac{(w-\mu_j)^2}{2\sigma_j^2}), \tag{3.3.1}$$

где $w$ — значение носителя распределенного параметра. Разделим плотность вероятности в нуле к плотности в математическом ожидании

$$\frac{p_j(\mathbf{w}|\mathcal{A})(0)}{p_j(\mathbf{w}|\mathcal{A})(\mu_j)} = \exp(-\frac{\mu_j^2}{2\sigma_j^2}), \tag{3.3.2}$$

и поставим следующую задачу оптимизации:

$$\xi = \arg\min_{j\in\mathcal{A}} \left|\frac{\mu_j}{\sigma_j}\right|, \tag{3.3.3}$$

где $\xi$ — индекс наименее релевантного, удаляемого параметра.

# 4 Предлагаемый метод определения релевантности параметров нейросети

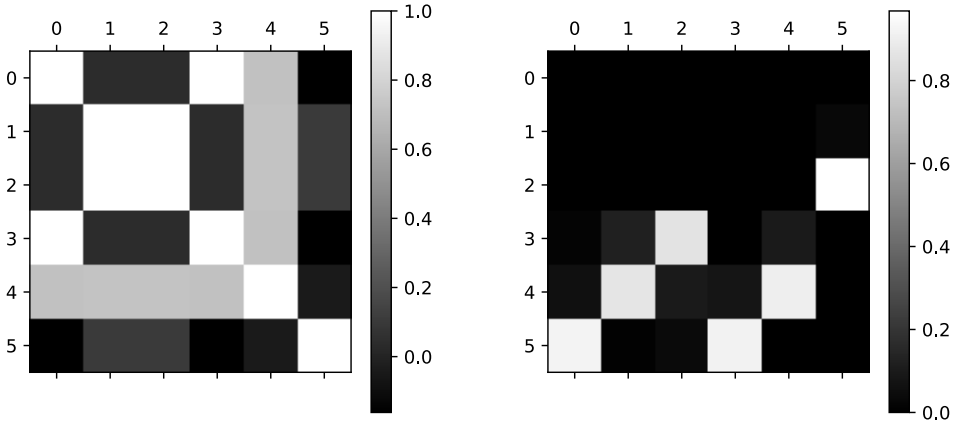
Предлагается метод основанный, на модификации метода Белсли. Пусть $\mathbf{w}$ — вектор параметров, доставляющий минимум функционалу потерь $\mathcal{L}_\mathcal{A}$ из (2.8) на множестве $\mathbb{W}_\mathcal{A}$, а $\mathbf{A}_{\text{ps}}$ соответствующая ему ковариационная матрица.

Выполним сингулярное разложение матрицы

$$\mathbf{A}_{\text{ps}} = \mathbf{U}\mathbf{\Lambda}\mathbf{V}^\mathsf{T}. \tag{4.1}$$

Индекс обусловленности $\eta_j$ определим как отношение максимального элемента к $j$-му элементу матрицы $\mathbf{\Lambda}$. Для нахождения мультикоррелирующих признаков требуется найти индекс $\xi$ вида:

$$\xi = \arg\max_{j\in\mathcal{A}} \eta_j. \tag{4.2}$$




(a) Матрица ковариации (b) Дисперсионные доли

Рис. 1: Илюстрация метода Белсли

Дисперсионный долевой коэффициент $q_{ij}$ определим как вклад $j$-го признака в дисперсию $i$-го элемента вектора параметра $\mathbf{w}$:

$$q_{ij} = \frac{u_{ij}^2/\lambda_{jj}}{\sum_{j=1}^{n} u_{ij}^2/\lambda_{jj}}. \tag{4.3}$$

Большие значение дисперсионных долей указывают на наличие зависимости между параметрами. Находим долевые коэффициенты, которые вносят максимальный вклад в дисперсию параметра $w_\xi$:

$$\zeta = \arg\max_{j \in \mathcal{A}} q_{\xi j}. \tag{4.4}$$

Параметр с индексом $\zeta$ определим как наименее релевантный параметр нейросети.

Проиллюстрируем принцип работы метода Белсли на примере. Рас-

Таблица 1: Илюстрация метода Белсли

| $\eta$ | $q_1$ | $q_2$ | $q_3$ | $q_4$ | $q_5$ | $q_6$ |
|---|---|---|---|---|---|---|
| 1.0 | $2 \cdot 10^{-17}$ | $4 \cdot 10^{-17}$ | $1 \cdot 10^{-16}$ | $2 \cdot 10^{-17}$ | $6 \cdot 10^{-17}$ | $3 \cdot 10^{-4}$ |
| 1.5 | $5 \cdot 10^{-17}$ | $9 \cdot 10^{-17}$ | $2 \cdot 10^{-16}$ | $5 \cdot 10^{-17}$ | $3 \cdot 10^{-20}$ | $3 \cdot 10^{-2}$ |
| 3.3 | $9 \cdot 10^{-18}$ | $1 \cdot 10^{-17}$ | $2 \cdot 10^{-17}$ | $9 \cdot 10^{-18}$ | $2 \cdot 10^{-19}$ | $9 \cdot 10^{-1}$ |
| $2 \cdot 10^{15}$ | $1 \cdot 10^{-2}$ | $1 \cdot 10^{-1}$ | $8 \cdot 10^{-1}$ | $2 \cdot 10^{-3}$ | $9 \cdot 10^{-2}$ | $1 \cdot 10^{17}$ |
| $8 \cdot 10^{15}$ | $6 \cdot 10^{-2}$ | $8 \cdot 10^{-1}$ | $9 \cdot 10^{-2}$ | $8 \cdot 10^{-2}$ | $9 \cdot 10^{-1}$ | $2 \cdot 10^{17}$ |
| $1 \cdot 10^{16}$ | $\mathbf{9 \cdot 10^{-1}}$ | $1 \cdot 10^{-2}$ | $4 \cdot 10^{-2}$ | $\mathbf{9 \cdot 10^{-1}}$ | $1 \cdot 10^{-3}$ | $5 \cdot 10^{-21}$ |

смотрим данные порожденные следующим образом:

$$\mathbf{w} = \begin{bmatrix} \sin(x) \\ \cos(x) \\ 2{+}\cos(x) \\ 2{+}\sin(x) \\ \cos(x) + \sin(x) \\ x \end{bmatrix}$$

с матрицей ковариации на рис. 1.a, где $x \in [0.0, 0.02, ..., 20.0]$.

В табл. 1 приведены индексы обусловленности и соответствующие им дисперсионные доли, которые также изображены на рис. 1.b. Согласно этим данным, максимальный индекс обусловленности $\eta_6 = 1.2 \cdot 10^{16}$. Ему соответствуют максимальные дисперсионные доли признаков с индексами 1 и 4, которые, как видно из построения выборки, коррелируют между собой.

# 5 Вычислительный эксперимент

Для анализа свойств предложенного алгоритма и сравнения его с существующими был проведен вычислительный эксперимент в котором параметры нейросети удалялись методами, которые были описаны в разделах 3.1—3.3 и методом Белсли.

В качестве данных использовались три выборки. Выборки Wine [11] и Boston Housing [10] — это реальные данные. Синтетические данные сгенерированы таким образом, чтобы параметры сети были мультикорелируемые. Генерация данных состояла из двух этапов. На первом этапе

генерировался вектор параметров $\mathbf{w}_{\text{synthetic}}$:

$$\mathbf{w}_{\text{synthetic}} \sim \mathcal{N}(\mathbf{m}_{\text{synthetic}}, \mathbf{A}_{\text{synthetic}}), \tag{5.1}$$

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

На втором этапе генерировалась выборка $\mathfrak{D}_{\text{synthetic}}$:

$$\mathfrak{D}_{\text{synthetic}} = \{(\mathbf{x}_i, y_i) | \mathbf{x}_i \sim \mathcal{N}(\mathbf{1}, \mathbf{I}), y_i = x_{i0}, i = 1 \cdots 10000\}. \tag{5.2}$$

В приведенном выше векторе параметров $\mathbf{w}_{\text{synthetic}}$ для выборки $\mathfrak{D}_{\text{synthetic}}$, наиболее релевантным является первый параметр, а все остальные параметры являются нерелевантными. Матрица ковариации была выбрана таким образом, чтобы все нерелевантные параметры были зависимы и метод Белсли был максимально эффективен.

Таблица 2: Описание выборок

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

Для алгоритмов тренировочная и тестовая выборки составили 80% и 20% соответсвенно. Критерием качества прореживания служит процент параметров нейросети, удаление которого не влечет значимой потери качества прогноза. Также критерием качества служит устойчивость нейросети к зашумленности данных.

Качеством прогноза $R_{\text{cl}}$ модели для задачи классификации является точность прогноза модели:

$$R_{\text{cl}} = \frac{\sum_{(\mathbf{x},y)\in\mathfrak{D}} [f(\mathbf{x}, \mathbf{w}) = y]}{|\mathfrak{D}|}, \tag{5.3}$$

Качеством прогноза $R_{\text{rg}}$ модели для задачи регрессии является среднеквадратическое отклонение результата модели от точного:

$$R_{\text{rg}} = \frac{\sum_{(\mathbf{x},y)\in\mathfrak{D}} (f(\mathbf{x}, \mathbf{w}) - y)^2}{|\mathfrak{D}|}, \tag{5.4}$$

**Wine.** Рассмотрим нейроную сеть с 13 нейронами на входе, 13 нейронами в скрытом слое и 3 нейронами на выходе.

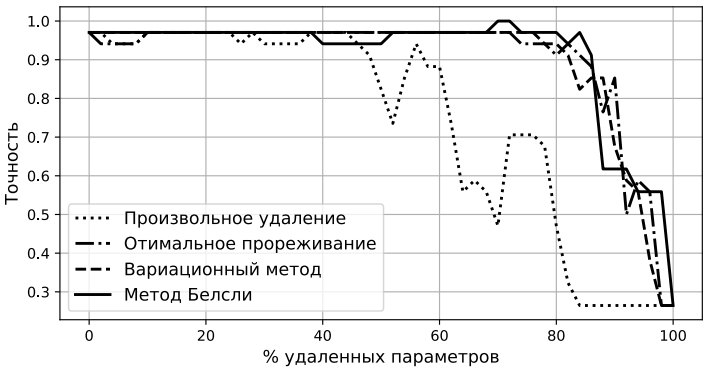
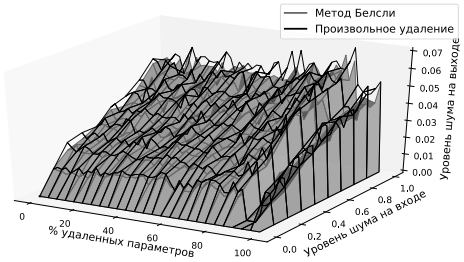
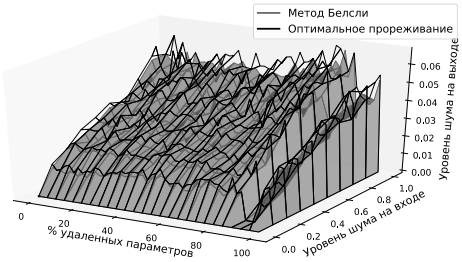


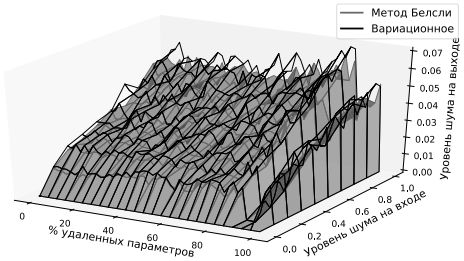

Рис. 2: Качество прогноза при удаление параметров на выборке Wine




(a) Произвольное удаление параметров




(b) Оптимальное прореживание




(c) Вариационный метод

Рис. 3: Влияние шума в начальных данных на шум выхода нейросети на выборке Wine

На рис. 2 показано как меняется точность прогноза $R_{\text{cl}}$ при удалении параметров указанными методами. Из графика видно, что метод оптимального прореживания, вариационный метод и метод Белсли позволяют удалить $\approx 80\%$ параметров и качество всех этих методов падает при удалении $\approx 90\%$ параметров нейросети.

На рис. 3 показаны поверхности изменения уровня шума ответов нейросети при изменении процента удаленных параметров и уровня шума входных данных для разных методов прореживания. На графиках показано, что при удалении параметров нейросети методом Белсли шум меньше, чем при удалении параметров другими методами, на это указывает то что поверхность которая соответствует методу Белсли ниже других поверхностей.

**Boston Housing.** Рассмотрим нейроную сеть с 13 нейронами на входе, 39 нейронами в скрытом слое и одним нейроном на выходе.

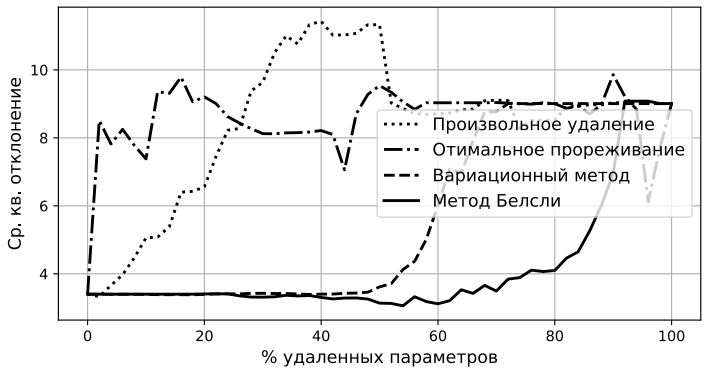
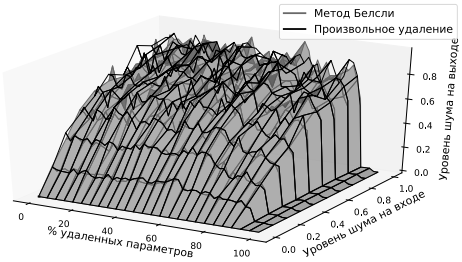


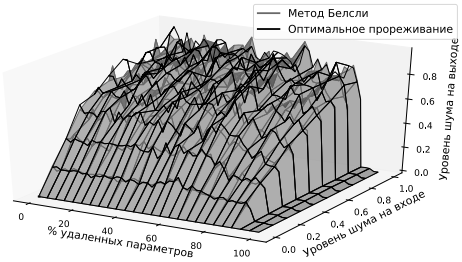

Рис. 4: Качество прогноза при удаление параметров на выборке Boston

На рис. 4 показано как меняется среднеквадратическое отклонение прогноза $R_{\text{rg}}$ от точного ответа при удалении параметров указанными методами. График показывает, что метод Белсли является более эффективным, чем другие методы, так-как позволяет удалить больше параметров нейросети без потери качества.

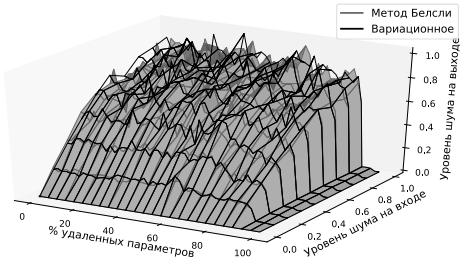
На рис. 5 показаны поверхности изменения уровня шума ответов нейросети при изменении процента удаленных параметров и уровня шума входных данных для разных методов прореживания. График показыва-




(a) Произвольное удаление параметров




(b) Оптимальное прореживание




(c) Вариационный метод

Рис. 5: Влияние шума в начальных данных на шум выхода нейросети на выборке Boston

ет, что уровень шума всех методов одинаковый, так как поверхности всех методов находятся на одном уровне.

**Синтетические данные.** Рассмотрим нейроную сеть с 100 нейронами на входе и одним нейроном на выходе.

На рис. 6 показано как меняется среднеквадратическое отклонение прогноза от $R_{\text{rg}}$ точного ответа при удалении параметров указанными методами. График показывает, что удаление параметров методом Белсли является более эффективным чем другие методы прореживания, так-как качество прогноза нейросети улучшается при удалении шумовых параметров.

На рис. 7 показаны поверхности изменения уровня шума ответов нейросети при изменении процента удаленных параметров и уровня шума входных данных для разных методов прореживания. На графиках показано, что при удалении параметров нейросети методом Белсли шум меньше, чем при удалении параметров другими методами, так-как по-

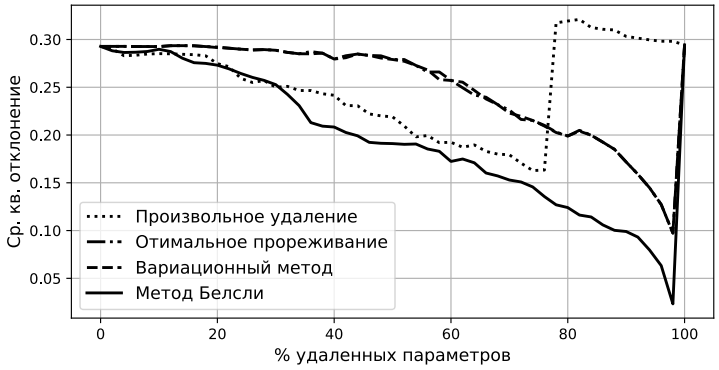
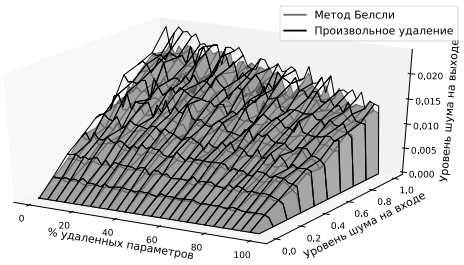
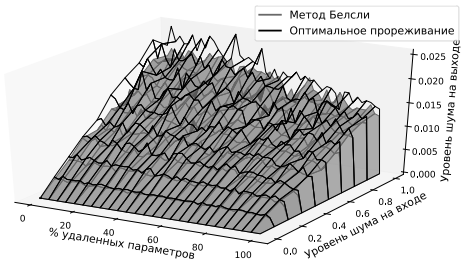


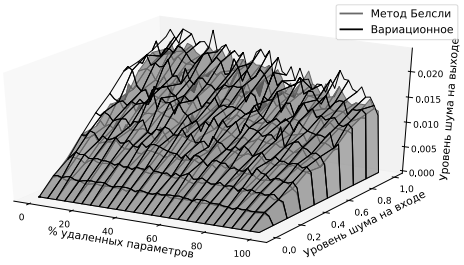

Рис. 6: Качество прогноза при удаление параметров на синтетической выборке




(a) Произвольное удаление параметров




(b) Оптимальное прореживание




(c) Вариационный метод

Рис. 7: Влияние шума в начальных данных на шум выхода нейросети на синтетической выборке

верхность которая соответствует методу Белсли ниже других поверхностей.

# 6 Заключение

В работе рассматривалась задача прореживания моделей нейросетей. Рассматривался метод оптимального прореживания и метод, основаный на вариационном подходе. Был предложен алгоритм прореживания, основаный на методе Белсли для удаления зависимых параметров модели. В ходе эксперимента было показано, что нейросети прорежены методом Белсли являются более устойчивы к шуму на входных данных. Качества прогноза нейросетей после прореживания методом Белсли не хуже качества прогноза нейросетей, полученных другими методами.

# Список литературы


[1] *Sutskever I., Vinyals O., Le Q.* Sequence to Sequence Learning with Neural Networks // Advances in Neural Information Processing Systems, 2014. Vol. 2. P. 3104–3112.

[2] *Maclaurin D., Duvenaud D., Adams R.* Gradient-based Hyperparameter Optimization Through Reversible Learning // Proceedings of the 32th International Conference on Machine Learning, 2015. Vol. 37. P. 2113–2122.

[3] *Luketina J., Berglund M., Raiko T., Greff K.* Scalable Gradient-based Tuning of Continuous Regularization Hyperparameters // Proceedings of the 33th International Conference on Machine Learning, 2016. Vol. 48. P. 2952–2960.

[4] *Molchanov D., Ashukha A., Vetrov D.* Variational Dropout Sparsifies Deep Neural Networks // Proceedings of the 34th International Conference on Machine Learning, 2017. Vol. 70. P. 2498–2507.

[5] *Neal A., Radford M.* Bayesian Learning for Neural Networks // PhD Thesis, Toronto, Ont., Canada, Canada, 1995. 195 p.

[6] *LeCun Y., Denker J. , Solla S.* Optimal Brain Damage // Advances in Neural Information Processing Systems, 1989. Vol. 2. P. 598–605.

[7] *Louizos C., Ullrich K., Welling M.* Bayesian Compression for Deep Learning // Advances in Neural Information Processing Systems, 2017. P. 3288–3298.

[8] *Graves A.* Practical Variational Inference for Neural Networks // Advances in Neural Information Processing Systems, 2011. P. 2348–2356.

[9] *Neychev R., Katrutsa A., Strijov V.* Robust selection of multicollinear features in forecasting // Factory Laboratory, 2016. Vol. 82. No. 2. P. 68–74.

[10] *Harrison D., D. Rubinfeld.* Hedonic prices and the demand for clean air, 1991. Available at: `https://www.cs.toronto.edu/~delve/data/boston/bostonDetail.html`.

[11] *Aeberhard, S.* Wine Data Set, 1991. Available at: `http://archive.ics.uci.edu/ml/datasets/Wine`.

[12] *Bishop C.* Pattern Recognition and Machine Learning. — Berlin: Springer, 2006. 758 p.